Confirmation Bias in the Utilization of Others' Opinion Strength

Andreas Kappes[1], Ann Harvey[2], Terry Lohrenz[2], Read Montague[2,3] & Tali Sharot[4]

[1] Department of Psychology, City, University of London
[2]Virginia Tech Carilion Research Institute
[3]Wellcome Trust Centre for Neuroimaging, University College London
[4] Affective Brain Lab, Department of Experimental Psychology, University College London

Abstract

Humans tend to discount information that undermines past choices and judgements. This confirmation bias has significant impact on domains ranging from politics to science and education. Little is known about the mechanisms underlying this fundamental characteristic of belief formation. Here, we report a novel mechanism underlying the confirmation bias. Specifically, we provide evidence for a failure to use the *strength* of others' disconfirming opinions to alter confidence in judgements, but adequate use when opinions are confirmatory. This bias is related to reduced neural sensitivity to the strength of others' opinions in the posterior medial prefrontal cortex (pMFC) when opinions are disconfirming. Our results demonstrate that existing judgements alter the neural representation of information strength, leaving the individual less likely to alter opinions in the face of disagreement.

People are more influenced when others express judgements with high confidence than low confidence[1–5]. All else being equal, if an eye witness is confident she observed Jim stabbing George, the jury would treat such testimony as strong evidence that Jim is guilty and would be more likely to convict Jim than if the eye witness was unsure it was Jim they observed. If a doctor is confident in her diagnosis, the patient is more likely to follow the recommended treatment. There are, however, many examples in which the strength of another's opinion is dissociated from the influence it exerts. For instance, over the last decade climate scientists have expressed greater confidence that climate change is man-made. Yet, the percentage of the population that believe this notion to be true has dropped over the same period of time[6]. While there are complex, multi-layered reasons for this specific trend, such examples may be related to a bias in the utilization of the *strength* of other's opinions.

Humans are inclined to discount information that contradicts past judgments – a phenomenon known as the confirmation bias[7–10]. It is unknown, however, whether people's sensitivity to the strength of new information is contingent on whether the information confirms or disconfirms a prior judgement. In other word, does it matter less whether another's opinion is strong or weak when it is disconfirmatory than when it is confirmatory? If a juror judges Jim to be innocent would it make a difference whether the prosecutor then presents a confident witness claiming otherwise or a less confident one?

Psychological theories of moral[11] and political[12,13] judgments suggest that people automatically reject information that does not fit their prior beliefs, only engaging in reasoning subsequently when justifying decisions to others. Recent computational models[8,10] offer a similar prediction. Specifically, that committing to a certain view, for example by voting, might cause a reduction in the neural sensitivity to subsequent disconfirming evidence. If indeed the sensitivity to disconfirming evidence is reduced as compared to confirming evidence, it is likely that the *strength* of the evidence matters less when it is disconfirming than confirming, thus having less impact.

One may also theorize, however, the opposite to be true. That is, disconfirming evidence may be processed with more scrutiny[14,15], perhaps due to surprise. Heightened processing of disconfirming information has been suggested by psychological theories that assume that increased attention is needed in order to reject incoming information[16,2]. Under this theory people may be *more* sensitive to the strength of disconfirming evidence as compared to confirming evidence, which may allow rationalization of the evidence as untrue or irrelevant.

Yet, a third class of models suggest that information is processed similarly regardless of whether it confirms or disconfirms a person's belief, but the former is given more weight when making subsequent judgments (e.g., override model[18,19] or value-shift model[8]). Override models, for example, suggests that people's current beliefs do not interfere with the initial processing of information, but exert influence when judgments are subsequently expressed[18,19]. Such theories may predict equal sensitivity to evidence strength whether it is confirming or disconfirming.

We hypothesized that if differential sensitivity to the strength of others' opinions exists based on whether it is confirming or disconfirming, it would likely be observed in markers of neural activity in the posterior medial prefrontal cortex (pMFC). The pMFC, which includes the dorsal anterior cingulate cortex and pre-supplementary motor area, has been implicated in error monitoring and performance monitoring, in particular when pMFC activity is then followed by performance adjustments[20–24]. Importantly, the pMFC has been shown to track post-decision information[25] and might signal when people should switch away from a previously chosen option[26,27]. It has been further suggested that individuals with impairments in this region may display cognitive inflexibility[28].

To test whether people differentially utilize a signal of the strength of others' opinions when it contradicts or aligns with a previous judgement, we combined functional magnetic resonance imaging (fMRI) with a behavioral task in which participants were asked to re-evaluate past decisions in light of the opinions of others.

Evidence is commonly defined as information indicating whether a belief is true. In the current task the post-decision evidence offered to participants was judgements of another individual. People frequently form their own beliefs based on the opinions of others, such as the opinions

of experts, friends, family and online users[29]. Opinions of others are especially susceptible to the confirmation bias[7], perhaps because they are relatively easy to dismiss as subjective. The signal of opinion strength provided to the participants was the wager another person put on their judgement, which serves as an incentive compatible proxy for confidence. We find that participants are less likely to utilize the strength of other's opinions to re-asses their judgement when it is contradictory. This bias was related to a failure to track the strength of contradictory opinions in the pMFC, leaving the individual unlikely to alter their judgments in the face of disagreement. The findings provide a novel mechanism underlying the confirmation bias.

Results

Participants arrived in the lab in pairs and were introduced to each other before retiring to individual cubicles. They then each made 175 binary judgements about the likely asking price of properties on a well-known international real estate website (e.g., 'is this property on the market for more or less than $1,000,000?') and wagered money on their judgments (on a scale from 1 cent to 60 cents). Wagering provides an incentive compatible proxy for confidence in a judgment. Each participant was then placed in one of two fMRI scanners facing each other with a glass wall dividing the two scanners.

In the scanner participants observed all stimuli again, reminded of their past judgement and wager, and were then presented with what they believed was the judgement of the other individual (post-decision information) and the wager of that individual (a proxy of opinion strength). On 10% of the trials the partner's judgment and wager was masked. Participants then inputted their final wager. Ten trials were selected randomly at the end of the study. If the participants' judgement was correct (i.e., fit the actual asking price on the market) they would receive the final amount they wagered as a bonus, if they were incorrect they would lose that amount. Unbeknownst to the participants the judgments and wagers they observed were not in fact of their task partner, but decided by an algorithm such that on half the trials the partner appeared to agree with the participant and on half the trials to disagree (see methods for details).

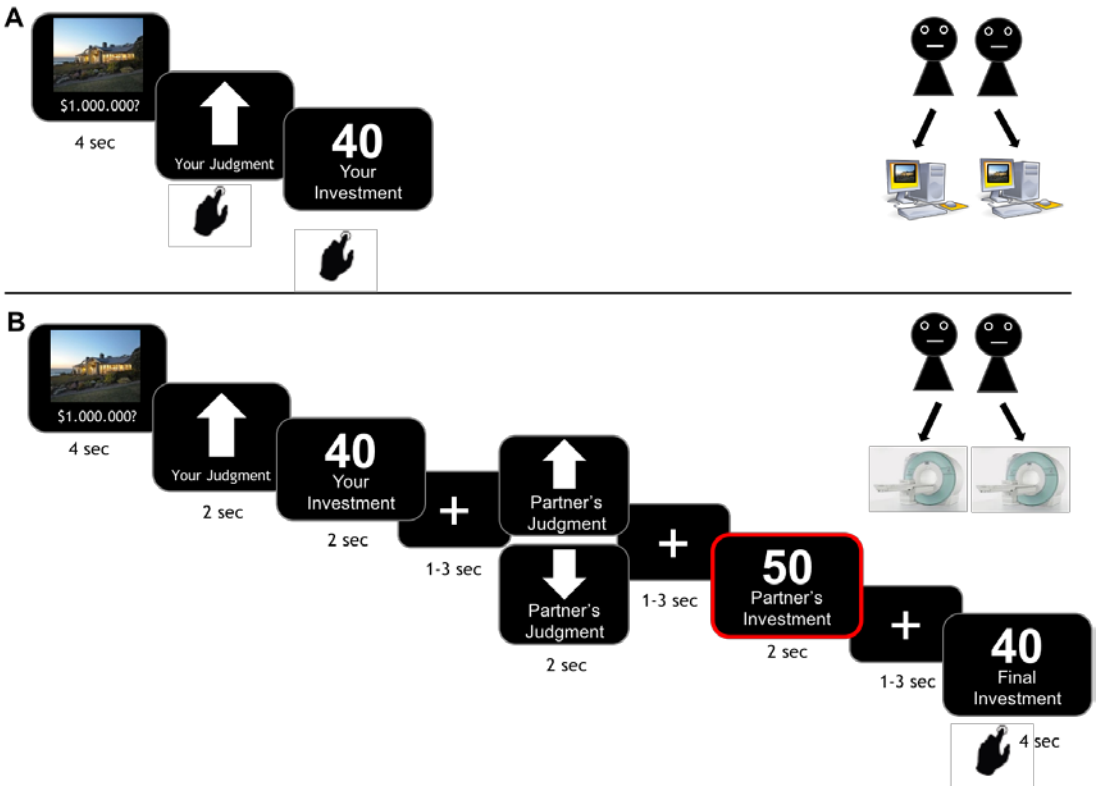

Figure 1. Experimental Paradigm. Pairs of participants completed a task that included two sessions. **(A)** In session one, participants were placed in individual cubicles and were presented with real estate photos

and prices. They were to indicate whether they believed the market price of the property on the real estate website was higher or lower than the one displayed. After making their judgment, they entered an amount between 1 cent and 60 cents to wager (invest) on their judgment. **(B)** Session two took place in two adjacent MRI scanners separated by a glass wall. On each trial participants were presented with the same photos and prices as in session one. They were reminded of their previous judgment and wager, followed by what they were led to believe was their partner's judgment and wager. They were then asked to enter a final wager. On half the trials the partner's judgment was the same as their own (i.e., confirmation) and on half the trials it was different (i.e., disconfirmation). The red outline is for demonstration purpose only - it indicates the time point of interest for fMRI analysis.

A confirmation bias in re-evaluating the accuracy of past judgements in response to another's opinion.

We first examined whether participants' responses were indicative of the classic confirmation bias. In accordance with the confirmation bias we found that confirmatory evidence (i.e., learning their partner agreed with them) had greater impact on participants' evaluation of their past judgement (as measured by change in wager) than disconformity evidence (learning their partner disagreed with them).

On trials when participants learned their partner agreed with them they increased their wager (M_{change} = 7.94 cents, significantly greater than zero $t(30) = 4.73$, $p < 0.001$) and on trials their partner disagreed with them they decreased their wager (M_{change} = 3.57 cents, significantly greater than zero $t(30) = 2.67$, $p = 0.004$). Importantly, the *magnitude* by which they altered their wager was significantly greater when their partner agreed with them than when they disagreed ($F(1,29) = 5.73$, $p = 0.008$, η^2partial = .19, **Figure 2A**). This pattern was observed in the majority of participants (**Figure 2B**), despite the fact that participants were always interacting with the same partner. In all analyses we control for participants' initial wager (see **Methods**).

On trials when the partner's opinion was not revealed, participants did not change their wager (M_{change} = 0.09, not different than zero: $t(30) = .07$, $p = 0.944$) (**Figure 2A**). The magnitude by which participants changed their wager on trials when their partner disagreed with them was not significantly different than when no opinion was provided ($F(1,29) = 2.62$, $p = 0.116$, η^2partial = 0.08). In contrast, when their partner agreed with them they increased their wager significantly more than when no information was provided ($F(1,29) = 26.15$, $p < 0.001$, η^2partial = .474). This pattern of results, which was replicated in an independent sample (**Supplementry Figure**), is consistent with a confirmation bias.

Participants utilize the *strength* of another's opinion when re-evaluating their judgments only when those opinions are confirmatory.

Thus far we have shown that opinions that support participants' prior judgements have greater impact on participants' re-evaluation of those judgements than those that contradict them. We next ask whether the *strength* of those confirming and disconfirming opinions matter. On each trial participants are exposed to their partner's wager, which provides a proxy of how confident their partner is on that specific trial (with high wager signalling greater confidence). The question is whether the partner's wager will differentially impact the participants' final wager on trials when the two agree and disagree.

We found a positive relationship between the partner's wager and participants' final wager, when the two agreed ($r_{partial}$ = 0.26, $p < 0.001$) but no significant relationship when the two disagreed ($r_{partial}$ = 0.05, $p = 0.17$), with the former relationship greater than the latter ($t(30) = 3.88$, $p = 0.001$, **Figure 2C**). This pattern, that was observed in the majority of participants (**Figure 2D**) and replicated in an independent sample (**Figure S1)**.

The results suggest that participants took into account the strength of another's opinion when re-evaluating their own judgement, but only when the opinion was confirmatory. Note, that we controlled for participants' initial wagers in this analysis (the more confident a subject was initially the less they updated their belief $r = -0.47$, $p < .001$).

One possibility is that participants paid less attention to their partner's wager when they disagreed with them. To test for this possibility we probed participants' memory for their partner's judgment and/or wager on 20 trials. There was no difference in the accuracy of

participants' memory of the partner's wager on trials in which the partner agreed or disagreed with them ($t(30) = .347$, $p = 0.73$). Thus, differential attention is an unlikely explanation. Moreover, there was no correlation between participants' memory accuracy of the partner's wager and the relation between participants' final wager and partner's wager when the partner agreed ($r_{partial} = -.47$, $p = 0.807$) or disagreed ($r_{partial} = -.074$, $p = 0.697$). Participants also recalled their partner's judgement better than chance ($t(30) = 12.91$, $p < 0.001$) in both conditions (agree, $t(30) = 2.68$, $p = 0.012$, disagree, $t(30) = 2.69$, $p = 0.012$) with no difference between the two conditions ($t(30) = 1.09$, $p = 0.32$). Thus, it is not the case that participants misremembered the partner as agreeing with them when in fact they disagreed with them. We conclude that it is unlikely that differential utilization of the strength of other's opinions is due to differential attention or memory. Instead, we speculate that participants are more likely to disregard their partner's opinion as invalid when it contradicts their own, treating fine-grained information about opinion strength as irrelevant.

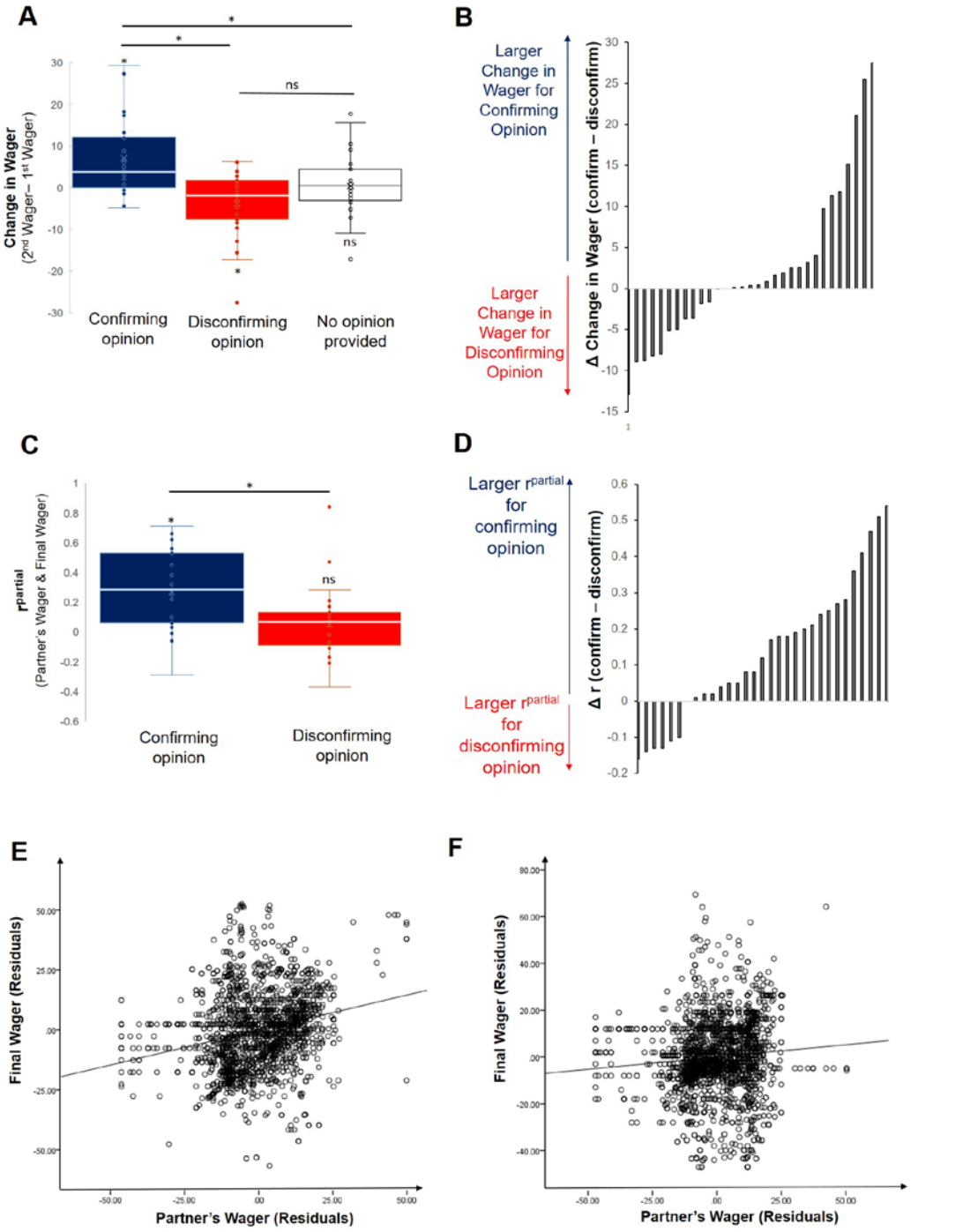

Figure 2. Participants neglect the strength of disconfirming, but not confirming, opinions. **(A)** The magnitude by which participants enhanced their wager after learning their partner confirmed their judgment was greater than the magnitude by which they decreased their wager after learning they disconfirmed (displayed are signed changes). When information about the partner's judgment was withheld there was no significant change in wager. **(B)** This pattern was observed in the majority of participants. **(C)** Participants were more likely to alter their wager in proportion to the partner's wager (controlling for initial wager) when the partner agreed with their judgment compared to when they disagreed. **(D)** This pattern was observed in the majority of participants. **(E)** For illustration purposes, we depicted the relationship between the partner's wager and participants' final wager across all trials, controlling for initial wager, when the partner agreed and **(F)** disagreed. Behavioural data in **(A & C**) are plotted as box plots for each condition in which horizontal lines indicate median values, boxes indicate 25–75% interquartile range and whiskers indicate 1.5x interquartile range; individual scores are shown separately as circles. * $p < .05$, ns = non-significant

Confirmation Bias is observed both when the partner is correct and incorrect.

The true value of the real-estate was known to us, as stimuli were extracted from well-known real-estate websites. We could thus examine if the confirmation bias was observed both on trials when the partner was correct and incorrect. Our analysis revealed it was.

First, absolute change in wager was greater on agree than disagree trials when partner was correct: (Mchange Agree = 7.5, Mchange Disagree = 3.5, $F(1,29) = 7.81$, $p = 0.009$) and incorrect (Mchange Agree = 8.05, Mchange Disagree = 0.5, $F(1,29) = 23.5$, $p < 0.001$).

Second, the correlation between the partner's wager and participant's final wager was greater on agree than disagree trials when the partner was correct (rpartial agree= 0.27, rpartial disagree= 0.11; $t(30) = 2.38$, $p = 0.024$) and incorrect (rpartial agree= 0.23, rpartial disagree= 0.06; $t(30) = 2.71$, $p = 0.01$).

Controlling for partners' accuracy (by calculating for each participant % of trials in which partner was accurate on agree trials minus on disagree trials, and adding this measure as a covariate) did not alter the confirmation bias. In particular, absolute change in participants' wager was greater on agree trials than disagree trials ($F(1,28) = 6.74$, $p = .015$); and correlation between the partner's wager and participants' final wager was greater on agree trials than disagree trials ($F(1,29) = 11.17$, $p = 0.002$).

Participants performed slightly better than chance (M correct = 52%, $p < 0.01$) and thus the partner was less likely to be correct on disagree than agree trials ($t(30) = 4.75$, $p < 0.001$). Participants seemed insensitive to their partners' accuracy. This was evident as the amount by which participants altered their wager was not different on trials in which their partner was correct (Mchange = 3.3) versus incorrect (Mchange = 3.45) ($t(30) = 0.19$, $p = 0.844$). This was true both for disagree trials (for trials in which the partner is correct Mchange = 3.6 and incorrect Mchange = 3.5, $t(30) = 0.10$, $p = 0.91$) and agree trials (correct Mchange = 7.5 and incorrect Mchange = 8.05, $t(30) = 0.96$, $p = 0.35$).

Together, these analyses show that the confirmation bias is not a function of partner's accuracy.

Reduced sensitivity to the strength of disconfirmatory (vs confirmatory) opinions in pMFC

Our behavioral results show that participants are more likely to incorporate the strength of another's opinion when evaluating the accuracy of their own judgement when that opinion aligns with their own. We next turned to our fMRI data to ask whether neural tracking of other's opinion strength was contingent on whether the opinion aligned or conflicted with one's judgement. We focused on the pMFC, which has been shown to track post-decision information. In particular, to signal the extent to which an initial decision is likely to be *in*correct given new information[30].

In our paradigm the participants learn whether their partner agrees with their judgement and then learn of their partner's wager. If a partner agrees with a participant's judgment and wagers the maximum amount that can be interpreted as a strong signal the participant is correct. However, if they agree but wager no money that is a weaker signal the participant is correct. Hence, one would expect a negative correlation between the partner's wager and activity in the pMFC at the time the partner's wager is observed, since the higher the partner's wager, the lower the likelihood that the judgment is incorrect. If a partner *disagrees with the* participant's judgment, however, and wagers the maximum amount, that can be interpreted as a strong signal the participant is incorrect. If they disagree but wager no money that is a weaker signal the participant is incorrect. Hence, one would expect a positive correlation between the partner's wager and activity in the pMFC at the time the partner's wager is observed since the higher the partner's wager, the higher the likelihood that the judgment is incorrect.

To test for the outlined interaction effect, we contrasted the BOLD parametric modulator tracking the partner's wager on agree and disagree trials. We found a significant effect in the pMFC (FWE cluster level corrected, $p < 0.0001$ after thresholding at $p < 0.0001$ uncorrected, k = 156, BA 6 and 8, peak voxel, MNI: 10, 24, 58) (**Figure 3A**). To tease apart the interaction effect we extracted the average betas in this cluster for each condition separately. We found that the interaction was characterized by a significant negative relationship between the

partner's wager and pMFC activity when the partner agreed with the participant (Beta = -.08, p < 0.001) and a non-significant positive relationship when the partner disagreed (Beta = .02, *p* = 0.19) (**Figure 3B, C**). The magnitude of these effects (i.e. comparing absolute betas in the two conditions across individuals) were significantly different from each other (t(30) = 2.37, p = 0.02). This suggests that while the pMFC tracks the strength of another's opinion when that opinion is confirmatory, it relatively fails to do so when that opinion is disconfirming.

We note that participants' own initial confidence was not tracked in the pMFC. Specifically, neither a model in which the participant's initial wager was the parameter modulating activity during the time participants observed their own wager, nor a model in which it was modulating activity at the time participants observed their partner's wager revealed effects in the pMFC (neither positive or negative effects on agree trials nor on disagree trials) even at a lenient threshold of p < 0.001 uncorrected.

An exploratory whole brain analysis revealed a second significant cluster. This was in the perigenual anterior cingulate cortex (pgACC) (BA 10, peak voxel in MNI space: 6, 52,14, k = 117, FWE cluster level corrected *p* < 0.0001 after thresholding at *p* < 0.0001 uncorrected) (**Figure 3E**). Extracting betas from this region revealed that the effect was due to BOLD signal tracking the partner's wager negatively when the partner agreed with the participant (Beta = -.07, p = 0.001), and positively when the partner disagreed (Beta = .06, *p* = 0.01) (**Figure 3F**). In contrast to our results in the pMFC, the magnitude of these effects were not significantly different from each other (t(30) = 0.27, p = 0.78), suggesting that the pgACC tracks both confirmatory and disconfirming information to a similar degree (in opposite directions). No voxels in the brain showed the inverse interaction effect.

The pMFC selectively mediates the utilization of other's opinion strength to alter one's own when there is agreement.

We next turn to ask whether the pMFC and/or the pgACC activity mediates the use of other's confidence when re-evaluating one's own confidence. In particular, we ask whether such a mediation is context specific, varying as a function of (dis)agreement.

To that end we tested for a "moderated mediation"[31] (see methods). A moderated mediation occurs when the effect of the independent variable (in our case the partner's wager) on the dependent variable (in our case the participants' final wager) via a mediator (in our case BOLD response) differs depending on a contextual factor - the moderator variable (in our case whether there is agreement/disagreement). In the first moderated mediation model we entered pMFC activity as a mediator and in a second model pgACC activity as a mediator. In order to examine the unique contributions of each region to behaviour, each of the moderated mediations (i.e. that of the pMFC and of the pgACC) were conducted while controlling for activity of the other region.

The first model (**Figure 3D**), in which pMFC activity was the mediator, revealed a significant moderated mediation. In particular, pMFC activity partially mediated the relationship between the partner's wager and the participant's final wager on agree trials (β = 0.006, t(30) = 2.07, p = 0.046, top of **Figure 3D**), but not on disagree trial (β = -0.0009, t(30) = 0.51, p = 0.61, bottom of **Figure 3D**), with the former mediation effect being significantly greater than the latter (β = -0.005, t(30) = 2.21, p = 0.035).

Consistent with the results reported in the previous section, the model highlighted a differential relationship between the partner's wager and pMFC activity on agree and disagree trials (β = 0.014, t(30) = 2.90, p = 0.007). In particular, the significant relationship between the partner's wager and pMFC activity on agree trials (β = -0.12, t(30) = 3.19, p = 0.003, top left of **Figure 3D**) was greater than the non-significant relationship on disagree trials (β = 0.005, t(30) = 0.93, p = 0.36, bottom left of **Figure 3D**). In contrast, the relationship between pMFC activity and participants' final wager did not differ on agree and disagree trials (β = 0.09, t(30) = 0.77, p = 0.448). In particular, there was a significant relationship between the two on agree trials (β = -0.26, t(30) = 2.68, p = 0.01, top right of **Figure 3D**) that was not significantly greater than the relationship on disagree trials (β = -0.19, t(30) = 1.03, p = 0.31 bottom right of **Figure 3D**).

Our second model, where pgACC was entered as a mediator, did not reveal a moderated mediation (β = -.001, t(30) = .08, p = 0.936). The pgACC did not mediate the relationship between the partner's wager and the participants' final wager on agree trials (β = 0.00025, t(30) = 0.168, p = 0.868) nor on disagree trials (β = -0.000019, t(30) = 0.006,p = 0.936).

Together, the fMRI results suggest that utilization of the strength of confirming opinions, but not disconfirming opinions, was mediated by the pMFC, but not pgACC. With the pMFC tracking the partner's wager more closely during agreement than disagreement.

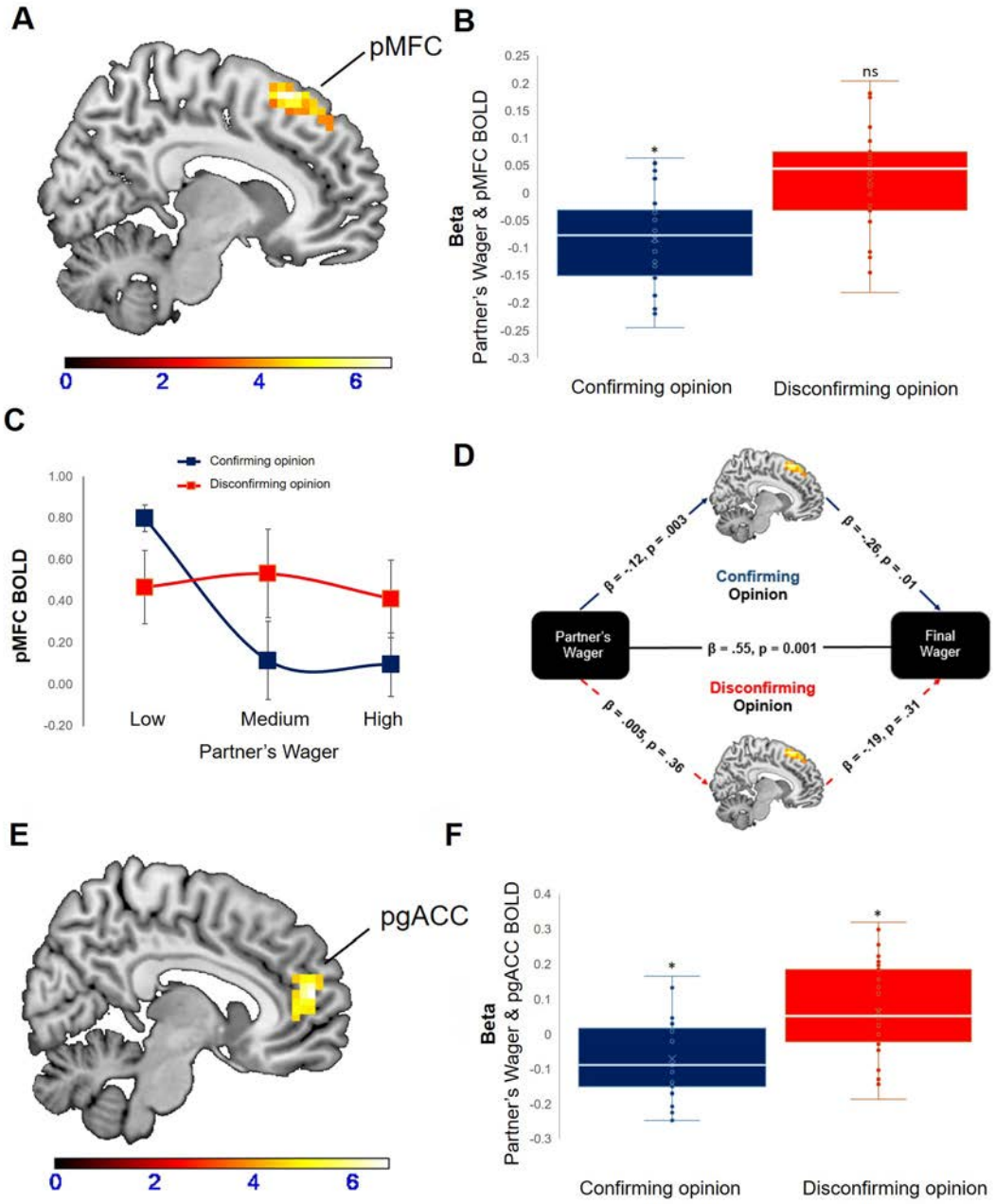

Figure 3. Reduced sensitivity to the *strength* of disconfirming (relative to confirming) opinions in the pMFC. A) An interaction effect in the pMFC between condition (agree/disagree) and a parametric modulator tracking the partner's wager at the time it is presented (k = 156, FWE cluster level corrected p < 0.0001). **B)** Extracting mean parametric betas across voxels in this cluster revealed the interaction was due to a significant negative correlation between pMFC activity and the partner's wager when the partner agreed with the participant and a non-significant positive correlation when they disagreed. The magnitude of these effects were significantly different from each other. **C)** For illustration purposes we display the mean BOLD activity across voxels in the pMFC cluster for trials in which the partner's wager was low (0-10), medium (20-30) and high (40- 50) separately for agree and disagree conditions. **D)** pMFC activity mediates the relationship between the partner's wager and final wager on agree trials but not disagree trials. **E)** An interaction effect in the pgACC between condition (agree /disagree) and a parametric modulator tracking the partner's wager (k = 117, FWE cluster level corrected p < 0.0001). **F)** Extracting mean parametric beta across voxels in this cluster revealed the interaction was due to a significant

negative correlation between pgACC activity and the partner's wager when the partner agreed with the participant and a significant positive correlation when they disagreed. The magnitude of these effects were not significantly different from each other. BOLD data in **(B & F)** are plotted as box plots for each condition in which horizontal lines indicate median values, boxes indicate 25–75% interquartile range and whiskers indicate 1.5x interquartile range; individual scores are shown separately as circles. * $p < .05$, ns = non-significant.

Discussion

The behavioural tendency to discount disconfirming information has significant implications for individuals and society as it can generate polarization and facilitate the maintenance of false beliefs[7,32,33]. Here, we characterize a novel mechanism underlying the confirmation bias. In particular we report a reduction in the use of the *strength* of others' disconfirming opinions to alter judgements. We further show that this bias is associated with reduced neural sensitivity to the strength of others' opinions in the pMFC.

Participants suitably increased their wager (which is a proxy of confidence strength) when their partner agreed with their judgement, decreased it when the partner disagreed and did not change it when the partner's opinion was unknown. Consistent with the confirmation bias, however, the impact of the partner's opinion was greater when it was confirmatory than disconfirmatory, as evident by the fact that the magnitude of wager increase when the partner agreed with the participant was greater than the magnitude of wager decrease when they disagreed.

Importantly, participants used the strength of their partner's opinion (i.e. the partner's wager) to re-assess the likelihood that their own judgment was correct when those opinions where confirmatory, but failed to do so when they were disconfirming. Utilization of the strength of confirming opinions, but not disconfirming opinions, was mediated by the pMFC, which tracked the partner's wager more closely during agreement than disagreement. These findings suggest that making a judgement diminishes the use of post-decision information strength selectively for contradictory information. The results of our memory checks suggest that this effect was not due to reduced attention or memory to disconfirming opinions. Rather, we speculate that contradictory opinions are more likely to be considered categorically wrong and thus the strength of those opinions are considered unimportant.

We focused specifically on a region of the frontal cortex, the pMFC, which is important for performance monitoring, especially in situations in which neural signal is followed by performance adjustments[20–24] and which tracks post-decision information[25]. Consistent with past results[25] we found an inverse relation between how strongly new information (in our case the partner's wager) supported a past decision and pMFC activity. This significant relationship, however, was observed only when the partner agreed with the participant, not when they disagreed. Moreover, the pMFC mediated the relationship between the partner's wager and the participants final wager when the two agreed, but not when they disagreed. Our whole-brain exploratory analysis identified another brain region that tracked the strength of other's opinions – the pgACC. The pgACC has been implicated in many functions including signalling conflict, prediction errors and affective processes[23,34–36]. In contrast to the pMFC, however, the efficacy by which the pgACC tracked the partner's wager did not differ as a function of agreement. Neither did we find that pgACC activity was mediating the influence of another's opinion strength on the participant's own on agree nor disagree trials. We thus conclude that the pMFC, but not pgACC, contributes to the confirmation bias in the use of the strength of others' opinion.

We designed a task that maximizes commitment to judgments by not allowing participants to alter their judgment, only the wager on it. This was due to past studies showing that confirmation biases are pronounced in such situations[7]. For example, in processing other's opinions about a product after it has been purchased or about a political candidate after a vote has been made. It is possible that a different pattern of results would emerge when participants are not committed to their original judgment (i.e., when a vote can be reversed or a product returned with minimal effort). Indeed, in a previous study in which participants could reverse their judgement and were incentivized for accurately assessing their past decisions a confirmation bias was not observed[25]. In that study the evidence available was not the opinion of another, but rather perceptual information. The former presumably is easier to dismiss as irrelevant (i.e.

one can easily conclude another individual is simply wrong). Because humans make the vast majority of decisions (including professional, personal, political and purchase decisions) based on information received from others, the identified bias in utilizing the strength of others' opinions is likely to have a profound effect on human behaviour.

The notion that the strength of disconfirming opinions is not necessarily proportionate to its impact on belief change is in accord with anecdotal and "real-world" observations in domains ranging from science to politics. The underlying process is remarkably flexible, with the neural circuitry involved switching on a trial by trials basis from high sensitivity to relative neglect, contingent on whether the opinion is confirmatory or disconfirming. This process may leave the individual less likely to alter opinions in the face of disagreement.

Methods

Participants

Forty-two participants (male = 20, female = 22, age = 18–38; *M* = 29.0, *SD* = 7.3 years) from Roanoke and Blacksburg, VA, area were recruited from a large database maintained by the Human Neuroimaging Laboratory. The sample size was determined using the effect size estimates from the pilot study (see SI). Data of five participants, who failed attention checks during the task (see below), were excluded leaving a sample of 37. In addition, fMRI data from six participants were not used because of insufficient coverage of the brain. Thus, fMRI and behavioural analysis are conducted on data from 31 participants. Replication study was approved by the ethics committee at UCL. fMRI study was approved by Virgina Tech Institutional Review Board.

Stimuli

We used real estate photos and prices from a Realty website. All photos depicted the exterior of a real estate in North America.

Procedure

We invited participants to play a real estate investment game (see **Figure 1**). Pairs of participants met each other immediately before the study and were given instructions. The study included two sessions, each consisting of 175 trials. In the first session participants were placed in individual cubicles. On each trial they were presented with a real estate photo and a possible price for 4 seconds. The price was either 20% higher or lower than the true asking price on the market. The participants' task was to decide whether the true price was higher or lower than the one displayed. After making their judgment, they entered an amount between 1 cent and 60 cents to wager on their judgment. If they were correct they could receive that amount, if they were incorrect they could lose that amount. Investments were made from a $6 endowment and ten trials were randomly selected at the end of the experiment for payment.

They were told in advance that in session two, which would take place in two MRI scanners separated by a glass wall, they would learn what their partner's judgment and wager was and their partner would learn of theirs. They will then have an opportunity to adjust their wager, but not their judgment. In order to ensure that participants did not hold back information or use their wagers strategically[37], they were told that on 10% of trials they would not be able to change their wager. Thus, they should always wager the sum they thought was most appropriate.

Note that using pilot data we estimated that the participants' initial wager would be around 31c on average (it was 32.66 in the main study). Thus, we allowed wagering from 1c to 60c such that on average participants would have as much room to up their wager after agreement as they would to lower their wager after disagreement if they so wished.

In the MRI scanner participants were presented on each trial again with a photo of a real estate and price for two seconds, followed by the presentation of their previous judgment for two seconds and previous wager for two seconds (**Figure 1**). Thereafter they were shown the judgment and wager of their partner for two seconds each. Finally, they had 4 seconds to enter their final wager.

In reality, we manipulated the input such that the partners' judgments confirmed that of the participant on half the trials (i.e., 75 trials), and contradicted it on the other half of the trials. On

15 trials, participants did not receive any information about either the partner's judgment or wager, but instead a row of Xs was displayed. Partner's wager was decided by a computerized script that drew randomly from a normal distribution with a mean that was either 10 cents higher or lower (*SD* = 5) than the participant's initial investment on that trial.

There were no systematic differences in participants' initial wager on trials in which the partner subsequently confirmed or disconfirmed judgments ($t(31) = .237$, $p = .814$; confirmation condition: $M = 32.28$, $SE = 2.21$, disconfirmation condition: $M = 32.18$, $SE = 2.22$), or in the partner's wager ($t(31) = .254$, $p = .80$; confirmation condition: $M = 29.99$, $SE = 1.56$, disconfirmation condition: $M = 29.82.18$, $SE = 1.74$). In all behavioral analyses we controlled for the participants' initial wager. Hence, the results reported cannot be attributed to systematic differences in either initial wager or partner's wager.

Attention check. To ensure that participants paid attention to the judgment of their partner, we probed participants' memory for the partner's judgment and wager immediately after they entered their final wager. This was done on average ten times for the partner's judgment and ten times for the partner's wager. Five participants whose memory of the partner's judgment was equal or lower than 50% (random guess is 50%) were excluded from all analysis.

fMRI data analysis

Image acquisition. The anatomical and functional imaging sessions were conducted on a 3 tesla Siemens Magnetom Trio scanners at Carilion Research Institute. High-resolution T1-weighted scans (1x1x1 mm^3) were acquired using an MP-RAGE sequence (Siemens, 176 sagittal slices)). Functional images were collected using echo-planar imaging with repetition time (TR) = 2,000ms and echo time (TE) = 25ms, flip angle = 90°, 37 slices, and voxel size = 3.4 × 3.4 × 4.0 mm. Functional data were first spike-corrected to reduce the impact of artifacts using AFNI's 3dDespike (http://afni.nimh.nih.gov/afni). Data were subsequently preprocessed with SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) for slice-timing correction using the first slice as the reference slice, motion correction, coregistration, gray/white matter segmentation, normalization to the Montreal Neurological Institute (MNI) template, and spatial smoothing using an 8mm full-width/half-maximum Gaussian kernel. Postprocessing voxels were 4 x 4 x 4 mm^3.

General linear model for standard fMRI analyses. Imaging analyses were conducted using SPM8. For each participant, the general linear model was used to model blood oxygen level-dependent (BOLD) signals during the task, incorporating an autoregressive [AR(1)] model of serial correlations and a high-pass filter at 1/128 s. The following regressors were included as stick functions, convolved with the SPM synthetic hemodynamic response function; one onset of (1) display of initial judgment and wager; (2) display of partner's judgment – separately for agree trials, disagree trials and no-information trials; (3) display of partner's wager - separately for agree trials, disagree trials and no-information trials, with the former two modulated by (4) the partner's wager; (5) display of screen prompting final wager - separately for agree and disagree trials; (6) attention check; and (7) fixation crosses. Six movement parameters were also included in the model.

Moderated Mediation analysis.

We set out to examine whether BOLD signal in the pMFC and/or pgACC mediates the effect of the partner's wager on the participant's final wager, and importantly whether this mediation is context specific (i.e., moderated). In other words, we tested if the mediation is different for agree and disagree conditions.

To that end we tested for a moderated mediation. A moderated mediation occurs when the effect of the independent variable (in our case the partner's wager) on the dependent variable (in our case final wager) via a mediator (in our case pMFC) differs depending on a contextual factor - the moderator variable (in our case whether there is agreement/disagreement).

First, following previous research[25,31,38,39], we extracted the trial-by-by-trial pMFC activation for each participant, using the pMFC cluster from the analysis displayed in Figure 3A as ROI. For each participant, we created a design matrix in which we modelled each presentation of the partner's wager (80 per condition) as a separate event (without parametric regressors attached to any of these events). In addition, we included regressors for (1) the display of initial judgment and wager, (2) display of partner's judgment (separately for agree trials, disagree trials and no-information trials), (3) display of screen prompting final wager (4) attention check and (5) fixation

crosses. Six movement parameters were also included in the model. Events were modelled as delta functions and convolved with a canonical hemodynamic response function to create regressors of interest. We then used this model to extract the BOLD signal on each trial when participants saw the partner's wager averaged across voxels in our ROI using the 'spm_summarise.m' function. BOLD signal for each presentation of the partner's wager as generated by this model was then used in our moderated mediation model. We repeated the exact same procedure for pgACC activation (ROI from Figure 3E).

We then created two moderated mediation models for each participant using the PROCESS macro for SPSS[40] - one included the signal extracted from pMFC as described above and the other from pgACC. Because we were interested in testing for unique contributions of each region to behaviour, each of the moderated mediation models (i.e. that of the pMFC and of the pgACC) were conducted while controlling for activity of the other region. In particular, using the Process toolbox a moderated mediation model (model 59) was fitted for each participant that provided the following:

1. estimates across all trials reflecting the relationship between: (i) partner's wager and final wager, (ii) partner's wager and ROI activity (iii) ROI activity and final wager.
2. the same estimates as above, but separately for only agree trials and only disagree trials.
3. estimates reflecting whether i, ii and iii in step 1 are each different for agree and disagree trials – this gives three moderation effects, each reflecting an interaction due to condition.
4. a mediation effect separately for only agree trials and only disagree trials reflecting an indirect effect between partner's wager and final wager via ROI activity.
5. estimate comparing the two indirect effects described in step 4, which reflects the moderated mediation effect.

Estimates across participants were then compared to zero using one sample t-tests.

Behavioral *Replication Study*
Before conducting our fMRI investigation, we piloted our experiment behaviorally. We tested 18 participants in pairs at the University College London. Data from one participant was lost due to a computer crash, leaving a final sample of 17 participants. The experimental paradigm was similar to the one reported in the main manuscript with the following exceptions. First, the experiment was not split into two sessions. On each trial participants observed the real estate and price, entered their judgment and wager, shown what they believe to be their partner's judgment and wager, and were asked to enter a final wager. Second, participants could wager between 1 and 99 pence. Third, participants were presented with the partner's wager that did not depend on their own wager, rather they saw a series of pre-selected wagers, ranging from 10 pence to 90 pence. Fourth, the total number of trials was only 75.

Acknowledgments. We thank Joe Marks, Filip Gesiarz, Chris Kelly, Elle Copland, Stephanie Lazzaro, Steve Fleming, and Yulin Wang for comments on previous versions of this manuscript. The research was funded by a Wellcome Trust Fellowship 214268/Z/18/Z to TS.

References

1. Bahrami, B. *et al.* Optimally Interacting Minds. *Science* **329**, 1081–1085 (2010).
2. Pulford, B. D., Colman, A. M., Buabang, E. K. & Krockow, E. M. The persuasive power of knowledge: Testing the confidence heuristic. *J. Exp. Psychol. Gen.* (2018).
3. Anderson, C., Brion, S., Moore, D. A. & Kennedy, J. A. A status-enhancement account of overconfidence. *J. Pers. Soc. Psychol.* **103**, 718–735 (2012).
4. Anderson, C. & Kilduff, G. J. Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance. *J. Pers. Soc. Psychol.* **96**, 491–503 (2009).
5. Moore, D. A. *et al.* Confidence Calibration in a Multiyear Geopolitical Forecasting Competition. *Manag. Sci.* **63**, 3552–3565 (2017).
6. Pew Research Center. The Politics of Climate. (2016).

7. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175 (1998).
8. Bronfman, Z. Z. *et al.* Decisions reduce sensitivity to subsequent information. *Proc R Soc B* **282**, 20150228 (2015).
9. Klayman, J. & Ha, Y.-W. Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* **94**, 211 (1987).
10. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Curr. Biol.* **28**, 3128-3135.e8 (2018).
11. Haidt, J. The new synthesis in moral psychology. *Science* **316**, 998–1002 (2007).
12. Taber, C. S. & Lodge, M. The illusion of choice in democratic politics: the unconscious impact of motivated political reasoning. *Polit. Psychol.* **37**, 61–85 (2016).
13. Mercier, H. & Sperber, D. Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* **34**, 57–74 (2011).
14. Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J. & Laurin, K. God and the government: testing a compensatory control mechanism for the support of external systems. *J. Pers. Soc. Psychol.* **95**, 18–35 (2008).
15. Westen, D., Blagov, P. S., Harenski, K., Kilts, C. & Hamann, S. Neural bases of motivated reasoning: an FMRI study of emotional constraints on partisan political judgment in the 2004 U.S. Presidential election. *J. Cogn. Neurosci.* **18**, 1947–1958 (2006).
16. Gilbert, D. T., Tafarodi, R. W. & Malone, P. S. You can't not believe everything you read. *J. Pers. Soc. Psychol.* **65**, 221–233 (1993).
17. Lovallo, D. & Kahneman, D. Delusions of success. *Harv. Bus. Rev.* **81**, 56–63 (2003).
18. Doll, B. B. *et al.* Reduced susceptibility to confirmation bias in schizophrenia. *Cogn. Affect. Behav. Neurosci.* **14**, 715–728 (2014).
19. Doll, B. B., Hutchison, K. E. & Frank, M. J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.* **31**, 6188–6198 (2011).
20. Yeung, N., Botvinick, M. M. & Cohen, J. D. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol. Rev.* **111**, 931–959 (2004).
21. Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624–652 (2001).
22. Botvinick, M. M., Cohen, J. D. & Carter, C. S. Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* **8**, 539–546 (2004).
23. Shackman, A. J. *et al.* The Integration of Negative Affect, Pain, and Cognitive Control in the Cingulate Cortex. *Nat. Rev. Neurosci.* **12**, 154–167 (2011).
24. Bartoli, E. *et al.* Temporal dynamics of human frontal and cingulate neural activity during conflict and cognitive control. *Cereb. Cortex* 1–15 (2017).
25. Fleming, S. M., Putten, E. J. & Daw, N. D. Neural mediators of changes of mind about perceptual decisions. *Nat. Neurosci.* 1 (2018).
26. Kolling, N. *et al.* Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* **19**, 1280 (2016).
27. Kolling, N., Behrens, T., Wittmann, M. K. & Rushworth, M. Multiple signals in anterior cingulate cortex. *Curr. Opin. Neurobiol.* **37**, 36–43 (2016).
28. Lak, A. *et al.* Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
29. Bonaccio, S. & Dalal, R. S. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organ. Behav. Hum. Decis. Process.* **101**, 127–151 (2006).
30. O'Connell, R. G. & Murphy, P. R. U-turns in the brain. *Nat. Neurosci.* **21**, 461–462 (2018).
31. Edelson, M., Dudai, Y., Dolan, R. J. & Sharot, T. Brain Substrates of Recovery from Misleading Influence. *J. Neurosci.* **34**, 7744–7753 (2014).
32. Quattrociocchi, W., Scala, A. & Sunstein, C. R. *Echo Chambers on Facebook*. (Social Science Research Network, 2016).
33. Taber, C. S. & Lodge, M. Motivated Skepticism in the Evaluation of Political Beliefs. *Am. J. Polit. Sci.* **50**, 755–769 (2006).
34. Krug, M. K. & Carter, C. S. Anterior Cingulate Cortex Contributions to Cognitive and Emotional Processing: A General Purpose Mechanism for Cognitive Control and Self-Control. in *Self Control in Society, Mind, and Brain* (eds. Hassin, R., Ochsner, K. & Trope, Y.) 3–26 (Oxford University Press, 2010).

35. Iannaccone, R. *et al.* Conflict monitoring and error processing: New insights from simultaneous EEG–fMRI. *NeuroImage* **105**, 395–407 (2015).
36. Holroyd, C. B. & Coles, M. G. H. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.* **109**, 679–709 (2002).
37. Hertz, U. *et al.* Neural computations underpinning the strategic management of influence in advice giving. *Nat. Commun.* **8**, 2191 (2017).
38. Garrett, N., Lazzaro, S. C., Ariely, D. & Sharot, T. The Brain Adapts to Dishonesty. *Nat. Neurosci.* **19**, 1727–1732 (2016).
39. Charpentier, C. J., Moutsiana, C., Garrett, N. & Sharot, T. The Brain's Temporal Dynamics from a Collective Decision to Individual Action. *J. Neurosci.* **34**, 5816–5823 (2014).
40. Hayes, A. F. Introduction to mediation, moderation, and conditional process analysis, New York. *Guilford Press* **43**, (2013).

Supplemental Figure

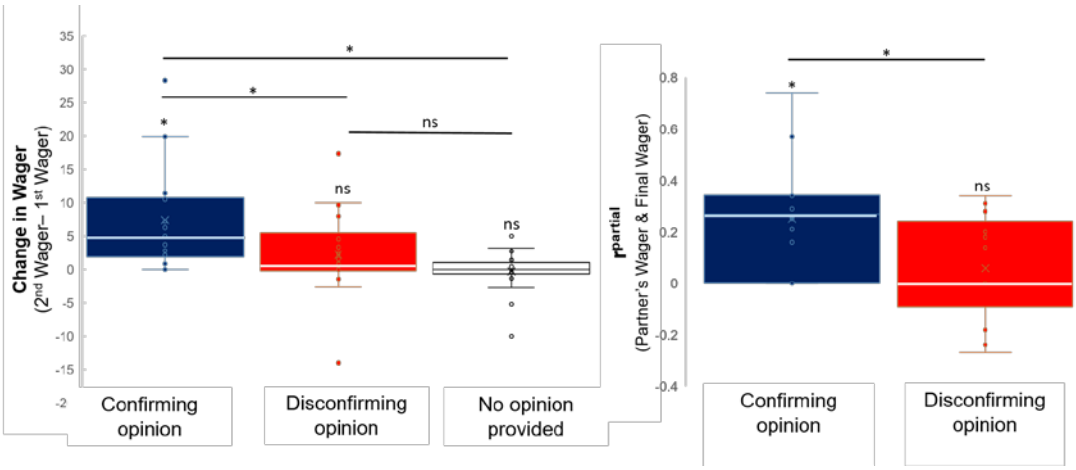

***Figure S1*. Participants relatively discount the strength of disconfirming opinions**. **(A)** When the partner's judgment confirmed the participant's, the participant increased their wager by 7.46 pence on average, which was significantly greater than zero ($t(16) = 3.77$, $p = 0.002$). When the partner's judgment disconfirmed the participant's, the participant decreased their wager by only 2.23 pence, which was not significantly different than zero ($t(16) = 1.37$, $p = 0.188$). The magnitude by which participants altered their wager was greater when the partner confirmed their judgment than when the partner disconfirmed it ($t(16) = 2.44$, $p = 0.026$). On trials when the partner's judgement was not revealed, participants decreased their wager by 0.38 cents on average, which was not different than zero ($t(16) = 0.386$, p = 0.705), significantly lower than trials on which the partner confirmed participants' opinion ($t(16) = 3.284$, $p = 0.005$) but not significantly higher than trials in which the partner disconfirmed ($t(16) = 1.129$, $p = 0.27$) participants' opinion. **(B)** When there was agreement in judgment, the greater the partner's wager the more participants would up their own wager (mean correlation between partner's wager and participant's change in wager, while controlling for participant's initial wager, $r_{partial} = 0.25$, $p = 0.001$). When the partner disagreed there was no relationship between the two ($r_{partial} = 0.05$, $p = 0.32$). The magnitude of these two relationships were significantly different ($t(12) = 2.66$, $p = 0.02$). Behavioural data in **(A & B)** are plotted as box plots for each condition in which horizontal lines indicate median values, boxes indicate 25–75% interquartile range and whiskers indicate 1.5x interquartile range; individual scores are shown separately as circles. * p < .05, ns = non-significant