TITLE

Confidence matching in group decision-making

AUTHORS

Dan Bang[1,2,3,4,*], Laurence Aitchison[5,6], Rani Moran[4,7,8], Santiago Herce Castañón[1], Banafshe Rafiee[9], Ali Mahmoodi[10], Jennifer Y. F. Lau[1,2,11], Peter E. Latham[6], Bahador Bahrami[12], and Christopher Summerfield[1]

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] Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom.
[6] Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom.
[7] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, United Kingdom.
[8] School of Psychological Sciences, Tel Aviv University, Ramat Aviv 69978, Israel.
[9] Department of Computing Science, University of Alberta, AB T6G 2S4, Canada.
[10] Bernstein Center Freiburg, University of Freiburg, 79104 Freiburg, Germany.
[11] Department of Psychology, King's College London, London SE5 8AF, United Kingdom.
[12] Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, United Kingdom.
[*] Email: danbang.db@gmail.com.

Most important decisions in our society are made by groups, from cabinets and commissions to boards and juries. 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], it can also 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]. However, solving this communication problem is 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 communicated confidence so that certainty and uncertainty is stated in approximately equal measure by each party. Combining the behavioural data with computational modelling, we show that this strategy is effective when group members have similar levels of expertise, and that it is robust when group members 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 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 handball referees who disagree about whether the ball crossed the goal line. Each referee states their opinion with a certain degree of confidence (**Figure 1A**, y-axis). This degree of confidence depends on the probability that their individual opinion is correct (**Figure 1A**, x-axis). The referees have, however, different subjective mappings (i.e., different functions mapping probability correct to confidence as indicated by the solid lines), with the blue referee biased towards higher confidence. Consequently, the group decision is, in this interaction, dominated by the blue referee, who is in fact less likely to be correct (dotted lines; their probability correct may vary because of differences in expertise or proximity to the incident). To avoid such miscommunication, the referees must align their subjective mappings so that their confidence is stated in a mutually consistent manner (**Figure 1B**).

It is, however, computationally difficult to reach the optimal solution. Without any prior interaction, the referees can only make guesses about their colleague's subjective mapping. But even with prior experience working together, estimating and adjusting to their colleague's subjective mapping is challenging, especially because the function mapping probability correct to confidence is not static but being adjusted in return[13,14]. Here we tested the hypothesis that people instead 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 vocabulary[15], and it has been proposed that mimicry can reduce miscommunication, by aligning agents' input-output functions[16,17].

We ran six behavioural experiments to test our hypothesis. In Experiment 1 (see **Methods**), pairs of participants (30 groups, tested in Iran) performed a psychophysical task (**Figure 1D**). 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 scale from 1 to 6. In the *social* condition (**Figure 1E**; EXP1-S: 160 trials, social task), participants performed the task together. Once both private responses had been registered, they were made public, and the private decision made with higher confidence was selected as the joint decision. Under this decision rule, the optimal strategy is to report 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, without seeing each other's responses.

Under our hypothesis, we would expect group members' confidence to be more similar when they performed the task together than when alone. Here we focus on convergence in mean confidence, but we found similar results for confidence variability and confidence distributions (**Supplementary Figure 1**). In line with our hypothesis, group members' mean confidence was correlated in the social condition only (**Figure 2A**, EXP1-I and EXP1-S) and the difference in their mean confidence was smaller in the social than in the isolated condition (**Figure 2B**, EXP1-I and EXP1-S; $t(29) = 4.195$, $p < .001$, paired) – regardless of the condition order (**Supplementary Figure 2**). We could rule out that group members simply converged onto a single value; for example, they may systematically have gravitated towards medium confidence to minimise conflict[18], or towards maximum confidence to dominate the joint decision[19]. As can be seen from the variability of data points *along* the diagonal in **Figure 2A** (EXP1-S), the convergence point varied considerably across groups.

We could also rule out that the convergence in mean confidence was driven by an underlying convergence in accuracy (fraction of correct individual decisions): in contrast to the results observed for mean confidence, the differences in accuracy were larger in the social than in the isolated condition ($t(29) = 2.083$, $p = .046$, paired). Overall, group members' difference in mean confidence did not scale with their difference in accuracy (**Figure 2C**). We found similarly-sized social effects (**Figure 2**) 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). Overall, the results were in line with our hypothesis that people *actively* match their confidence during group decision-making – here 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).

Confidence matching should also have testable consequences for group performance. Intuitively, the strategy seems sensible 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**, right panel). 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 model[4] to simulate how joint accuracy (fraction of correct joint decisions) varies with differences in expertise and mean confidence (see **Methods**). **Figure 3B** shows landscapes of expected joint accuracy as a function of the mean confidence of *simulated* group members with equal expertise (left panel) and unequal expertise (right panel, here member 2 has higher expertise).

In each landscape, confidence matching corresponds to the diagonal. For group members of equal expertise, confidence matching improves joint accuracy (black dot is on diagonal in **Figure 3B**, left panel). However, when one group member is more of an expert than the other, confidence matching reduces joint accuracy compared to the optimal solution (black dot is off diagonal in **Figure 3B**, right panel). Consistent with this predicted pattern of results, we observed empirically that dissimilar group members were the furthest from reaching an optimal level of group performance (**Figure 3D;** here including data from EXP5-S and EXP6-S). The results show that confidence matching may be one cause of the common finding that group performance depends on the similarity of group members' expertise[10–12].

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 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? As can be

seen in the right panel of **Figure 3B**, points above the diagonal are associated with *higher* values than those along the diagonal, whereas points below the diagonal are associated with *lower* values than those along the diagonal. In other words, confidence matching – that is, moving towards the diagonal – should be *costly* for well-calibrated groups but *beneficial* for poorly calibrated groups. To test this prediction, we conducted Experiment 4 (see **Methods**) in which we manipulated group calibration, by pairing naïve participants with computer-generated partners.

Participants (*N* = 38, tested in the UK) performed the isolated task (EXP4-I: 240 trials), providing an estimate of their baseline confidence, and then performed the social task (EXP4-S: 4 x 240 trials) over four blocks. For each block, they were paired with a simulated partner, but told that they were paired anonymously with another participant. We varied the accuracy (low or high) and the mean confidence (low or high) of the four partners – creating two poorly calibrated and two well-calibrated groups per participant (see **Methods**). The results were in line with our prediction. First, consistent with confidence matching, the difference in mean confidence was smaller in the social blocks than prior to interaction (**Supplementary Figure 8**; $|c_{\text{participant}}^{\text{social}} - c_{\text{partner}}| < |c_{\text{participant}}^{\text{isolated}} - c_{\text{partner}}|$: $t(151) = -5.066$, $p < .001$, paired). Second, the extent to which joint accuracy was higher than expected prior to interaction depended on initial group calibration: it was higher than expected for poorly calibrated groups but lower than expected for well-calibrated groups (**Figure 3D**).

The reason confidence matching is sub-optimal is that group members may end up using the same confidence to indicate different values of probability correct (**Figure 3A**, right panel). A pressing question is whether confidence matching is robust to financial incentives for reporting confidence in an *objectively* accurate manner. In Experiment 5 (see **Methods**), participants (*N* = 20) responded on a probability scale and were rewarded according to a proper scoring rule[20] – in the isolated task (EXP5-I: 160 trials) and in the social task (EXP5-S: 160 trials). Under this scoring rule, participants would maximise their earnings by indicating "70%" when they believed that they had a 70% probability of being correct and so forth. Participants still matched their confidence: their mean confidence was correlated in the social condition only (**Figure 2A**) and the difference in their mean confidence was smaller in the social than in the isolated condition (**Figure 2B**; $t(9) = 2.158$, $p = .045$, paired). The presence of confidence matching – and hereby the absence of a relationship between relative confidence and relative expertise (**Figure 2C**) – was very surprising as participants interacted anonymously and thus were not under any social pressure to conform.

We have presented confidence matching as an *active strategy* for negotiating individual influence on group decisions. An obvious test of this hypothesis is to see whether confidence matching can be found in the absence of group decisions. In Experiment 6 (see **Methods**), we compared the social task with a task where participants *observed* their partner's response after having made their own response but where no joint decision was selected. Participants (*N* = 24) performed the isolated task (EXP6-I: 160 trials), and then the social task (EXP6-S: 160 trials) and the observe task (EXP6-O: 160 trials), each time paired anew with another anonymous partner. While group members' mean confidence was correlated in both the observe and the social condition (**Figure 2A**), the strength of this relationship was stronger in the social condition and the difference in their mean confidence was smaller in the social than in the observe condition (**Figure 2B**; $t(22) = 2.100$, $p = .047$, two-sample, using the data from the isolated task to normalise the data from the social and the observe tasks). The results indicate that confidence matching reflects a mixture of 'context-general' behavioural imitation and 'context-specific' strategic thinking.

We have in six independent experiments provided *aggregate* evidence for confidence matching: we have shown, at the individual level, that group members' mean confidence is more similar during

group decision-making than in control conditions, and, at the group level, that group performance follows the pattern expected under confidence matching rather than the optimal solution. Finally, we considered whether confidence matching also can be observed at short time scales. An obvious way to match your partner's confidence is to keep a running estimate of their confidence and then adapt your own accordingly: 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 built a temporal difference learning model[21] which sought to minimise the distance between its own mean confidence and its estimate of the partner's mean confidence on the basis of recent trials (see **Methods**). The model, which can account for convergence in mean confidence (**Figure 4A**), makes predictions about the trial-by-trial data. In particular, a participant's current confidence should depend on their partner's recent confidence (**Figure 4B**) – a pattern which we observed empircally, extending three trials back into time (**Figure 4C**). The results show that confidence matching happens at short time scales and suggest that short-range temporal dependencies may underlie the observed aggregate results.

In conclusion, confidence matching may be a sensible strategy for group decision-making. First, the strategy is computationally inexpensive. People do not need to infer latent states or functions but only need 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 traits[22]. 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. The resulting "equilibrium" may not be perfect but it does not require that people have insight into their own or others' expertise; an insight that cannot be taken for granted[23,24].

Our study has 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 probability correct for report[3,25,26]. Our results show that this variation can also be systematic, here driven by the recent history of social interaction (in **Supplementary Figure 9C** we show how history effects can be confused with noisy read-out as inferred from standard measures of metacognition[25,27]). At the aggregate level, under- and overconfidence is often been attributed to limitations on the way in which the human mind represents and processes uncertainty[28]. Our results raise the intriguing possibility that these biases are at least in part of a social nature – reflecting social norms or social strategies[29]. We have argued that the observed social effects operate at the level of report – the function mapping probability correct to confidence – but it remains to be seen whether social interaction also can change the internal estimate of probability correct.

We suggested that it was too difficult for group members to find the optimal strategy in our task and that they therefore used a heuristic one. An alternative explanation is that group members had a different objective in mind: for example, they may have tried to maintain *equal* influence on the group decision[24], perhaps to avoid conflict[30] or to diffuse responsibility [31]. These social hypotheses can be tested by changing the decision weights assigned to group members (e.g., such that one must report higher confidence to maintain equal influence) and/or by introducing asymmetric payoffs (e.g., such that taking responsibility for difficult decisions is highly rewarded).

We usually assume that "speaking the same language" facilitates effective communication. We have shown, in the case of confidence, that this perceived wisdom is typically true when individuals with similar levels of expertise compare their opinions. However, we have also shown that, without the right precautions, "speaking the same language" can be detrimental when comparing the opinions of individuals with different levels of expertise. This finding is relevant to contemporary debates

concerning topics from climate change[32,33] to economic and geopolitical forecasting[34,35] and the value of expert opinions in public debates.

METHODS

Participants

Sample-sizes were chosen based on earlier studies[36]. 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). In Experiments 1 to 3, participants were recruited and took part in pairs; they knew each other beforehand. In Experiments 4 to 6, participants were recruited individually and took part as part of a large group. Experiment 4 involved deception; participants were debriefed afterwards, 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; 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 the 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. We used three versions of this task in our experiments. In an *isolated* version of the task (isolated task), participants performed the task on their own. After having made their response, participants received feedback about the accuracy of their decision and continued to the next trial. In a *social* version of the task (social task), participants performed the visual task as part of a pair. Once both individual responses had been registered, they were made public and the individual decision made with higher confidence was automatically selected as the joint decision. In the case of a confidence tie (i.e., different decisions but same confidence), one of the two individual decisions were randomly selected. Participants received feedback about the accuracy of both the individual decisions and the joint decision before continuing to the next trial. Participants were instructed to make as many correct joint decisions as possible. In an intermediate version of the task (observe task), the individual responses were made public but no joint decision was selected. Participants received feedback about the accuracy of the individual decisions before continuing to the next trial. Each participant had their own display monitor and response device. The stimulus has been described in detail elsewhere[36]. Experiments were implemented using the Cogent 2000 toolbox (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-by-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 score[20] where participants on each trial accrued rewards as a function of their decision accuracy and their confidence: $£5 * (1 - (correct - confidence)^2)$ where $correct$ indicates the decision accuracy (0: incorrect; 1: correct) and $confidence$ indicates the chosen probability. Participants were paid the sum of their average trial-by-trial earnings in the isolated and in the social 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 the 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

For the robust regression analyses shown in **Figure 2A** and **Supplementary Figure 1A**, the labelling of group members as 1 and 2 was arbitrary. We therefore repeated the analysis 10^5 times, each time randomly re-labelling the group members as 1 and 2. The displayed *p*-value shows the average *p*-value for the slope of the best-fitting line across these regressions. We complemented the standard parametric tests in the main text with a permutation-based approach. Our general approach was to create for each measure of interest, ϑ, a distribution under the null hypothesis, $p(\vartheta)$, by randomly re-pairing group members and re-computing the measure of interest for each set of re-paired group members (10^6 sets). Here the null hypothesis is that the observed value (e.g., average difference in mean confidence in an experiment) is not specific to the *true* pairing of group members. In contrast, under our hypothesis, we would expect the observed value to be specific to the true pairing of group members: it is the result of dynamic interaction between group members and shuffling the data breaks this relationship. To test whether we could rule out the null hypothesis, we asked whether the observed value was smaller than 95% of the values from its corresponding null distribution (i.e., $p < .05$, one-tailed). All permutation tests were consistent with the results reported in the main text: the observed values were only specific to the true pairing of group members in the social task. We show all null distributions in **Supplementary Figure 3.**

Computational model

We developed a simple model (i) to unpack 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 7**). On each trial, an agent receives noisy sensory evidence, x, sampled from a Gaussian distribution,

$x \in N(s, \sigma^2)$, whose mean, s, is given by the stimulus, and whose standard deviation, σ, specifies the level of sensory noise. As in our task, s is drawn uniformly from the set, $s \in \{-.15, -.07, -.035, -.015, .015, .035, .07, .15\}$. The sign of s 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, $z = x$. 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 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 $z = P(s > 0|x)$[4]. The agent maps the internal estimate onto a response, r, by applying a set of thresholds, $r = f(z)$. 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 the time respond "-6", 2% of the time respond "-5" and so forth)[4] – using maximum entropy distributions with a fixed mean or a participant's observed response distribution (see below). Note that, for different levels of sensory noise, different thresholds must be used to generate the same response 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 quantify 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, σ_1 and σ_2. We then derived their joint accuracy under different pairs of confidence distributions, each associated with a specific mean. We limited our analyses to maximum entropy distributions (see **Supplementary Figure 4**); while there are many distributions that can generate a given mean, this is not the case when considering one family of distributions. 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 this procedure.

Comparing observed and optimal joint accuracy

We used our model to quantify how far each group in our experiments was from reaching optimal performance (**Supplementary Figure 7**). We first fitted our model to the data of each participant by searching for the sensory noise 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 good fits 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 the thresholds were fitted 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 (using

the same procedure as in **Figure 3B**) 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**). See **Supplementary Notes** for control analyses.

Computer-generated partners

We used our model to make the simulated 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 (σ) to specify the mean accuracy of the 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 partners: the mean confidence was about 2.2 for the low-confidence partners and about 4.2 for the high-confidence partners – the choice of confidence distributions was informed by data from earlier experiments. 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), 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 generated the required response distribution as specified above. To mimic lapses of attention and response errors, we randomly selected a response (from a uniform distribution over 1 to 6) on 5% of the trials (12 out 240 trials). We also varied the agents' reaction time (randomly sampled from a uniform distribution over 2 to 5 seconds), so that participants on some of the trials had to wait for the public display after having made their own response. Statistical tests showed that we obtained the 2-by-2 differences between participants and partners in terms of accuracy and mean confidence (see **Supplementary Table 1** for test statistics).

Joint accuracy expected prior to interaction

In Experiment 4, to compute the joint accuracy expected prior to interaction, $a_{\text{joint}}^{\text{isolated}}$, we played out responses of a given simulated partner against those of the participant from the isolated task – with joint decisions selected as in the social task. We estimated $a_{\text{joint}}^{\text{isolated}}$ across 10^4 iterations as the partner's responses were subject to sensory noise and the 5% lapse rate. This procedure allowed us to test whether the observed joint accuracy, $a_{\text{joint}}^{\text{social}}$, was higher or lower than expected prior to interaction, $a_{\text{joint}}^{\text{isolated}}$.

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 2** for average responses).

Learning model

We furnished our signal-detection model with a simple learning rule to provide a process account of how confidence matching arises (**Figure 4**). The agent updates on each trial t its mean confidence, c^o, as a mixture of its own mean confidence and its estimate of the partner's mean confidence, c^p, as follows: $c^o_{t+1} := c^o_t + \gamma(c^p_t - c^o_t)$ where γ describes the rate of adaptation and $(c^p_t - c^o_t)$ describes the mismatch between the agent's mean confidence and its estimate of the partner's mean confidence. The agent updates on each trial t its estimate of the partner's mean confidence as follows: $c^p_{t+1} := c^p_t + \alpha(\hat{c}^p_t - c^p_t)$, where α describes the rate of learning, $\hat{c}^p_t$ is the partner's confidence on trial t and $(\hat{c}^p_t - c^p_t)$ is a prediction error. The initial values of c^o and c^p reflect the agent's baseline mean confidence and its expectation for the partner's baseline mean confidence. The agent uses c^o to update the function, $r = f(z)$, which governs the mapping from the agent's internal estimate of the evidence strength onto a response, r_t. In our simulations, we assumed that that a pair of agents had the same levels of sensory noise ($\sigma = .10$); 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); that the learning rate was fixed ($\alpha = .12$) for both agents; and that they used the same degree of adaptation ($\gamma_1 = \gamma_2 = .20$; this value was chosen as it generated a degree of serial-dependence in trial confidence similar to that observed in our data). In each simulated experiment, the agents performed 160 trials, with stimuli drawn as in our task. 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.

Data availability

The behavioural data is available here: https://github.com/danbang/article-confidence-matching.

Code availability

Analyses and simulations were conducted in MATLAB (2015b). All code is available upon request from the corresponding author (D.B.: danbang.db@gmail.com).

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. Schelling, T. C. *The Strategy of Conflict*. (Harvard University Press, 1980).
19. Mahmoodi, A., Bang, D., Ahmadabadi, M. N. & Bahrami, B. Learning to make collective decisions: the impact of confidence escalation. *PLoS One* **8,** e81195 (2013).
20. Brier, G. W. Verification of forecasts expressed in terms probability. *Mon. Weather Rev.* **78,** 1–3 (1950).
21. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).
22. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27,** 415–444 (2001).
23. 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).
24. Mahmoodi, A. *et al.* Equality bias impairs collective decision-making across cultures. *Proc. Natl. Acad. Sci.* **112,** 3835–3840 (2015).
25. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious. Cogn.* **21,** 422–430 (2012).
26. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice.

Nat. Neurosci. **16,** 105–110 (2012).
27. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8,** (2014).
28. Harvey, N. Confidence in judgment. *Trends Cogn. Sci.* **1,** 78–82 (1997).
29. Shea, N. *et al.* Supra-personal cognitive control and metacognition. *Trends Cogn. Sci.* **18,** 186–193 (2014).
30. Bazerman, M. H., Curhan, J. R., Moore, D. A. & Valley, K. L. Negotiation. *Annu. Rev. Psychol.* **51,** 279–314 (2000).
31. Forsyth, D. R., Zyzniewski, L. E. & Giammanco, C. A. Responsibility diffusion in cooperative collectives. *Personal. Soc. Psychol. Bull.* **28,** 54–65 (2002).
32. 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).
33. 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).
34. Morreau, M. Grading in groups. *Econ. Philos.* **32,** 323–352 (2016).
35. Kent, S. *Sherman Kent and the Board of National Estimates: Collected Essays*. (History Staff, Center for the Study of Intelligence, Central Intelligence Agency, 1994).
36. 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).

ACKNOWLEDGEMENTS

This work was supported by the Calleva Research Centre for Evolution and Human Sciences at Magdalen College (D.B. and J.Y.F.L.), the Gatsby Charitable Foundation (L.A. and P.E.L.), the DAAD (A.M.), the Wellcome Trust (S.H.C.: 099741/Z/12/Z), and the European Research Council (B.B.: 309865-NeuroCoDec; C.S.: 281628-URGENCY). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

D.B., J.Y.F.L, B.B. and C.S. conceived the study and designed the experiments. D.B., S.H.C., B.R. and A.M. performed the experiments. D.B., L.A., R.M., P.E.L. and C.S. developed the models and the simulations. D.B. analysed the data and performed the simulations. D.B., L.A., R.M., S.H.C., P.E.L, B.B. and C.S. interpreted the results. D.B. drafted the manuscript. D.B., L.A., R.M., P.E.L., B.B. and C.S. wrote the manuscript.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to D.B. (danbang.db@gmail.com).

COMPETING INTERESTS

The authors declare no competing interests.

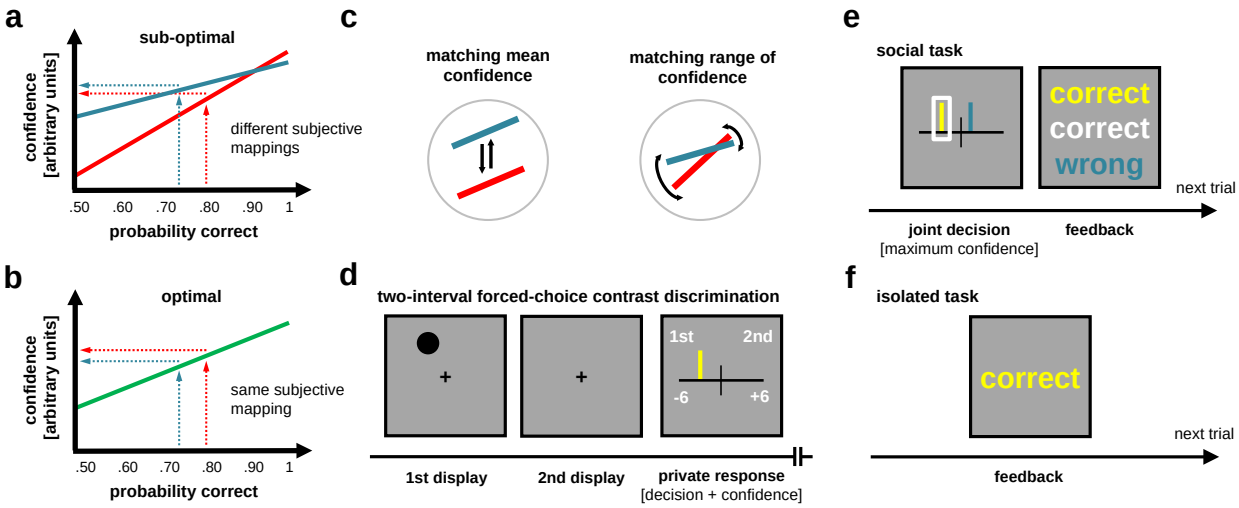

Figure 1 | Theoretical and experimental framework. a, Communication problem. Two handball referees disagree about whether the ball crossed the goal line. They have different functions (solid lines) mapping some internal estimate of the probability that their individual opinion is correct (x-axis) to confidence (y-axis). Here the blue referee expresses higher confidence but is less likely to be correct (dotted lines). **b**, Optimal solution. To maximise the probability that the group makes the correct decision, the referees must align their subjective mappings (green line). **c**, Confidence matching. The intercept (*left*) and the slope (*right*) of the referees' subjective mappings would change under confidence matching. **d**, Psychophysical task. Participants viewed two displays, each containing six contrast gratings (here dots). In one of the 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 private responses. Participants received feedback 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.

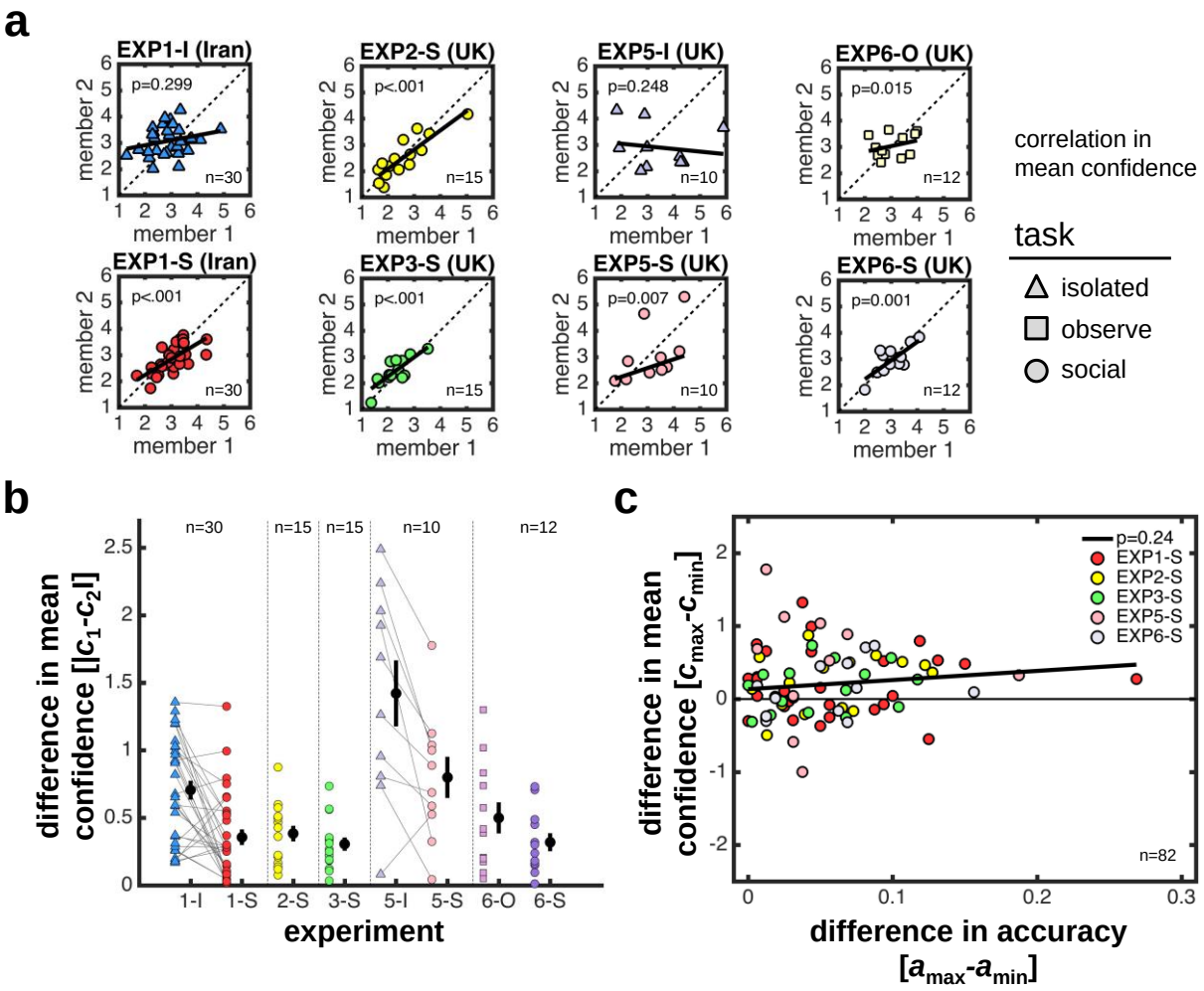

Figure 2 | Behavioural evidence for confidence matching. a, Correlation in mean confidence. The axes show group members' mean confidence, c_1 and c_2. **b**, Convergence in mean confidence. The y-axis shows the absolute difference between group members' mean confidence, $|c_1\text{-}c_2|$. **c**, Relative confidence does not scale with relative expertise. The axes show the difference in accuracy (x-axis, $a_{\max} - a_{\min}$) and in mean confidence (y-axis; $c_{\max} - c_{\min}$) between the more accurate (max) and the less accurate (min) group member. In all panels, each dot is a group. In panel B, the black dots are the group average. Error bars are 1 SEM. The lines connect group data when the same pairing of group members was used in two conditions. In panels A and C, the solid line is the best-fitting line of a robust regression. The *p*-value indicates the significance of its slope. In panel A, the *p*-value was calculated using a permutation procedure described in the **Methods**. In panel C, all *p*-values > .25 when analysis was done separately for each experiment. Triangle: isolated task. Circle: social task. Square: observe task.

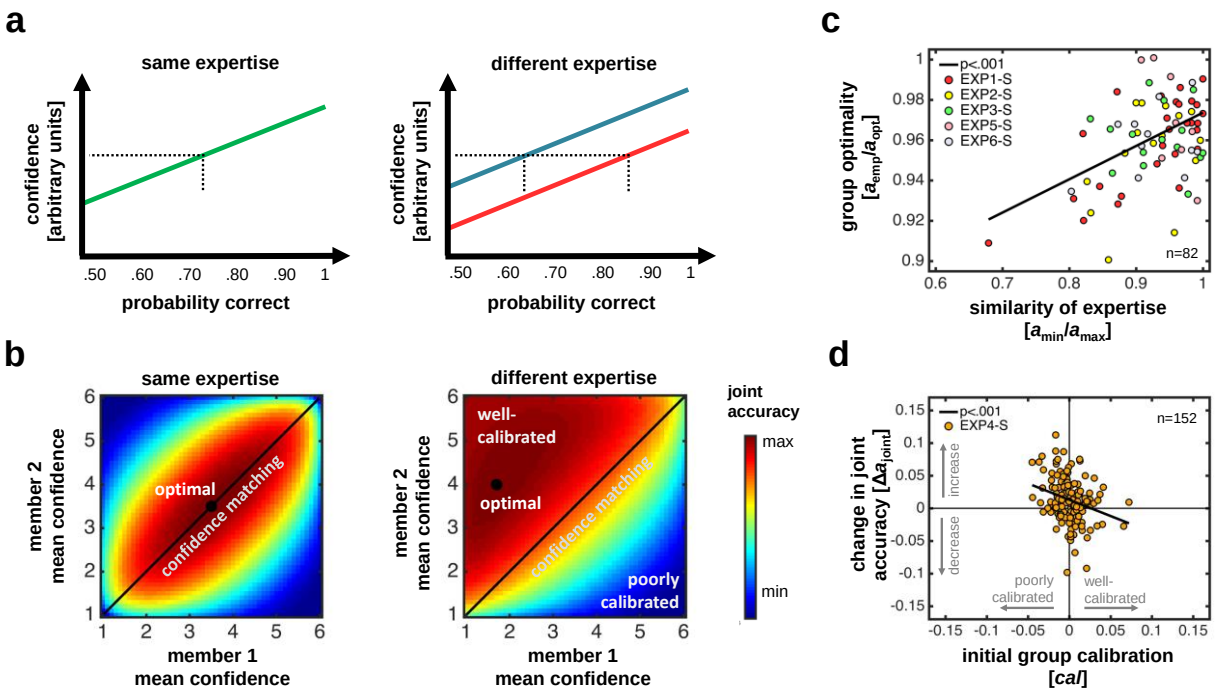

Figure 3 | Confidence matching is sub-optimal. a, Relative expertise and group performance. If the referees sampled from similar distributions of probability correct (*left*), then their subjective mappings should converge under confidence matching. If they sampled from different distributions of probability correct (*right*), then their subjective mappings should not converge and the less competent one should exert too much influence. **b**, Confidence landscapes. Under our model, group members with similar levels of accuracy (*left*) maximise joint accuracy (black dot: optimal) when their mean confidence is matched. For group members with different levels of accuracy (*right*), joint accuracy reaches its maximum when the more competent is the more confident. **c**, Optimality scales with similarity. The x-axis shows the ratio of the accuracy of the less accurate group member to that of the more accurate group member, $a_{\min}/a_{\max}$. The y-axis shows the ratio of the observed joint accuracy to that expected under the optimal solution, $a_{\text{emp}}/a_{\text{opt}}$. **d**, Confidence matching helps poorly calibrated groups but hurts well-calibrated groups. The x-axis shows a measure of group calibration prior to interaction: $cal = (a^{\text{isolated}}_{\text{participant}} - a_{\text{agent}}) * [(c^{\text{isolated}}_{\text{participant}} - c_{\text{agent}})/(c^{\text{isolated}}_{\text{participant}} + c_{\text{agent}})])$, where a is accuracy and c is mean confidence. This measure is positive when the difference in accuracy and in mean confidence have the same sign. The y-axis shows the difference between the observed joint accuracy and that expected prior to interaction: $\Delta a_{\text{joint}} = a^{\text{social}}_{\text{joint}} - a^{\text{isolated}}_{\text{joint}}$, where $a^{\text{isolated}}_{\text{joint}}$ was estimated by playing out the responses of a virtual partner against those recorded from a participant in the isolated task. In panels C and D, each dot is a group. The line is the best-fitting line of a robust regression. The *p*-value indicates the significance of its slope.

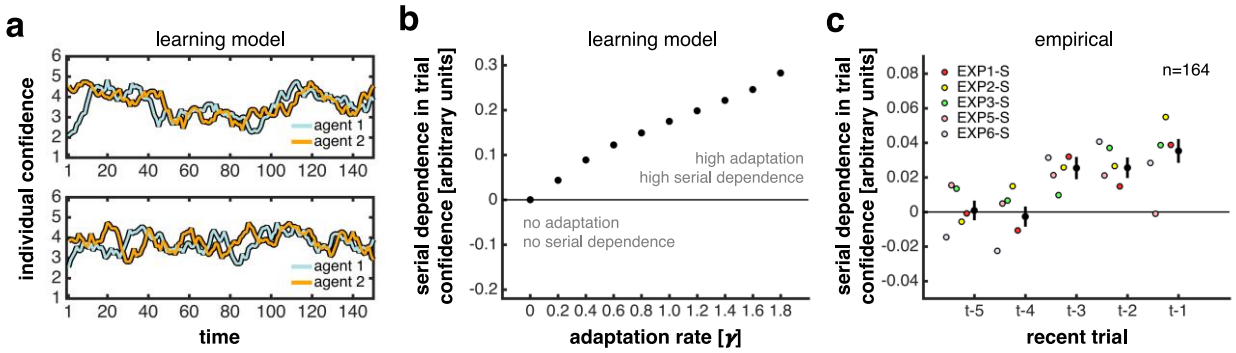

Figure 4 | Confidence matching at short time scales. a, Temporal difference learning model. Each agent keeps a running estimate of its partner's mean confidence and adapts its mapping from probability correct to confidence accordingly. Each plot show how the trial confidence of a pair of agents evolves over time and confirm that the learning mechanism can cause a convergence in mean confidence. The data was smoothed using a sliding average. **b**, Model predicts short-range serial dependence. The x-axis shows the degree to which each agent adapts its subjective mapping to its partner. The y-axis shows coefficients from a linear regression measuring the degree to which the agent's confidence on trial t depended on its partner's confidence on trial $t-1$. The higher the degree of adaptation, the higher the social influence. We included the stimulus ($t-1$ and t) and the agent's own confidence ($t-1$) as nuisance predictors. We simulated 10^5 simulated experiments for each degree of adaptation. **c**, Short-range serial dependence in the empirical data. Same analysis as in panel B, but now going 5 trials back into time. We tested significance by comparing the coefficients pooled across participants to zero (trial $t-3$ to $t-1$: all $t(163) > 3.900$, all $p < .001$, one-sample t-test, null: 0). We note that the degree to which participants influenced each other was correlated and that there was no short-range serial dependence in the isolated task (**Supplementary Figure 9**). In panels B and C, the black dots are the simulation/group average. Error bars are 1 SEM. In panel C, the coloured dots show the group average in each experiment.