**Title**

Confidence matching in group decision-making

**Authors**

Dan Bang1,2,3,4,\*, Laurence Aitchison5, Rani Moran6, Santiago Herce Castañón1, Banafshe Rafiee7, Ali Mahmoodi8, Jennifer Lau2,9, Peter E. Latham5, Bahador Bahrami10, and Christopher Summerfield1

**Affiliations**

1 Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, United Kingdom.

2 Calleva Research Centre for Evolution and Human Sciences, University of Oxford, Oxford OX1 4AU, United Kingdom.

3 Interacting Minds Centre, Aarhus University, 8000 Aarhus, Denmark.

4 Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom.

5 Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom.

6 School of Psychological Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.

7Department of Computing Science, University of Alberta, AB T6G 2S4, Canada.

8 Bernstein Center Freiburg, University of Freiburg, 79104 Freiburg, Germany.

9 Department of Psychology, King’s College London, London SE5 8AF, United Kingdom.

10 Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, United Kingdom.

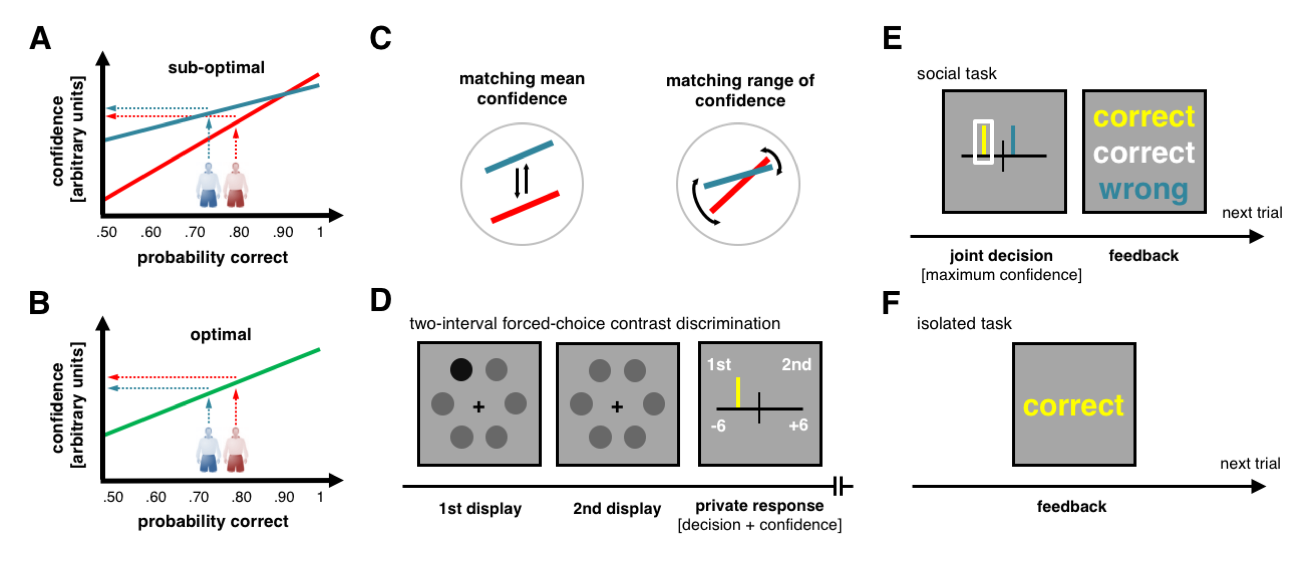
\* Correspondence to: [danbang.db@gmail.com](mailto:danbang.db@gmail.com).

**Most important decisions are made by groups of people. When disagreement arises, opinions expressed with higher confidence tend to carry more weight**1,2. **Although an individual’s degree of confidence often reflects the probability that their opinion is correct**3,4**, the intensity with which they convey this information can vary with task-irrelevant psychological, social, cultural and demographic factors**5–9**. Therefore, to combine their opinions optimally,** **group members must adapt to each other’s individual biases and express their confidence according to a common metric**10–12**. Solving this communication problem is, however, computationally difficult. Here we show that pairs of individuals making group decisions meet this challenge by using a heuristic strategy that we call *confidence matching*: they match their confidence such that certainty and uncertainty are stated in approximately equal measure by each party. Combining the behavioural data with model simulations, we show that this strategy works when group members have similar levels of expertise, or have no insight into their relative levels of expertise. Confidence matching is, however, sub-optimal and can cause miscommunication about who is more likely to be correct. This herding-like behaviour is one reason why groups can fail to make good decisions**10–12**.**

To illustrate the communication problem inherent to group decision-making, consider two football referees who disagree about whether the ball crossed the goal line. Each referee states their individual opinion with a certain level of confidence (**Figure 1A**, y-axis). This level of confidence is a function of some internal estimate of the probability that their individual opinion is correct (**Figure 1A**, x-axis). The referees have, however, different subjective mappings (solid lines), with the blue referee generally more biased towards high confidence. As a consequence, the group decision is in this interaction dominated by the blue referee who is in fact less likely to be correct (dotted lines). To avoid such miscommunication, the referees must align their subjective mappings so that their confidence is reported in a mutually consistent manner (**Figure 1B**).

It is, however, effectively impossible to reach this solution instantly. Without prior interaction, the referees can only make guesses about their colleague’s subjective mapping. Further, it is difficult to achieve over time. To estimate and adjust to their colleague’s subjective mapping, the referees would need extensive experience and feedback. Even then, especially because their colleague’s subjective mapping is being adjusted in return, the learning process may place too high demands on their working memory and depth of inference13,14. Here we tested the hypothesis that people solve the communication problem using a heuristic strategy: they seek to align their *unobservable* subjective mappings by matching their *observable* confidence (**Figure 1C**). Indeed, individuals tend to mimic each other’s communicative behaviours, such as vocabulary15, and it has been proposed that mimicry can reduce miscommunication, by aligning agents’ input-output functions16,17.

We ran six behavioural experiments to test our hypothesis. In Experiment 1, pairs of participants (30 groups, tested in Iran) performed a psychophysical task (**Figure 1D** and **Methods**). On each trial, they privately indicated which of two visual displays they thought contained a faint target, and how confident they felt about this decision on a discrete scale from 1 to 6. In the *social* condition (**Figure 1E**; EXP1-S: 160 trials, social task), participants performed the task together. The private responses were made public, and the individual decision made with higher confidence was selected as the joint decision. Under this decision rule, participants must report their confidence in a way that maximises the probability that the group makes the correct decision. In the *isolated* condition (**Figure 1F**; EXP1-I: 160 trials, isolated task), participants performed the task alone. Half of the groups performed the social condition first; the other half performed the isolated condition first.



**Figure 1.** Theoretical and experimental framework. **(A)** Communication problem. Two football referees disagree about whether the ball crossed the goal line. They have different subjective mappings (blue and red lines) from their internal estimates of uncertainty (x-axis) onto confidence reports (y-axis). In this interaction, the blue referee expresses higher confidence but is less likely to be correct (dotted lines). We assume a continuous confidence scale and linear functions for ease of exposition. **(B)** Optimal solution. To minimise miscommunication, the referees must align their subjective mappings (green line) such that their confidence reflects a common metric. This solution is optimal in the sense that it maximises the probability that the group makes the correct decision. **(C)** Confidence matching. The intercept (*left*) and the slope (*right*) of the referees’ subjective mappings would automatically change under confidence matching. **(D)** Psychophysical task. Participants viewed two consecutive displays, each containing six contrast gratings (here dots). In one of the two displays, there was a higher contrast target (darker dot). Participants responded by moving a marker along a scale with a fixed midpoint. The response sign indicated the decision (1st or 2nd display), and the absolute response value indicated the confidence (1 to 6 in steps of 1). **(E)** The social task. Participants’ private responses (colour-coded) were shared, and the response made with higher confidence was automatically selected as the joint decision (white box). Confidence ties were resolved by randomly selecting one of the two private responses. Participants received feedback (colour-coded) about the accuracy of each decision before continuing to the next trial. **(F)** The isolated task. Participants performed the task on their own, without any social interaction.

Under our hypothesis, we would expect group members’ confidence to be more similar when they performed the task together as compared to alone. Because of the boundedness of the confidence scale, the summary statistics that we can use to describe group members’ confidence are not fully independent. Here we focus on convergence in mean confidence, but we found similar results for the variance of confidence and confidence distributions (**Supplementary Figure 1**). In line with our hypothesis, the magnitudes of group members’ mean confidence were correlated in the social condition only (**Figure 2A**) and the differences in their mean confidence were smaller in the social than in the isolated condition (**Figure 2B**; *t*(29) = 4.195, *p* < .001, paired) – regardless of the condition order (**Supplementary Figure 2**). Since confidence correlates with accuracy18, one concern is that convergence in mean confidence can reflect a convergence in accuracy (fraction of correct individual decisions); maybe because group members matched the effort invested into the task19. However, this explanation was not consistent with the data: the differences in accuracy were in fact larger in the social than in the isolated condition (*t*(29) = 2.083, *p* = .046, paired; no effect of condition order).

Does the observed confidence matching reflect convergence onto a single fixed point? For example, group members may have gravitated towards medium confidence to minimise conflict20, or towards maximum confidence to dominate the joint decision21. Under our hypothesis, the convergence point would not be the same for all groups (e.g., scale centre or scale extremity) but dynamically vary across groups as a function of group members’ individual biases. We would therefore expect the differences in mean confidence to be smaller for *actual* groups compared to *shuffled* groups (formed by randomly re-pairing group members) as shuffling removes group-specific convergence points (**Figure 2C**). Building on this logic, we used a permutation-based approach to test for dynamic convergence (see **Methods**). In line with our hypothesis, the differences between group members’ mean confidence were smaller than expected under shuffled data in the social condition only (isolated: *p* = .405; social: *p* < .001). We found comparable social effects (**Figure 2**; permutation: all *p* < .001) in two additional experiments (see **Methods**) where participants had more task experience (EXP2: 15 groups, 384 trials, social task, tested in the UK) and used a continuous scale (EXP3: 15 groups, 384 social trials, social task, tested in the UK). The results were overall consistent with our hypothesis that people *actively* match their confidence during group decision-making – regardless of cultural context (Iran or UK), task experience (160 or 384 trials) and low-level factors such as the nature of the scale (discrete versus continuous).

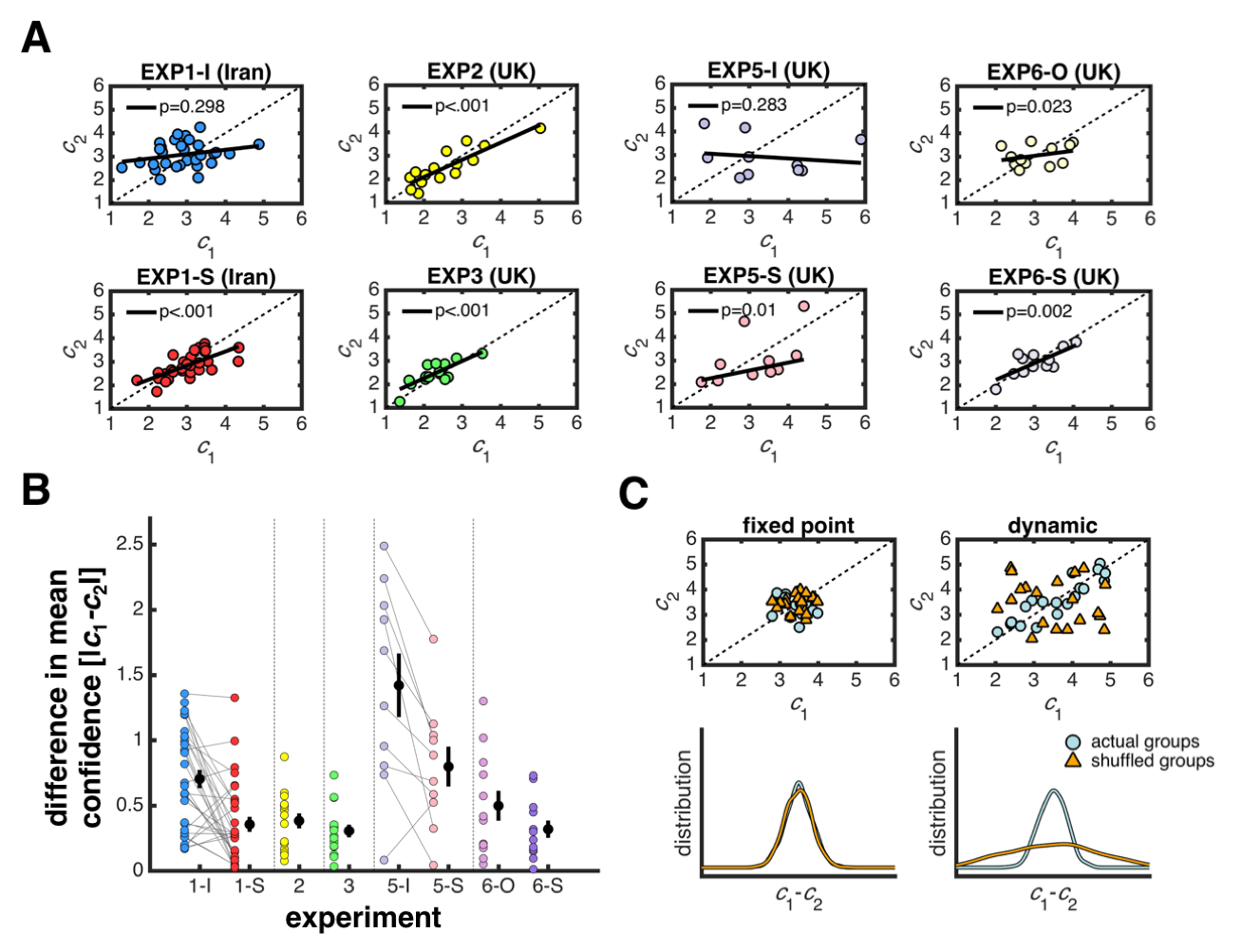
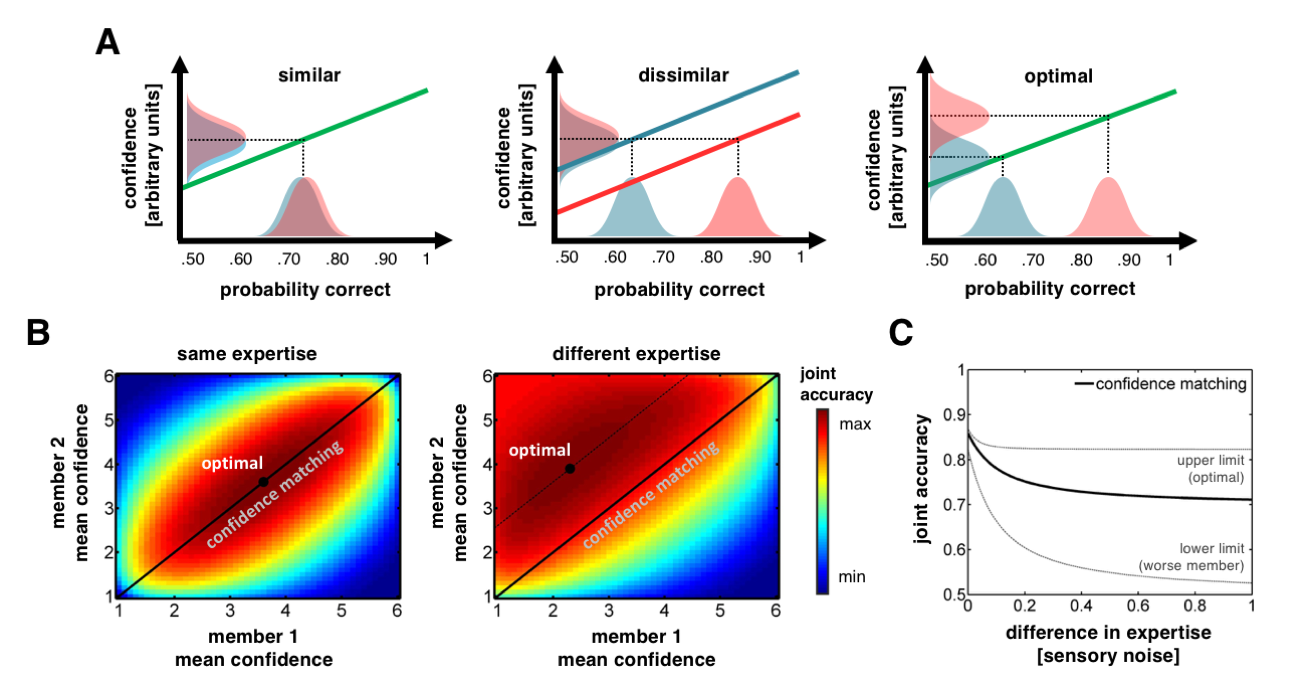


Figure 2. Behavioural evidence for confidence matching

. (**A**) Correlations in mean confidence. The axes show group members’ mean confidence, *c*1 and *c*2. Each dot is a group. Each line is the best-fitting line of a robust regression; the *p*-value indicates the average *p-*valueof its slope across 105 regressions, for each randomly sorting group members into 1 and 2. (**B**) Convergence in mean confidence. The y-axis shows the absolute difference between group members’ mean confidence. Each black dot is data averaged across groups. Each coloured dot is a group; the lines connect group data when the same pairing of group members was used in multiple conditions. Error bars are 1 SEM. (**C**) Shuffling data removes group-specific convergence points. (*left*) If the convergence point is shared across groups (here medium), then randomly re-pairing group members does not change the distribution of differences in mean confidence. (*right*) In contrast, if the convergence point is dynamically negotiated between group members, then randomly re-pairing group members will induce larger differences in mean confidence.

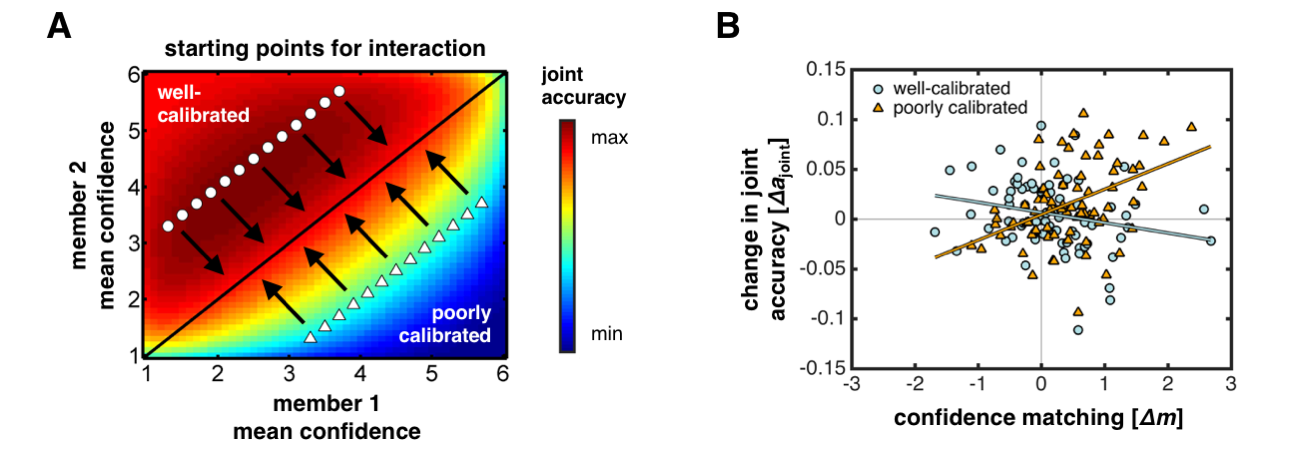
How does confidence matching affect group performance? Confidence matching clearly makes sense when group members have similar levels of expertise (**Figure 3A**, left panel), but we would expect it to be costly compared to the optimal solution when group members have different levels of expertise (**Figure 3A**, middle and right panel). Intuitively, if one group member is better than the other, then pooling their opinions with equal weight should lead to sub-optimal group decisions. To quantify this intuition in the context of our task, we used a signal detection model4 to simulate how joint accuracy varies with differences in expertise and mean confidence (see **Methods**). Using **Figure 3A** to illustrate our model,higher expertise means that the distribution of probability correct is shifted to the right and higher mean confidence means that the confidence distribution is shifted up (we used maximum entropy distributions in our simulations but the results should generalise to other reasonable distributions).



**Figure 3.** Optimality of confidence matching. **(A)** The effect of confidence matching depends on relative expertise. (*left*) If the referees sampled from similar distributions of probability correct, then their subjective mappings should converge under confidence matching. (*middle*) However, if one referee was more of an expert than the other, so that they sampled from different distributions of probability correct, then their subjective mappings would not align under confidence matching and the less competent referee would exert far too much influence. (*right*) If their subjective mappings were optimally aligned, then their confidence would preserve their relative expertise and the less competent referee would prevail less often. **(B)** Confidence landscapes. (*left*) For agents with similar levels of sensory noise, joint accuracy (colour scale) reaches its maximum (black dot) when their mean confidence is matched. (*right*) This is not the case for agents with different levels of sensory noise. **(C)** Cost of confidence matching. The x-axis shows the difference between agents’ sensory noise. The y-axis shows the joint accuracy under different scenarios. Solid line: average along the diagonal in a landscape. Upper grey line: optimal solution in a landscape. Lower grey line: lower bound, determined by the accuracy of the worse member.

**Figure 3B** shows landscapes of joint accuracy (fraction of correct joint decisions) as a function of mean confidence for two *simulated* agents with equal expertise (left panel) and unequal expertise (right panel). In each landscape, confidence matching corresponds to the coordinates along the diagonal. For agents of equal expertise, confidence matching can improve joint accuracy – as long as they do not only use the extremes of the scale (black dot on middle of diagonal in **Figure 3B**, left panel). However, when one agent is more of an expert than the other, confidence matching incurs a relative cost to joint accuracy (the optimal solution given by black dot has higher accuracy than any point along diagonal in **Figure 3B**, right panel). This pattern is summarised in **Figure 3C** which shows the joint accuracy expected under confidence matching as a function of differences in expertise. In a control analysis (see **Methods**) we show that the divergence between the joint accuracy observed in our experiments and that expected under the optimal solution was indeed larger for dissimilar group members as expected under confidence matching. The results show that confidence matching may be one cause of the common finding that group accuracy is positively correlated with the similarity of group members’ expertise10–12.

Is confidence matching ever helpful for group members with different levels of expertise? Groups are usually made up of individuals with different levels of expertise and varying mean confidence. A group can be said to be *well-calibrated* when its better member is in fact the more confident and *poorly calibrated* when its worse member is the more confident. Both types of group are likely to arise as people move between tasks and contexts. How do they fare under confidence matching? Based on our simulations, confidence matching should be costly for well-calibrated groups (joint accuracy decreases as circles move towards diagonal in **Figure 4A**) but beneficial for poorly calibrated groups (joint accuracy increases as triangles move towards diagonal in **Figure 4A**). To test this prediction empirically, we conducted Experiment 4 in which we directly manipulated group calibration, by pairing naïve participants with computer-generated (virtual) agents.



**Figure 4.** Confidence matching in stereotypical social scenarios. **(A)** Effect of confidence matching depends on group calibration in simulation. The symbols denote pairs of agents who are well-calibrated (circles) or poorly calibrated (triangles) in terms of their sensitivity and baseline confidence. The arrows indicate the movement through the landscape expected under confidence matching. Well-calibrated groups incur a cost, whereas poorly calibrated groups benefit. **(B)** Results of Experiment 4. The x-axis shows the degree of confidence matching; positive (negative) values indicate matching (anti-matching). The y-axis shows the change in joint accuracy compared to baseline: positive (negative) values indicate that the observed joint accuracy was higher (lower) than expected prior to interaction. The symbols denote group calibration as in panel A. Each dot is a group (4 dots per participant). Each line shows the simple effect of confidence matching (robust regression) for each type of group calibration. See **Supplementary Table 1** for test statistics. We found similar results using a mixed-effects model with a random intercept for each participant.

Participants (*N* = 38, tested in the UK) sat at private work stations in a computer lab (see **Methods**). Participants first performed the isolated task (EXP4-I: 240 trials), providing an estimate of their baseline confidence. They then performed the social task (EXP4-S: 4 x 240 trials) over four blocks. For each block, they were told that they were paired (anonymously) with one of the other participants. In reality, they were, for each block, paired with a virtual partner. We used our model to create the four virtual partners, tuning the parameters of each virtual partner to a participant’s baseline data. We varied the accuracy (low or high) and the mean confidence (low or high) of the virtual partners in a 2-by-2 within-subject design – creating two poorly calibrated and two well-calibrated groups per participant (see **Methods**).

To test our prediction, we first quantified the degree to which participants matched their partner’s mean confidence, , as: where serves as a normalising constant. The higher is, the higher the degree of confidence matching. We note that the values pooled across participants and conditions were positive (*t*(151) = 5.066, *p* < .001, one-sample, null: 0), providing further evidence for our hypothesis. We then estimated the difference between the observed joint accuracy and that expected prior to interaction as: . The latter value, , was estimated by playing out the responses of a given virtual partner against those recorded from a participant in the isolated task (see **Methods**). We then used (robust) multiple linear regression to test the effect of confidence matching (), group calibration (; poorly versus well-calibrated) and their interaction () on changes in joint accuracy (). In line with our prediction, the effect of confidence matching depended on group calibration (**Figure 4B**; : *t*(148) = 4.68, *p* < .001). Confidence matching had a negative effect for well-calibrated groups (blue line; : *t*(*148*) = -2.06, *p* = .041, simple effect), but a positive effect for poorly calibrated groups (red line; : *t*(148) = 4.42, *p* < .001, simple effect).

How do financial incentives affect confidence matching? In our experiments, participants may have been reluctant to break ‘status quo’ when this behaviour was not rewarded22. To test the robustness of confidence matching, we conducted Experiment 5 where participants (*N* = 20) responded on a probability scale and where the responses (individual versus joint) were submitted to a scoring rule23 (see **Methods**). Under this rule, participants would maximise their monetary earnings by reporting their confidence in a way that reflected their expertise – regardless of whether they performed the isolated task (EXP5-I: 160 trials) or the social task (EXP5-S: 160 trials). Participants sat at private work stations in a computer lab and were paired anonymously. Even then, and at the expense of their earnings (**Supplementary Figure 8**), they matched their confidence during group decision-making: the magnitudes of their mean confidence were correlated in the social condition only (**Figure 2A**), the differences in their mean confidence were smaller in the social than in the isolated condition (**Figure 2B**; *t*(9) = 2.158, *p* = .045, paired) and smaller than expected under shuffled data in the social condition only (isolated: *p* = .812; social: *p* = .071).

How general is confidence matching? One possibility is that confidence matching is not specific to group decision-making but that it can arise from mere exposure to others’ confidence24. To test the specificity of confidence matching, we conducted Experiment 6 (see **Methods**) where we compared the social task (EXP6-S: 160 trials) with a task where participants *observed* their partner’s response after having made their own response but where no joint decision was selected (EXP6-O: 160 trials). Participants sat at private work stations in a computer lab; they first performed the isolated task (EXP6-I: 160 trials) and performed the social and the observe tasks with a new (anonymous) partner. While the magnitudes of group members’ mean confidence were correlated in both the observe and the social condition (**Figure 2A**), the differences in their mean confidence were smaller in the social than in the observe condition (**Figure 2B**; *t*(22) = 2.100, *p* = .047, two-sample, using group members’ mean confidence from the isolated task to normalise the measures) and smaller than expected under shuffled data in the social condition only (observe: *p* = .171; social: *p* = .002). The results showed that confidence matching is most pronounced in group contexts where individual confidence has consequences for the group outcome.

How do people match their partner’s confidence? An obvious way is to keep a running average of their confidence, giving more weight to recent reports: if you think their confidence is higher than yours, you increase yours; if you think it is lower, you decrease yours. To formalise this intuition, we equipped our model with a learning rule that seeks to minimise the distance between the agent’s mean confidence and its estimate of the partner’s mean confidence (see **Methods**). Indeed, the model shows that confidence matching can arise from trial-by-trial dynamics (see convergence in mean confidence in **Figure 5B**, top panel). The model makes two predictions that can be compared against our data. First, a participant’s current confidence should be influenced by their partner’s recent confidence (**Figure 5A**). Second, while the magnitudes of their mean confidence may be variable, the difference in their mean confidence should be stable (**Figure 5B**, bottom panel). In this way, confidence matching may be thought of as a dance: the group members occupy different positions in confidence space but stay close to each other. In support of the model as a possible process account of confidence matching, both predictions were borne out in the data (**Figure 5C-D**; partner influence: *t*(81) = 6.504, *p* < .001, one-sample, null: 0; mean > distance: *t*(81) = 6.876, *p* < .001, paired). We note that interdependence in confidence on its own does not imply confidence matching; participants could increase or decrease their confidence at the same time without ever converging towards one another. However, the results show that the reverse does not hold: a simple trial-by-trial model of confidence matching implies interdependence in confidence.

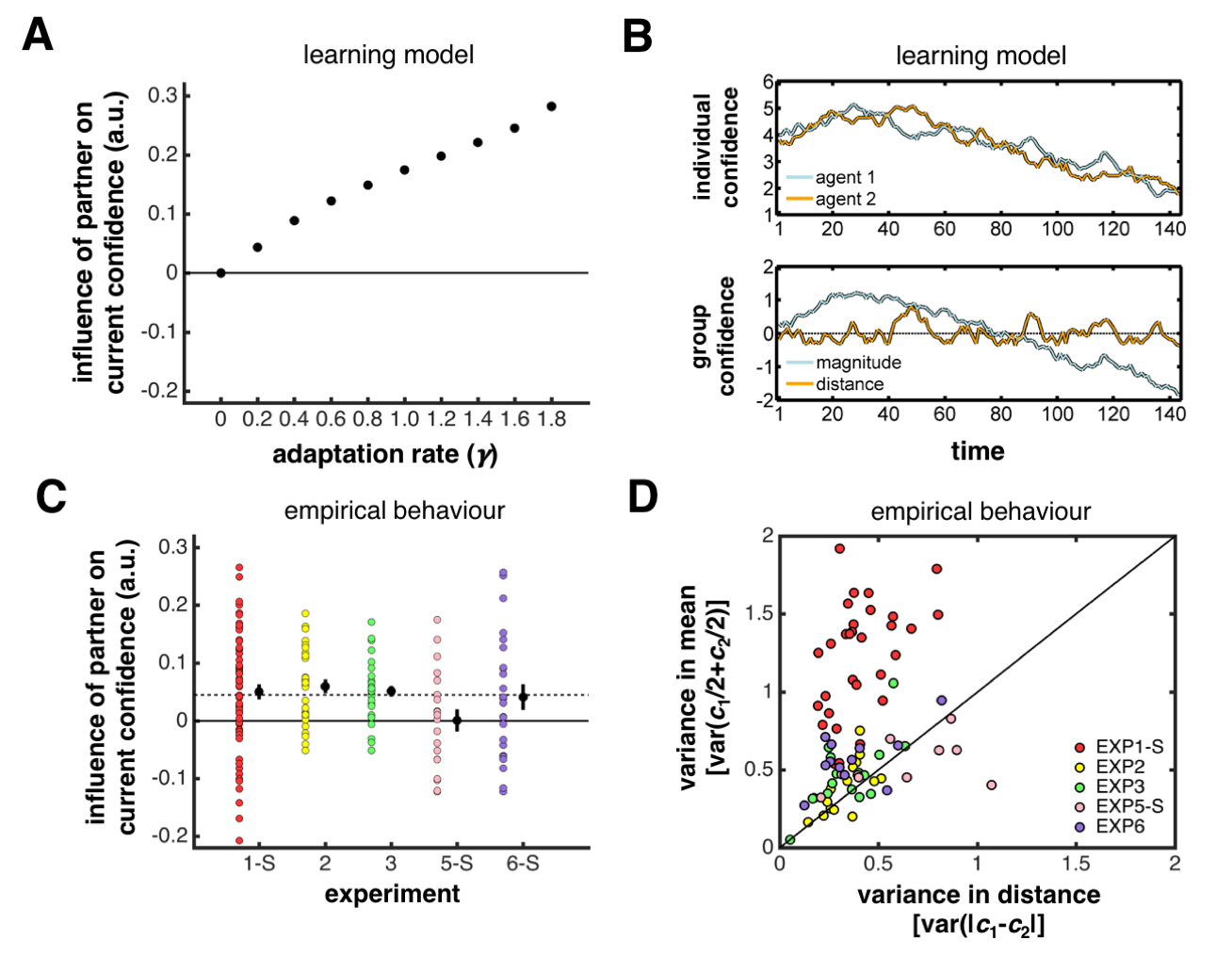


Figure 5. Trial-by-trial confidence matching

. **(A)** Model: interdependence in confidence. Coefficients (y-axis) from a multiple linear regression encoding the degree to which an agent’s confidence on trial is influenced by its partner’s confidence on trial . The agent uses prediction-error learning to estimate its partner’s mean confidence and uses this estimate to adapt its subjective mapping; the adaptation rate (; x-axis) describes the degree of adaptation. We included the stimulus ( and ) and the agent’s own confidence () as nuisance predictors. Each black dot is data averaged across 105 simulated experiments. Error bars are 1 SEM. **(B)** Model: dynamics of confidence. The model predicts (*top*) that adaptation () causes a convergence in mean confidence and (*bottom*) that the *magnitudes* of group members’ mean confidence, , is more variable than the *differences* in their mean confidence, . We created the time series using a sliding window, at each time point computing the measure of interest as the average from trial to trial *t* where (we used a fairly large for ease of illustration). The plots show data from one simulated experiment where the agents’ adaptation rate was set to . **(C)** Behaviour: interdependence in confidence. Same analysis as in panel (A). We note that the interdependence extends several trials back in time (**Supplementary Figure 9A**)andthat the degree to which participants influenced each other was correlated (**Supplementary Figure 9B**). The dashed line indicates the average coefficient observed across experiments and participants. Each black dot is data averaged across participants. Each coloured dot is a participant. Error bars are 1 SEM. **(D)** Behaviour: dynamics of confidence. The variance of the magnitudes of group members’ mean confidence, , is higher than the variance of their differences in mean confidence, var. Here and were created using a sliding window where . Importantly, this analysis is biased against the model prediction: when and are created by sampling uniformly from the range 1 to 6, then the expected variance is higher for var than for . Each dot is a dyad.

In summary, confidence matching may be a good strategy for reducing miscommunication among individuals. First, confidence matching is computationally inexpensive. People do not have to infer latent states or functions but only have to track observable behaviours. Second, the strategy fares best when people have similar levels of expertise. Fortunately, that is often the case, as we tend to associate with friends, partners or colleagues with whom we are likely to share traits25. Lastly, even when people differ in expertise, the strategy helps when the less competent is the more confident. In such cases, confidence matching prompts people to report their confidence in a way that better reflects their relative levels of expertise. This “equilibrium” does not require that they have insight into their own or others’ expertise; an insight that cannot be taken for granted22,26.

Our findings have implications for theories of confidence. At the single-trial level, variation in confidence for a constant stimulus is usually assumed to reflect noise – either in the encoding of the sensory evidence or in the read-out of some internal estimate of the evidence strength for report3,27,28. Our findings show, however, that this variation can be systematic, here driven by the recent history of social interaction (in **Supplementary Figure 9C** we show how such history effects can be confused with noisy read-out). At the aggregate level, under- and overconfidence has traditionally been attributed to limitations on the way in which the human mind represents and processes uncertainty29. Our findings, however, raise the possibility that these biases are at least in part of a social nature – reflecting social norms moulded through patterns of social interaction30.

Our study provides an important anchor point for future research. In our task, optimality is well-defined: group members must report their confidence in a way that maximises the probability that the group makes the correct decision. There are, however, different possible explanations of the observed departure from optimality. The one that we have favoured is that optimality is hard for group members to achieve – it requires that they simultaneously adjust their subjective mappings to a single point in a high-dimensional space – and that they therefore use a heuristic strategy – a strategy which sometimes takes them *close* to optimality. An alternative explanation is that group members had a different definition of optimality in mind: they may have sought to report their confidence in a way that maximises *equal* influence on the group decision22, perhaps to avoid confrontations31 or to diffuse responsibility for difficult decisions32. These hypotheses can be tested by changing the decision weights assigned to group members (e.g., such that one has to report higher confidence to ensure equal influence) and having asymmetric payoffs (e.g., such that taking responsibility for difficult decisions is highly rewarded).

We usually assume that “speaking the same language” facilitates successful communication. Our study highlights, however, that we might at times be intending very different meanings despite using a shared set of expressions. This misalignment may arise whenever communication involves an arbitrary mapping between intended meanings and their expression. Our study is relevant to many high-risk settings. For example, communication about the risks of climate change33,34 and the uncertainty associated with financial and geopolitical forecasting35,36 can easily go wrong when people do not communicate information according to a common metric.

**METHODS**

*Participants*

Participants (aged 18-40) were recruited from participant pools at the University of Tehran (EXP1: *N* = 60, all male) and the University of Oxford (EXP2: *N* = 30, all male; EXP3: *N* = 30, all male; EXP4: *N* = 38, 19 females; EXP5: N = 20, 13 females; EXP6: *N* = 24, 14 females). For Experiments 1 to 3, participants were recruited in pairs (i.e., they knew each other beforehand). For Experiments 4 to 6, participants were recruited individually. Experiment 4 involved deception; participants were debriefed after the study, with no one having noticed the deception or deciding to leave the study. All participants reported normal or corrected-to-normal vision. All participants provided informed consent and were reimbursed for their participation; in Experiment 5, participants could earn an additional performance-based bonus. The experiments were approved by the Ethics Committee at the Faculty of Electric Engineering, University of Tehran, and the Central University Research Ethics Committee, University of Oxford.

*Task*

All experiments were based on the same task (two-interval forced-choice contrast discrimination; **Figure 1D**). On each trial, participants were presented with two consecutive viewing displays, each containing six vertically oriented Gabor patches. In one of the two displays, the contrast level of one of the six Gabor patches (the target) was increased by adding one of four values (.015, .035, .07, .15) to its baseline contrast (.10). After the two displays, participants were presented with a horizontal line bisected at its midpoint. A vertical marker was placed on top of the midpoint. The marker could be moved along the line by up to six steps on either side of the midpoint; the left-side steps were negative values (-6 to -1), whereas the right-side steps were positive values (1 to 6). The sign of the response indicated the decision (negative: 1st; positive: 2nd), and its absolute value indicated the confidence (1: “unsure”; 6: “certain”). We use *response* and *confidence* to refer to signed and unsigned values, respectively. In a *social* version of the task (social task), participants performed the visual task as part of a pair. The individual responses were first made public. The individual decision made with higher confidence was then automatically selected as the joint decision. In the case of a confidence tie (different displays but same confidence), one of the two individual decisions was randomly selected. Participants received feedback about the accuracy of each decision before automatically continuing to the next trial. Participants were instructed to make as many correct joint decisions as possible. In an *isolated* version of the task (isolated task), participants performed the visual task without any social interaction. In an intermediate version of the task (observe task), individual responses were made public but no joint decision was selected. Each participant had their own display monitor and response device. The stimulus has been described in detail elsewhere37. Experiments were run using Cogent 2000 (<http://www.vislab.ucl.ac.uk/cogent.php/>) for MATLAB.

*Procedure*

In Experiment 1,pairs of participants performed the social and the isolated task. The order of the two tasks was counterbalanced. There were 320 trials, divided into two blocks (social: 160 trials; isolated: 160 trials). In Experiment 2, pairs of participants performed the social task only. There were 384 trials, divided into three blocks. In Experiment 3, pairs of participants performed the social task only. In contrast to the other experiments, confidence was indicated on a continuous scale. There were 384 trials, divided into three blocks. In Experiment 4, participants sat at private work stations in a computer lab. The experiment consisted of two sessions. In the first session, participants performed the isolated task. In the second session, participants performed the social task over four blocks. For each block, they were told that they were paired anew with one of the other participants present in the room. In reality, they were, for each block, paired with a computer-generated agent; each agent was tuned to the participant to reflect a 2 x 2 within-subject design. The order of the four conditions (agents) was counterbalanced across participants. There were 1160 trials, divided into five blocks (isolated: 200 trials; social: 4 x 240 trials). In Experiment 5, participants sat at private work stations in a computer lab. They performed first the isolated task and then the social task. In contrast to the other experiments, responses were made on a probability scale and submitted to a strictly proper scoring rule. We used a variant of the Brier score23 where participants on each trial accrued rewards as a function of the accuracy of their decision and their confidence: ) where indicates the decision accuracy (0: incorrect; 1: correct) and indicates the chosen probability. In the isolated task, the individual response was submitted to the scoring rule. In the social task, the joint response was submitted to the scoring rule. Participants were paid the sum of their average trial-by-trial earnings in each task. There were 320 trials, divided into two blocks (isolated: 160 trials; joint: 160 trials). In Experiment 6, participants sat at private work stations in a computer lab. They first performed first isolated task and then the social and the observe task, each time paired anew with another participant. The order of the social and the observe tasks was counterbalanced across participants. There were 480 trials, divided into three blocks (isolated: 160 trials; observe: 160 trials; joint: 160 trials).

*Statistical tests*

Sample-sizes were chosen based on earlier studies37. To test for dynamic convergence – that is, between-group differences in the convergence point – we complemented our standard parametric tests with a permutation-based approach. Our general procedure was to create for each measure of interest, , a distribution under the null hypothesis,, by randomly re-pairing group members and re-computing the measure of interest for each set of re-paired group members (106 sets) – with the null hypothesis being that there is a fixed convergence point across groups. To test whether we could rule out the null hypothesis, we would ask whether the observed (average) value for a given measure of interest (e.g., differences in mean confidence) was smaller than 95% of the (average) values from its respective null distribution (i.e., *p* < .05, one-tailed). We show the null distributions for statistical inference in **Supplementary Figure 3.**

*Computational model*

We developed a simple model (i) to intuit how joint accuracy (fraction of correct joint decisions) varies with differences in expertise and mean confidence (**Figure 3**) and (ii) to establish an optimal benchmark against which empirical group performance could be compared (**Supplementary Figure 8**). On each trial, an agent receives noisy sensory evidence, *x*, sampled from a Gaussian distribution, , whose mean, , is given by the stimulus, and whose standard deviation, , specifies the level of sensory noise. As in our task, is drawn uniformly from the set, . The sign of indicates the target display (negative: 1st; positive: 2nd) and its absolute value indicates the contrast added to the target. The agent uses the raw sensory evidence as its internal estimate of the evidence strength, . The internal estimate thus ran from large negative values, indicating a high probability that the target was in the first display, through values near 0, indicating high uncertainty, to large positive values, indicating a high probability that the target was in the second display. We chose this formulation for mathematical simplicity but note that our analyses would show the same results for any model in which the internal estimate is a monotonic function of the sensory evidence, including probabilistic estimates such 4. The agent maps the internal estimate onto a response, , by applying a set of thresholds, . The position of the thresholds in *z*-space determines the proportion of times that each response is made. As in our task, the sign of the response indicates the decision (negative: 1st; positive: 2nd), and its absolute value indicates the confidence. Our general approach was to set the thresholds in *z*-space so as to generate a specified distribution over responses (e.g., 5% of “-6”, 2% of “-5”, 10% of “-4”, and so on)3,4 – using a pre-specified set of distributions for simulations (maximum entropy distributions) and the distributions observed in our experiments for model fitting. Note that, for different levels of sensory noise, different thresholds must be used to generate the same distribution. The level of sensory noise determines the agent’s expertise and the set of thresholds determines the agent’s mean confidence. See **Supplementary Methods** for model details.

*Confidence landscapes*

We used our model to create heat maps which show how joint accuracy (fraction of correct joint decisions) varies as a function of the mean confidence of a given pair of agents (**Figure 3**). For each pair of agents, we first specified their respective levels of sensory noise, and . We then derived their joint accuracy under different pairs of confidence distributions, each associated with a specific mean. While there are obviously many distributions that can generate a given mean, this is not the case when considering one family of distributions. We therefore limited our analyses to maximum entropy distributions, running from mean 1 to 6 in steps of .1 (see **Supplementary Figure 4**). Before deriving joint accuracy, we transformed each confidence distribution (1 to 6) to a response distribution (-6 to -1 and 1 to 6) by assuming symmetry around 0 – this transformation was needed to place the thresholds in *z*-space and generate both decisions and confidence. See **Supplementary Methods** for details about derivation of individual and joint accuracy and how the maximum entropy distributions were created.

*Control analysis: observed compared to optimal joint accuracy*

We used our model to estimate how far each group in our experiments was from reaching optimal performance (**Supplementary Figure 8**). We first fitted our model to the data of each participant by searching for the sensory noise ( ranged from .001 to 1 in steps of .001) that minimised the squared error between the observed accuracy (fraction of correct individual decisions) and that derived from the model. For each step of the search, we set the thresholds in *z*-space so as to generate the participant’s response distribution observed across stimuli and then derived their accuracy. Our model thus has only one free parameter (sensory noise) as the thresholds are determined by a participant’s observed response distribution. Despite having only one free parameter, our model provided a good fit to the individual data. We show empirical and model psychometric functions and response distributions for *each* stimulus in **Supplementary Figures 5-6**; especially the latter fits are reassuring as we fitted the model using a participant’s response distributions observed *across* stimuli. We next computed a confidence landscape for each pair of participants using their fitted levels of sensory noise (see details above) and used it to identify the joint accuracy expected under the optimal solution (maximum value in a landscape; a landscape for each group is shown in **Supplementary Figure 7**). Although the more accurate group members were on average more confident than their less accurate partners (more dots above horizontal line in **Supplementary Figure 8A**; *t*(81) = 4.336, *p* < .001, one-sample, null: 0), their differences in mean confidence did not scale with their differences in accuracy as we would expect under the optimal solution and were too small for group members with very different levels of accuracy (**Supplementary Figure 8A**, *r*(80) = .090, *p* = .421, Pearson). Consistent with our hypothesis, the ratio of the observed joint accuracy to that expected under the optimal solution was smaller for group members with very different levels of accuracy (**Supplementary Figure 8B**; *r*(80) = .682, *p* < .001, Pearson). See the **Supplementary Text** for details about this comparison.

*Virtual partners*

We used our model to create the virtual partners in Experiment 4. We varied their mean accuracy (low or high) and their mean confidence (low or high) in a 2-by-2 within-subject design. We first fitted our model to each participant’s data from the isolated task; this was done while they were waiting to start the social task. We used the fitted sensory noise () to specify the mean accuracy of the virtual partners: sensory noise was 50% higher than the fitted noise for the low-accuracy partners and 50% lower than the fitted noise for the high-accuracy partners. We used two custom confidence distributions to specify the confidence of the virtual partners: the mean confidence was about 2.2 for the low-confidence partners and about 4.2 for the high-confidence partners. We transformed the confidence distributions (1 to 6) to response distributions (-6 to -1 and 1 to 6) by assuming symmetry around 0. To generate the trial-by-trial responses of a given partner, we first created the trial-by-trial sequence of stimuli to be shown to the participant. We then created a trial-by-trial sequence of random values (sensory evidence), with each drawn from a Gaussian distribution whose mean was given by the stimulus on the corresponding trial and whose standard deviation was given by the level of sensory noise. Next, we transformed the sequence of random values into trial-by-trial responses by applying – post-hoc – a set of thresholds that gave rise to the specified response distribution. To mimic lapses of attention and response errors, we then randomly selected a response (from a uniform distribution over 1 to 6) on 5% of the trials (12 out 240 trials). In addition, we varied the agents’ reaction time (randomly sampled from a uniform distribution over 2 to 5 seconds), so that participants had to wait for their partner on some of the trials. Statistical tests showed that we obtained the 2-by-2 differences between participants and virtual partners in terms of mean accuracy and mean confidence (see **Supplementary Table 2** for test statistics).

*Joint accuracy expected prior to interaction*

For experiment 4 we computed the joint accuracy expected prior to interaction, , by playing out responses of a given virtual partner for whom we knew the generative model against those of the participant in the isolated task – with joint decisions selected as in the social task. This procedure allowed us to test whether the observed joint accuracy, , was higher or lower than expected prior to interaction, . To control for fluctuations in individual performance between conditions, we normalised each measure by the participant’s and the partner’s accuracy: and . We used the normalised values to compute the change in joint accuracy: . We estimated across 104 iterations as the partner’s responses varied between iterations due to sensory noise and the 5% lapse rate.

*Questionnaires*

In Experiment 4, participants completed a questionnaire about their partner in each social block. They were asked to indicate: (1) whether they thought the partner was male or female; (2) how much they liked the partner; (3) how well they performed as a group; (4) whether the partner was more accurate than they were; and (5) whether the partner was more confident than they were. Interestingly, participants displayed the stereotype that females (males) are less (more) confident and they liked the high-accuracy but low-confidence partners the most (see **Supplementary Table 3** for average responses).

*Learning model*

We furnished our model with a simple learning rule to provide a process account of how confidence matching arises (**Figure 5**). The agent updates on each trial its mean confidence, , as a mixture of its own mean confidence and its estimate of the partner’s mean confidence, , as follows: where describes the rate of adaptation and describes the mismatch between the agent’s mean confidence and its estimate of the partner’s mean confidence. The agent updates on each trial its estimate of the partner’s mean confidence as follows: ), where describes the rate of learning,is the partner’s confidence on trial and ) is a prediction error. The initial values of and reflect the agent’s baseline mean confidence and its expectation for the partner’s baseline mean confidence. The agent uses to update the function, , which governs the mapping from the agent’s internal estimate of the evidence strength onto a response, . In our simulations, we assumed that that a pair of agents had the same levels of sensory noise (); that their mapping functions were updated so as to maintain maximum entropy over confidence (i.e., we set the thresholds in *z*-space using a set of maximum entropy distributions running from mean 1 to 6 in steps of .001); and that the learning rate was fixed () for both agents. In each simulated experiment, the agents performed 160 trials, with stimuli drawn as in our task. To control for random variation due to sensory noise, we averaged across 105 simulated experiments. The agents’ baseline mean confidence and its expectation for the partner’s baseline mean confidence were for each simulated experiment sampled uniformly from the range 2 to 5.

*Code availability*

Analyses and simulations were conducted in MATLAB (2015b). The code is available upon request.

**Acknowledgements**

This work was supported by the Calleva Research Centre for Evolution and Human Sciences, Magdalen College (DB and JL); the Gatsby Charitable Foundation (LA and PEL); the DAAD (AM); the Wellcome Trust (SHC: 099741/Z/12/Z); the European Research Council (BB: 309865-NeuroCoDec; CS: 281628-URGENCY).

**REFERENCES**

1. Laughlin, P. R. & Ellis, A. L. Demonstrability and social combination processes on mathematical intellective tasks. *J. Exp. Soc. Psychol.* **22,** 177–189 (1986).

2. Zarnoth, P. & Sniezek, J. A. The social influence of confidence in group decision making. *J. Exp. Soc. Psychol.* **33,** 345–366 (1997).

3. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90,** 499–506 (2016).

4. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLOS Comput. Biol.* **11,** e1004519 (2015).

5. Ais, J., Zylberberg, A., Barttfeld, P. & Sigman, M. Individual consistency in the accuracy and distribution of confidence judgments. *Cognition* **146,** 377–386 (2016).

6. Fleming, S. M. & Dolan, R. J. Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Conscious. Cogn.* **19,** 352–363 (2010).

7. Broihanne, M. H., Merli, M. & Roger, P. Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Financ. Res. Lett.* **11,** 64–73 (2014).

8. Mann, L. *et al.* Cross-cultural differences in self-reported decision-making style and confidence. *Int. J. Psychol.* **33,** 325–335 (1998).

9. Niederle, M. & Vesterlund, L. Gender and Competition. *Annu. Rev. Econom.* **3,** 601–630 (2011).

10. Bahrami, B. *et al.* Optimally interacting minds. *Science* **329,** 1081–1085 (2010).

11. Bang, D. *et al.* Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Conscious. Cogn.* **26,** 13–23 (2014).

12. Koriat, A. When are two heads better than one and why? *Science* **336,** 360–362 (2012).

13. Devaine, M., Hollard, G. & Daunizeau, J. The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? *PLoS Comput. Biol.* **10,** (2014).

14. Yoshida, W., Seymour, B., Friston, K. J. & Dolan, R. J. Neural mechanisms of nelief inference during cooperative games. *J. Neurosci.* **30,** 10744–10751 (2010).

15. Pickering, M. J. & Garrod, S. Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* **27,** 169-190-226 (2004).

16. Friston, K. J. & Frith, C. D. Active inference, communication and hermeneutics. *Cortex* **68,** 129–143 (2015).

17. Friston, K. J. & Frith, C. D. A duet for one. *Conscious. Cogn.* **36,** 390–405 (2015).

18. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8,** (2014).

19. Jackson, J. M. & Harkins, S. G. Equity in effort: an explanation of the social loafing effect. *J. Pers. Soc. Psychol.* **49,** 1199–1206 (1985).

20. Schelling, T. C. *The Strategy of Conflict*. (Harvard University Press, 1980).

21. Mahmoodi, A., Bang, D., Ahmadabadi, M. N. & Bahrami, B. Learning to make collective decisions: the impact of confidence escalation. *PLoS One* **8,** e81195 (2013).

22. Mahmoodi, A. *et al.* Equality bias impairs collective decision-making across cultures. *Proc. Natl. Acad. Sci.* **112,** 3835–3840 (2015).

23. Brier, G. W. Verification of forecasts expressed in terms probability. *Mon. Weather Rev.* **78,** 1–3 (1950).

24. Sherif, M. A study of some social factors in perception. *Arch. Psychol.* **187,** 5–61 (1935).

25. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27,** 415–444 (2001).

26. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329,** 1541–1543 (2010).

27. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21,** 422–430 (2012).

28. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16,** 105–110 (2012).

29. Harvey, N. Confidence in judgment. *Trends Cogn. Sci.* **1,** 78–82 (1997).

30. Shea, N. *et al.* Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18,** 186–193 (2014).

31. Bazerman, M. H., Curhan, J. R., Moore, D. A. & Valley, K. L. Negotiation. *Annu. Rev. Psychol.* **51,** 279–314 (2000).

32. Forsyth, D. R., Zyzniewski, L. E. & Giammanco, C. a. Responsibility diffusion in cooperative collectives. *Personal. Soc. Psychol. Bull.* **28,** 54–65 (2002).

33. Taylor, A. L., Dessai, S. & de Bruin, W. B. Communicating uncertainty in seasonal and interannual climate forecasts in Europe. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **373,** 20140454 (2015).

34. Budescu, D. V., Broomell, S. & Por, H.-H. Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychol. Sci.* **20,** 299–308 (2009).

35. Morreau, M. Grading in groups. *Econ. Philos.* **32,** 323–352 (2016).

36. Kent, S. *Sherman Kent and the Board of National Estimates: Collected Essays*. (History Staff, Center for the Study of Intelligence, Central Intelligence Agency, 1994).

37. Bahrami, B. *et al.* What failure in collective decision-making tells us about metacognition. *Philos. Trans. R. Soc. B Biol. Sci.* **367,** 1350–1365 (2012).