nature communications

Article

https://doi.org/10.1038/s41467-024-55202-0

Asymmetric projection of introspection reveals a behavioural and neural mechanism for interindividual social coordination

Received: 28 June 2023

Accepted: 29 November 2024

Published online: 20 January 2025

Check for updates

Kentaro Miyamoto[1] ✉, Caroline Harbison[2,3,8], Shiho Tanaka[1,8], Marina Saito[1], Shuyi Luo[2,3], Sara Matsui[1], Pranav Sankhe[2], Ali Mahmoodi[2], Mingming Lin[1], Nadescha Trudel[2,3,4,5], Nicholas Shea[6,7] & Matthew F. S. Rushworth[2,3]

When we collaborate with others to tackle novel problems, we anticipate how they will perform their part of the task to coordinate behavior effectively. We might estimate how well someone else will perform by extrapolating from estimates of how well we ourselves would perform. This account predicts that our metacognitive model should make accurate predictions when projected onto people as good as, or worse than, us but not on those whose abilities exceed our own. We demonstrate just such a pattern and that it leads to worse coordination when working with people more skilled than ourselves. Metacognitive projection is associated with a specific activity pattern in anterior lateral prefrontal cortex (alPFC$_{47}$). Manipulation of alPFC$_{47}$ activity altered metacognitive projection and impaired interpersonal social coordination. By contrast, monitoring of other individuals' observable performance and outcomes is associated with a distinct pattern of activity in the posterior temporal parietal junction (TPJp).

Imagining how we ourselves may behave and how others may behave next makes cooperation between people possible and is essential for the social coordination of behaviour[1–3]. The psychological and neural mechanisms associated with how people estimate the skills of others have been investigated previously[4–8]. However, these studies typically focused on learning about others' abilities by observing the actions they made and their outcomes[8]. On the basis of these observations, and by following various heuristics, it is possible to estimate how well another person might perform in the future[7]. However, this capacity for social learning may be complemented by another ability; when we need to estimate another's ability, especially for dealing with novel problems that we have not experienced ourselves, we may try to imagine solving the problem ourselves and simulate how other people might cope with it.

Achieving this prospectively, before actually executing the task in reality, requires metacognition[9–11]. We previously found that activity in the anterior lateral prefrontal cortex (area 47; alPFC$_{47}$) predicts the chances of successfully performing a perceptual task oneself and, moreover, disrupting alPFC$_{47}$ activity diminished the accuracy of task performance predictions prior to task execution without actually affecting the performance of the task itself[12]. Here we examine whether a similar prospective metacognitive process operates when predicting the performances of others and, if it exists, how it might operate. In the present study, we hypothesised the existence of a psychological process entailing the projection of introspection of one's own task performance abilities onto others: we refer to this process as 'social metacognition.' Our proposal thus bears similarities to "experience sharing" or "simulation theory" accounts of social cognition which

[1]Laboratory for Imagination and Executive Functions, RIKEN Center for Brain Science, Wako, Japan. [2]Department of Experimental Psychology, University of Oxford, Oxford, UK. [3]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK. [4]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK. [5]Wellcome Centre for Human Neuroimaging, University College London, London, UK. [6]Institute of Philosophy, School of Advanced Study, University of London, London, UK. [7]Faculty of Philosophy, University of Oxford, Oxford, UK. [8]These authors contributed equally: Caroline Harbison, Shiho Tanaka. ✉e-mail: kentaro.miyamoto.wg@riken.jp

emphasise that our ability to infer another person's thoughts, and therefore, to predict their behaviour, is greater when we ourselves have had similar experiences to the other person in the past or when we are in some way more similar to the other person[13–15]. One's estimate of one's own behaviour provides an important reference point or anchor for estimating the thoughts or predicting the behaviour of another person if we perceive them to be similar to ourselves but not otherwise[16].

To test this hypothesis and identify the neural mechanisms that might enable social metacognition, we asked participants to predict the chances of success in a difficult perceptual task for one of two potential social partners and to compare this estimate with the chances that one might be successful oneself. If social metacognition depends on the projection of one's own abilities onto the other individual, then it should be more accurate when the partner performs the task as well as or worse than oneself; when dealing with worse or similarly performing others than, own performance provides an important reference point; we are able to simulate it[16] by reference to our own behaviour. Even though there may be some trials in which the partner performs incorrectly when we perform correctly, the combination of a simulation of what we would do in the same situation with an estimate of the difficulty of the trial provides the basis for simulating what the partner would do[5]. By contrast, when a partner is better at a task than oneself, then one's own introspective performance estimates may be of limited utility in estimating the other's likely performance (Fig. 1a) because we are unable to simulate much of what a better partner might do by reference to our own behaviour. For example, we are unable to simulate a scenario in which we would know which choice would be correct when the perceptual judgement is difficult, and we ourselves could not select the correct choice reliably. Analogously, beyond the social domain, our ability to imagine and simulate actions is anchored to the actions that we are capable of making[17,18], and we can imagine actions that we are incapable of only slowly and inaccurately. It is difficult to extrapolate from our own performance to identify situations in which a highly performing other person might succeed or fail if those situations are indistinguishable from our personal vantage point because we would probably fail in both cases. Thus, we focused on this possible asymmetry in social metacognition between our ability to make prospective estimates about the abilities of partners who are, in general, better than us or worse than us.

In contrast to previous studies on social cognition, or studies investigating a social belief constructed by learning of outcomes obtained by others[5,19], in the setting of our experiment, participants first needed to evaluate the difficulty of a task themselves. To do this, they relied on their own perception of the task first. It is for this reason that the presence of a better partner with a higher chance of success in the difficult task will not necessarily improve the overall performance of the current social task; participants themselves cannot tell how difficult the challenge posed by any task trial is for the partner.

First, we used this new approach to record activity from alPFC$_{47}$ and other brain regions. We then examined the impact of disrupting alPFC$_{47}$ activity with transcranial magnetic stimulation (TMS) to assess alPFC$_{47}$'s causal importance for forming prospective estimates of other's abilities.

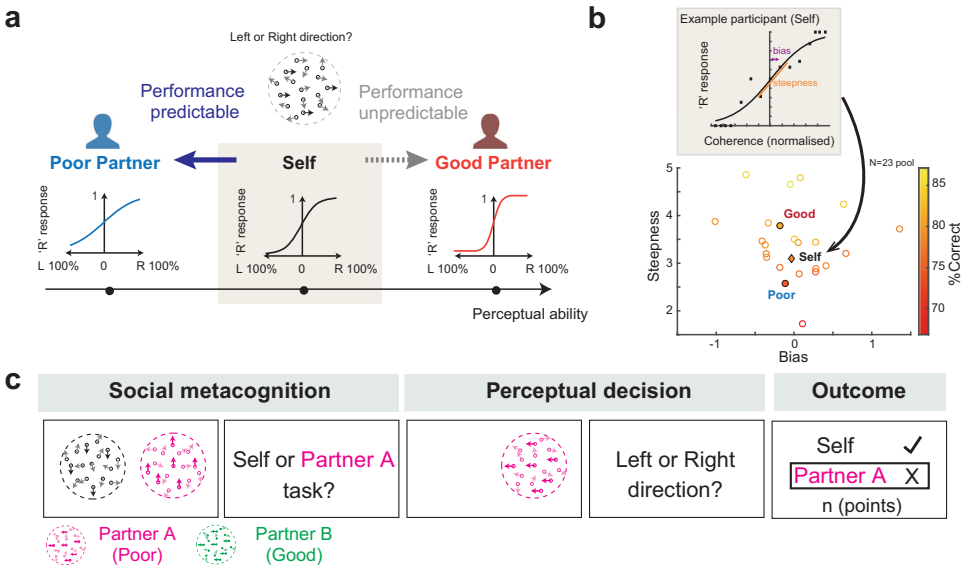

Fig. 1 | Prospective social metacognition task: asymmetric introspection for performance prediction between worse and better partners. a Prediction of perceptual performance by a poorer-skilled person than the self is possible by adapting and adjusting predictions about performance that might be made for the same problem when performed by the self (left, Poor Partner). However, this is not possible to the same degree when the partner is a better-skilled person (right, Good Partner). **b** Participants first observed and learned how well two other persons performed a random-dot motion (RDM) task: one performed the task better (Good Partner), and the other performed the task worse (Poor Partner) than the participants themselves. Good and Poor Partner performances were selected from a pool of behavioural data obtained from 23 other people who had performed the same task previously[12]. Good and Poor Partners were picked so as to ensure that steepnesses of the slopes of their psychometric functions for the RDM task were, respectively, greater and lower than for the participant. **c** Example task sequence. Participants were asked to make a prospective decision in the social metacognitive judgement stage: they decided whether they wanted to perform one task themselves and to receive a reward based on their own performance (left white RDM), or wanted to choose the other task, to be performed by a partner, with their reward a function of the partner's performance (right coloured RDM) in the following perceptual decision stage. Two partner identities were specified by two colours, green and pink (different colours were assigned to the Good and Poor Partner with counterbalancing across participants; participants were not explicitly informed that one partner performed better and one performed worse than the participants themselves). Even when they chose the partner task, the participant was also required to perform the same task themselves, and performance feedback for self-performance was presented in the final outcome stage of the trial along with performance feedback for the partner. When the self task [or partner task] was chosen then only the outcome for self [partner] performance contributed to accumulated total reward even though the partner's [self] outcome was visible.

Results

Metacognitive judgements on self and other's probabilities

Participants ($N = 27$) performed a social metacognitive matching task employing random dot motion (RDM) stimuli. Participants made left or right key presses depending on the direction in which most dots moved (the "coherent" dot direction). Because it was easier to judge the direction of dot motion when coherence was high, participants were more likely to make correct judgements when coherence was high. Therefore, the coherence of the synchronised dot motions indirectly determined the "probability" of whether a reward would be received as a function of each participant's ability. Prior to making this perceptual decision, during a metacognition stage, participants had the opportunity to choose which one of two decision problems would be attempted (Fig. 1c). One option was a problem to be solved by the participants themselves (self-task). If they chose this task, the reward outcome depended on their own performance on the problem. The other option was a problem to be solved by their partner. If they chose this task, the reward outcome depended on their partner's performance on the problem (partner's task). Importantly, the problems for the self and the partner had different coherence levels, and these changed independently across trials. The reward for correct performance when either the participant themselves or the partner performed the task was the allocation of a single point to the participant. Thus, on each trial, participants were simply incentivized to select the player, either themselves or the partner, who was most likely to perform their task correctly.

During the task participants encountered two different partners (although only one partner was presented on each trial). One of the partners, in general, performed the task very well, while the other performed very poorly relative to the participants themselves. Prior to this main task, the participants observed and learned about how the two partners performed the RDM with different coherence levels. To construct the two partners, we used data from various pairs of human participants who performed the same task in a previous study[12] (Fig. 1b). The partners were selected so that one partner performed better and the other partner performed worse than the participant. We chose the better and worse performing partners on the basis of the slopes of their psychometric functions, after fitting them to performance in the RDM task. The potential partner slopes were then compared with the participants' own psychometric slopes (Fig. 1b, "Methods"). Participants were not instructed about the skill levels of these two partners before the test; that is, they did not know before the start of the test that one was better and one worse than themselves. All they were told was that one partner identity was indicated by pink dots and the other partner identity was indicated by green dots and the link between dot colour and partner identity would remain constant during the experiment.

Our hypothesis was that participants would be able to predict the poor partner's performance for a particular RDM by using their metacognitive skills; they would make a metacognitive estimate of their own ability to solve the RDM and this insight into their own likely performance for the same RDM coherence level might be used to derive an estimate of the poor partner's performance. This would be possible because the participants themselves would be able to make accurate estimates of the difficulty of all the trials that the poor partner could perform and also of many of the trials that exceeded the partner's ability to perform well. In contrast, we predicted that participants would be less able to predict the good partner's performance because they would have only limited metacognitive insight into the good partner's performance. The participant's limited ability to estimate the difficulty of trials that were both within and beyond the good partner's ability would make it difficult for the participant to estimate precisely which trials the partner would perform well. Thus, we expect that a comparison of brain activity associated with evaluating the poor partner's probability of success and the good partner's probability of success should reveal the neural substrate for social prospective metacognition when this is achieved by projecting one's own metacognitive insights into one's own performance levels onto other people.

For all the combinations resulting from the eight different levels of coherence for the self and for the partner's tasks, we calculated the proportion of trials that a self-task option was chosen by the participants as opposed to a good or poor partner option (Fig. 2a). Logistic multiple regression analyses revealed that participants' preferences for the self-task option increased in proportion to the motion coherence (which systematically relates to probability of success; see Methods for conversion of RDM coherence into probability for the participant, or partner, to make a correct motion direction judgement) for the self-task (good partner: $\beta_{self} = 0.98 \pm 0.082$ [mean±s.e.m.]; $p < 0.001$ [t test against zero]; poor partner: $\beta_{self} = 0.99 \pm 0.11$; $p < 0.001$) and decreased in proportion to the probability that the partner would judge their problem correctly (good partner: $\beta_{self} = -0.39 \pm 0.080$; $p < 0.001$; poor partner: $\beta_{self} = -0.66 \pm 0.12$; $p < 0.001$; Fig. 2b).

The size of the regression coefficient for the probability of correct performance on the self-task option (self p(correct)) reflects the influence the participant's estimate of their own probability of performing the trial correctly has on their task selection: whether they choose to tackle a problem or whether they choose for the partner to tackle a problem. This estimate had a similar impact on decisions taken by the participants when they were deciding between themselves and the good partner or between themselves and the poor partner; self p(correct) had a similar influence on decisions involving either the good or poor partner ($t_{25} = 0.056$, $p = 0.95$). In a complementary fashion, the influence of the probability that the partner would be correct (partner p(correct)) on whether or not the participant should choose to perform their task should be negative and, as noted, this was indeed the case. However, remarkably, the size of the negative regression coefficient for partner p(correct) was significantly larger for the poor partner than for the good partner ($t_{25} = 2.75$, $p = 0.010$). Thus, differences between good and poor partner trials are not general in nature and they are not related to a difference in how participants take their own potential performance into consideration. Instead, differences in poor and good partner trials are specifically related to how participants take into consideration poor partner performance versus good partner performance.

The differential effects of partner identity on the size of the negative regression coefficient for partner p(correct) was also apparent in a regression analysis performed on behavioural data taken from Good and Poor Partners together but employing a binary regressor encoding the partner's identity (Poor Partner: 1, Good Partner:−1) and its interaction with self probability and with the partner's probability (the interaction between the partner's identity and partner's coherence: $\beta = -0.10 \pm 0.047$; $p = 0.032$). This analysis was, in fact, conducted prior to the aforementioned separate regression analyses (see "Methods" in detail; Supplementary Fig. 1d). These results suggest that when participants made social metacognitive judgements and chose whether to tackle a problem themselves or to let the partner tackle the problem, they were more informed by the probability of the poor partner performing correctly than the good partner performing correctly. This is consistent with the participant having less insight into which problems can really be solved by the good, as opposed to the poor partner. A further indication that the good and poor partners were treated differently was that feedback about the success or failure of the good partner's and poor partner's choices exerted different types of influence over subsequent social metacognitive judgements made by the participant. That is, the participants' decisions were affected by the previous outcomes of self and poor partner choices, but not of good partner choices (Supplementary Fig. 1a). Moreover, participants learned the relationship between outcomes in the most

https://doi.org/10.1038/s41467-024-55202-0

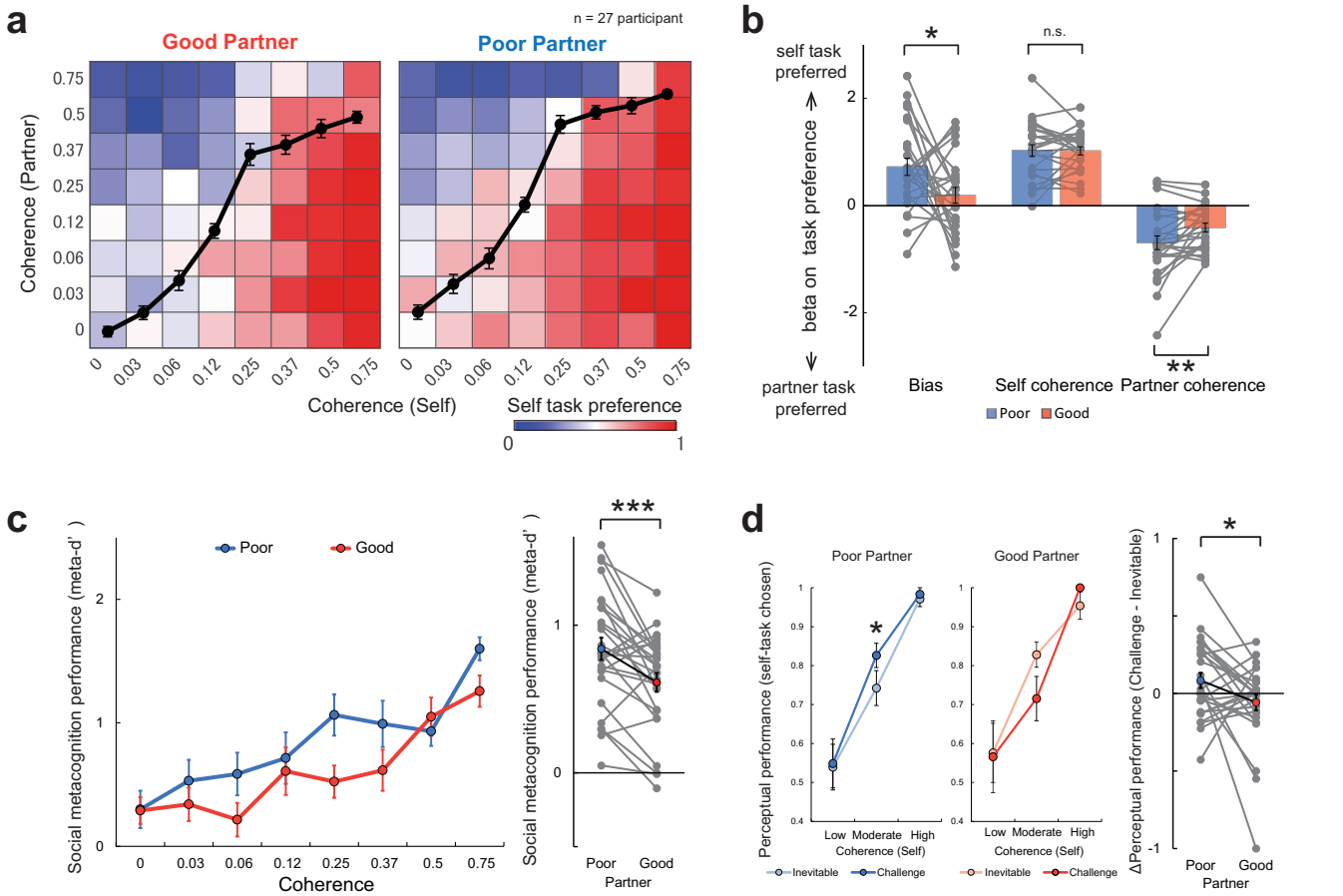

Fig. 2 | More appropriate task selection occurs between the self and a poor partner than between the self and a good partner. a The proportion of trials on which the self task was chosen by participants ($N = 27$) for each pair of coherences (one for self and one for partner's task) for trials with the Good (left) or Poor Partner (right). Participants' choices of the self-task in the social metacognitive judgement stage increased when either the self-task coherence increased, or the partner's task coherence decreased. The overlaid black line indicates when self and other performances were equated. **b** Influence on the choice made in the social metacognitive judgement stage by self and partner's task coherence ($N = 26$ in regression analysis). Multiple regression analyses were performed separately for trials in which the participant was paired with the poor (blue) and good partner (red). The participants generally preferred to choose the self-task on trials when they were paired with a poor rather than good partner (left, bias). Comparable significant positive effects of self-task coherence on self-task preference were observed in trials in which the participant was paired with either a poor or good partner (middle, self-coherence). However, the negative effect of the partner's task coherence was significantly larger for the poor than the good partner. $^{*}p = 0.042$, $^{**}p = 0.010$, paired t test. **c** Optimality of social metacognition when choosing whether tasks should be performed by the self or the partner ($N = 27$). The meta-d' is larger (judgement is more optimal) when coordinating with the Poor as opposed to the Good Partner. $^{***}P = 0.00067$, paired t test. **d** Perceptual decision performance on the self-task was higher when the self-task had been paired with a Poor Partner's task that was, on average, more likely to yield reward (Challenge trials) than when the self-task was paired with a Poor Partner's task less likely, on average, to yield reward (Inevitable trials) ($N = 27$). In contrast, a similar effect was not observed when metacognitive judgements were made between the self and the good partner's task. $^{*}P = 0.045$, paired t-test. The mean and SEM error bars are displayed in panels (**a**–**d**).

recent trial and p(correct) for the poor partner but not for the self or good partner ($\beta_{\text{outcome} \times \text{poor partner p(correct) [trial n-1]}} = -0.16 \pm 0.06$; $p = 0.014$) (Supplementary Fig. 1b). In addition to this approach, we also used an alternative approach for ascertaining the p(correct) estimate held by participants for themselves and for partners, but the conclusions drawn with these additional approaches remained the same (Supplementary Fig. 2a, b). More generally, the higher probability of real outcomes for the chosen task than expected outcomes for the unchosen task suggests that the participants could prospectively estimate the self and the partners' performance (Supplementary Fig. 2c).

The accuracy of metacognitive decision-making is typically described by a type II signal detection theory-based index (meta-d')[20]. We calculated meta-d' based on the equilibrium line for the performance of the self and partner as shown by the black trace in Fig. 2a (see Methods for more details of calculation of meta-d' in a social task). The meta-d' for optimal judgements of whether participants themselves or the poor partners should tackle a problem is significantly higher than that for judgements of whether the partners themselves or the good partners should tackle a problem across the different levels of self-probability (two-way ANOVA [8 self-probability levels × 2 partner identity]: main effect of partner, $F_{1,26} = 14.88$, $p = 0.0007$; main effect of self-probability, $F_{7182} = 10.50$, $p < 0.0001$; interaction, $F_{7182} = 1.34$, $p = 0.21$. Problems with Poor Partner vs. Good Partner, $t_{26} = 3.85$, $p = 0.00067$, effect size Cohen's $d = 0.61$) (Fig. 2c). Importantly, the asymmetric social metacognition abilities indicated by the meta-d' levels cannot be explained by differences in choice bias between conditions with good and poor partners (Supplementary Fig. 2d, e). Another possible explanation for the asymmetric social metacognition prediction is the participants' limited ability to predict the probability of success for difficult tasks with low RDM coherence as compared to easy tasks with high RDM coherence. If the participant cannot assess differences between two difficult trials with low coherences that are beyond their ability to discriminate, then this inability will similarly limit how well they can construct a model of how the Good Partner might do so. We examined whether participants could reliably

distinguish between two tasks that both had low coherence levels; we assessed how reliably participants could distinguish between the tasks and whether they could reliably pick the slightly easier one (Supplementary Fig. 3). We found that participants are worse at judging low coherence levels and they even worse when these judgements are made in the context of discrimination between stimuli that are also close together in coherence level ($\beta_{\text{interaction(left} \times \text{left-right)}} = 0.78 \pm 0.083$, $t_{22} = 7.50$, $p = 1.6 \times 10^{-7}$). These observations suggest that the participants' poorer performance in the metacognition task with Good Partners might be attributed to their inability to model and predict the Good Partner's performance in trials with low coherence. Additional analyses are presented in Supplementary Fig. 4c.

Another possible interpretation that might be put forward might be that participants' judgements about both the Poor and Good partners simply reflected their estimates of how well they themselves would do on the task. However, according to this view, while metacognitive judgements should be worse for Good Partners in difficult trials, it also suggests that they should also be worse for Poor Partners in easy trials (because these trials would sometimes be performed well by the Poor Partner and sometimes poorly by the Poor Partner but they would nearly always be performed well by the participant themselves). However, this was not the case; there was no subset of trials on which judgements involving good partners were performed better than judgements involving poor partners when we divided trials into high-coherence easy trials and low-coherence difficult trials (Supplementary Fig. 4d and see also Supplementary Fig. 4e). In summary, metacognitive decision accuracy was higher when participants were making choices to coordinate their own and a poor partner's performances than when they were making choices to coordinate their own and a good partner's performances.

If social metacognitive judgements gave participants the opportunity to select problems for themselves that they correctly and appropriately realised they were likely to succeed in performing, then this should be apparent when comparing the rates at which participants performed self-decisions correctly on two types of trials we refer to as 'challenge' and 'inevitable' trials[12]. In challenge trials, the partner's option was, on average, linked to a higher probability of reward than the self option. Inevitable trials are ones on which the partner's option was, on average, linked to a lower probability of reward than the self option. In metacognitive judgement tasks that lack any social component but allow participants to decide which problems to tackle, it has been shown that self-task performance on challenge trials is indeed better than on inevitable trials[12]. Better performance on challenge trials may be a consequence of metacognitive insight into slight changes in stimulus features that make a particular stimulus easier than might be expected on average given the coherence level. If the participants can estimate the likely performance of a partner correctly in a similar manner, then a similar challenge trial advantage should be seen; participants should choose the self-task in challenge trials only when they are very confident. Accordingly, self-task performance in challenge trials should be higher than that in inevitable trials. The increase in the performance of challenge trials compared with inevitable trials was larger in the task with the poor as opposed to the good partner ($t_{26} = 2.09$, $p = 0.045$) (Fig. 2d). If the participants decide to take the self-task on challenge trials simply because they are more motivated, then performance after choosing the self task is not expected to be different between challenge and inevitable trials; a higher motivation level will not necessarily lead to greater ability and a higher performance level specifically for a poor partner's trials but not for a good partner's trials. This means that participants are better able to select challenging trials based on insight into their own and into the poor partner's performances than when they make judgements based on estimates of the good partner's performance (see also Supplementary Fig. 1c).

These results show consistently that participants can estimate how others perform a task when the others perform similarly or more poorly than themselves, perhaps by introspection into their own performance levels. However, participants are less able to predict how well better performers will do on any given trial.

Metacognitive evaluation and matching of the evidence for self and poor partner in the anterior lateral prefrontal cortex

To search for neural activity linked selectively to self-probability estimation relative to both the good and poor partners' probabilities, we sought brain activity modulated more significantly by self p(correct) during social prospective metacognitive judgements (fMRI-GLM1; see "Methods" for details) (Fig. 3a). Activity correlated with self-probability when the self-option was chosen was most prominent in a left anterior lateral prefrontal area ([x,y,z] = [−42, 38, −8], $Z = 4.03$, cluster size = 147 voxels; $p < 0.05$, cluster-level corrected [$z > 3.1$]). The cluster, with a peak in area 47, overlapped with a cluster that we identified as a site for prospective metacognitive comparison between self p(correct) and the probability of receiving a reward associated with an environmental stimulus ('external probability') in a previous study ($alPFC_{47}$: [x, y, z] = [−38, 40, −10])[12]. The previous study found that $alPFC_{47}$ activity rose during a metacognitive judgement in proportion to self p(correct), but it was more concerned with metacognitive assessment of one's own potential performance levels rather than external probabilities – the probability that a stimulus would yield reward if chosen. Thus, we focused on activity in $alPFC_{47}$ and examined whether $alPFC_{47}$ activity carried any information about another person's chances of performing a trial correctly – partner p(correct). We were particularly interested in whether $alPFC_{47}$ carried estimates of partner p(correct) in the same way that it carried estimates of self p(correct) or if partner p(correct) estimates had only a minimal impact on $alPFC_{47}$ activity as had been the case for external probabilities.

First, we confirmed that activity in $alPFC_{47}$ increased with self p(correct) when the self-option task was chosen. To avoid any circularity in the analysis, mean regression weight relative to social metacognition onset was tested not just in a region of interest [ROI] that corresponded to the one first defined by Miyamoto and colleagues[12] but also with the same time window of 4–11 s. We then performed a further post-hoc analysis to test whether effects were present in the first half or the second half of the 4–11 s time period. We found that there was a significant effect in the window where two peaks reside at early [4–7.5 s] and late [7.5–11 s] periods ($\beta_{\text{self}} = 0.048 \pm 0.014$, $t_{26} = 3.34$, $p = 0.0025$, t test against zero). By contrast, $alPFC_{47}$ did not correlate significantly with partner p(correct) when both good or poor partners were considered together ($\beta_{\text{partner}} = -0.0082 \pm 0.018$, $t_{26} = -0.43$, $p = 0.66$) (ROI-GLM1; see "Methods" for details) (Fig. 3b; see also Supplementary Fig. 5a). However, a significant difference in the effect (self vs. partner, $t_{26} = 2.36$, $p = 0.025$) suggests that $alPFC_{47}$ was modulated by the difference between self p(correct) on trials where the self task was ultimately chosen and partner p(correct) on trials where the partner's task was ultimately chosen. Further analysis revealed that the self p(correct) versus partner p(correct) effect was apparent when judgements were made between the self and a poor partner (self vs. poor partner, $t_{26} = 3.35$, $p = 0.0024$) but not with a good partner (self vs. good partner, $t_{26} = 0.57$, $p = 0.56$) (Supplementary Fig. 5b). Activity in $alPFC_{47}$ increasingly reflected the key decision variable that should determine the metacognitive judgement, self p(correct) – partner p(correct), when making judgements between the self and a poor partner but not between the self and a good partner (self – poor partner vs. self – good partner; early phase [4-7.5 s]: $t_{26} = 0.92$, $p = 0.36$; late phase [7.5–11 s]: $t_{26} = 2.52$, $p = 0.018$). Once the haemodynamic lag is taken into account, this suggests that a signal associated with evidence for selecting the self-task is accumulated when it can be compared with evidence for selecting the poor partner's task, which can be estimated by extrapolation from self-introspection.

https://doi.org/10.1038/s41467-024-55202-0

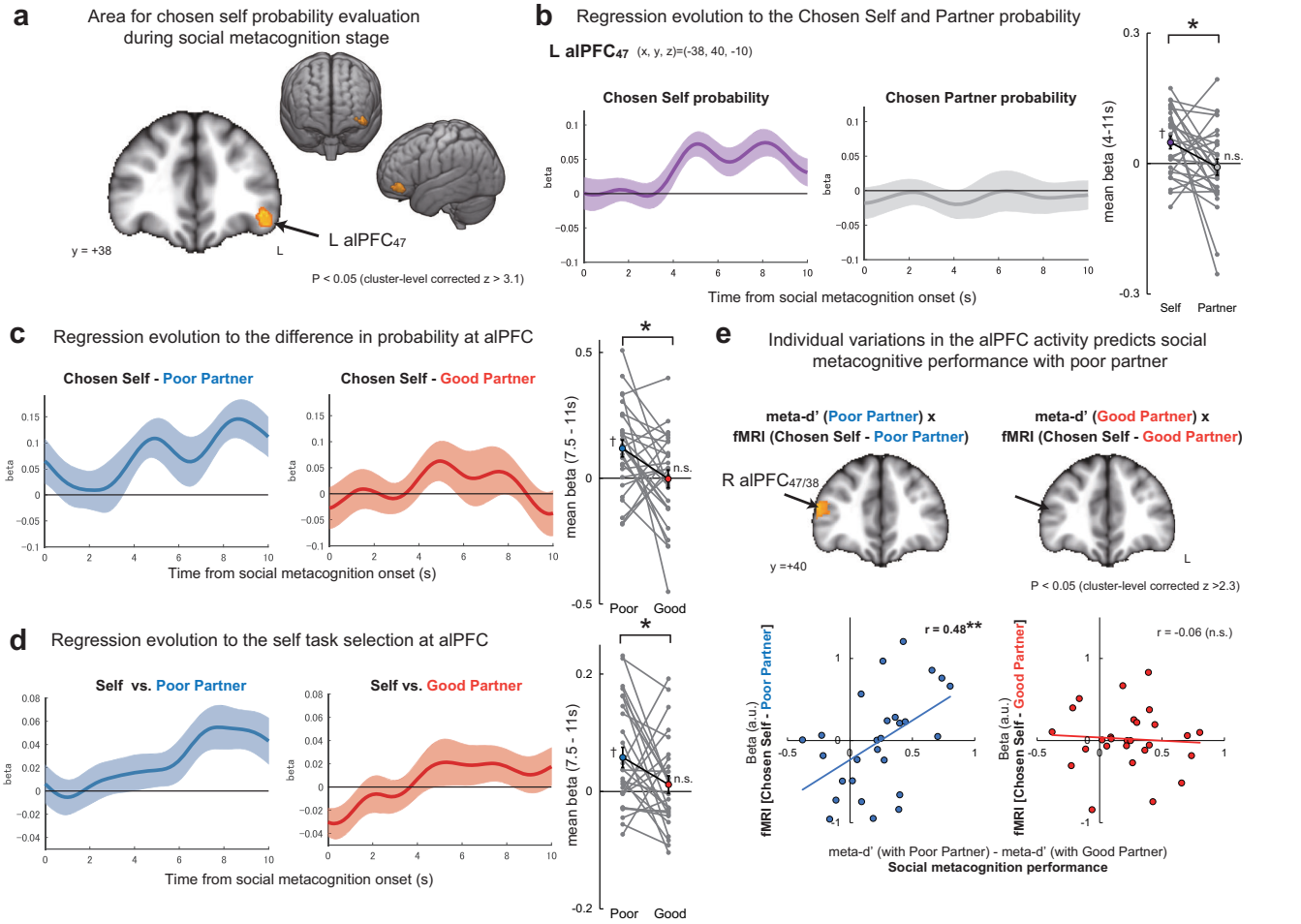

Fig. 3 | The anterior lateral prefrontal cortex (alPFC$_{47}$) is active during prospective social metacognitive comparison between the self and the poor partner. a Activity in left alPFC$_{47}$ reflected probability of reward associated with the self task when it was chosen during the metacognitive judgement stage ($N = 27$; whole-brain effects family-wise error cluster-corrected; $z > 3.1$ and $p < 0.05$, two-sided). **b** Evolution of regression weights across time indexing impact on the BOLD signal of self (purple) and partner (grey) when it was chosen on left alPFC$_{47}$: ROI ([x, y, z] = [−38, 40, −10] based on previous work[12]) (left). Beta values for self-task coherence, when it was chosen, and partner's task coherence, when it was chosen, were significantly different. *$p = 0.025$, paired t test. †$p = 0.0025$, t-test against zero (right). **c** Evolution of the regression weights at alPFC$_{47}$ as a function of the coherence of the self or partner's task when each was chosen. Data are illustrated separately for trials with the Poor (blue) and Good Partner (red) (left). Mean beta values (in the late BOLD response period) were significantly different between trials with the Poor and Good Partner. *$p = 0.018$, paired t test. †$p = 0.0016$, t test against zero (right). The mean and SEM error bars are displayed in panels (**b**, **c**). **d** Evolution of BOLD regression weights at alPFC$_{47}$ in response to selection of the self task in trials with a poor (blue) or good partner (red) (left). Mean beta values, reflecting late response after social metacognition stage onset, were significantly different between trials with poor and good partners. *$p = 0.037$, paired t-test. †$p = 0.0023$, t test against zero (right). **e** Individual variation in activity in the vicinity of alPFC$_{47}$ in response to 'self task coherence − partner's task coherence' was associated with individual variation in social metacognition performance (meta-d') for trials with the Poor (left) but not the Good Partner (right). The focus of activity in alPFC$_{47/38}$ ([x, y, z] = [38, 20, −14], $Z = 3.55$; $p < 0.05$, cluster-level corrected [$z > 3.1$], one-sided) was within 20 mm of right alPFC$_{47}$[12]. The effect is illustrated at $z > 2.3$ and $p < 0.05$ for ease of visualisation.

This suggests that alPFC$_{47}$ plays a role in metacognitively comparing self p(correct) and partner p(correct) only when the participants can imagine and estimate the performances of both the self and the partner (Fig. 3c and Supplementary Fig. 5b).

These results were reproduced when we used two complementary analyses. In the first, we used an alternative approach for ascertaining the p(correct) estimate held by participants and/or partners (Supplementary Fig. 5c, d). Moreover, the activity in alPFC$_{47}$ during the late phase predicted participants' selection of the self task option against a poor partner ($\beta_{\text{self vs poor partner}} = 0.058 \pm 0.017$, $t_{26} = 3.28$, $p = 0.0023$) but not against a good partner ($\beta_{\text{self vs good partner}} = 0.011 \pm 0.015$, $t_{26} = 0.77$, $p = 0.44$; difference in the effect between poor and good partners: $t_{26} = 2.19$, $p = 0.037$) (ROI-GLM2; see "Methods" for details) (Fig. 3d and Supplementary Fig. 5b). In the second complementary approach, we focused on a subset of trials in which the difference between the participant and each of the partner's performance levels were matched; the difference in the alPFC's signal between trials with Poor and Good Partners was reproduced even when we focused on trials in which the coherence level of the self and partners' stimuli were both low (coherence: 0, 0.03, 0.06, 0.12) and so the difference in performance between the participant and the partner was of a comparable size for both good and poor partners ($t_{26} = 2.07$, $p = 0.048$) (Supplementary Fig. 5e; for the feature of trials we used for the analysis, see also Supplementary Fig. 2f). This observation suggests that the difference in alPFC's activity on good and poor partner trials cannot be simply explained by a greater similarity between performance levels between the self and one of the partners.

Next, we searched across the whole brain for any activity predicting variation in the sensitivity of social metacognitive judgement across participants by employing an analysis of covariance (ANCOVA) (fMRI-GLM2, fMRI-GLM3; see "Methods" for details). We found that individual variation in activity in response to self p(correct) against

poor partner p(correct) at the border of the right alPFC$_{47}$ and temporal pole close to area 38[21] was correlated with individual variation in social metacognitive accuracy (meta-d'$_{Poor}$) ([x, y, z] = [38, 20, −14], Z = 3.55; p < 0.05, cluster-level corrected [z > 3.1] within 20 mm of the peak at alPFC$_{47}$; hereafter we refer to this area as alPFC$_{47/38}$) (Fig. 3e, top left panel). In contrast, there was no brain activity in which the individual variation in response to self p(correct) versus good partner p(correct) predicted individual variation in meta-d'$_{good}$ (Fig. 3e, top right panel). The significant difference between the predictability of individual variation in social metacognitive performance (meta-d'$_{poor}$ − meta-d'$_{good}$) by individual variation in alPFC$_{47/38}$ activity for comparisons with a poor partner (r = 0.48, p = 0.0099) and a good partner (r = −0.065, p = 0.74) (poor vs. good: ΔFisher's z = 2.03, p = 0.021) confirmed that alPFC$_{47}$ is involved in prospective metacognition in a social setting only when participants are able to use their own metacognitive insights to predict other's performance (Fig. 3e bottom panels).

Estimation of perceptual skills based on social heuristics at the posterior temporoparietal junction

Our behavioural results demonstrated that participants could predict a good partner's performance even though their ability to do so was generally worse than their ability to predict a poor partner's performance (Fig. 2b, c). We, therefore, hypothesised that there should be some neural substrate to estimate and evaluate others' performance, and thus, we searched for brain areas in which activity was correlated with partners' p(correct) during the social metacognition stage of each trial. To search for neural activity linked selectively to partner p(correct) estimation, we sought brain activity modulated more significantly by both partners' (good and poor partners') p(correct) during prospective metacognitive judgements (fMRI-GLM1). A cluster of activity arose in correlation with partner p(correct) in the ventral temporoparietal junction (peak at [x, y, z] = [48, −72, 14], Z = 3.55, cluster size = 180 voxels; p < 0.05, cluster-level corrected [z > 3.1]) (Fig. 4a). The posterior temporoparietal junction (TPJp), which lies in the vicinity of the cluster, is known as a locus for social cognition[2,8,22–24], is often active when thinking about other agents[24]. We found that the right TPJp ([x, y, z] = [52, −54, 28]) was active in proportion to partner p(correct) (Z = 2.50, p = 0.006). TPJp was more active when the partner task option was chosen when the participant played with a good partner, and it was more active when the partner task option was rejected if the participant played with a poor partner (Supplementary Fig. 5f). Short interviews after fMRI scanning (see also Fig. 4d left) confirmed that almost all participants (n = 26 out of 27) correctly understood which partner was the better one. Thus, we hypothesised that TPJp contributes to the selection of either the self or the partner task in proportion to the strength of belief in the partner's perceptual skill. To test this hypothesis, we split the trials during fMRI into two groups that we termed 'intuitive' partner choice trials ('chosen good partner' and 'unchosen poor partner' trials) and counter-intuitive partner choice trials ('unchosen good partner' and 'chosen poor partner' trials). The first group of trials are intuitive in the sense that the agent that is selected is the agent that is, on average, the better one while the second group of trials are counter-intuitive in the sense that the agent that is selected is the agent that is, on average, the worse one. We then compared the evolution of the effect in TPJp (Fig. 4b) in the two types of trials. TPJp was active in intuitive trials ($\beta_{intuitive}$ = 0.056 ± 0.021, t_{26} = 2.68, p = 0.012) but not in counter-intuitive trials ($\beta_{counter\text{-}intuitive}$ = −0.0020 ± 0.018, t_{26} = −0.11, p = 0.91), and the difference between the two trial types was significant (t_{26} = 2.12, p = 0.043). At the level of the whole-brain analysis, the cluster including the ventral temporoparietal junction survived cluster-correction for the contrast of intuitive partner trials (peak at [x, y, z] = [48, −70, 14], Z = 3.57, cluster size = 92 voxels; p < 0.05, cluster-level corrected [z > 3.1]) but did not reach cluster-significance for counter-intuitive trials (Fig. 4a).

TPJp was active in correlation with the partner's p(correct). We hypothesised that the participants' beliefs about the abilities of the social partners – their p(correct) – are formed and adjusted as participants observe the trial-by-trial outcomes of the partners' performances (see also Supplementary Fig. 1a). In fact, the activity of TPJp during the social metacognition stage was modulated by the outcome of the partner's performance one trial before (which was observed by participants; Supplementary Fig. 5g). The effect was more substantial for Good Partner than Poor Partner (t_{26} = 2.48, p = 0.019), suggesting that participants rely more on trial-by-trial learning to predict the Good partner's performance than the Poor partner's performance and that this is mediated by TPJp. To reveal how these social beliefs are developed, updated and, utilised for social prospective metacognition, first, we examined the relationship between individual variations in the influence of observations of feedback about the self's/partner's successes or failures on subsequent social metacognitive choices ('outcome effect' evaluated by $\beta_{Self(n-1)}$, $\beta_{Partner(n-1)}$ in Supplementary Fig. 1a) and individual variations in TPJp activity in response to intuitive partner choice as compared to counter-intuitive partner choice ($\beta_{intuitive}$ − $\beta_{counter\text{-}intuitive}$ in Fig. 4b). We found that the extent of the influence of the partner's outcomes on subsequent behaviour (red and blue bars in Supplementary Fig. 1a) was significantly correlated with the TPJp activity in intuitive choice trials (r = −0.38, p = 0.046). However, there was no corresponding relationship between the influence of self-outcomes (grey bar in Supplementary Fig. 1a) and TPJp activity on intuitive choice trials (r = 0.066, p = 0.74) (Fig. 4c). There was a significant effect of partner's outcome (β = −0.25 ± 0.11, p = 0.039) with no significant effect of self outcome (β = 0.096 ± 0.12, p = 0.46) on TPJp activity when a multiple regression analysis including these two variables as predictors was performed. These results suggest that variations in participants' tendencies to pursue strategies of increasing [or decreasing] their preferences to pick partners after the partner had succeeded [or failed] were correlated with individual variation in TPJp activation levels for intuitive choices. However, by contrast, any partner preference changes that occurred as a function of self-outcomes were not correlated with the TPJp activity.

We also investigated participants' subjective understanding of the relationship between the self p(correct) and the partner p(correct). For each participant, good and poor partners were selected from a data pool collected in advance so as to make the distance in performance between the self and good partner and the distance between the self and poor partner approximately equal. Thus, if a participant has perfect introspection into the relative levels of performance they themselves achieve (self p(correct)) and others achieve (partner p(correct)), the subjective evaluation of the location of the self should be the midpoint between good and poor partners (Fig. 4d, left panel). Strikingly, TPJp ($\beta_{intuitive}$) was more active for participants with a poor subjective evaluation (large deviation from the midpoint) (r = 0.58, p = 0.00096) (Fig. 4d, right panel). These observations suggest that TPJp tracks others' performance levels on the basis of observations of their choice outcomes. By contrast, however, an understanding of one's own ability in relation to those of others may depend on a different process. That participants relied more on recent observations of the good partner's task outcomes is consistent with another feature of the behavioural results: that the participants could not predict the Good partner's performance based on the projection of self metacognition. By integrating our results with the existing body of work on TPJp and social decision-making[2,8,22–24], we propose that TPJp updates beliefs about other agents in social contexts on the basis of learning from observation. In line with this account, we found that individual variation in intuitive choices, which accord with social beliefs, is related to individual variation in TPJp activity; across participants,

https://doi.org/10.1038/s41467-024-55202-0

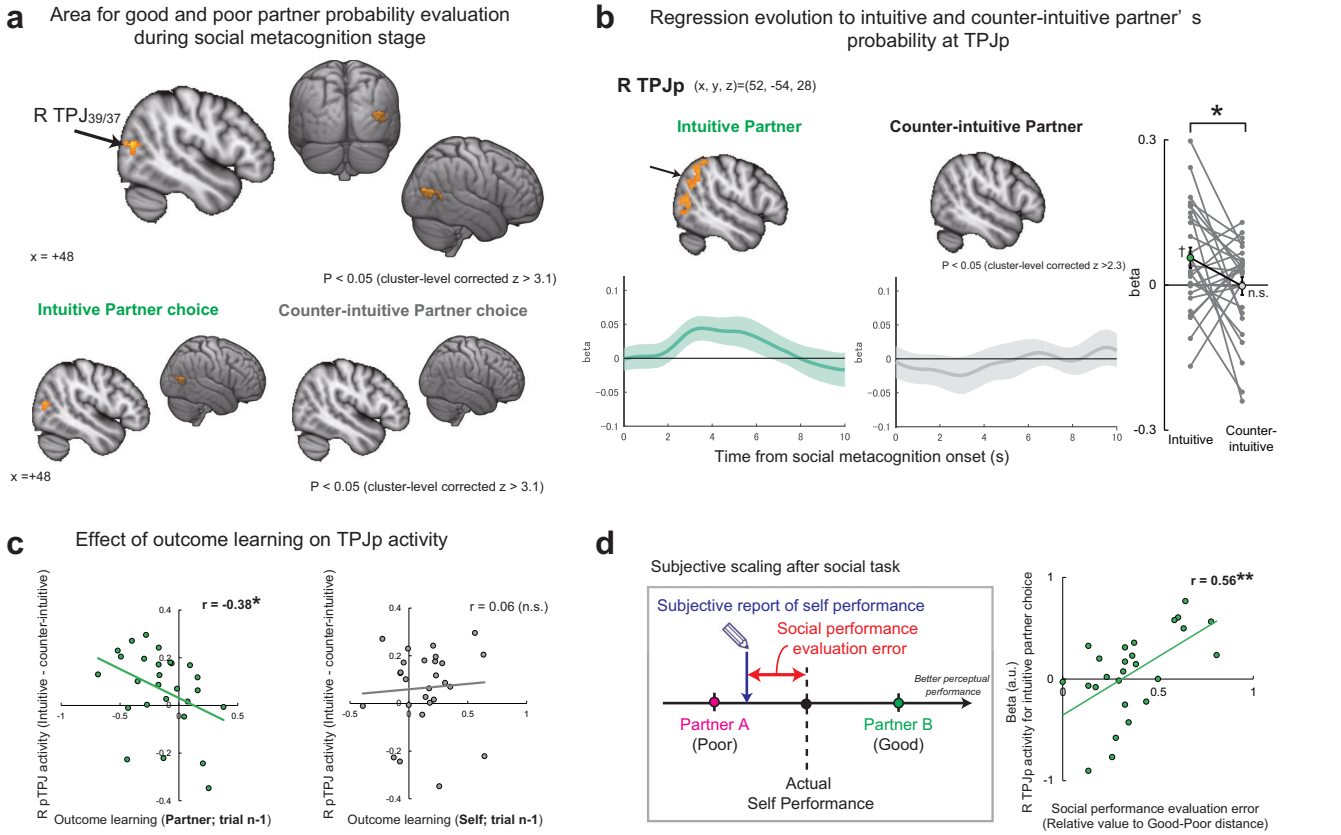

Fig. 4 | The posterior temporoparietal junction (TPJp) activity reflects observations of the partner's actions and outcomes. a Activity in the right posterior temporoparietal junction reflected the probability of reward associated with the partner's task during the metacognitive judgement stage (top). The activity was significantly positive when participants chose the good partner to perform their task or the poor partner not to perform their task (Intuitive partner choice; bottom left). It was not significant, however, when participants chose for the good partner not to perform their task or for the poor partner to perform their task (counter-intuitive partner choice; bottom right) ($N = 27$; whole-brain effects family-wise error cluster corrected with $z > 3.1$ and $p < 0.05$, two-sided). **b** Evolution of regression weights across time indexing impact of partner's task coherence when intuitive choices (green) and counter-intuitive choices (grey) were made on neural activity of left TPJp ([x, y, z] = [52, − 54, 28]; coordinates taken from a previous study[22]) (left). Mean beta values in response to the partner's task coherence were significantly different when intuitive and counter-intuitive choices were made: $^{*}p = 0.043$, paired t test. †p = 0.012, t test against zero (right). The mean and SEM error bars are displayed. **c** TPJp activity in response to the partner's task coherence on intuitive as opposed to counter-intuitive choices was modulated by the impact that the partner's recent outcomes had on the participant's task choices (when the partner's outcome effect has a negative beta [x-axis] then this indicates that if the partner had been successful recently, then the partner was more likely to be chosen subsequently (and self task performance was less likely to be chosen) (left). The effect was specific to the partner's outcomes and was not seen for the participant's own outcomes (right). There was a significant effect of the partner's outcome ($\beta = -0.25 \pm 0.11$, $p = 0.039$, two-sided) with no significant effect of self outcome ($\beta = 0.096 \pm 0.12$, $p = 0.46$) on TPJp activity in a multiple regression analysis including these two variables. $^{*}p = 0.046$. **d** Social performance evaluation error: large deviation from the midpoint corresponding to actual self-skill during the subjective report of self-performance was correlated with TPJp activity in response to the partner's task coherence when intuitive choices were made. $^{**}p = 0.00096$.

increased TPJp activation levels are correlated with an increased tendency to pursue strategies of increasing [or decreasing] preferences to pick partners after the partner has been observed to succeed [or fail].

Causal evidence of the contribution of the anterior lateral prefrontal cortex in interpersonal social coordination

Finally, to evaluate the causal role of alPFC in metacognitive projection, we disrupted alPFC$_{47}$ activity with continuous theta-burst transcranial magnetic stimulation (cTBS) and examined its causal impact on social prospective metacognitive performance (Fig. 5a). We compared the effect of cTBS to left alPFC$_{47}$ at the same point that we had investigated with fMRI ([x, y, z] = [− 38, 40, − 10][12]) (Fig. 3b) with a no-stimulation control condition and stimulation of a control region under the vertex within participants ($N = 21$).

Targeted disruption of alPFC$_{47}$ altered patterns of preference for self or partner probability options during social metacognitive judgements (Fig. 5b and Supplementary Fig. 6a). We evaluated the accuracy of metacognitive judgements and self/partner task selection by calculating meta-d' based on the equilibrium line for performance by the self and partner. The meta-d' measures for the partner were affected by alPFC$_{47}$ disruption, and the size of the impact was different for good partners and poor partners (two-way ANOVA [stimulation, no-stimulation × good, poor partner]: main effect of stimulation, $F_{1,20} = 4.96$, $p = 0.037$; main effect of partner, $F_{1,20} = 11.28$, $p = 0.0031$; interaction, $F_{1,20} = 9.03$, $p = 0.0080$) (Fig. 5c). Post-hoc simple main effect tests revealed that alPFC$_{47}$ disruption reduced the meta-d' associated with the poor partner (stimulation vs. no-stimulation, $F_{1,20} = 5.32$, $p = 0.0095$) but not the good partner (stimulation vs. no-stimulation, $F_{1,20} = 0.038$, $p = 0.54$). The difference in the meta-d' between poor and good partners without stimulation (poor vs. good, $F_{1,20} = 21.63$, $p = 1.54 \times 10^{-4}$, effect size Cohen's $d = 0.83$) disappeared with stimulation (poor vs. good, $F_{1,20} = 1.48$, $p = 0.23$, effect size Cohen's $d = 0.28$). It is important to note, however, that although alPFC$_{47}$ disruption had these different effects on poor and good partner trials, there was no difference in the impact that alPFC$_{47}$ disruption had on self-related metacognition in the poor partner and good partner trials (Fig. 5d). In summary, alPFC$_{47}$ disruption particularly impaired metacognitive assessment of the poor partners' performances.

https://doi.org/10.1038/s41467-024-55202-0

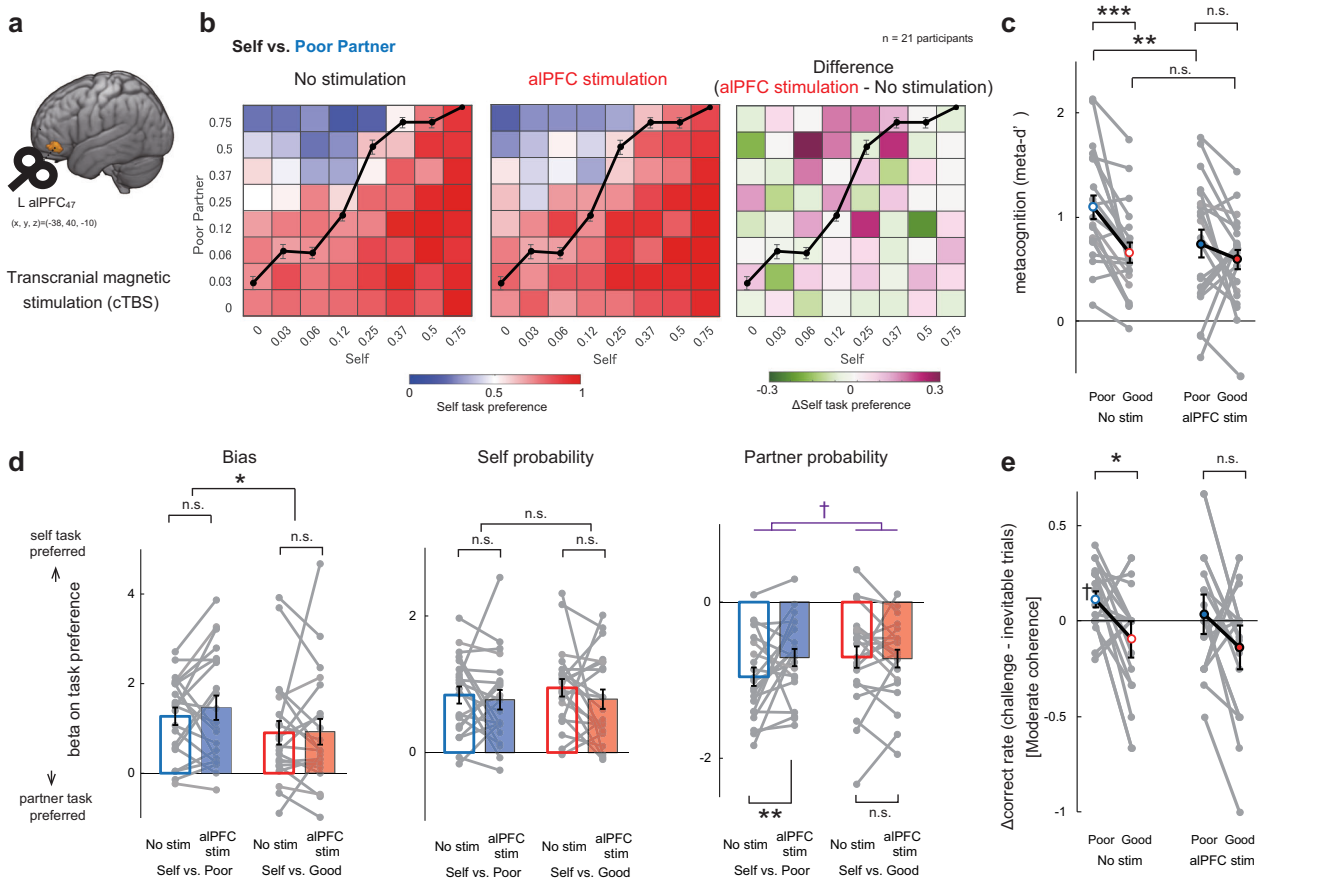

Fig. 5 | Causal impact of aIPFC$_{47}$ disruption is selective for metacognitive judgements when coordinating with the poor partner. a TMS (cTBS) ($N = 21$) was directed to left aIPFC$_{47}$ ([x, y, z] = [−38, 40, −10] based on previous work[12]) or vertex (control) on separate days. **b** Preference for choosing the self task in the social metacognitive judgement stage when coordinating with a poor partner after aIPFC$_{47}$ stimulation (centre) and no-stimulation (left). The overlaid black line indicates when performances by the self and partner are equated with one another based on the first simple RDK task (i.e., baseline perceptual performance) (left). The differences between aIPFC$_{47}$ stimulation and no-stimulation (right) show that, after aIPFC$_{47}$ cTBS, participants chose the self-task more often even when the poor partner's task was optimal. **c** Impact of aIPFC$_{47}$ stimulation on social metacognition as quantified by meta-d'. aIPFC$_{47}$ disruption impaired meta-d' on Poor, but not Good, Partner trials ($N = 21$). The significant difference in meta-d' between Good and Poor Partners without stimulation disappeared with stimulation. $^{**}p = 0.0095$, $^{***}p = 1.54 \times 10^{-4}$ (two-sided). **d** Impact of aIPFC$_{47}$ cTBS on social metacognitive judgements as a function of constant bias (left), self-task probability (middle), and partner's task probability (right) on trials with the Poor Partner (blue) and Good Partner (red) ($N = 21$). aIPFC$_{47}$ disruption changed the regression coefficient on Poor, but not Good, Partner trials. The significant difference in the coefficient between Good and Poor Partners without stimulation disappeared with stimulation. $^{**}p = 0.011$, $^{*}p = 0.023$. †$p = 0.049$, significant interaction in two-way ANOVA (stimulation, no-stimulation × Good, Poor Partner). **e** Causal impact of aIPFC$_{47}$ cTBS on participants' assessment of the merits of tackling challenge trials as opposed to inevitable trials for the moderate coherence condition (see Fig. 2d). In the absence of aIPFC$_{47}$ cTBS, on Poor Partner trials, participants performed challenge trials better than inevitable trials, but this was not the case on Good Partner trials. This difference, however, was abolished by aIPFC$_{47}$ cTBS. We demonstrated that this was the case when analysing data from 19 out of 21 participants who chose the self-task on trials adjacent to the black line in panel (**b**). †p < 0.05. $^{*}p = 0.031$ (t test against zero with Bonferroni correction).

No similar change in accuracy was observed when cTBS was applied to the control vertex site. The difference in meta-d' between poor and good partners was consistent irrespective of whether vertex stimulation was applied (two-way ANOVA [stimulation, no-stimulation × good, poor partner]: main effect of stimulation, $F_{1,20} = 0.010$, $p = 0.97$; main effect of partner, $F_{1,20} = 7.64$, $p = 0.011$; interaction, $F_{1,20} = 0.00014$, $p = 0.99$) (Supplementary Fig. 6b). In contrast to the causal impact on the accuracy of metacognitive judgements for self/partner task selection by aIPFC$_{47}$ disruption, actual perceptual performance of the RDM task was not affected by either aIPFC$_{47}$ disruption or vertex stimulation (two-way ANOVA [stimulation, no-stimulation × good, poor partner]. aIPFC$_{47}$: main effect of stimulation, $F_{1,20} = 0.30$, $p = 0.58$; main effect of partner, $F_{1,20} = 0.0052$, $p = 0.94$; interaction, $F_{1,20} = 0.22$, $p = 0.64$. vertex: main effect of stimulation, $F_{1,13} = 0.024$, $p = 0.87$; main effect of partner, $F_{1,20} = 0.54$, $p = 0.46$; interaction, $F_{1,20} = 0.12$, $p = 0.72$) (Supplementary Fig. 6c). This suggests that aIPFC$_{47}$ plays a specific role in prospective evaluations of social probability rather than perceptual decision per se.

The pattern of changes is different in trials with a poor partner as opposed to a good partner. Logistic multiple regression analyses (see also Fig. 2b) revealed that aIPFC$_{47}$ cTBS did not change the constant bias to prefer the self-task option with a poor partner compared to a good partner (two-way ANOVA [stimulation, no-stimulation × good, poor partner]: main effect of stimulation, $F_{1,20} = 0.34$, $p = 0.56$; main effect of partner, $F_{1,20} = 6.0$, $p = 0.023$; interaction, $F_{1,20} = 0.49$, $p = 0.49$) (Fig. 5d, left). In addition, aIPFC$_{47}$ cTBS did not change the size of the regression coefficient for the probability of correct performance on the self-task option (self p(correct)) when participants were coordinating with either the good or poor partner (two-way ANOVA [stimulation, no-stimulation × good, poor partner]: main effect of stimulation, $F_{1,20} = 0.81$, $p = 0.37$; main effect of partner, $F_{1,20} = 0.66$, $p = 0.42$; interaction, $F_{1,20} = 0.35$, $p = 0.56$) (Fig. 5d, middle). However, the sizes of the regression coefficient for the partner were affected by aIPFC$_{47}$ disruption and the size of the impact was different for good partners and poor partners (two-way ANOVA [stimulation, no-stimulation × good, poor partner]: main effect of stimulation, $F_{1,20} = 3,48$,

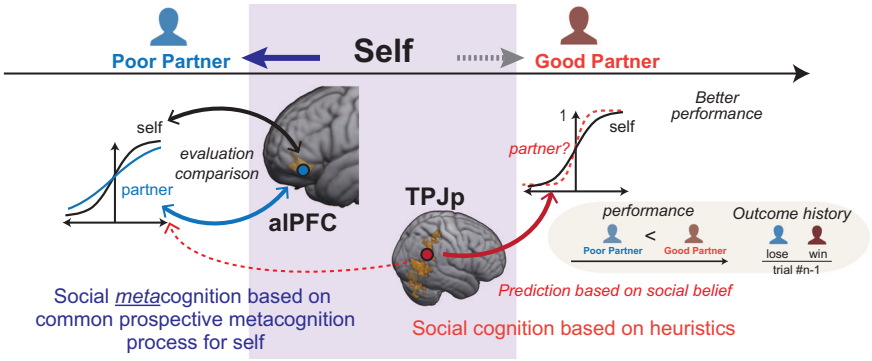

Fig. 6 | Proposed mechanisms of social metacognition in aIPFC$_{47}$ and TPJp. An aIPFC$_{47}$ mechanism operates by projecting onto others the participant's own metacognitive introspections about how well they themselves might perform and a TPJp mechanism operates by tracking average levels of performance of different individuals by observation of their actions and outcomes.

$p = 0.076$; main effect of partner, $F_{1,20} = 1.36$, $p = 0.25$; interaction, $F_{1,20} = 4.38$, $p = 0.049$) (Fig. 5d, right). Post-hoc simple main effect tests revealed that the difference in this regression coefficient between poor and good partners without stimulation (poor vs. good, $F_{1,20} = 7.66$, $p = 0.011$) disappeared with stimulation (poor vs. good, $F_{1,20} = 0.03$, $p = 0.86$); aIPFC$_{47}$ cTBS led to poor partner performances having less influence on participants' choices about who should tackle a decision making problem, but it did not change the influence of good partner performances. This asymmetric impairment between tasks with poor and good partners suggests that aIPFC$_{47}$ contributes to comparisons of probability based on inferences about other decision-makers that are extrapolated from metacognitive insights into what one's own performance would be with a similar coherence level. However, the asymmetry may also be related to the fact that participants could reliably distinguish between two tasks that both had high coherence levels (Supplementary Fig. 3) and this may have been useful in predicting especially the poor partner's performance when aIPFC$_{47}$ was not disrupted. When cTBS was applied to the vertex as a control site, no similar asymmetric changes in the regression coefficient were observed (interaction between stimulation conditions and partner identity in two-way ANOVA for the regression coefficient of partners, $F_{1,20} = 1.59$, $p = 0.22$).

Finally, to evaluate the degree that the participants use insight into their own performance and into the partner's performances when selecting trials that are particularly challenging to perform by the participants themselves, we carried out an analysis of challenge and inevitable trial performance (see also Fig. 2d) but now with and without aIPFC$_{47}$ disruption. We analysed data of 19 out of 21 participants who chose the self task both immediately above and below the equilibrium line for the performance of the self and partner consistently across all four conditions (stimulation, no-stimulation × good, poor partner). Note that this analysis could not be performed on the data from the remaining two participants. The benefit furnished by metacognitive insight on challenge trials, as opposed to intuitive trials, was diminished by the aIPFC$_{47}$ and the effect was greater for poor partner than good partner trials. In the no-stimulation condition, the benefit (Δperformance [challenge − inevitable trials]) was significantly larger when coordinating with a poor partner as opposed to a good partner (paired t test, $t_{18} = 2.66$, $p = 0.031$, Bonferroni corrected), whereas the difference disappeared after aIPFC$_{47}$ cTBS ($t_{18} = 1.74$, $p = 0.19$) (Fig. 5e). In conjunction, these results suggest that aIPFC$_{47}$ plays a causally essential role in predicting partners' performances based on projection of introspection into one's own likely performance.

In summary, the present study revealed an interaction between two neural mechanisms for estimating other people's abilities. An aIPFC$_{47}$ mechanism operates by projecting onto others the participant's own metacognitive introspections about how well they themselves might perform (Fig. 6, left). This complements the better known and more widely studied TPJp mechanism that operates by tracking average levels of performances of different individuals and making choices as a function of these average estimates (Fig. 6, right).

Discussion

When we collaborate with other people to tackle a novel problem, we need to estimate who will do what best. To work this out, we can observe how the other person performs and use metacognition to estimate our own ability to perform. The two processes have been linked to TPJp[2,8,22–24] and, recently, to aIPFC$_{47}$[12] respectively. The current results, however, suggest that we can also employ metacognition to introspect into our own abilities and project them onto others to estimate how they might perform instead. However, while such projection of self-introspection or 'social metacognition' may help us to estimate how partners with poorer skills than us are likely to behave, it fails to provide reliable predictions for better-skilled partners for whom we cannot make a metacognitive model by reference to our own abilities. Lower metacognitive decision accuracy for Good Partners may be due to variable and inconsistent prediction of perceptual performance for difficult trials with low coherences. Being able to make predictions for such trials would be necessary to predict the Good Partner's performance based on a projection from self-introspection. In this study, we found that aIPFC$_{47}$ is active during the comparison of one's own chances of success with a poorer partner's chances. Activity in aIPFC$_{47}$ reflects the evidence for taking the self task and the degree to which this is the case increases over time in a manner consistent with an evidence accumulation process[25]; models of evidence accumulation processes first proposed in the context of sensorimotor decision-making can be used to explain what might superficially appear to be quite distinct cognitive processes such as those deployed during social interaction[26]. The elevation of the BOLD signal in the human aIPFC$_{47}$ is analogous to increasing neuronal firing reported in the macaque parietal cortex during perceptual decisions, which is accounted for by the evidence accumulation model[25]. Subjective confidence in the current perceptual decision being taken has been found to be reflected by neuronal responses in the intraparietal sulcus[27] and supplementary eye field[28] in macaques: the more the neurons were active, the more often the animals reported higher confidence.

However, in the current experiment, a similar pattern does not seem to emerge when the comparison is made with a better partner. While aIPFC$_{47}$ mediates a process of social prediction via metacognitive insight into one's own likely decision-making ability, a separate social prediction system linked to TPJp, makes a different type of contribution; it tracks the actions and outcomes of both types of partners, both better and worse players. The claim that well-

established TPJp-based mechanisms for social cognition are complemented by alPFC$_{47}$-dependent social metacognitive processes was also supported by a selective impairment of social metacognition-based performance with poorer partners when cTBS was applied to alPFC$_{47}$. This is analogous to the way in which disruption of areas in which neural activity has been linked to evidence accumulation during perceptual decision-making disrupts perceptual decision-making[29]. Notably, however, alPFC$_{47}$ cTBS only disrupted social and metacognitive evidence accumulation but had no impact on perceptual decision-making itself. In summary, alPFC$_{47}$ activity emerges only when participants are making metacognitive judgements, but it is true when they make such judgements about themselves and when they make them about others. alPFC$_{47}$ activity does not appear when participants make equally difficult judgements about the same range of probabilities but on the basis of assessments of external contingencies as opposed to metacognitive judgements. In other words, the patterns of behaviour and neural activity are associated with metacognitive judgement, but the metacognitive judgements can be directed by the participants towards themselves or towards others.

The alPFC$_{47}$ region identified here is the same one that has previously been linked with the estimation of what Miyamoto and colleagues[12,30] referred to as 'internal probabilities'. They argued that in addition to estimating the chance that a choice will lead to a desired outcome (external probability), it is equally important to know one's chances of making the choice correctly (internal probability). They gave the example that we might estimate our ability to drive to a new restaurant without a GPS (internal probability) as well as estimate the likelihood that the restaurant is open (external probability). Miyamoto and colleagues reported that alPFC$_{47}$ accumulated evidence in favour of taking a choice defined by its internal probability as opposed to one defined by its external probability. Disruption of neural activity in alPFC$_{47}$ by TMS changed the way internal probability is evaluated during metacognitive judgements made prior to perceptual decisions. No disruption occurred when TMS was applied to a control site. It is possible that TMS at the alPFC$_{47}$ and the control site differed in terms of their side effects. However, we note that the TMS protocol we used was a theta-burst TMS protocol and behavioural data collection did not occur at the time of TMS delivery but, instead, only after it had been completed when any side effects would have elapsed (see "Methods"). A related observation is that prospective evaluation skills are also altered by the application of transcranial alternating current stimulation (tACS) over the same area[31]. Human neuroimaging studies on metacognition repeatedly report that, within the frontal cortex, especially around the anterior cingulate cortex, domain-general neural correlates of confidence are found across various tasks[32]. The frontopolar cortex, in particular, has been associated with both introspection ability[10] and coding of counterfactual reward[33] and is considered to play a key role in higher-order cognition. These studies focusing on retrospective or ongoing introspection did not identify alPFC$_{47}$ as a critical locus for metacognition; this suggests a special role for alPFC$_{47}$ in underpinning prospective metacognition. Differential encoding of retrospective confidence reports concerning the self and others have been reported previously: TPJ and dorsomedial prefrontal cortex were active, especially in correlation with estimations of other's confidence but not with self-confidence[5]. Activity in alPFC$_{47}$, however, was not identified. This suggests that alPFC$_{47}$ is recruited only when coordination of introspection about the self and other are required just as previously alPFC$_{47}$ was active when participants compared the internal probabilities of success associated with their choices with external probabilities.

In contrast, we found that TPJp was more active in participants with a poor subjective sense of their own performance abilities in relation to other people's; these participants exhibited larger errors when positioning the poor and good partner's skills in relation to their own. TPJp was also more active when participants made intuitive' partner choices (when they chose the good partner or when they did not choose the poor partner) as opposed to counter-intuitive partner choices (the good partner was not chosen or the poor partner was chosen). We identified six participants who reported either that both partners were better or worse than they were themselves. In each case, TPJp activity associated with this belief was positive, suggesting that TPJp estimates and compares the skills of multiple other people independently of self-performance estimates. TPJp has been implicated in the building of beliefs about the accuracy of information provided by others during social exploration and learning[8,34]. The present study also suggests that TPJp is active in correlation not only with a learned belief (or heuristic) about others' performance but also with updating of the belief based on learning[35] from recent performance of the task by others. We propose that alPFC$_{47}$ plays a complementary and critical role in imagining others' skills by projecting and adjusting estimates of performance based on introspection of one's own likely performance, considered in the light of knowledge about others obtained by observation and encoded in TPJp.

Methods

Participants

Thirty participants took part in the functional MRI experiment (Experiment 1). Participants were excluded because of premature termination of an experimental session ($N = 3$) (final sample: 27 participants; 13 female; age (mean ± SD), 24.5 ± 3.7). Thirty-two participants took part in the TMS experiment (Experiment 2). Participants were excluded because of premature termination of an experimental session ($N = 7$) and extreme behavioural bias during the main experiment due to poor understanding of the task rule ($N = 4$) (final sample: 21 participants; 14 female; age (mean ± SD), 23.6 ± 4.6). The sample size for both Experiment 1 and 2 after excluding participants of premature termination was comparable with fMRI experiments in Miyamoto et al.[12] ($N = 23$), where meta-d' was statistically significant when compared against a null hypothesis value of zero with effect size Cohen's $d = 1.29$. The observed effect sizes of the same meta-d' in the present study were Cohen's $d = 2.07$ (meta-d' between the self and Poor Partner), $d = 1.79$ (meta-d' between the self and Good Partner) in the fMRI experiment, and Cohen's $d = 2.09$ (meta-d' between the self and Poor Partner), $d = 1.50$ (meta-d' between the self and Good Partner) in no-stimulation control conditions of the TMS experiment. The study was approved by the Research Ethics Second Committee at RIKEN (Experiment 1: Wako3 2021-28(5)) and the Central Research Ethics Committee at the University of Oxford (Experiment 2: R60547/RE001). All participants gave informed consent.

Experimental procedure

We conducted two experiments. The first experiment assessed the neural correlates of social metacognitive decisions with fMRI, while the second experiment probed the causal contribution of alPFC$_{47}$ in prospective social metacognition using transcranial magnetic stimulation (TMS) in a continuous theta burst (cTBS) design. The two experiments used different samples of participants.

Experiment 1. Participants took part in one session lasting approximately 4 hours in total, including 2 h of magnetic resonance imaging (MRI) scanning. Participants received 1500 JPY per hour and a bonus based on task performance (accumulated across sessions: 1200–1600 JPY per session). Participants first performed a simple RDM task to measure the baseline perceptual performance outside the scanner. Next, based on the measurement, we picked a better-skilled partner and a poorer-skilled partner from the pool of data on past participants' responses to the same RDM stimuli[12] and the participants observed how accurately each of the two partners could perform the task. Then, they performed the main prospective social metacognition task inside the MRI scanner. Each fMRI scanning session included 195 trials and lasted for 60–65 min. After scanning, participants had a short verbal

interview, conducted by an experimenter, about the features of the two partners and filled in a questionnaire that asked about participants' subjective understanding of the relationship between the self p(correct) and the partners' p(correct) (see Fig. 4d).

Experiment 2. The second experiment included four sessions: a behavioural task session (Session 1; 1 h), a structural MRI session (Session 2; 30 min) and two cTBS TMS sessions (Sessions 3 and 4; each 2 h). In the first session, the participants performed a simple RDM task to measure baseline perceptual performance. In addition, we assessed participants' motor thresholds, which determined the intensity of cTBS stimulation that was used in later cTBS sessions (see section 'Transcranial magnetic stimulation (TMS)' for more details). To predict participants' tolerance and comfort with the stimulation protocol in Sessions 3 and 4, we first applied a milder stimulation protocol over the targetted anterior frontal region, 'a taster session' during Session 1. The taster session included a stimulation protocol comprising a 10 s train of cTBS with the stimulator output set to 20%. One participant who reported an uncomfortable sensation dropped out after the taster session and the rest of the 31 participants, who did not feel discomfort, moved to the main experiments. Session 2 served to acquire structural MRI scans that would guide the neuronavigation localisation of the TMS target areas in the subsequent two sessions. Session 3 and 4 consisted of two blocks: a stimulation block and a no-stimulation block. For each block, participants performed a shortened version of the main experimental task used in Experiment 1. Each block lasted for 30 min (232 trials). Stimulation was applied before one block ("TMS block"), but not the other ("control block") within each session 3 and 4. The stimulation order within the session was counterbalanced across participants. The difference between Session 3 and 4 was the stimulation site: the stimulation site was either centred on $aIPFC_{47}$ [MNI x/y/z- coordinate: −38, 40, −10][12] or vertex [MNI x/y/z- coordinate: 0, −34, 72], with cTBS being applied immediately before the start of the "TMS block". Further counterbalancing meant that some participants performed aIPFC sessions first and some performed vertex sessions first. As a result of the various types of counterbalancing, participants performed sessions in the following orders (5 participants: Session 3, aIPFC; Session 4, vertex; TMS block before control block. 7 participants: Session 3, vertex; Session 4, aIPFC; TMS block before control block. 5 participants: Session 3, aIPFC; Session 4, vertex; TMS block after control block. 4 participants: Session 3, vertex; Session 4, aIPFC; TMS block after control block). The participants took at least a 30-minute break from the end of the "TMS block" to the start of the "control block" to decrease the possibility of any remaining effects of TMS. To avoid any possible long-term effects of learning across sessions, even though the two TMS sessions were at least one week apart from each other, participants learned two new partners at the beginning of each session. The assignment of two colours (pink or green) to the good or poor partner was counterbalanced across participants and sessions.

Behavioural tasks

Experiments 1 and 2 used the same behavioural tasks. First, the participants performed a simple RDM task. In the task, participants were required to judge whether the majority of a total of 100 small dots moved leftwards or rightwards by pressing the left or right arrow key on a keyboard. Each RDM stimulus was presented for 1.5 s, and the movements of the dots were ambiguous (0, 3, 6, 12, 25, 37, 50, or 75% denote the different coherence levels). We plotted the proportion of the trials that the participant judged as 'rightward direction' against motion coherence from 75% right motion to 75% left motion (logarithmic transformation was applied to the value of coherence) and fitted a sigmoid function as follows where x is the logarithmic transformed signed motion coherence and y(x) is the proportion of trials for which 'rightward direction' was chosen when coherence x was employed (see also the inset panel of Fig. 1b):

$$\ln\left(\frac{y(x)}{1-y(x)}\right) = \alpha + \beta x \quad (1)$$

We quantified individual participants' baseline perceptual performances by bias (α) and sensitivity (β) and plotted these together with another 23 participants' data for the same task that we had collected previously[12]. Before the experiments, we calculated β (sensitivity) of data sets drawn from a pool of 23 candidates[12]. The standard deviation of the β was 0.68. We then picked a better and poorer partner from the pool based on this variance. That is, we first quantified each individual participant's baseline perceptual performances in terms of bias ($\alpha_{participant}$) and sensitivity ($\beta_{participant}$); then we picked from the pool a good partner whose β is the closest to $\beta_{participant} + 0.68$ and a poor partner whose β is the closest to $\beta_{participant} - 0.68$. An important feature of the task is that we showed actual behavioural responses of other participants collected in our previous study[12] by using exactly the same RDM stimuli as Good and Poor Partner's data. When we gave instructions on the task rule, we explicitly explained this fact. Our participants, therefore, thought that they were considering real patterns of behaviour generated by real people (as indeed they were). Due to the presentation of real empirical data and clear instructions, we found that all participants reported that they did indeed believe that this was the case. For example, they endorsed the statements "The person indicated by green dots was better at performing the task than the person indicated by pink dots" or "The person indicated by pink dots was more reliable."

Next, the participant performed the RDM task again, but in one-half of the trials, the stimulus was pink, and in the other half, the stimulus was green. The good partner was assigned to pink, and the poor partner was assigned to green for half of the participants, and vice versa for the other half. In each trial, after the participant indicated the motion direction of the RDM stimulus, the participant was given feedback about their performance (correct or incorrect) as well as the performance by the partner associated with the colour to exactly the same RDM stimulus at the level of each random and coherent dot's distribution and movement.

Then participants performed the main social metacognition task inside an MRI scanner. The main task comprised two stages: each trial comprised a social metacognitive judgement followed by a perceptual decision and a final outcome phase (Fig. 1c). In the metacognitive judgement stage, participants had to choose one of the two RDM stimuli that were presented simultaneously. One RDM on the left side represented the self-task, and the other RDM on the right side represented the partner's task. As noted, the two different partners were indicated by pink and green RDM stimuli. All possible combinations of eight levels of self-task coherence [self p(correct)] and eight levels of partner task coherence [partner p(correct)] were offered during the metacognitive judgement stage either with the good or poor partner. The participant's goal was to pick the task that was most likely to be performed correctly by the designated player; participants were incentivized by the award of a single point whenever either they themselves or their partner performed the task correctly. Participants were given a cash bonus depending on performance as a further incentive. In the Metacognition stage, each RDM stimulus was moving upward or downward for 1.5 s. After the disappearance of the stimuli, participants chose the task they wanted to have performed (either by themselves or by their partner) in the subsequent perceptual decision stage by pressing a button with their right hand. After a stimulus onset asynchrony (SOA) (Experiment 1, 2.5–8.5 s [Poisson distribution, mean of 4.5 s]; Experiment 2, 1 s; note that we did not have to control for the BOLD response in the second experiment and therefore SOAs are shorter, moreover given the limited duration of cTBS effects it was important to collect trials more quickly in Experiment 2), participants

moved into a perceptual decision stage where the same stimulus that they chose in the social metacognitive judgement stage appeared again for 1.5 s. This time, however, the direction of the dot motion was rotated by +/− 90 degrees. For example, if they selected the external probability option in the first stage and the stimulus was moving upwards they could not know until the second stage perceptual decision whether the stimulus would be moving leftwards or rightwards. After the disappearance of stimuli, participants were asked to answer if the RDM stimulus was moving leftward or rightward by pressing a button. The rotation of the stimulus was introduced to prevent participants from making a perceptual decision about motion direction during the social metacognitive judgement phase of the trial instead of during the subsequent perceptual decision phase of the trial. However, we wanted the participants to estimate and compare the utility of choosing either the self or partner's options to make an optimal metacognitive judgement. In the experiment, we rotated the direction of the stimulus chosen in the metacognitive judgement phase of the trial either clockwise or anticlockwise randomly when it appeared at the perceptual decision phase of every trial. Therefore, participants could not predict the motion direction in the perceptual decision stage from that in the metacognitive judgement stage[36]. Even when they chose the partner's task, participants were asked to judge and report the motion direction for the same RDM stimulus. After an SOA (Experiment 1, 2.5–8.5 s; Experiment 2, 1 s), outcome feedback appeared for 1 s. If participants chose the self-task and judged the motion direction correctly, a reward (white 'tick' symbol on the upper centre of the screen indicated success) was given while, otherwise, no reward (white 'X' symbol on the upper centre of the screen indicated failure) was given. Simultaneously, the partner's success or failure was indicated by 'tick' or 'X' respectively in the partner's colour on the lower centre of the screen, but this had no impact on the reward if the participant had chosen to perform their own task themselves. When, however, participants chose the partner's task, a reward was given based on the partner's performance for the RDM stimulus (coloured 'tick' and 'X' symbols lower on the centre of the screen indicated success and failure, respectively). Performance by the self for the same RDM stimulus was presented simultaneously on the upper centre of the screen but this did not determine the reward in the trial.

The total number of rewarded trials also appeared on the bottom of the screen during the feedback period. Based on the number of rewarded trials, participants received a monetary bonus reward after the experiment. After 1 s of inter-trial interval (ITI), the next trial started. Eye positions were monitored in Experiment 1 with an eye tracker (iRecHS2 system). We used eye-tracking data to confirm that all participants engaged in performing the task during fMRI scanning.

FMRI data acquisition and data processing

Imaging data in Experiment 1 were acquired with a Siemens Prisma 3 T MRI using a multiband T2*-weighted echo planar imaging sequence with an acceleration factor of three and a 64-channel head-coil. Slices were acquired parallel with the PC-AC line to reduce signal dropout in the prefrontal cortex. Other acquisition parameters included 2.4 × 2.4 × 2.4 mm voxel size, TE = 30 ms, TR = 1250 ms, 64° flip angle, a 192 mm field of view and 63 slices per volume. A structural scan was obtained with slice thickness = 0.7 mm; TR = 2180 ms, TE = 2.95 ms and 0.7 × 0.7 × 0.7 mm voxel size. Imaging data were analysed using in-house pre-processing programmes and FMRIB's Software Library (FSL)[37]. Preprocessing stages included motion correction based on Analysis of Functional NeuroImages (AFNI), removal of the spike or the step noise caused by participant motion, slow drift, and physiological noise (heartbeat and respiration recorded during scanning), slice-timing correction, brain extraction, high-pass filtering and spatial smoothing using full-width half maximum of 5 mm. Images were co-registered to an individuals' high-resolution structural image and then nonlinearly registered to the MNI template using 12 degrees of freedom. In Experiment 2, we obtained a structural scan with slice thickness = 1 mm; TR = 1900 ms, TE = 3.97 ms and 1 × 1 × 1 mm voxel size with a larger field of view covering the nose tip and both ears, which serve as the landmarks for frameless stereotactic neuronavigation (see the next 'Transcranial magnetic stimulation (TMS)' section).

Transcranial magnetic stimulation (TMS)

TMS was applied using a Magstim Rapid stimulator which was connected to a 50 mm Figure-8 coil[38] in the postero-anterior direction[39]. In Session 1 of Experiment 2, we assessed participants' active motor threshold (AMT) for the left M1 'hot spot', which is the scalp location where TMS evoked the largest motor evoked potential (MEP) amplitude in right first dorsal interosseous (FDI)[40] (mean ± SD: 51.7 ± 9.0% stimulator output). The AMT was defined as the minimum stimulation intensity sufficient to evoke an MEP in contralateral FDI in at least 50% of trials when participants exerted a small constant force between index finger and thumb (20% of maximum contraction force). Electromyographic (EMG) activity in the right FDI was recorded with bipolar surface Ag-AgCl electrode montages. Responses were band-pass filtered between 10 and 1000 Hz, with additional 50 Hz notch filtering, sampled at 5000 Hz, and recorded using a D440 Isolated EMG amplifier (Digitimer), a Hum Bug 50/60 Hz Noise Eliminator (Quest Scientific), a CEDmicro1401 Mk.II A/D converter, and PC running Spike2 (Cambridge Electronic Design).

The region of interest was left alPFC (Session 3 or 4) with MNI x/y/z-peak coordinates (− 38, 40, − 10), which was identified by the previous fMRI experiment (Experiment 1; see Figs. 3a and 5a). We used the same coordinate for left alPFC stimulation. To stimulate the vertex, the coil was placed over MNI x/y/z-peak coordinates (0,− 34, 72). No neural activity with any relation to either internal or external probability was found at this vertex location suggesting that it was an appropriate control site. The location was projected onto the high-resolution, T1-weighted MRI brain scan of each participant using frameless stereotactic neuronavigation (Brainsight; Rogue Research). We used a standard continuous theta-burst stimulation (cTBS) protocol to stimulate alPFC and vertex: 600 pulses were administered in bursts of three pulses at 5 Hz (total stimulation duration was 40 s). TMS coils were held in place tangentially to the skull by an experimenter during stimulation. For each participant, stimulation intensity was determined by 80% of the AMT[41]. The use of such a low subthreshold intensity (80% AMT) had the advantage of ensuring decreased spread of stimulation away from the targeted site and enabled us to focus on the alPFC site. In the cTBS protocol, behavioural data collection occurred following stimulation. Side effects such as headaches can continue for a few hours following stimulation in some participants and differ by stimulation site. However, our participants did not report any such headaches after the cTBS. Any side effects would have elapsed during the behavioural experiment if it had happened. A quantitative review article of the magnitude and time course of cortical excitability changes induced by cTBS provided the information that cTBS applied for 40 s decreases cortical excitability for up to 50 min with a mean maximum depression of − 22.81 ± 2.86%[42]. In our TMS protocol, if TMS was delivered at the beginning of the first session before the task, the task in the second session started 60 min after the TMS as it took around 30 min to finish the task in the first session and participants took a rest for 30 min until the beginning of the second session. Thus, TMS should have no effect on the control part of the session even when that was the second part of the testing schedule. Moreover, as in our previous studies[6,12,34], the order of TMS and matched control session was counterbalanced across participants.

Data analysis

Behavioural data. To evaluate performance during the metacognitive judgement stage, we employed an analysis based on signal detection theory[20]. Specifically, for each good and poor partner, we classified the

metacognitive judgement trials into trials in which it was optimal for participants to choose the self-task and trials in which it would be optimal to choose the partner's task. For each participant, if the probability of reward determined by the partner's performance for the partner's task option was higher than the probability of reward that would be expected given the baseline level of perceptual performance of the self-task option (obtained during the first simple RDM task), then such trials were categorised as partner task optimal trials. If not, they were categorised as the self task optimal trials. Based on the proportion of trials in which they chose the self-task option when the self option was optimal (Hit trials) and when the partner's option was optimal (False alarm [FA] trials), we calculated meta-d' (type-II d-prime). We calculated meta-d' separately for trials with different coherence levels in the self-task to control for the difference in baseline perceptual performance (Fig. 2c left panel) and compared their average across partner's conditions and brain stimulation conditions (Fig. 2c right panel and 5c).

First, to evaluate the effects (on task selection in the initial metacognitive judgement phase of each trial) of reward probability associated with both the self and the partner's task option as well as the partner identity (Poor Partner: 1, Good Partner: −1) and its interaction with the reward probability, we employed logistic multiple regression analyses for both trials paired with a poor partner and trials paired with a good partner together as shown below (Supplementary Fig. 1d).

$$\ln\left(\frac{y(n)}{1-y(n)}\right) = \alpha + \beta_{self}\, x_{self}(n) + \beta_{partner}\, x_{partner}(n) + \beta_{partnerID}\, x_{partnerID}(n) + \beta_{self*partnerID}\, x_{partnerID}(n)\cdot x_{self}(n) + \beta_{Partner*partnerID}\, x_{partnerID}(n)\cdot x_{partner}(n) + \beta_{self_{outcome}} x_{self_{outcome}}(n-1) + \beta_{poorp_{outcome}} x_{poorp_{outcome}}(n-1) + \beta_{goodp_{outcome}} x_{goodp_{outcome}}(n-1) + \beta_{ID*self_{outcome}} x_{partnerID}(n)\cdot x_{self_{outcome}}(n-1) + \beta_{ID*poorp_{outcome}} x_{partnerID}(n)\cdot x_{poorp_{outcome}}(n-1) + \beta_{ID*goodp_{outcome}} x_{partnerID}(n)\cdot x_{goodp_{outcome}}(n-1) \quad (2)$$

As the effects of reward probability associated with both the self and the partner's task option were found to be affected by the partners' identity, we employed logistic multiple regression analyses separately for trials paired with a poor partner and trials paired with a good partner as shown below (Figs. 2b and 5d).

$$\ln\left(\frac{y(n)}{1-y(n)}\right) = \alpha + \beta_{\text{self}}\, x_{\text{self}}(n) + \beta_{\text{partner}}\, x_{\text{partner}}(n) \quad (3)$$

Dependent variable $y(n)$ denotes the task chosen during the metacognitive judgement stage (self task = 1; partner's task = 0) at trial #n. Independent variables x_{self} (n) and $x_{partner}(n)$ denote the reward probability of the self task and partner's task at trial #n, respectively. The reward probability here is denoted by the baseline performance of the self for each motion coherence condition during the first simple RDM task. We also measured the influence of outcome in the most recent and second most recent trials (trials #n-1 and #n-2) by adding binary independent variables separately for the self, poor and good partners (see Supplementary Fig. 1 more in detail). All the independent variables were normalised (mean of zero and standard deviation of one) within each session before including them in the analysis.

Functional MRI data

Whole-brain analysis. We used FSL FEAT for first-level analysis. First, data was pre-whitened with FSL FILM to account for temporal auto-correlations. Temporal derivatives were included in the model. We used three fMRI general linear models (fMRI-GLM1, 2, 3) to analyse fMRI data across the whole brain. Results were calculated using FSL's FLAME 1 + 2 with a cluster-correction threshold of $z > 3.1$ and $p < 0.05$, two-tailed (fMRI-GLM1).

To analyse BOLD changes across participants (fMRI-GLM1), a second-level analysis was applied (FLAME1 + 2). We also used two covariate fMRI analyses (fMRI-GLM2, 3) during which we associated a covariate with particular regressors in the second level (FLAME 1 + 2).

All whole-brain GLMs shared the following features: we included all three phases of a trial (social metacognitive judgement, perceptual decision, and outcome) into the fMRI-GLMs. Each phase included a constant regressor, which was the onset of each phase with a fixed duration of 1.5 s for social metacognitive judgement and perceptual decision and a duration of 1 s for the outcome phase. Parametric regressors were modelled as stick functions (i.e., duration of zero) time-locked to the relevant phase onset as below. All parametric regressors were normalised before inclusion into the analysis. In addition, all GLMs contained one regressor time-locked to all button presses, modelled as a stick function, at the first-level fixed-effect analysis stage.

fMRI-GLM1

First, we tested for neural correlates of self and partner's option probabilities during the social metacognitive judgement stage (Figs. 3a and 4a). We included the following regressors, along with the constant regressor coding the phase of social metacognitive judgement in each trial:

Chosen self-task coherence (poor partner),
Chosen partner's task coherence (poor partner),
Unchosen self-task coherence (poor partner),
Unchosen partner's task coherence (poor partner),
Chosen self-task coherence (good partner),
Chosen partner's task coherence (good partner),
Unchosen self task coherence (good partner),
Unchosen partner's task coherence (good partner)

We set these parametric regressors separately for trials with a good partner and trials with a poor partner. Coherence values were first transformed logarithmically (see also Fig. 2b). Then each regressor was normalised before inclusion into the analysis (mean of zero and standard deviation of one). If participants chose the self-task on trial #n, then the coherence value of the self-task option and the partner's task option were coded as chosen self-task coherence and unchosen partner's task coherence, respectively. These variables were time-locked to the onset of the social metacognitive judgement stage when participants chose the self-task. The chosen partner's task coherence and unchosen self-task coherence, as well as regressors related to the unavailable partner, were not defined for those trials. If participants chose the partner's task on trial, the coherence values of the partner's task option and the self-task option were coded as the chosen partner's task coherence and unchosen self-task coherence, respectively. These variables were time-locked to the onset of the social metacognitive judgement stage when participants chose the partner's task. Chosen self-task coherence and unchosen partner's task coherence, as well as regressors related to the unavailable partner, were not defined for those trials. To identify neural activity that reflected the chosen self-task coherence, we calculated the sum of 'chosen self-task coherence (poor partner)' and 'chosen self-task coherence (good partner)' (Fig. 3a). We also derived the contrasts to identify neural activity that reflected partner's task coherence in three different ways as follows: (1) partner's task coherence = 'chosen partner's task coherence (poor partner)' + 'chosen partner's task coherence (good partner)' + ' unchosen partner's task coherence (poor partner)' + 'unchosen partner's task coherence (good partner)'; (2) intuitive partner's task coherence = 'chosen partner's task coherence (good partner)' + ' unchosen partner's task coherence (poor partner)'; (3) counter-intuitive partner's task coherence = 'unchosen partner's task coherence (good partner)' + 'chosen partner's task coherence (poor partner)' (Fig. 4a, b).

In order to capture activity related to making decisions about the directions of stimuli, in the perceptual decision stage of each trial, we included the following regressors:

Chosen self-task coherence,

Unchosen self-task coherence,
Chosen poor partner's task coherence,
Unchosen poor partner's task coherence,
Chosen good partner's task coherence,
Unchosen good partner's task coherence

Coherence values were first transformed logarithmically. Then each regressor was normalised before inclusion into the analysis (mean of zero and standard deviation of one). These variables were time-locked to the onset of the perceptual decision-making stage.

In order to capture activity related to the outcome of each decision, the outcome phase included the following regressors:

Outcome of chosen self task [1 (correct) or 0 (incorrect)],
Outcome of chosen poor partner's task [1 (rewarded) or 0 (unrewarded)],
Outcome of chosen good partner's task [1 (rewarded) or 0 (unrewarded)]

Then each regressor was normalised before inclusion into the analysis (mean of zero and standard deviation of one). These variables were time-locked to the onset of the outcome stage. The outcome variable was defined for the task chosen: for example, if participants chose the self task, then good and poor partners' outcome were not defined.

fMRI-GLM2: covariate analysis in trials with a poor partner

Next, we were interested whether signals associated with the contrasts 'chosen self task coherence – chosen poor partner's task coherence' during the social metacognitive judgement stage covaried with individual meta-d' for the trials with a poor partner (Fig. 3e, left). We included the meta-d' values as covariates at the second stage of group analysis when averaging across participants (FLAME 1).

fMRI-GLM3: covariate analysis in trials with a good partner

Next, we were interested whether signals associated with the contrasts 'chosen self task coherence – chosen good partner's task coherence' during the social metacognitive judgement stage covaried with individual meta-d' for the trials with a poor partner (Fig. 3e, right). We included the meta-d' values as covariates at the second stage of group analysis when averaging across participants (FLAME 1).

Region of interest (ROI) analyses

We calculated ROIs with a radius of three voxels (size = 33 voxels) that were centred on the alPFC$_{47}$ and TPJp: the MNI coordinates of these areas were [− 38, 40, − 10] and [52, − 54, 28] which were drawn from previous studies[12,22]. The selected ROI was transformed from MNI space to subject space and the pre-processed BOLD time courses were extracted for each participant's session. Time courses were averaged across volumes, then normalised and oversampled by a factor of 20 for visualisation. ROI-GLMs were applied to each time point to derive beta weights per time point for each regressor. For analyses across conditions, we used the same principle as applied to the whole-brain fMRI-GLM1: we averaged across the group. For all ROI analyses, regressors were normalised (mean of zero and standard deviation of one).

For the time course analysis ROI-GLM1, we used the same parametric predictors described in the whole-brain fMRI analysis conducted with fMRI-GLM1 for the metacognitive judgement stage (Fig. 3b, c). We also time-locked the time courses to the same phase on sets as described in fMRI-GLM1.

For the time course analysis ROI-GLM2, we used parametric predictors for the metacognitive judgement stage as follows (Fig. 3d):

Self-task coherence (poor partner),
Partner's task coherence (poor partner),
Self task coherence (good partner),
Partner's task coherence (good partner),
Chosen task (self: 1, poor:0) (poor partner),
Chosen task (self: 1, poor:0) (good partner)

Then each regressor was normalised before inclusion into the analysis (mean of zero and standard deviation of one). Regressors related to unavailable partners in a trial were not defined for the trial.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data collected in this study have been deposited in the RIKEN CBS Data Sharing Platform (https://neurodata.riken.jp/) with an identifier (https://doi.org/10.60178/cbs.20241115-001). Source data are provided in this paper.

Code availability

Code scripts generated in this study have been deposited in Zenodo (https://zenodo.org/) with an identifier (https://doi.org/10.5281/zenodo.14174382).

References

1. Heyes, C., Bang, D., Shea, N., Frith, C. D. & Fleming, S. M. Knowing ourselves together: The cultural origins of metacognition. *Trends Cogn. Sci.* **24**, 349–362 (2020).
2. Wittmann, M. K., Lockwood, P. L. & Rushworth, M. F. S. Neural mechanisms of social cognition in primates. *Annu. Rev. Neurosci.* **41**, 99–118 (2018).
3. Behrens, T. E., Hunt, L. T. & Rushworth, M. F. The computation of social behavior. *Science* **324**, 1160–1164 (2009).
4. Wittmann, M. K. et al. Self-other mergence in the frontal cortex during cooperation and competition. *Neuron* **91**, 482–493 (2016).
5. Bang, D., Moran, R., Daw, N. D. & Fleming, S. M. Neurocomputational mechanisms of confidence in self and others. *Nat. Commun.* **13**, 4238 (2022).
6. Wittmann, M. K. et al. Causal manipulation of self-other mergence in the dorsomedial prefrontal cortex. *Neuron* **109**, 2353–2361 (2021).
7. Park, S. A., Miller, D. S. & Boorman, E. D. Inferences on a multidimensional social hierarchy use a grid-like code. *Nat. Neurosci.* **24**, 1292–1301 (2021).
8. Trudel, N., Lockwood, P. L., Rushworth, M. F. S. & Wittmann, M. K. Neural activity tracking identity and confidence in social information. *Elife* **12**. https://doi.org/10.7554/eLife.71315 (2023).
9. Miyamoto, K. et al. Causal neural network of metamemory for retrospection in primates. *Science* **355**, 188–193 (2017).
10. Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J. & Rees, G. Relating introspective accuracy to individual differences in brain structure. *Science* **329**, 1541–1543 (2010).
11. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
12. Miyamoto, K. et al. Identification and disruption of a neural mechanism for accumulating prospective metacognitive information prior to decision-making. *Neuron* **109**, 1396–1408 (2021).
13. Mitchell, J. P. Inferences about mental states. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **364**, 1309–1316 (2009).
14. Tamir, D. I. & Mitchell, J. P. Neural correlates of anchoring-and-adjustment during mentalizing. *Proc. Natl. Acad. Sci. USA* **107**, 10827–10832 (2010).
15. Thornton, M. A. & Mitchell, J. P. Consistent neural activity patterns represent personally familiar people. *J. Cogn. Neurosci.* **29**, 1583–1594 (2017).
16. Tamir, D. I. & Mitchell, J. P. Anchoring and adjustment during social inferences. *J. Exp. Psychol. Gen.* **142**, 151–162 (2013).
17. Parsons, L. M. Imagined spatial transformations of one's hands and feet. *Cogn. Psychol.* **19**, 178–241 (1987).
18. Petit, L. S., Pegna, A. J., Mayer, E. & Hauert, C. A. Representation of anatomical constraints in motor imagery: mental rotation of a body segment. *Brain Cogn.* **51**, 95–101 (2003).

19. Jiang, S., Wang, S. & Wan, X. Metacognition and mentalizing are associated with distinct neural representations of decision uncertainty. *PLoS Biol.* **20**, e3001301 (2022).
20. Maniscalco, B. & Lau, H. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn.* **21**, 422–430 (2012).
21. Petrides, M. & Pandya, D. N. Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human and the macaque brain and corticocortical connection patterns. *Eur. J. Neurosci.* **11**, 1011–1036 (1999).
22. Mars, R. B. et al. Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cereb. Cortex* **22**, 1894–1903 (2012).
23. Konovalov, A., Hill, C., Daunizeau, J. & Ruff, C. C. Dissecting functional contributions of the social brain to strategic behavior. *Neuron* **109**, 3323–3337 (2021).
24. Saxe, R. Uniquely human social cognition. *Curr. Opin. Neurobiol.* **16**, 235–239 (2006).
25. Shadlen, M. N. & Kiani, R. Decision making as a window on cognition. *Neuron* **80**, 791–806 (2013).
26. Wolpert, D. M., Doya, K. & Kawato, M. A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**, 593–602 (2003).
27. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
28. Middlebrooks, P. G. & Sommer, M. A. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75**, 517–530 (2012).
29. Jeurissen, D., Shushruth, S., El-Shamayleh, Y., Horwitz, G. D. & Shadlen, M. N. Deficits in decision-making induced by parietal cortex inactivation are compensated at two timescales. *Neuron* **110**, 1924–1931 (2022).
30. Miyamoto, K., Rushworth, M. F. S. & Shea, N. Imagining the future self through thought experiments. *Trends Cogn. Sci.* **27**, 446–455 (2023).
31. Soutschek, A., Moisa, M., Ruff, C. C. & Tobler, P. N. Frontopolar theta oscillations link metacognition with prospective decision making. *Nat. Commun.* **12**, 3943 (2021).
32. 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).
33. Boorman, E. D., Behrens, T. E., Woolrich, M. W. & Rushworth, M. F. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743 (2009).
34. Mahmoodi, A. et al. Causal role of a neural system for separating and selecting multidimensional social cognitive information. *Neuron* **111**, 1152–1164 (2023).
35. Cortese, A., De Martino, B. & Kawato, M. The neural and cognitive architecture for learning from a small sample. *Curr. Opin. Neurobiol.* **55**, 133–141 (2019).
36. Bennur, S. & Gold, J. I. Distinct representations of a perceptual decision and the associated oculomotor plan in the monkey lateral intraparietal area. *J. Neurosci.* **31**, 913–921 (2011).
37. Smith, S. M. et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208–S219 (2004).
38. Johnen, V. M. et al. Causal manipulation of functional connectivity in a specific neural pathway during behaviour and at rest. *ELife* **4**. https://doi.org/10.7554/eLife.04585 (2015).
39. Derosiere, G., Vassiliadis, P. & Duque, J. Advanced TMS approaches to probe corticospinal excitability during action preparation. *Neuroimage* **213**, 116746 (2020).
40. Rossini, P. M. et al. Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee. *Clin. Neurophysiol.* **126**, 1071–1107 (2015).
41. Rossi, S., Hallett, M., Rossini, P. M. & Pascual-Leone, A. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin. Neurophysiol.* **120**, 2008–2039 (2009).
42. Wischnewski, M. & Schutter, D. J. Efficacy and time course of theta burst stimulation in healthy humans. *Brain Stimul.* **8**, 685–692 (2015).

Acknowledgements

We would like to thank Dr Ken-ichi Ueno and Dr Chisato Suzuki for supporting the MRI scan and data preprocessing at the RIKEN Centre for Brain Science, and Dr Marco Wittmann and Dr Tomoko Yamagata for helpful discussion. This research was supported in part by MEXT/JSPS KAKENHI Grants JP22K18665, JP22H00092 and JP23H03843, and by AMED under Grant JP24wm0525034 and JP23wm0625001 to K.M., by RIKEN Junior Research Associate Programme to S.T., by JSPS Fellowships (DC1, PD) to M.S., Wellcome Trust (221794/Z/20/Z) and MRC (MR/P024955/1) grants to M.F.S.R., and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 681422 (MetCogCon) to N.S. All collaborators of this study have fulfilled the criteria for authorship required by Nature Portfolio journals and have been included as authors, as their participation was essential for the design and implementation of the study. Roles and responsibilities were agreed among collaborators ahead of the research.

Author contributions

K.M. and M.F.S.R. designed the research. C.H., S.T., M.S., S.L., S.M. and P.S. conducted experiments and collected data. K.M., N.T., N.S. and M.F.S.R. conceived behavioural and neural analyses. K.M. and M.S. conducted data analyses. A.M. and M.L. developed the setup for experiments. K.M., N.T., N.S. and M.F.S.R. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-55202-0.

Correspondence and requests for materials should be addressed to Kentaro Miyamoto.

Peer review information *Nature Communications* thanks Xiaohong Wan, and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.