Title:

Distributed neural replay of decision confidence mediates informational conformity

Authors:

Ali Mahmoodi[1,2,*]

Hamed Nili[3]

Carsten Mehring[1,2,7]

Bahador Bahrami [4, 5,6,7]

Affiliation:

[1] Bernstein Centre Freiburg, University of Freiburg, Hansastrasse 9a, 79104, Freiburg, Germany

[2] Faculty of Biology, University of Freiburg, Freiburg, Germany

[3] FMRIB, Wellcome Centre for Integrative Neuroimaging, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK

[4] Department of Psychology, Royal Holloway, University of London, Egham, UK

[5] Faculty of Psychology and Educational Sciences, Ludwig Maximilian University, Munich, Germany

[6] Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin Germany

[7] These authors contributed equally to this work

[*] Corresponding author email: ali.mahmoodi1367@gmail.com

Abstract:

We occasionally receive conflicting views from others. To maximize accuracy, we should exercise informational conformity by changing our mind proportional to our confidence about our initial view. This account predicts that neural correlates of confidence in the private decision should be replayed as the private and social information are integrated. In a perceptual estimation task (N=120), influence from others was proportional to private confidence. Human fMRI (N=20) showed that consistent with the replay hypothesis, confidence covaried with temporally distributed activity during private estimate (Precuneus and dorsal anterior cingulate cortex, dACC), social change of mind (dACC) and social outcome (dorsolateral prefrontal cortex and dACC). During social change of mind and only when paired with alleged human (but not with computer) partners, left temproparietal junction carried information about participants' social use of confidence. Our study reveals the neuronal substrates of the role of confidence in computational implementation of informational conformity.

Introduction:

Our decisions are often accompanied by a sense of confidence. Occasionally other people confront us with their opinion or information that contradict our decisions. In these situations, we sometimes decide to insist on our opinion and other times accept that others might know better. We intuitively follow our subjective sense of confidence when we navigate between these opposite social behaviours. We may revise our initial choice depending on our certainty about our decision and our estimate of the accuracy of others. How we come to a balanced comparison between our own, subjective sense of uncertainty and others' disagreeing decisions communicated to us, is an open and fundamental question in social and cognitive science. Here we offer an empirical framework to understand the mechanism(s) of changes of mind through social influence by disentangling the behavioural and neuronal substrates of subjective confidence.

It has been proposed that confidence is used to guide our own future behaviour[1]. For example, in one study, in a multi-stage decision-making paradigm, participants made serially-dependent perceptual decisions in every trial. The study found that participants adjusted the balance between their speed and accuracy in the second stage according to their confidence in the first stage of the sequence[2]. Adapting this approach to social influence, we begin by noting that joint decision making is often similarly a multi-stage affair (see Figure 1): individuals make a private decision first (left panel), then share their possibly conflicting choices with one another (middle panel) and eventually combine them to converge to a revised decision. We therefore hypothesise that confidence in private opinions is involved in the process of group decision making[3–5]. If individuals exercise informational conformity, where the aim is to boost accuracy[6], individuals should strike a balance between their subjective uncertainty (i.e. their confidence) with that of social information (other's decisions) when integrating their personal estimate with social information. An observation consistent with this account is that an individual aiming for informational conformity would take less influence from others when she is more confident about her own decisions[5,7–9]. We therefore hypothesise that under informational conformity, confidence affects the social influence an individual may take from others when changing her mind in response to others' opinions (Figure 1A). This suggests that at the neural level, one should be able to find the correlates of the individual's confidence (in a decision privately made earlier, Figure 1 left panel) replayed at a later, social stage (red arrow connecting the left and right panels) when the individual examines whether to revise her opinion in response to others' decisions.

Numerous previous works help us predict candidate brain regions in which we should see the correlates of confidence at the private decision-making stage. A variety of areas within prefrontal, parietal and occipital cortices have been linked to confidence[10] including dorsolateral prefrontal cortex[11,12], perigenual anterior cingulate cortex[13] and ventromedial prefrontal cortex for value-based[14,15] and perceptual decisions[16]. Importantly, all of these studies exclusively focused on isolated

individuals making private decisions[17–19] thus offering a clear set of target areas for neural activations in the private stage of our framework (Figure 1, left panel).

At the social stage (i.e. where participants could revise their decision given the social information), we hypothesise that the decision confidence is not estimated again in the same neural circuitry as described above but replayed in a set of other brain areas that contribute to the integration of private opinion with social opinion(Figure 1, right panel). We predict that confidence replay will be observed in the brain areas involved in social cognition and performance monitoring which, importantly, would not necessarily show correlations with confidence at the initial private decision-making phase. A recent study found deviating from a recommendation (i.e. "dissent") was associated with increased activity in the dorsal anterior cingulate cortex (dACC)[20]. Building up on this finding, we note that in order to maximise accuracy, participants should dissent from others' recommendation when they are confident about their own choice. We set out to examine (1) if dACC activity is correlated with confidence of a subject's initial decision and (2) if this activity can be dissociated from dissent. Circumstantial evidence supporting our hypothesis comes from the literature showing that ACC is involved in decision-making process[21] and comparison between options[22], processing the cost and benefit of an option[23,24], and conflict detection in social contexts[25,26].

Using a paradigm adapted from a recent study[27], participants first made a perceptual estimate and reported their confidence. Then after observing a partner's estimate, they were given a chance to revise their estimate. Behaviourally, we found that participants' reported confidence at the private stage was correlated with their change of mind in response to social information. In line with our neural replay prediction, we found that at the time of making the revised estimate, the dACC BOLD response was modulated by the confidence that participants had reported earlier for their initial private estimate. We also found that patterns of activity in the left temporoparietal junction (lTPJ) at the same later social stage, carried information about the deviance between confidence and social influence only when participants interacted with a human but not a computer partner. In some trials the partner had the chance to revise their estimate in response to the participant's estimate. At the time of revealing partners' revised estimate, BOLD activity in the dorsolateral prefrontal cortex (BA46) was correlated with participants' previously reported confidence. Finally, we found that the BOLD activity in the dorsolateral prefrontal cortex (BA9) was associated with the participants' influence over their partners.

Results:

Participants (N=60 for behavioural exp1; N=40 for behavioural part of exp2; N=20 for fMRI part of exp2) were invited to the lab and were told that they will work with another partner/partners in the experiment (see Methods for details). In the behavioural experiment 1, participants came to the lab one at a time. Half of the participants (N=30) were told that their partner is a computer. The other half (N=30) were told that their partner is another human placed in another room. In the fMRI experiment (exp2), three participants came to the lab at the same time and after briefly meeting each other, one participant was placed in the scanner and the other two performed the behavioural part of the experiment in separate rooms. In this experiment, participants were told that in some blocks they will work with a computer and in other blocks they will work with the other participants they had just met. At the beginning of each trial, a photo of one of the other participants or of a computer was displayed to indicate the partner with whom the participant was working with in the current block. In reality, unknown to the participants, in both experiments (exp1 and exp2) all partners' responses were generated by a computer algorithm.

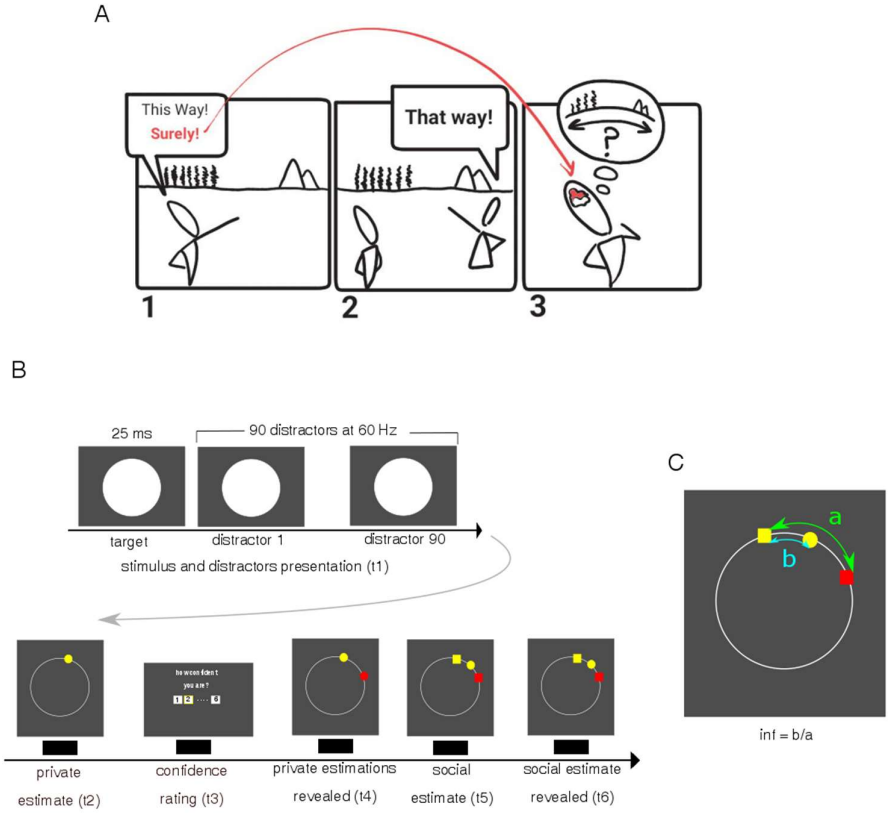

Figure1 1: Experimental task. ***a*** *multi-stage decision making frame work for social interactive decision making.* ***b*** *Participants first observed a series of dots on the screen (t1). They were required to indicate where they saw the very first dot that appeared on the screen (t2) and then declare their numerical confidence (t3). After making their individual estimates, they were presented with the estimate of a partner (red dot) concerning the same stimulus (t4). After observing the partner's choice, in some trials the participants and in the remaining trials the partner was given a second chance to revise their initial estimate (t5). Afterwards they were briefly presented with the initial and the second*

choices (t6). ***c*** *Influence was computed as the angular displacement toward the partner's choice divided by their initial distance from each other's choice (b/a).*

In each trial, the participants made a perceptual estimate about the location of a target on the screen (Figure 2) which was followed by rating their confidence on a scale of 1 (not confident at all) to 6 (very confident). After stating their initial private estimate, the participants saw their partner's estimate of the location of the same stimulus. Next, the participants either revised their own estimate or observed the partner revising theirs. Participants were required to put their second estimate somewhere between their own first estimate and that of their partner's (Figure 1). Using this constraint guaranteed that the amount of change made in the second stage was solely due to observing the partner's estimate rather than any random change of mind due to being given a second chance[2]. The partner's initial estimate was generated by sampling randomly from a distribution centred on the correct answer (see Methods).

Behavioural results

For all behavioural results we will report the aggregated data from exp1 and exp2, as there was no significant difference between the experiments for any of the presented results. We computed the influence that participants took from their partners as the ratio of the angular displacement between their initial and final estimate toward their partner's estimate divided by their initial angular distance from their partner (Figure 1C). Participants' average influence taken from their partner was .38±.15 (mean ± std dev). This indicated that participants gave a slightly but significantly higher weight to their own opinion as values below .5 indicates that participants weighed their partner's decision less than their own (sign rank test against .5, $W = 582$, $p<.0001$). Consequently, not only this result replicates earlier findings on egocentric discounting in advice taking[28], it also rejects the null hypothesis that participants combined the two judgements randomly.

Given previous findings on human ability to distinguish between their correct and incorrect responses[29], we tested whether participants' confidence was correlated with their absolute error. In experimental setups like here identification error is not categorical but graded. Therefore, the correlation between confidence and absolute error is a measure of metacognitive sensitivity in our task. Using a linear mixed model we found that participants' trial-by-trial absolute error in their first estimate (defined as the angular distance between their first estimate and the true location of the target) was negatively correlated with their reported confidence (linear mixed model, parameter estimate -.41, 95% CI [-.38 -.44], $t(26998) = -9.48$, $p <.0001$, see Supplementary Material for the details of the model).

We then went on to investigate the potential correlation between participants' trial by trial confidence and the amount of influence that participants took from their partners. To investigate the relationship between influence and confidence, we employed a linear mixed model. Confidence had a significant negative effect on influence (parameter estimate -.23, 95% CI [-.17 -.29], t(13468) = -7.41, 1×10^{-1} , see Supplementary Material for the details of the model). Consistent with our prediction, this result suggests that change of mind is correlated with participants' subjective sense of confidence.

Thus, our results indicated that Participant's revised decisions are consistent with confidence-weighted informational conformity when presented with others' opinion during social decision-making. But as expected, the influence confidence correlation (ICC), which we defined as the within-subject correlation between confidence and inverse influence varied across participants to a great degree (Figure 2). On the other hand, metacognitive sensitivity, which we defined as the within-subject correlation between confidence and inverse error, varied across individuals as well[30] (Figure 2). One potential possibility is that, those holding more reliable confidence report (i.e. higher metacognitive ability), would later rely more on their confidence for change of mind (i.e. higher negative influence confidence association). In other words, participants with high perceptual metacognition will have high ICC as well. A significant correlation between the perceptual metacognition and ICC (Figure 2, Pearson correlation coefficient, r = .23, p = .01) confirmed our prediction. Note that in both cases, ICC or metacognition, more positive values indicate more confidence-weighted conformity or more reliable confidence report, respectively.

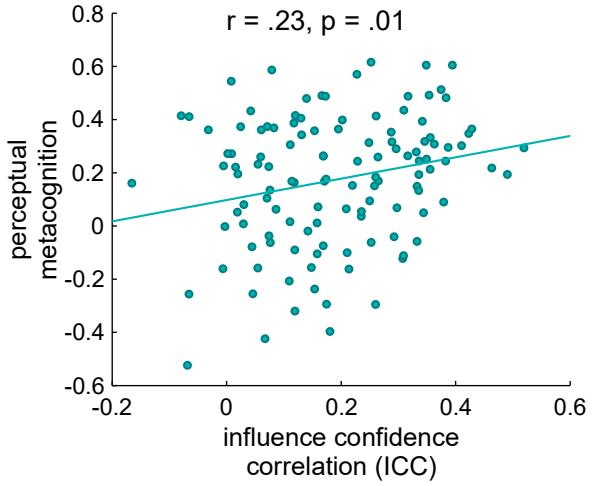

Figure 2: Behavioural results. Metacognition (i.e. the correlation between confidence and error) vs. the influence confidence correlation (ICC). Every dot is a single participant, the line was obtaind by linear regression.

Neuroimaging results

We started by investigating the neural correlates of confidence at the private stage. We reasoned that comparison of our findings to numerous previous works that have asked the same question in the context of other perceptual and value-based decisions could serve a good benchmark. We employed

a general linear model approach (GLM1 see Methods). The times t1-t6 (see Figure 1B) were defined as the onsets in our GLM, and the participants' confidence was introduced as a parametric modulator (see Methods for details). All analyses were corrected for multiple comparison using $P < 0.001$, family-wise error (FWE) cluster size-corrected $P < 0.05$. We found that confidence was positively correlated with the BOLD response in the left post central gyrus at the time of stimulus presentation (t1), (Figure S1, peak coordinates [-50 -28 50], k = 28, t_{peak}(19) = 5.43, p = .04). At the time of the private estimate (t2), we found a negative correlation between confidence and the activity at precuneus (Figure S2, peak coordinates [6 -48 50], k = 129, t_{peak}((19) = 4.93, p <.0001). At the time of the confidence rating (t3), we found a positive correlation between confidence and the right lingual gyrus (Figure S3, peak coordinates [10 -84 -6], k = 82, t_{peak} (19) = 6.99, p <.0001). These findings closely replicated a number of key previous findings[31] despite various differences in task and experimental setups and therefore, encouraged us to proceed with our main hypotheses.

Replay of confidence in the dorsal anterior cingulate cortex

Consistent with our hypothesis, at the social stage (i.e. revision time, t5), the BOLD response at dorsal portion of the anterior cingulate cortex was positively correlated with private confidence (Figure 3A peak coordinates [2 -16 42], k = 59, t_{peak}((19) = 6.26, p <.0001). To rule out potential confounds, we regressed out reaction time, influence, the angular distance between participants' initial estimate and their partner's estimate, and the angular distance between participants' revised estimate and their partner's first estimate (which is a measure of dissent). Previous studies showed that BOLD activity of the dACC was negatively correlated with confidence at the private decision-making[11,32]. We defined a spherical (r = 10mm) Region of Interest (ROI) centred at MNI coordinates identified from previous literature ([0, 17, 46])[11,32]. Consistent with previous reports, we found that activity in this ROI was negatively correlated with confidence at the time of first estimate (t2) (p<.05, Figure 3B). This ROI was created based on previous studies. Nonetheless, we also tested for any negative effect of confidence in the same dACC blob which was positively correlated with confidence at the social stage (Figure 3). Critically, the same dACC blob which was positively correlated with confidence at the social stage, was negatively correlated with confidence at the private stage (p <.05). As expected, the time course of dACC activity showed that, activity in this area was initially negatively correlated with confidence at the private stage (consistent with earlier findings[11,32]) and then become positively correlated with confidence at the social stage (Figure 3C).

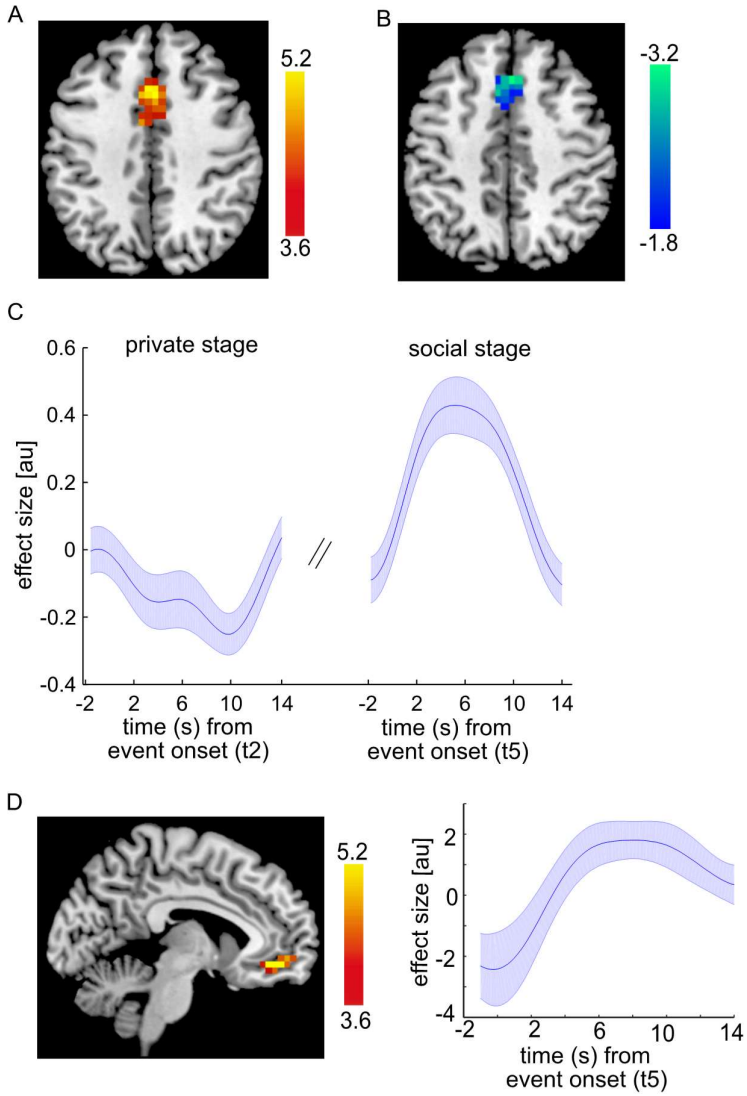

Figure 3: Whole brain analysis of activity related to confidence and influence. Threshold at p<.05, FWE corrected for multiple comparisons, cluster defining threshold p<.0001, N = 20 subjects. ***A*** *activity of the dorsal anterior cingulate cortex was signficantly modulated by confidence at the social decision stage. Reaction time and the distance between the estimate made by the two players were regressed out.* ***B*** *At the time of private estimate (t2), activity of the dACC defined from previous literature was negatively correlated with confidence* ***C*** *time course sampled from the dACC based on the ROI from previous literature at the private stage and the dACC cluster which was significantly modulated by the confidence at the social stage.* ***D*** *left panel, activity of the ventromedial prefrontal cortex was signficantly modulated by influence that participants took from their partner in each trial at the time of making the second estimate. Right panel, time couse sampled from a cluster in vmPFC which was significantly modulated by the influence that participants took from their partners at the time of making second estimate.*

Influence signal in the ventromedial prefrontal cortex

We then searched for the neural correlates of the influence that participants took from their partner. Consistent with previous studies[33], activity in the ventromedial prefrontal cortex (vmPFC) at the time of making the second estimate (t5) was positively modulated by the amount of influence that participants took from their partners (Figure 3D peak coordinates [-10 44 -10], k = 26, t(19) = 5.49, p = .04). Notably, activity of vmPFC was only correlated with the influence that participants took from their partners but not the replay of confidence. To prove this, we did an additional analysis by including

confidence in the regressor but not influence. We found that activity of the vmPFC was not correlated with confidence.

Distinct neural correlates of deviance in the temporoparietal junction between human and computer interactions

So far, our results showed that behaviourally, participants' revision in response to their partner's estimate was negatively correlated with their previously reported confidence; meanwhile, at the neural level we observed that activity of the dACC at the time of revision was correlated with the participant confidence. However, we also note that confidence in our private choice may not be the only factor affecting the influence participants took from their partners. For example, if we hold a high opinion of our collaborator's accuracy, we may conform to them even if we were very confident. Moreover, normative factors such as reciprocity[27] and flattery[34] affect the balance between social and private information. Thus, the influence we take from others might not always be consistent with our private confidence. Therefore, we argue that the deviance between confidence and influence is an indirect indicator of the cognitive mechanism by which social information impacts the joint decision. The neural correlates of such deviance is unknown. An important constraining factor in the search for this neural signature is the finding that obligation to norms such as reciprocity is specific to human-human but not human-computer interactions[27]. Therefore, we predicted the neural correlates of deviance might be different when participants cooperated with a human or computer partner and would require theory of mind (ToM)[35]. We expected to observe strongest brain responses to high deviance between confidence and influence in ToM network consisting of medial prefrontal cortex[36], and temporoparietal junction (TPJ) when participants interacted with human partners.

We computed a trial by trial measure of deviance between confidence and influence as follows: in the interest of clarity, lets first define insistence as "the degree to which a participant insists on her decision" as 1 minus influence. Then we note that the absolute difference between insistence and confidence shows how much a participant has deviated from taking influence based on her confidence. Given this description, we define deviance at each trial as the absolute difference between insistence and confidence as follows:

$$deviance_t = |insistence_t - conf_t| = |1 - inf_t - conf_t|$$

In which *t* indicates the trial number and *inf* and *conf* are variables obtained behaviourally on each trial. Univariate analysis did not identify any brain area whose activity was modulated by the deviance at the time of social estimate (t5 in figure 1). We therefore tested for the effect of deviance on multivariate activity patterns[37,38] . To this end, we used representational similarity analysis (RSA). In RSA, the representational content of brain responses can be compared with behavioural

measurements or cognitive or computational models. This would be achieved by computing the similarities (e.g. correlations) between their representational dissimilarity matrices (RDMs). An RDM is a square symmetric matrix with zero diagonals in which off-diagonal elements indicate the dissimilarity between activity patterns or behavioural variables associated with two different experimental conditions. We constructed a behavioural RDM based on the absolute pairwise differences between the single-trial deviances and measured its correlation with the brain RDM for each searchlight sphere. To construct brain RDMs, we compared single-trial activity patterns and followed the standard procedure provided in the RSA toolbox[38] (see Methods). We adopted a whole-brain searchlight mapping approach[37,39]. In each participant, we computed the correlation between brain RDM and behavioural RDM for human and computer conditions separately. We then tested for the difference of the two correlations. We found that a cluster in the left temporoparietal junction (lTPJ) carried information about deviance that was different for human and computer conditions (Figure 4A, peak coordinates [-46 -48 18], k = 56, $t_{peak}(19)$ = 5.3, p = .004). This result could be interpreted in two ways: on one hand, it is possible that the lTPJ carries *some* information about deviance when working with any partner (animate or artificial) but these might constitute different patterns for human and computer. A second possibility is that lTPJ carries information about deviance only when working with one type of partner (i.e. human or computer). To distinguish these possibilities, we looked for the effect of deviance on response patterns separately under human and computer conditions. There was a significant effect of deviance on the activity pattern on the same lTPJ cluster in the human condition (Figure 4B, peak coordinates [-46 -48 14], k = 31, $t_{peak}(19)$ = 4.8, p = .04). There was no effect of deviance on the activity pattern neither in the lTPJ nor in any other brain area in the computer condition. Therefore, this result suggests that the activity pattern in the lTPJ carries information about the deviance only when interacting with a human partner. We found no area whose activity pattern correlated with the deviance when we pooled across both human and computer conditions.

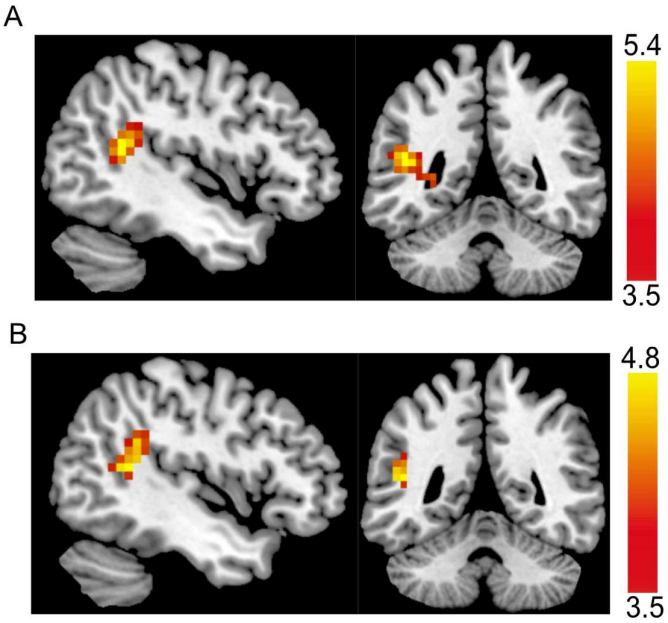

*Figure 4: **A** at the time of social decision, the pattern of activity in the lTPJ is different between human and computer conditions. **B** The pattern of activity in the lTPJ carries information about deviance only in the human condition.*

Dorsolateral prefrontal cortex (BA9) encodes participants' influence over their partner

Humans' monitor their influence over others[41] and this influence might affect their evaluations of their own accuracy[27]. Therefore, we looked for brain areas whose activity was correlated with participants' influence over their partners at the time of showing their partner's revised estimate (t6). To investigate the neural correlate of such monitoring process we carried out univariate and multivariate analysis. Our multivariate (but not univariate) analysis showed that activity patterns in a cluster at the dorsolateral prefrontal cortex (BA9) carried information about participants' influence over their partners (Figure 5A, peak coordinates [38 32 30], k = 61, t_{peak} (19) = 5.4, p = .002) when their partner's revised decision was revealed (t6). BA9 has been implicated in error processing[42]. One interpretation of this result is that participants may employ their influence over others as a feedback to evaluate their performance. This strategy might be more prevalent among less confident individuals as more confident individuals may not pay attention to their influence over others. If this is the case, we would expect that less confident individuals pay more attention to their influence over others. To directly test this hypothesis, we analyzed participants rating of their own accuracy on a scale from 1 (very low accuracy) to 10 (very high accuracy) at the end of each experimental block. We also extracted the average similarity values (correlation between brain and behaviour RDMs in the BA9) and defined it as the influence signal strength. Participants who have monitored their influence over their partners would elicit a large influence signal strength. We found a positive correlation between participants' average perceptual error and influence signal strength (Pearson r = .58 p = .007, Figure 5B) and a negative correlation between their subjective performance rating and the influence signal strength

(Pearson r = -.45 p = .04, Figure 5C) and also between their average confidence and the influence signal strength (Pearson r = -.44 p = .05).

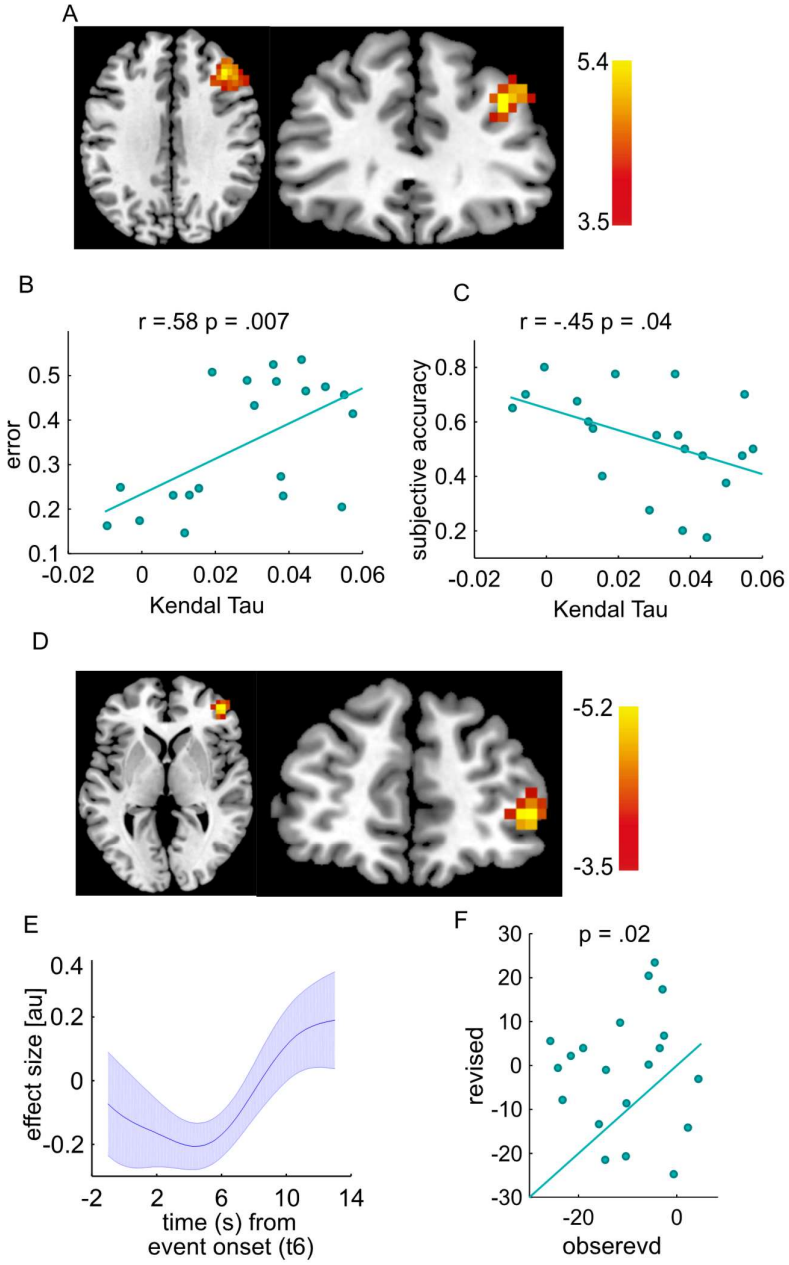

Figure 5: **A** In trials in which the partner made the social choice, at the time of second estimation revelation (t6), the pattern of activity in the dlPFC (BA9) was correlated with the participants' influence over their partners. **B-C** participants' average error and their estimate of their own performance (Y-axis of panels B and C, respectively) is depicted against strength of the coding of the participants' influence over their partners (X-axis). Every dot is a participant, the line was obtaind by linear regression. **D** at the time of second estimation revelation (t6), activity of a cluster within dlPFC (BA46) was negatively correlataed with their previosuly reported confidence. **E** time course of activity of BA46 at the time of second estimation revelation. **F** correlates of confidence at BA46 is significantly larger (more negative beta values) in trials in which the participants observed their partner's revised estimate compared to trials which they revised the estimate. The line indicates the identity line.

Replay of confidence when partners' revised estimation was revealed

Previous studies suggested that confidence reflects the posterior probability that a decision is correct[43] and is used to shape a global estimation of accuracy (i.e. estimating one's accumulative accuracy during an entire task)[44]. On the other hand, as posited above, social signals (participants' influence over their partners) may affect participants' assessment of their accuracy[27]. Therefore, it is likely that when their

influence over their partners is revealed, participants may replay their previously reported confidence in order to update such global estimate of their performance. We also predicted that neural representation of confidence might be different depending on the availability of a social outcome signal (i.e. influence over partner). Consistent with our prediction, in trials where the partner made the social decision, at the time of revealing the partner's revised estimate (t6) the activity in dlPFC (BA46) was negatively correlated with confidence (Figure 5D-E, peak coordinates [46 48 2], k = 26, t_{peak} (19) = 4.1, p = .03). We also sought whether confidence was represented in the brain when the participants made the revised decision. At the time of revised decision presentation (this time in trials in which participants made the revision), again the dACC was the only area whose activity was correlated with confidence in trials in which the participant made the revised decision (peak coordinates [6 4 38], k = 32, t_{peak} (19) = 4.2, p = .02). Critically, the activity of the BA46 was not correlated with confidence in these trials. A direct comparison showed that the response of the BA46 to confidence was significantly higher (i.e. more negative) in trials in which the partner made the revision compared to trials in which the participants made the revision (Figure 5F sign rank test W = 46 p = .02).

Discussion:

Confidence plays an important role in our daily interaction with others[8]. In this study, we investigated the contribution of human confidence when participants performed a perceptual task together with other agents (human or computer). The information coming from others may lead us to revise our opinions and to do that, we may consider our own level of certainty. We observed that confidence predicted the extent to which participants were influenced by their partners' estimates once they were given a chance to revise their initial decisions. Moreover, we showed that confidence information was replayed multiple times in different stages of the task possibly serving different cognitive functions (Figure 6): at the time of stimulus presentation, BOLD signal in the post central gyrus was positively correlated with confidence. At the time of making a private estimate, activities in dACC and precuneus were negatively correlated with confidence. Activity of the right lingual gyrus was positively correlated with confidence during confidence rating. When a *social estimate* was made (revision), activity of the dACC was positively correlated with confidence. When social outcome, that is to say, the revised estimate was revealed to the participant and the partner, activity in the dACC was positively correlated with confidence only in trials which the participant made the social estimate; in trials which the partner made the social estimate, however, activity in the dlPFC was negatively correlated with confidence.

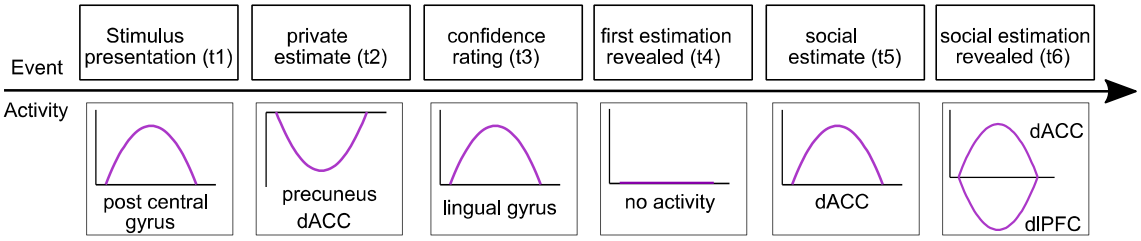

Figure 6: summary of the neural correlates of confidence in different stages of our task. Upward and downward graphs show negative and positive modulations of average BOLD signal with confidence, respectively.

Our neuroimaging results using human fMRI show that at the time of the social estimate (i.e. revision), the replay of confidence was represented in the dorsal portion of the anterior cingulate cortex (dACC). Some studies found that activity of the dACC is negatively modulated by confidence[32,45,46], when making decisions or when anticipating the outcome of the decisions[47–49]. Consistent with these studies, we also found that at the time of making the private estimate, activity of the dACC was negatively correlated with confidence (Figure 3B). During the social estimate (i.e. revision time) when confidence did not need to be estimated again, confidence information was replayed in the opposite polarity. We therefore speculate that these two different coding schemes for confidence in the dACC may serve different computational purposes benefitting the individual and social stages of decision-making.

The rich literature of research on the dACC reflects the fact that this area's role in the humans decision-making remains controversial[50,51]. A recent theory has proposed that dACC encodes variables which are relevant to decision-making[21]. This theory suggests that dACC tracks task-relevant information to guide appropriate action. Consistent with this hypothesis, we saw that dACC is involved in decisions in social[52] and non-social[22] contexts. Our behavioural data shows that confidence is correlated with the magnitude of revision at the social stage (i.e. partner's influence on the participants). Taken together, we suggested that the dACC should have access to the previously reported confidence to contribute to the computations involved in decision-making. Consistent with this account, our result show that the dACC replays confidence information when the alternatives (estimates made by self and others) are compared and the social decision is made.

Interestingly, a recent study found that activity of the dACC increased when participants decided to dissent (i.e. deviating from a collaborator's recommendation) and the activity of the vmPFC increased as participants conformed to a collaborator's decision[20]. Both of these findings are consistent with our results. However, invoking the conflict monitoring theory of the dACC[57] Qi et al.[20] concluded that the increased activity of the dACC might be related to being in conflict with others. It is important to note that our behavioural results show that people take less advice from their partners (i.e. showing a stronger tendency to dissent) when they are more confident of their opinions. Therefore, confidence and dissenting from others' recommendations are inevitably intertwined. This raises the question

whether the previous report suggesting dACC's association with dissent from others' opinion may have in fact been reflecting the individual's high confidence and conviction in their personal opinion rather than being a signature of dissent per se. To find out which of the variables modulates the activity of the dACC, we included the initial disagreement between participants' estimate and their partners' estimate and the degree of disagreement between participants' revised estimate and their partner's estimate in our GLM. In either case, dACC was modulated by the confidence while neither the initial disagreement nor the final (i.e. disagreement after revised decision) explained any of the variance in the activity of this area in our experiment. Thus, we propose that, the dACC signal is associated with participants' confidence, not with dissent or disagreement. The result of Qi et al. may then be explained by the strong association between confidence and dissent pointing to directions for future research about the contributions of dACC to social decision-making.

Activity of the vmPFC has been suggested to be associated with updating one's own preferences after observing others' preferences or choices[40,53,54] and is also involved in persuasive message processing[55,56]. In line with these studies, we also found that the influence that participants take from their partners was correlated with the activity of the vmPFC.

Outside of informational conformity, normative factors such as reciprocity[27] and flattery[34] affect the balance between social and private information and could lead to higher deviance between confidence and influence. However, normative factors are confined to human-to-human interactions and do not extend to human-to-nonhuman interactions (at least not yet). Therefore, we expected the neural correlates of deviance to be different between these two sorts of interactions (i.e. human-human and human-computer). Consistent with our prediction, we found that patterns of activity in the lTPJ were correlated with deviance only for human-human interactions and were different between human to human and human to computer interactions (Figure 4). This result suggests that the lTPJ, an area well-known for its role in TOM and social interaction and representing others' beliefs[58], might influence social decision-making by introducing normative factors in the process of balancing between own and other humans' decisions.

We found two neighbouring but separate clusters within dlPFC that showed specific responses during the final stage of the trial when social outcome was revealed. If the social decision was made by the partner and therefore the outcome revealed the participant's influence over their partner, multivariate patterns of activity in Brodmann area 9 (BA9) reflected the magnitude of this influence. In addition, activity in BA46 was negatively correlated with confidence. BA9 has been previously implicated in responding to errors made by self[42] and others[59]. Brodmann area 46 has been reported to be negatively correlated with confidence[11,60]. Our findings are consistent with both of these notions. A recent study suggested that confidence is used to make a global self-performance estimation[44]. It was also

suggested that the degree to which others are influenced by us affects our global self-performance estimation[27]. One may speculate that after disclosure of their influence over their partner, the participants' BA46 replays confidence to assess self-performance by combining influence and confidence or assessing influence based on self-confidence. This speculation is supported by the observation that there is a connection between the participants' confidence and tracking their influence over their partner: high confident individuals had weaker coding of their influence over others (Figure 5 B, C). However, future studies are needed to make conclusive statement about distinct role of confidence and influence and their interaction in forming a global self-performance estimate.

To conclude, we showed that confidence information is replayed dynamically and flexibly in various anatomical locations in the human brain over several different temporal stages of the social decision-making. Our findings help understand the neurobiological substrates of informational conformity and help identify the computational characteristics of how the subjective sense of confidence in private decisions is integrated in this social process.

Methods:

Participants: In total, 120 healthy adult participants (60 females, mean age ± std:25 ± 3) participated in the two experiments after having given written informed consent. Each participant participated in only one of the two experiments. The respective experimental procedures were approved by the ethics committee of the University of Freiburg (exp1) and the ethics committee at the University College London (UCL) (exp2).

Experimental paradigm: Participants were presented with a sequence of 91 visual stimuli consisting of small circular Gaussian blobs (r = 5mm) in rapid serial visual presentation on the screen. The first stimulus was presented for 30ms while every other stimulus was presented for 15ms each. Participants' task was to identify the location of the first stimulus. Participants were required to wait until the presentation of all stimuli were finished, and then indicate the location of the target stimulus using a mouse in exp1 (Figure 1B) or a keyboard in exp2. The reported location was marked by a yellow dot. After participants reported their initial estimate, they were required to report their confidence about their estimate on a numerical scale from 1 (low confidence) to 6 (high confidence). In the fMRI experiment, this stage was followed by a blank jitter randomly drawn from a uniform distribution from 1.5-4.5 seconds. Afterwards, participants were shown the estimate of their partners about the same stimulus for 1.5 seconds by a small red dot on the screen (plus a jitter time randomly drown from a uniform distribution from 1.5-4.5 seconds for the fMRI experiment). Then, either the participant revised her estimate or observed the partner revise theirs. After the second estimate was made, all estimates were presented to the participants for 3 seconds (plus a jitter time randomly drown from a uniform distribution from 1.5-4.5 seconds for the fMRI experiment). In this stage, the first estimate was shown by a hexagon (for the behavioural experiment) or by a dot with a different colour (for the fMRI experiment) to be distinguished from the second estimate which was shown by a circle (Figure 1 B). Participants were told that their payoff would be calculated based on the accuracy of their first and second estimates. However, everyone was given a fixed amount at the end of the experiment. In the fMRI experiment, 10 participants dot colour was yellow and their partner's dot colour was red. For the remaining 10 participants, the colours were reversed. Further details of the experimental paradigm are described in our previous study[27].

In experiment 1, half of the participants were told that they will do the experiment with another partner located in an adjacent room. The rest of the participants were told that they will do the experiment with a computer algorithm (see below). Each participant completed 330 trials in which in half of them they announced the second estimate. We carried out experiment 1 as part of a previous study[27] and re-analysed the data here.

In experiment 2, three participants came to the MRI facilities at the same time. After reading the task instructions, one participant was scanned and the other two carried out the behavioural task outside the scanner. In this experiment, participants were told that they will play with four different partners: two human partners (the two they met before the experiment) and two computer partners which were controlled by the algorithm described above. Participants completed 4 blocks of the experiment each consisting of 30 trials. In each block they only worked with one partner. At the beginning of each trial, a photo of the partner they work with was shown to the participants. Photos of two different computers with different colours (counterbalanced across participants) represented the two computer partners. In reality, and unknown to the participants, all partners' estimates were generated by a computer algorithm. The partners only differed in the way they generated their second choice which is beyond the scope of this study. All results presented here are qualitatively the same regardless of the alleged identity of the partners (human or computer) and the algorithm used for the second estimate. The behavioural participants completed the task with a mouse, however, the fMRI version of the task was adopted to be completed with keyboard.

All experiments were performed using Psychophysics Toolbox[61] implemented in MATLAB (Mathworks). The behavioural data were analysed using MATLAB.

Debriefing: After each session of the experiment, all participants were debriefed to assess to what extent they believed the cover story. We interviewed them with indirect questions about the cover story and all participants stated that they believed they were working with other human participants in neighbouring experimental rooms (if they were told that their partner is a human partner).

Constructing computer partners: The estimates of computer partners were calculated as in our previous study[27]. The error distribution of the computer algorithm's first choices was modelled from participants' actual estimation errors during a pilot experiment carried out as part of our previous study[27]. Ten participants performed an experiment identical to experiment 1. We aggregated errors of all participants (N =10) and fitted the concentration parameter kappa of a von Mises distribution centred on the target, yielding the value kappa = 7.4. Then in each trial we drew the first choice of the computer partner from this distribution. We suspected that participants' assessment of their partners' performance may be strongly influenced by the few trials with high confidence (confidence level of 5 or 6). To avoid this potential problem, the partner's first choice was not taken from the von Mises distribution in high confidence trials but randomly drawn from a uniform distribution centred on the participants' choice with a width of +/- 20 degrees in the behavioural experiment and +/- 50 degrees in the neuroimaging experiment.

MRI data acquisition. Structural and functional MRI data were obtained using a Siemens Avanto 1.5 T scanner equipped with a 32-channel head coil at the Birkbeck-UCL Centre for Neuroimaging. The

echoplanar image) sequence was acquired in an ascending manner, at an oblique angle (≈ 30°) to the AC–PC line to decrease the impact of a susceptibility artefact in the orbitofrontal cortex with the following acquisition parameters: 44 volumes of 2 mm slices, 1 mm slice gap; echo time = 50 ms; repetition time = 3,740 ms; flip angle = 90°; field of view = 192 mm; matrix size = 64 × 64. A structural image was obtained for each participant using MP-RAGE (TR = 2730 ms, TE = 3.57 ms, voxel size = 1 mm^3, 176 slices).

fMRI data analysis. Imaging data were analysed using Matlab (R2016b) and Statistical Parametric Mapping software (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK). Images were corrected for field inhomogeneity and corrected for head motion. They were subsequently realigned, coregistered, normalized to the Montreal Neurological Institute template, spatially smoothed (8 mm FWHM Gaussian kernel), and high filtered (128 s) following SPM12 standard preprocessing procedures.

GLM1: The design matrix for this GLM included 6 events. These were the times of stimulus representation (t1), making the first (private) estimate (t2), reporting the confidence (t3), showing the first estimates (t4), making the second (revised) estimate (t5), showing the revised estimates (t6). Furthermore, regressors for t1, t2, and t3 were parametrically modulated by subject's reported confidence. The regressors for t4 included confidence as parametric modulator together with the angular distance between the participant's own and the partner's first estimate. The regressor for t5, included the parametric modulators confidence, angular distance between the participant's own and the partner's first estimate, angular distance between participant's second and the partner's first estimate, and the influence that participants took from their partners. The regressor for t6, included the parametric modulators confidence and participants' influence over their partner. For events in which the duration depended on the participants' reaction time (t2, t3 and t5), the natural logarithm of the reaction time i.e. log(RT) was included as the parametric modulator. Parametric modulators were not orthogonalized to allow the regressors to compete for explaining the variance.

Representational Similarity Analysis (RSA): We followed the standard procedure available in the RSA toolbox[38]. For each participant, we first computed a behavioural RDM. It was defined as the absolute value of the difference of deviances for all pairs of trials. This matrix served as a model RDM for following analysis. For brain RDMs, we first ran a single trial GLM at the time of making the second estimate (t5) which resulted in a beta value per voxel for each trial in which the participant made the second estimate. Then for each voxel within a brain mask we defined a spherical ROI (radius = 10 mm) and analysed data from its 100 closest neighbours. The brain RDM was obtained by computing the pairwise Euclidean distances between the 100-dimensional activity patterns of all trials. Euclidean distance was computed for the t-statistics comparing each trial to the baseline (i.e. univariate noise normalisation). Next, we obtained a correlation map by computing the correlation between the

behavioural and the brain RDMs for each participant using Kendall's tau-a. These correlation maps were submitted to a two-sided t-test for group inference. To compare this effect for human and computer conditions, we computed the correlation maps separately for human and computer conditions and tested for their difference. These maps were used for group-level inference. To investigate the neural basis of influence at the time of showing the revised estimate(t6), we computed our behaviour RDM using participants trial by trial influence over their partners using the same procedure explained above. The remaining analyses were the same as the aforementioned RSA analyses (see above). Note that we performed the same analysis to test for any multivariate effects of confidence at the social stages (t5 and t6) and influence that participants took from their partner (t5), but we did not find any significant cluster in our whole brain analysis.

Authors Contributions: A.M., H.N., C.M., and B.B., designed the experiments. A.M. collected the behavioural data, A.M. and B.B collected the neuroimaging data. A.M. carried out the data analysis. A.M., H.N., C.M., and B.B interpreted the results. A.M. drafted the manuscript and all authors contributed to the final manuscript.

Acknowledgments:

This work was supported by a PhD scholarship from the Graduate School Scholarship Program of the German Academic Exchange Service (DAAD) to (AM), Humboldt Foundation (BB) and NOMIS Foundation (BB), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 309865 - acronym: NEUROCODEC ; grant agreement No. 819040 - acronym: rid-O) (BB) the German Research Foundation (DFG, grant no INST 39/1014-1 FUGG) (CM) and the ''Struktur- und Innovationsfonds Baden-Württemberg (SI-BW)'' of the state of Baden-Württemberg (CM).

Declaration of conflict of interest: The authors declare no conflict of interest.

Data availability: The behavioral and neural data that support the findings of the experiment 2 and the code that was used to generate the findings and to conduct the experiments of this study will be provided to all readers upon request. We cannot make the raw data from experiment 1 publicly available because consent for sharing of the data was not explicitly obtained from participants.

References:

1. Folke, T., Jacobsen, C., Fleming, S. M. & De Martino, B. Explicit representation of confidence informs future value-based decisions. *Nat. Hum. Behav.* **1**, 0002 (2017).
2. Van den Berg, R. *et al.* A common mechanism underlies changes of mind about decisions and confidence. *Elife* **5**, (2016).
3. Mahmoodi, A., Bang, D., Ahmadabadi, M. N. & Bahrami, B. Learning to make collective decisions: the impact of confidence escalation. *PloS One* **8**, e81195 (2013).
4. Bang, D. *et al.* Confidence matching in group decision-making. *Nat. Hum. Behav.* **1**, 0117 (2017).
5. Bahrami, B. *et al.* Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
6. Cialdini, R. B. & Goldstein, N. J. Social influence: Compliance and conformity. *Annu Rev Psychol* **55**, 591–621 (2004).
7. 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).
8. Bang, D. & Frith, C. D. Making better decisions in groups. *R. Soc. Open Sci.* **4**, 170193 (2017).
9. 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).
10. Rouault, M., McWilliams, A., Allen, M. G. & Fleming, S. M. Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personal. Neurosci.* **1**, (2018).
11. Morales, J., Lau, H. & Fleming, S. M. Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. *J. Neurosci.* **38**, 3534–3546 (2018).
12. Lau, H. C. & Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc. Natl. Acad. Sci.* **103**, 18763–18768 (2006).
13. Bang, D. & Fleming, S. M. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc. Natl. Acad. Sci.* **115**, 6082–6087 (2018).
14. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105 (2013).

15. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18**, 1159 (2015).

16. Gherman, S. & Philiastides, M. G. Human VMPFC encodes early signatures of confidence in perceptual decisions. *bioRxiv* 224337 (2017).

17. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13**, 572–586 (2012).

18. Fleming, S. M. & Dolan, R. J. The neural basis of metacognitive ability. *Phil Trans R Soc B* **367**, 1338–1349 (2012).

19. Vaccaro, A. G. & Fleming, S. M. Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. *Brain Neurosci. Adv.* **2**, 239821281881059 (2018).

20. Qi, S., Footer, O., Camerer, C. F. & Mobbs, D. A Collaborator's Reputation Can Bias Decisions and Anxiety under Uncertainty. *J. Neurosci.* **38**, 2262–2269 (2018).

21. Heilbronner, S. R. & Hayden, B. Y. Dorsal anterior cingulate cortex: a bottom-up view. *Annu. Rev. Neurosci.* **39**, 149–170 (2016).

22. Wunderlich, K., Rangel, A. & O'Doherty, J. P. Neural computations underlying action-based decision making in the human brain. *Proc. Natl. Acad. Sci.* **106**, 17199–17204 (2009).

23. Walton, M. E., Kennerley, S. W., Bannerman, D. M., Phillips, P. E. M. & Rushworth, M. F. S. Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Netw.* **19**, 1302–1314 (2006).

24. Klein-Flugge, M. C., Kennerley, S. W., Friston, K. & Bestmann, S. Neural Signatures of Value Comparison in Human Cingulate Cortex during Decisions Requiring an Effort-Reward Trade-off. *J. Neurosci.* **36**, 10002–10015 (2016).

25. Izuma, K. & Adolphs, R. Social manipulation of preference in the human brain. *Neuron* **78**, 563–573 (2013).

26. Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A. & Fernández, G. Reinforcement learning signal predicts social conformity. *Neuron* **61**, 140–151 (2009).

27. Mahmoodi, A., Bahrami, B. & Mehring, C. Reciprocity of social influence. *Nat. Commun.* **9**, 2474 (2018).

28. Yaniv, I. & Kleinberger, E. Advice taking in decision making: Egocentric discounting and reputation formation. *Organ. Behav. Hum. Decis. Process.* **83**, 260–281 (2000).

29. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).

30. Fleming, S. M. & others. Relating introspective accuracy to individual differences in brain structure (vol 329, pg 1541, 2010). *Science* **336**, 670–670 (2012).

31. Molenberghs, P., Trautwein, F.-M., Böckler, A., Singer, T. & Kanske, P. Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study. *Soc. Cogn. Affect. Neurosci.* **11**, 1942–1951 (2016).

32. Fleming, S. M., Huijgen, J. & Dolan, R. J. Prefrontal contributions to metacognition in perceptual decision making. *J. Neurosci.* **32**, 6117–6125 (2012).

33. Cascio, C. N., Scholz, C. & Falk, E. B. Social influence and the brain: persuasion, susceptibility to influence and retransmission. *Curr. Opin. Behav. Sci.* **3**, 51–57 (2015).

34. Park, S. H., Westphal, J. D. & Stern, I. Set up for a Fall: The Insidious Effects of Flattery and Opinion Conformity toward Corporate Leaders. *Adm. Sci. Q.* **56**, 257–302 (2011).

35. Frith, C. D. Interacting Minds--A Biological Basis. *Science* **286**, 1692–1695 (1999).

36. Amodio, D. M. & Frith, C. D. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* **7**, 268–277 (2006).

37. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* (2008). doi:10.3389/neuro.06.004.2008

38. Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).

39. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* **103**, 3863–3868 (2006).

40. Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J. & Frith, C. D. How the opinion of others affects our valuation of objects. *Curr. Biol.* **20**, 1165–1170 (2010).

41. Hertz, U. *et al.* Neural Computations Underpinning The Strategic Management Of Influence In Advice Giving. *bioRxiv* 121947 (2017).

42. Chevrier, A. D., Noseworthy, M. D. & Schachar, R. Dissociation of response inhibition and performance monitoring in the stop signal task using event-related fMRI. *Hum. Brain Mapp.* **28**, 1347–1358 (2007).

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

44. Rouault, M., Dayan, P. & Fleming, S. M. Forming global estimates of self-performance from local confidence. *Nat. Commun.* **10**, (2019).

45. Fleming, S. M., van der Putten, E. J. & Daw, N. D. Neural mediators of changes of mind about perceptual decisions. *Nat. Neurosci.* **21**, 617–624 (2018).

46. Qiu, L. *et al.* The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biol.* **16**, e2004037 (2018).

47. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49**, 757–763 (2006).

48. Volz, K. G., Schubotz, R. I. & von Cramon, D. Y. Predicting events of varying probability: uncertainty investigated by fMRI. *Neuroimage* **19**, 271–280 (2003).

49. Krain, A. L. *et al.* An fMRI examination of developmental differences in the neural correlates of uncertainty and decision-making. *J. Child Psychol. Psychiatry* **47**, 1023–1030 (2006).

50. Kolling, N. *et al.* Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* **19**, 1280 (2016).

51. Shenhav, A., Cohen, J. D. & Botvinick, M. M. Dorsal anterior cingulate cortex and the value of control. *Nat. Neurosci.* **19**, 1286 (2016).

52. Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P. & O'Doherty, J. P. Neural Mechanisms Underlying Human Consensus Decision-Making. *Neuron* **86**, 591–602 (2015).

53. Bartra, O., McGuire, J. T. & Kable, J. W. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).

54. Izuma, K., Saito, D. N. & Sadato, N. Processing of the incentive for social approval in the ventral striatum during charitable donation. *J. Cogn. Neurosci.* **22**, 621–631 (2010).

55. Falk, E. B., Berkman, E. T., Mann, T., Harrison, B. & Lieberman, M. D. Predicting persuasion-induced behavior change from the brain. *J. Neurosci.* **30**, 8421–8424 (2010).

56. Falk, E. B., Morelli, S. A., Welborn, B. L., Dambacher, K. & Lieberman, M. D. Creating buzz: the neural correlates of effective message propagation. *Psychol. Sci.* **24**, 1234–1242 (2013).

57. Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S. & Cohen, J. D. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* **402**, 179 (1999).

58. Samson, D., Apperly, I. A., Chiavarino, C. & Humphreys, G. W. Left temporoparietal junction is necessary for representing someone else's belief. *Nat. Neurosci.* **7**, 499 (2004).

59. Wittmann, M. K. *et al.* Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91**, 482–493 (2016).

60. Fleck, M. S. Role of Prefrontal and Anterior Cingulate Regions in Decision-Making Processes Shared by Memory and Nonmemory Tasks. *Cereb. Cortex* **16**, 1623–1630 (2005).

61. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).

Supplementary Material:

Supplementary Methods:

Correlation between error and confidence:

To investigate potential correlation between confidence and error we applied the following linear mixed model:

$$C_t = \beta_{1s} + \beta_{2s} \times e_t \quad (1)$$

C_t and e_t correspond to the participants' confidence and error in trial t, respectively. The intercept (β_{1s}) and all slope (β_{2s}) were allowed to vary across participants by including random effects of the form $\beta_{ks} = \beta_{k0} + b_{ks}$ where $b_{ks} \sim N(0, \sigma^2)$.

Correlation between influence and confidence:

To investigate potential correlation between confidence and error we applied the following linear mixed model:

$$I_t = \beta_{1s} + \beta_{2s} \times C_t \quad (1)$$

I_t and C_t correspond to the participants' influence and confidence in trial t, respectively. The intercept (β_{1s}) and all slope (β_{2s}) were allowed to vary across participants by including random effects of the form $\beta_{ks} = \beta_{k0} + b_{ks}$ where $b_{ks} \sim N(0, \sigma^2)$.

Supplementary Figures:

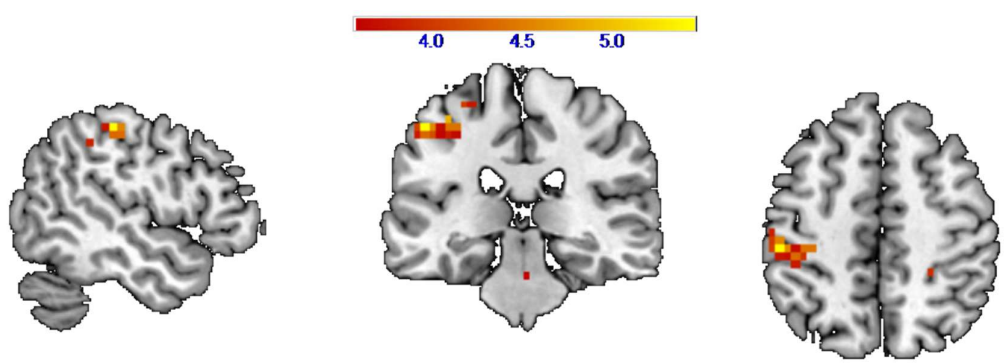

Figure S1: At whole brain, activity of left postcentral gyrus at the time of stimulus presentation (t1) was significantly modulated by the confidence. Threshold at p<.05, FEW corrected for multiple comparisons, cluster defining threshold p<.0001, N = 20 subjects.

Figure S2: At whole brain, activity of precuneus at the time of first estimate (t2) was significantly negatively modulated by the confidence. Threshold at p<.05, FEW corrected for multiple comparisons, cluster definding threshold p<.0001, N = 20 subjects.

Figure S3: At whole brain, activity of the right linual gyrus at the time of confidence rating (t3) was significantly modulated by the confidence. Threshold at p<.05, FEW corrected for multiple comparisons, cluster definding threshold p<.0001, N = 20 subjects.