Dissecting functional contributions of the social brain to strategic behavior

Konovalov, Arkady ; Hill, Christopher ; Daunizeau, Jean ; Ruff, Christian C

DOI: https://doi.org/10.1016/j.neuron.2021.07.025

Dissecting Functional Contributions of the Social Brain to Strategic Behavior

Arkady Konovalov[1,3,4,5], Christopher Hill[1,5], Jean Daunizeau[2], Christian C. Ruff[1,4]

[1] Zurich Center for Neuroeconomics (ZNE), Department of Economics, University of Zurich, Zurich 8006, Switzerland
[2] Université Pierre et Marie Curie, Paris, France, Institut du Cerveau et de la Moelle épinière, Paris, France, INSERM UMR S975, Paris, France
[3] Lead contact
[4] Correspondence: arkady.konovalov@uzh.ch; christian.ruff@uzh.ch
[5] These authors contributed equally

Summary

Social interactions routinely lead to neural activity in a "social brain network" comprising, among other regions, the temporoparietal junction (TPJ) and the dorsomedial prefrontal cortex (dmPFC). But what is the function of these areas – are they specialized for behavior in *social contexts* or do they implement computations required for dealing with any *reactive process*, even non-living entities? Here, we use fMRI and a game paradigm separating the need for these two aspects of cognition. We find that most social-brain areas respond to both social and non-social reactivity, rather than just to human opponents. However, the TPJ shows a dissociation from the dmPFC: Its activity and connectivity primarily reflect context-dependent outcome processing and reactivity detection, while the dmPFC engagement is linked to implementation of a behavioral strategy. Our results characterize an overarching computational property of the social brain but also suggest specialized roles for subregions of this network.

Introduction

Strategic interactions lead to activity in the so-called "social brain" (Adolphs, 2009), a network comprising the temporo-parietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), precuneus, and right and left temporal poles (rTP and lTP, extending into the superior temporal sulcus (STS)). This activity is usually thought to reflect the need to simulate mental states of the interaction partners, called "mentalizing" or "theory of mind" (ToM) (Baron-Cohen et al., 1985; Carrington and Bailey, 2009; Coricelli, 2005; Fletcher et al., 1995a; Frith and Frith, 2005; Gallagher and Frith, 2003; Koster-Hale and Saxe, 2013; Saxe, 2006a, 2006b; Saxe and Kanwisher, 2003; Schaafsma et al., 2015; van Veluw and Chance, 2014). One region in particular - the right TPJ (rTPJ) - might be crucial for this function (Carter and Huettel, 2013; Carter et al., 2012; Decety and Lamm, 2007; FeldmanHall et al., 2013; Frith and Frith, 2001; Morishima et al., 2012; Santiesteban et al., 2012; Saxe, 2006a; Tso et al., 2018; Young et al., 2010a) since it is activated during strategic interactions for which mentalizing is important, such as bluffing in poker (Carter et al., 2012), manipulation in the trust game (Bhatt et al., 2010), detecting false advice (Behrens et al., 2008; Diaconescu et al., 2017), consensus decision-making (Suzuki et al., 2015), and strategic thinking in general (Hampton et al., 2008; Coricelli and Nagel, 2009; Hill et al., 2017; Ogawa and Kameda, 2019).

What functional contributions do social-brain areas make in these contexts? The theories put forward to answer this question can be grouped into two families. The first is *social-oriented*, focusing on the idea that these areas may implement cognitive functions that are unique to social contexts, such as animacy detection (Schultz et al., 2005), theory of mind (Saxe, 2010), or social context construction (Carter and Huettel, 2013). These theories thus assume that, in analogy to how some visual areas evolved to represent faces (Kanwisher et al., 1997) or body parts (Downing et al., 2001), the

evolutionary pressure for social behavior shaped brain processes dealing with the *presence of other humans.*

The second family is *process-oriented.* It assumes that the social brain may be recruited during strategic interactions because this type of situation requires particular domain-general cognitive operations. Again, different theories focus on different operations, such as attentional re-orientation (Cabeza et al., 2012), multimodal integration (Decety and Lamm, 2007), updating of internal predictive models (Geng and Vossel, 2013), response inhibition (Kolodny et al., 2017), or reactive control (Knyazev et al., 2019). However, all theories assume that both social and non-social situations may require these operations for non-habitual, attentionally demanding behavioral control. This is often the case when one interacts with other humans but can also be required in non-social environments (e.g., invalid cues or no-go responses).

These two perspectives are not mutually exclusive. For instance, the social-context construction account proposes that in non-social contexts, the rTPJ could still be engaged in constructing a "virtual" social context (Carter and Huettel, 2013), as perhaps reflected in the human tendency to anthropomorphize non-animate processes (Chaminade et al., 2007; Mar et al., 2007; Cullen et al., 2014). Thus, much of the debate about social brain functions has moved from examining "pure" social specificity (Mitchell, 2008; Young et al., 2010b) to defining the predictive problem that can co-occur with the presence of a social context (Koster-Hale and Saxe, 2013). However, this is not clear-cut even for mentalizing (Frith and Happé, 1994; Heyes, 2014; Schaafsma et al., 2015).

One of the crucial problems that needs to be addressed during strategic interactions is *reactivity of the environment*. Formally, reactivity refers to any situation where environmental events (for instance, actions of another person) are influenced by the individual's behavior. This is particularly clear for learning processes in strategic

interactions. For example, cognitive hierarchy theory (Camerer et al., 2004) assumes that players employ different levels of strategic sophistication to predict the opponent's actions and win the game. To defeat an opponent who relies on first-order beliefs (i.e., tracks choice frequencies to predict "what will the opponent do?"), one must mentalize at the level of second-order beliefs (i.e., simulate "what does my opponent think I will do?"). These strategies often necessitate learning and belief-updating processes to manipulate the frequency and sequence of one's own actions to influence the opponent's strategy (Camerer et al., 2004; Devaine et al., 2014a, 2014b; Hampton et al., 2008; Zhu et al., 2012). Such processes can be characterized in a reinforcement learning framework as updates of action values upon observing the resulting outcomes (Burke et al., 2010; Lockwood et al., 2016; Zhu et al., 2012). Other studies have focused on how people update the estimated probabilities of future opponent choices, based on the opponent's past actions and inferred opponent beliefs (Devaine et al., 2014a; Forgeot d'Arc et al., 2020; Hampton et al., 2008; Hill et al., 2017). Both these types of learning mechanisms are thought to be implemented by the social brain network and more generic reward-processing regions, with slight variations depending on what is being updated by the learner (reward prediction error, action value update, or belief update) (Lockwood et al., 2020; Ruff and Fehr, 2014). However, the exact functional specialization of specific regions remains unclear, and we know little about how these mechanisms are employed in social versus non-social contexts, or under different computational demands (in our case, dealing with reactive environments).

For example, while rTPJ activity appears causally necessary for the ability to track the effect of one's own actions onto an opponent in a competitive social context (Hill et al., 2017), computations related to reactivity may also be needed in non-social contexts. Humans routinely interact with non-social objects that react according to the rules of

biology (e.g., animals during hunting and farming, bendy trees during climbing) or natural physics (e.g., moving across differentially slippery or bouncy surfaces). It could be efficient for the brain to employ generic neural computations for processing all types of reactivity (Blakemore et al., 2001, 2003; Dehaene, 2005; Dehaene and Cohen, 2007; Grossman et al., 2000). However, most previous studies did not test whether activity in social brain regions reflects specifically the social nature of interaction or whether it instantiates general-purpose computations dealing with the reactivity of the environment (but see Ogawa and Kameda, 2019 and Devaine et al., 2014a).

Here we set out to explicitly test the general predictions of these families of models against one another. We did so by measuring BOLD activity in the social brain network during repeating strategic choices that were matched across many different dimensions (visual input, reward-relevant stimulus action associations, motor output) but that differed in whether choices were taken in a *social context* (i.e., against human opponents) and whether the (social or non-social) opponent showed *reactivity* to actions.

We had participants play a standard, robust, well-defined game (matching pennies; **Figure 1A** and **Methods**) framed to require them to either learn and predict a human opponent's next choice of a card (social context, N = 31 participants) or the next draw from a virtual card deck (non-social context, N = 29 participants). In both contexts, the opponents or card decks were in fact simulated by the same two artificial, computer-generated opponents: An active "*learner*" reacting to how the subject behaves (reflecting a "fictitious play" mechanism commonly used to describe behavior in strategic learning situations (Camerer, 2003)) and a passive "*sequencer*" producing sequences that crucially did not react to the subject's choices but required sequential prediction (a common motor and perceptual task) ((Clegg et al., 1998); **Figure 1** and **Methods**).

This approach allowed us to decompose activity in the social brain network into components reacting to context (social vs non-social), opponent algorithm (reactive vs non-reactive), and interactions of both factors.

Results

Behavior: Context cues strategy use

We tested whether participants' choices against the *learner* or the *sequencer* algorithm led to different behavioral success in the *social* versus *non-social context*. We found an interaction effect in both the fMRI dataset (N = 60) and a pilot behavioral dataset (N = 20) (see **Methods** and **Figure 2A**, mixed-effects regression, fMRI: $z = 2.27$, $p = 0.02$, behavioral: $z = 2.57$, $p = 0.01$, combined: $z = 3.2$, $p = 0.001$). Against the *learner*, subjects performed better in the social context ($t(78) = 3.4$, $p = 0.001$), while performance against the *sequencer* was the same in both contexts ($t(78) = 0.05$, $p = 0.956$). Sequencer and learner performance did not differ in both non-social ($t(37) = 1.335$, $p = 0.19$) and social ($t(41) = -1.49$, $p = 0.14$) conditions. This lack of differences in performance between the two algorithms was systematic and not due to clustering of subjects in groups with different performance profiles: Reward-rate differences between both algorithms were distributed unimodally and symmetrically, and were well fit by a Gaussian with mean 0 and s.d. 0.11 (see **Figure S1**). Only for 13 subjects out of 60 did we find a significant difference at $p < 0.05$ (two-sample proportion test, 5 better against sequencer and 8 better against learner). Additionally, the results confirmed that the task was not too easy for the subjects, since the average performance level across subjects was below 60% and no subjects reached the optimal level of performance (80% due to the opponent's decision noise, see **Methods**).

Thus, the results suggest that a social context cues a specific computational strategy that benefits dealing with a reactive opponent, which is consistent with previous behavioral findings (Devaine et al. 2014a, Forgeot d'Arc 2020). To investigate what strategy people may have been cued to use in the social context, we fitted a model to subjects' choices that allowed us to tease apart the influence of context and opponent reactivity, as well as the impact of own/other's choices. Our experiment was designed so that each opponent algorithm was best countered with a distinct strategy that guaranteed optimal performance. Against the *learner*, the optimal strategy was to alternate one's choice on each trial, which makes choices perfectly anti-correlated, sets the frequency of each choice to 0.5, and prevents the 0-ToM learner algorithm from exploiting the subject's choices (allowing the subject to win 80% of trials, see **Methods** for details). Against the *sequencer* algorithm we employed, the optimal strategy was to switch one's choice when the algorithm repeated its outcome twice, thereby introducing an anti-correlation between the subject's current choice and the opponent's choice two trials back (and to a lesser degree also with the subject's choice two trials back). This strategy also predicted the algorithm's choice in 80% of trials (see **Methods** for details).

To identify our subjects' use of these strategies in a unified framework, we employed a regression model in which the choice on each trial was determined by a linear combination of the subject's own choices and the algorithm's choices in the two preceding trials (linear Volterra decomposition truncated at order 2) (Barraclough et al., 2004; Devaine et al., 2014a; Lee et al., 2004). This unified model can quantify and compare the use of other- and self-referential thinking (first-order and second-order beliefs) without the need to apply two structurally different learning models optimized for each algorithm type. Specifically, the logistic regression predicted the subject's current choice in trial t by their choices in trials $t-1$ and $t-2$ (self-referential inputs),

and the opponent's choices at $t-1$ and $t-2$ (other-referential inputs, see **Methods** for details). In this framework, the optimal strategy against the *learner* is to maximize the negative weight on your own choice at $t-1$ (i.e., to switch choices; we label this coefficient κ), whereas the optimal strategy against the *sequencer* is to maximize the negative weight on the opponent's choice at $t-2$ (i.e., to choose the opposite of what the opponent played two trials before; we label this coefficient γ). For ease of presentation, we invert the coefficient signs so that successful strategy implementation results in a positive correlation with reward rate.

Subjects' choices matched the model predictions very well: 71.5% (s.d. = 6%) of choices were correctly captured by the model (balanced classification accuracy). Importantly, this model performed better than standard specialized strategic-learning models such as Q-learning, win-stay-lose-shift, influence learning, and Markov matrix-based sequence learning (see **Supplementary Materials, Figure S1**).

To quantify the importance of strategy use for behavioral success, we regressed individual reward rates on all individual coefficients, separating the data by algorithm type (**Figure 2B**, top panel). As expected given our algorithm design, the κ-weight correlated with performance against the *learner* ($r(78) = 0.77$, $p < 0.001$, **Figure 2B,** bottom left panel) whereas the γ weight correlated with performance against the *sequencer* ($r(78) = 0.86$, $p < 0.001$, **Figure 2B,** bottom right panel; the weight on own choice two trials back also correlates with reward rate because is trivially correlated with the opponent's choice two trials back). The regression also confirmed that the increased performance against the *learner* in the social context was accompanied by a significant increase in the κ parameter (**Figure 2C**, mixed effects regression, social vs non-social, $z = 4.6$, $p < 0.001$) and a decrease in the γ parameter (**Figure 2C**, mixed effects regression, $z = 2.76$, $p = 0.006$). Against the *sequencer*, the social context manipulation did not lead to

any significant difference between the weights (**Figure 2C**, mixed effects regression, z = 0.28, p = 0.78 for κ, and z = 0.45, p = 0.65 for γ). Confirming the link between behavioral success and context-dependent application of a specific strategy, the subjects who showed the largest difference in reward rates between the two opponents also showed the largest difference between the κ and in γ parameters ($r(59) = 0.82$, $p < 0.001$, Figure S1).

These results suggest that subjects assigned more weight to second-order beliefs – modeled as the κ parameter – when competing in a social situation (**Figure 2C**). The better performance against the *learner* algorithm in the social context appears to reflect use of a specific model-based strategy that is well suited against this algorithm. Thus, presence of a social context might cue dedicated neural computations instantiated by the social brain. We examined this possibility in more detail with the fMRI analysis described in the next section.

fMRI: Activity in the social brain reflects both context and algorithm type

To test whether activity in the social brain contains information just about social context, or reactive/non-reactive properties of the opponent algorithm, or both, we carried out extensive region-of-interest (ROI) analyses. We defined a-priori ROIs of the social-brain network using the automated meta-analytic tool Neurosynth (Yarkoni et al., 2011), resulting in functional masks for right and left TPJ, right and left temporal pole, dmPFC and precuneus (**Figure 3A**, **Table S1**, **Methods**). Given previous results suggesting a role for reward-outcome-processing regions in strategic behavior (Hill et al., 2017; Ruff and Fehr, 2014; Zhu et al., 2012), we also included the nucleus accumbens as the region most reliably activated by reward (Knutson and Gibbs, 2007).

In line with previous results, we focused our analyses on outcome signals, that is, the feedback generated by wins and losses. Such outcome signals are ubiquitous across the brain, possess high signal-to-noise ratio, and are necessary for learning in both non-social (Daw et al., 2012) and social (Hampton et al., 2008; Hill et al., 2017; Zhu et al., 2012) contexts. These three qualities make these signals especially likely to encode properties of our task, such as the social context and the algorithm type, as well as the interactions of these conditions with reward (which is only revealed in the feedback stage and shapes subsequent behavior). However, for completeness, we also report the results of these analyses for the choice stage (see **Figure S2**).

We quantified the effects of context, algorithm, and outcome (as well as their interactions) on the feedback-related ROI BOLD signals using a linear mixed-effects model (**Figure 3B, Methods**). This analysis confirmed that all areas reacted strongly to outcome (winning versus losing, all $p < 0.001$, see **Figure 3B** for t-statistics). Most crucially, however, almost all social-brain regions were activated more strongly when participants played against the *learner* opponent versus when they competed against the *sequencer* (all $p(FDR) < 0.05$; see **Figure 3;** see **Figure S3** for beta plots for each region by condition). The only two exceptions were the precuneus ($t(538) = 0.92$, $p = 0.36$) and the nucleus accumbens ($t(538) = -0.65$, $p = 0.51$). Remarkably, *social context* per se did not lead to differential activations across the network (for the precuneus, $t(178) = -1.69$, $p = 0.09$, for all other regions $t < 0.69$, $p > 0.49$). This suggests that dealing with reactivity of the environment is a core computational contribution of the social brain to the control of behavior.

Our analysis also revealed a functional dissociation in the response profile of different areas in the social brain network. The right TPJ exhibited a complex pattern: Not only was it the region most responsive to the learner versus sequencer algorithm ($t(538)$

= 5.7, $p < 0.001$), but it also showed an interaction of social context with algorithm ($t(535) = -2.06$, $p = 0.04$) and reward ($t(535) = 2.94$, $p = 0.003$). This pattern appears clearly reminiscent of the interaction effect observed in the analysis of the behavioral data (**Figure 2A**). No such interaction effects were observed in most of the other areas, including dmPFC ($t < 1.69$, $p > 0.1$), which only responded to opponent type. Only the precuneus ($t(535) = 4.02$, $p < 0.001$) and the nucleus accumbens ($t(535) = 2.38$, $p = 0.02$) showed a different type of interaction between social context and reward.

It is important to note that the context manipulation was between-subject, so the corresponding analysis had lower statistical power than the algorithm manipulation. However, the weak effect size suggests that statistical inference about the presence of a social context effect would not have changed even if we had collected a larger sample. Specifically, the betas extracted from the main 7 ROIs used in all analyses show effect sizes <0.13 for the difference between the social and non-social context. Using 60 subjects and a within-subject design would only increase the power to detect these effect sizes to 0.17 and still not yield any significant effect. The only exception is the precuneus ($p = 0.23$, effect size of 0.31); however, this difference could be also driven by the interaction with outcome that we report.

Participants clearly believed the social context framing, as ascertained by post-experiment questionnaires (see **Methods**). Nevertheless, as an additional robustness check, we repeated our analyses and excluded 14 subjects who were less convinced by the social/non-social framing as indicated by their Social Belief Index (SBI, see **Methods**). We again found no difference in BOLD activity between the social and non-social contexts in all 7 ROIs ($p(FDR) > 0.5$), confirming that the lack of difference in the BOLD signal in these regions was not caused by a potential failure of the deception protocol. We also found no significant correlation between the social belief index (SBI) and performance,

for both the sequencer (r = -0.1, t(58) = -0.78, p = 0.43) and learner (r = 0.18, t(58) = 1.4, p = 0.17) opponents (and no correlations when looking specifically at the social/non-social groups). This suggests that our results, and in particular the lack of a clear neural social context effect, is not due to differences in beliefs about the opponent across the two contexts.

We obtained similar results in an exploratory full-brain contrast analysis (**Figure 3C, Table S2, Table S3**): Activity in the ToM network was not preferentially driven by the social context but was rather related to the reactivity of the environment (Learner > Sequencer contrast). However, in the rTPJ, the neural computations required to solve this problem appeared to be cued by social context (**Figure 3C**, interaction between the context and outcome) (t(56) = -5.1, MNI peak x = 31, y = 21, z = 50). The only region that had a stronger response to winning against the sequencer was outside the social-brain network, in a region of the intraparietal sulcus (IPS, t(59) = 4.5, MNI peak x = -42, y = -42, z = 48) previously found to be involved in sequential prediction (Konovalov and Krajbich, 2018).

To consolidate our results, we ran a robust out-of-sample machine learning analysis of ROI BOLD activations on the run/opponent/outcome level (Pereira et al., 2009). This analysis predicted the experimental conditions (algorithm type and context) using neural beta coefficients for winning versus losing outcomes, averaged across all voxels (weighted by signal intensity) within the ROIs as model features.

In line with the results of the main analysis, outcome-related activity in the social brain contained information about both the social context and the algorithm (**Figure 3D**, algorithm: 63.3%, p = 0.008 context: 61.6%, p = 0.021), with notable differences between different areas. The right TPJ was the only region where both labels could be decoded significantly (algorithm: 59.1%, p = 0.03, context: 65%, p = 0.02). For the dmPFC (and

both temporal poles), we could decode only the algorithm but not the context (context: 45%, p = 0.71, algorithm: 63.3% p = 0.003). Conversely, precuneus activity could only be used to decode the context but not the algorithm (context: 65%, p = 0.03, algorithm: 56% p = 0.09). Neither the social context nor the algorithm could be decoded from outcome-related activity in the nucleus accumbens (context: 51.6%, p = 0.41, algorithm: 51.6% p = 0.33).

These results further suggest that the social brain network mainly implements computations to deal with the reactive properties of the environment, but the rTPJ plays a special related to cuing of such computational strategies by social context.

fMRI: Activity in the social brain network reflects distinct computational specialization

Having confirmed that neural signals in the ToM network are modulated by both social context and algorithm type, we examined model-based computations related to valuation and learning. To do so, we first confirmed that the BOLD signal observed at the choice stage (not the outcome stage, as analyzed above) encoded our model's predicted choice values. Consistent with previous findings (Bartra et al., 2013; Clithero and Rangel, 2013), the model-predicted value of the chosen option during choice was represented in the vmPFC (**Table S3**, peak at MNI x, y, z = -2, 56, -5, t(59) = 6.1, p(SVC) = 0.0002 using the mask from Bartra et al (2013)). This provides further neural validation of our behavioral model.

Next, we examined different learning signals derived from the literature that have been proposed to index different learning processes that may guide behavior in strategic interactions (Cooper et al., 2012; Diaconescu et al., 2017; McClure et al., 2003; Zhu et al., 2012). This analysis was motivated by the fact that in the present context with repeated

interactions and no prior knowledge about the opponents, the participants need to infer the strategy of the opponent from their observations. We selected several well-established generic variables capturing learning about other's actions in such situations, signals reflecting updating of action values, and a dynamic index of the opponent's reactivity to the participant's actions. We indexed these processes by four model-based regressors as parametric modulators of BOLD activity during the feedback stage: (1) a reward prediction error (RPE) that quantifies the deviations of the obtained from the expected reward, (2) an action prediction error (APE) that indexes how much the observed opponent action differs from the expected probability of this action, (3) a dynamic block-level estimate of observed opponent reactivity, estimated by modelling how much the opponent's current choices are guided by the participant's past actions (using the same model also employed to model the participant's actions, but now fit to the opponent's actions), and (4) the absolute strength of the choice value update (for detailed descriptions and definitions see **Methods**).

The standard reward prediction error signal was expressed in all ROIs of the social brain **(Figure 4A**). This measure is by construction strongly correlated with reward outcome (win/loss), so this result relates to the outcome encoding already described in our model-free analyses (**Figure 3B**). However, we also found a significant difference in encoding of RPE between the two contexts in the precuneus ($t(51) = 2.2$, $p = 0.03$). In the rTPJ, BOLD activity related to RPE only in the social context (but the difference between the contexts was not significant). This result mirrors the model-free activity pattern in the rTPJ, with more marked responses to outcome in the social context (**Figure S3**).

The analysis of the APE revealed further dissociation between the regions of the social brain. Across both contexts, the dmPFC, left temporal pole, and precuneus showed a significant correlation with this predictor ($p(FDR) < 0.05$). However, numerically the

dmPFC showed stronger encoding of the action PE in the non-social condition (**Figure 4A**, $t(30) = -3.2$, $p(FDR) = 0.01$), while the precuneus showed stronger correlation with APE in the social condition (**Figure 4A**, $t(30) = -3$, $p(FDR) = 0.02$). In both cases, however, the difference between the two contexts was not significant. Note that both regions show a negative correlation with APE; an increase in activity in these ROIs thus signals confirmation that the chosen strategy was right rather than an error signal indexing the deviation of the observed from the expected action.

Activity in rTPJ did not show encoding of APE, in neither context, but was the only region correlating significantly with the reactivity measure ($t(59) = 3.8$, $p(FDR) < 0.001$). Splitting the analysis by context, we found evidence for reactivity detection in left TPJ and both temporal poles specifically in the non-social context (**Figure 4A**). These results suggest that the TPJ and the TPs might engage in reactivity detection most strongly in situations in which such reactivity is surprising (i.e., in the non-social context). Please note that these analyses are again consistent with our model-free findings that these areas also show a stronger response to the (more reactive) learner algorithm in the non-social context (Figure 3B).

Finally, the analysis showed that the absolute action value update does not correlate with activity in the areas of the social brain, but rather in the nucleus accumbens ($t(59) = 4.27$, $p(FDR) < 0.001$), for both social and non-social contexts (**Figure 4A**).

Overall, these results are generally consistent with the preceding model-free analyses. Most notably, they suggest that the dmPFC is more involved in engagement in a specific strategy to predict opponent actions, whereas the rTPJ seems most concerned with context-specific signaling of the opponent's degree of reactivity.

Additionally, we tested whether the outcome-related neural signals in the social-brain regions detected by our analyses are indeed relevant for behavior, in particular for

the use a specific strategy when playing against a reactive or passive opponent. In the behavioral analysis, we had found that the subject performed better against the learner in the social context. Now, we investigated whether the individual subjects' κ and γ weights while playing against the two types of algorithms correlated with outcome-related BOLD activity in the pre-defined social ROIs. For each of these regions, we ran a mixed-effects model regressing choice weights (κ and γ) in each run against each opponent on BOLD activity during the feedback period (**Figure 4B**). Our results confirm that neural computations in the social-brain network relate to the incorporation of second-order beliefs into behavior: BOLD activity in several regions correlated with κ specifically against *learner* opponents (see **Figure 4B** for t-statistics). Specifically, during interactions with the learner in the social context, dmPFC activity was positively correlated with κ ($t(89) = 2.7$, $p = 0.008$) and negatively correlated with γ ($t(89) = -2.3$, $p = 0.02$) (**Figure 4C**), suggesting that this region plays a role in increasing the weight on the strategy that is optimal against the reactive algorithm (supporting the idea that the mPFC might guide choice during strategic interactions (Hampton et al., 2008; Hill et al., 2017)).

Moreover, the correlation between κ and BOLD activity was indeed specific for the social versus non-social context, in both the dmPFC (interaction terms in the mixed effects regressions, $t(174) = 2.25$, $p = 0.03$) and the precuneus ($t(174) = 2.48$, $p = 0.01$). Thus, while the rTPJ may be specialized for social-context cueing of strategies that are optimal against a reactive algorithm (e.g., detecting reactivity when this is surprising), the dmPFC may be more involved in implementing these strategies to control behavior when this is necessary (e.g., against the reactive learner opponent).

fMRI: The rTPJ shows stronger functional coupling with reward regions when playing against reactive opponents

To study functional integration in the social brain network, we carried out exploratory functional connectivity analyses (see **Methods**) of how the rTPJ interacted with other social-brain regions and nucleus accumbens in the different experimental contexts. We chose the rTPJ as the seed region since it was the only area that showed a significant interaction between the algorithm and context, in close similarity to the pattern of social-context cuing observed in the behavioral data. This suggests that the rTPJ may play a special role for guiding the strategy use implemented in the social brain network based on context and outcome signals. The analysis thus comprehensively considered all the experimental factors affecting brain activity (win vs loss, learner vs sequencer algorithm, and social vs non-social context; see **Figure 4D**).

During win vs loss outcomes, connectivity with the rTPJ was indeed increased for most social-brain regions, including the dmPFC (beta = 0.12, t(59)= 4.44, p(FDR) = 0.0004) and the right temporal pole (beta = 0.12, t(59)= 3.84, p(FDR) = 0.001), independent of algorithm type and context (t(59) < 1.3, p(FDR) > 0.31) (**Figure 4D**). This confirms the hypothesized tight integration in the social brain network, with the rTPJ communicating behaviorally-relevant outcome information to the interconnected areas.

No such general win- vs loss-related increases were evident for connectivity of the rTPJ and nucleus accumbens (beta = 0.01, t(59)= 0.42, p(FDR) = 0.75). However, the rTPJ showed differential connectivity with the nucleus accumbens when winning versus losing against a learner opponent (interaction of outcome and opponent. t(59) = 2.15, p(unc) = 0.03, p(FDR) = 0.31). Post-hoc analysis indicated that this result was primarily driven by loss outcomes (**Figure 4D**, last two columns, beta = 0.07, t(59)= 3.55, p(FDR) = 0.007). If rTPJ activity relates to use of mentalizing (as indexed by κ), and nucleus

accumbens activity integrates strategy signals in the values of choice options, these connectivity results may reflect specific rTPJ communication with nucleus accumbens when loss outcomes signal the need to update wrong mentalizing-based predictions of opponent choices.

Discussion

Our study was motivated by long-standing proposals that the social brain network has evolved to specifically support social interactions (Atzil et al., 2018; Frith, 2007; Frith and Frith, 2001), and by fMRI findings that mere presence of social context often triggers increased BOLD responses in social-brain regions (Frith, 2007; Lockwood et al., 2020; Tso et al., 2018; Van Overwalle, 2009). On the other hand, an increasing number of studies have linked these activations to specific computational operations, as formalized in computational models originating from decision and learning frameworks originally developed for non-social contexts (Konovalov et al., 2018; Lockwood et al., 2020). However, the specific functional contributions of the TPJ, precuneus, and dmPFC to guide behavior in these contexts remain a debated topic. The fundamental question still stands whether these areas are indeed specialized for social behavior, or whether they share their function with other non-social processes and computations (Lockwood et al., 2020; Ruff and Fehr, 2014).

Our results provide a new angle for the interpretation of results from standard laboratory Theory-of-Mind tasks, as often used to measure activity in social brain regions when the subject has to reason about mental states of other people (Carrington and Bailey, 2009, 2009; Saxe, 2006b; van Veluw and Chance, 2014), take other people's perspective (Krach et al., 2008; Tusche et al., 2016), or infer social constellations (Hooker et al., 2010; Janowski et al., 2013; Vanderwal et al., 2008). While these tasks often do not

employ direct social interactions/reactivity per se, they often imply reactivity that needs to be simulated (e.g., when a person has to judge someone's response to another human) or a degree of animacy that is attributed to non-social objects (Blakemore et al., 2003; Cross et al., 2016). Based on our new approach, it may be interesting to carefully design non-social control conditions (and potentially computational models) that could provide a more mechanistic view of how reactivity processing may underlie many of the brain responses observed during such classical laboratory tasks.

It is important to note that some of the situations that involve activation of the social brain do not necessarily involve a reactive environment. Notable examples include observational learning (Burke et al., 2010; Charpentier et al., 2020; Collette et al., 2017; Cooper et al., 2012; Hill et al., 2016), or learning of other people's preferences (Garvert et al., 2015), abilities and values (Boorman et al., 2013; FeldmanHall et al., 2017), and generosity (Stanley, 2016). One potential way to reconcile these results with our findings is to assume that the brain may simulate reactivity in many types of social choices (for instance, by imagining the other person's reaction to the individual's choices). Such a role for reactivity simulation in eliciting activity in the social brain may resemble the findings that in perceptual and motor brain systems, mere simulation of percepts or actions also leads to comparable activity as processing of observed or executed percepts and actions (Calvo-Merino et al., 2005; Hesslow, 2002; Kan et al., 2003; Lotze et al., 1999). Other possibilities are that reactivity processing is just a part of a more complex computation that is yet to be identified, or that parts of the social brain may not only process reactivity but also other aspects of social interactions, as suggested by our findings on functional specialization of the different areas (see Figure 4A).

We also recognize that our study addresses a general question about the social brain in just one specific experimental game setting (matching pennies, a task commonly

used to study strategic interactions) and against two specific types of algorithms. However, our experimental setting and algorithms reflect prototypical behavioral patterns observed in studies of strategic interactions (Camerer, 2003) and are designed to mirror behaviors that are ubiquitous in everyday life (Boylan and El-Gamal, 1993; Spiliopoulos, 2013). While it thus remains to be established whether our findings will also apply to other social settings (such as collaborative behavior), our study illustrates a general approach that may be extended to studies of other types of social interactions. We thus hope that our study not only establishes that the functions of different social brain areas can be quite different (at least during strategic competition as often encountered in everyday life), but that it also motivates further studies of how the specific computational roles of different social brain regions may be expressed in different settings.

Functional dissociation of the social brain areas

The social brain network may be less unitary that commonly thought, as its regions are always not engaged to a similar degree. Of course, many regions (specifically, TPJ and dmPFC) show similar response profiles during processing of self-other representations, social goals and values, and other related functions (Carter and Huettel, 2013; Decety and Lamm, 2007; Morishima et al., 2012; Ogawa and Kameda, 2019; Van Overwalle, 2009). However, many studies still detect a unique functional role for the TPJ (Donaldson et al., 2015; FeldmanHall et al., 2013; Hampton et al., 2008; Hill et al., 2017; Krall et al., 2015; Lee and McCarthy, 2016; Ogawa and Kameda, 2019; Santiesteban et al., 2012, 2015). For dmPFC, substantial evidence points to a critical role of this brain region in selecting between alternative models and responses even in non-social contexts (Hill et al., 2016; Kolling et al., 2016; Kovach et al., 2012; O'Reilly et al., 2013; Rushworth et al., 2004, 2011).

Moreover, the precuneus - another core social-brain area identified by meta-analyses (Cavanna and Trimble, 2006) - is also involved in a wide number of contexts, such as value-based decisions (Bartra et al., 2013), memory (Fletcher et al., 1995b) spatial navigation (Epstein, 2008; Hebscher et al., 2018), and default-mode activity (Utevsky et al., 2014). While all these findings suggest that TPJ, dmPFC, and precuneus may make distinct functional contributions to the control of behavior, their exact functional characterization in social contexts has remained an open question.

Our results provide crucial new information, as they support a clear functional dissociation between the rTPJ, dmPFC, and precuneus: While almost all regions of the social brain responded to the reactivity in the environment, the rTPJ response to outcomes differed between the social and non-social context, and showed stronger connectivity with nucleus accumbens when participants acted in a reactive environment. By contrast, activity in both dmPFC and precuneus reflected the behavioral strategies used against the reactive opponent, while rTPJ activity did not. This difference in response profile was also mirrored in the model-based analysis, which showed that the dmPFC encoded mainly signals confirming how correct the prediction of an opponent's action was, whereas the TPJ reacted most strongly to reactivity in opponents for which this was surprising. Our results thus suggest a specific dissociation of social brain function in terms of detecting reactivity, computing action value updates, and representing action prediction errors (Devaine et al., 2014a; Forgeot d'Arc et al., 2020; Hampton et al., 2008; Hill et al., 2017). This appears consistent with previous proposals that in social interactions, TPJ might be encoding contextual updates, temporary goals, and other instantaneously relevant variables, while dmPFC may serve as a hub that stores longer-term representations of other individuals and guides behavioral strategy in general (Jamali et al., 2021; Matsuzaka et al., 2012; McDonald et al., 2020; Van Overwalle,

2009; Venkatraman et al., 2009). In the next two sections, we will focus on these putative functional roles of the dmPFC and TPJ.

The rTPJ encodes context-dependent social outcomes

Our results suggest that the TPJ differentially processes outcomes and updates the associated beliefs depending on the social context. This implies a bottom-up processing role for the rTPJ in the social network, with this region processing trial-by-trial updates rather than long-term behavioral strategies (Hampton et al., 2008; Hill et al., 2017; Mengotti et al., 2017; Ong et al., 2021). Thus, our findings further refine previous proposals on the functional contributions of the rTPJ in social and non-social contexts. That is, our results are potentially inconsistent with the hypothesis that social context merely signals increased complexity and thus the need to deploy additional resources, as suggested by both the nexus model (Carter and Huettel, 2013) and the contextual-updating model (Geng and Vossel, 2013) of the rTPJ. The nexus model predicts that activity in the rTPJ should be mainly driven by the social context rather than by computational demands (which are carefully controlled across the two types of opponents we used). The contextual updating theory predicts that rTPJ activity should reflect a domain-general computation (updating of a contextual model) that is similarly present in all conditions of our experiment; this theory would therefore not predict any differential activity between social and non-social contexts.

Functional specialization of the rTPJ was also evident at the level of connectivity (see also Hill et al., 2017): The rTPJ was the only region showing increased connectivity with the nucleus accumbens when subjects faced the *learner* compared to the *sequencer*. While our connectivity analyses cannot establish directionality, this result may indicate that the nucleus accumbens may draw on input from the rTPJ to update its expectation

during outcome processing and/or that socially-relevant computations in rTPJ may be triggered by prediction-error signals in ventral striatum (Smith et al., 2014). Additionally, we observed an outcome-by-context interaction in connectivity between the left and right TPJ (however, this result did not reach significance under FDR correction), and a triple (outcome by context by algorithm) interaction in the dmPFC-rTPJ connectivity, further supporting a functional dissociation between the regions in the social network.

dmPFC engagement reflects top-down strategy implementation

Unlike the rTPJ and the precuneus, overall activity in the dmPFC was hardly sensitive to social context and mainly reflected the opponent type. This observation is compatible with prior accounts of this region's involvement in strategic decision making. In humans, the dmPFC has been involved in hierarchical beliefs in the beauty context game (Coricelli and Nagel, 2009), assessing (Jamali et al., 2021) and learning based on other's beliefs (Zhu et al., 2012), integrating social information in economic settings (De Martino et al., 2013, 2017), and use of second-order learning (Bhatt and Camerer, 2005; Bhatt et al., 2010; Hampton et al., 2008; Hill et al., 2017). In monkeys, this region has been found to encode signals that correlate with a monkey's tendency to anticipate the learning of a computer opponent in a non-social context. Only monkeys showing these signals were found to be able to out-compete a simple 0-ToM algorithm (Seo and Cai, 2014) (but note that monkeys may obviously show less sophisticated "theory of mind"-like behavior in competitive tasks compared to humans). Our "non-social learner" condition essentially mimics this study in humans, and it found that the dmPFC tracked the accuracy of opponent-actions predicted based on a specific strategy. Thus, our results

further consolidate the importance of the dmPFC for sustaining such behavioral strategies.

Specifically, our results support the idea that dmPFC might be controlling the sustained use of a specific behavioral strategy, as suggested by previous findings that activity in this area correlates with an index of how strongly participant's employed a particular mentalizing strategy (Hampton et al., 2008). Note that activity in the rTPJ did not reflect these long-term variables and mainly showed transient responses to changing outcome events. This fundamental distinction between the two social-network regions may resemble functional separation in attentional networks, with parietal regions more engaged in transient bottom-up processing and frontal regions engaged in sustained top-down attentional sets (Buschman and Miller, 2007; Katsuki and Constantinidis, 2014; Ruff et al., 2008). Anatomically, this distinction may also reflect that the rTPJ is more closely interconnected with the perceptual areas in the parietal cortex (Carter and Huettel, 2013), while the dmPFC is more densely connected with areas tied to cognitive control (Taren et al., 2011) as well as action selection and policy evaluation (Hill et al., 2016; Kolling et al., 2016; Kovach et al., 2012; O'Reilly et al., 2013; Rushworth et al., 2004, 2011).

Functional dissociation in the social brain: Potential clinical implications

Deficits characterized by ToM dysfunction are core symptoms of various disorders, in particular autism spectrum disorder (Baron-Cohen, 2000; Baron-Cohen et al., 1999; Kana et al., 2014, 2015; Marsh and Hamilton, 2011; Murdaugh et al., 2014). In principle, the computational theory-of-mind approach employed here could be applied to better understand the neurocognitive characteristics of individuals with such disorders. In fact, a behavioral variant of our approach already showed that, in people with autism,

cognitive computations employed to deal with reactivity are insensitive to context cues (i.e. when playing against a reactive artificial agent framed as a human opponent, neurotypical subjects outperform participants with autism spectrum disorder) (Forgeot d'Arc et al., 2020).

False-positive perception of animacy could be the byproduct of perceiving reactivity in the environment in response to one's actions. Such mechanisms may also be at work in gambling addiction: We know that many such patients suffer from the delusion that they can influence the random process (Toneatto et al., 1997). Our fMRI results provide a potential neural basis for such links between perceived reactivity and social interpretation.

ACKNOWLEDGEMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 725355, BRAINCODES).

AUTHOR CONTRIBUTIONS

C.H., J.D., and C.R. designed the experiment. C.H. programmed and conducted the experiment. All authors designed the models and analyses. A.K. and C.H. performed the data analysis. All authors contributed to the manuscript. C.R. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

MAIN FIGURES TITLES AND LEGENDS:

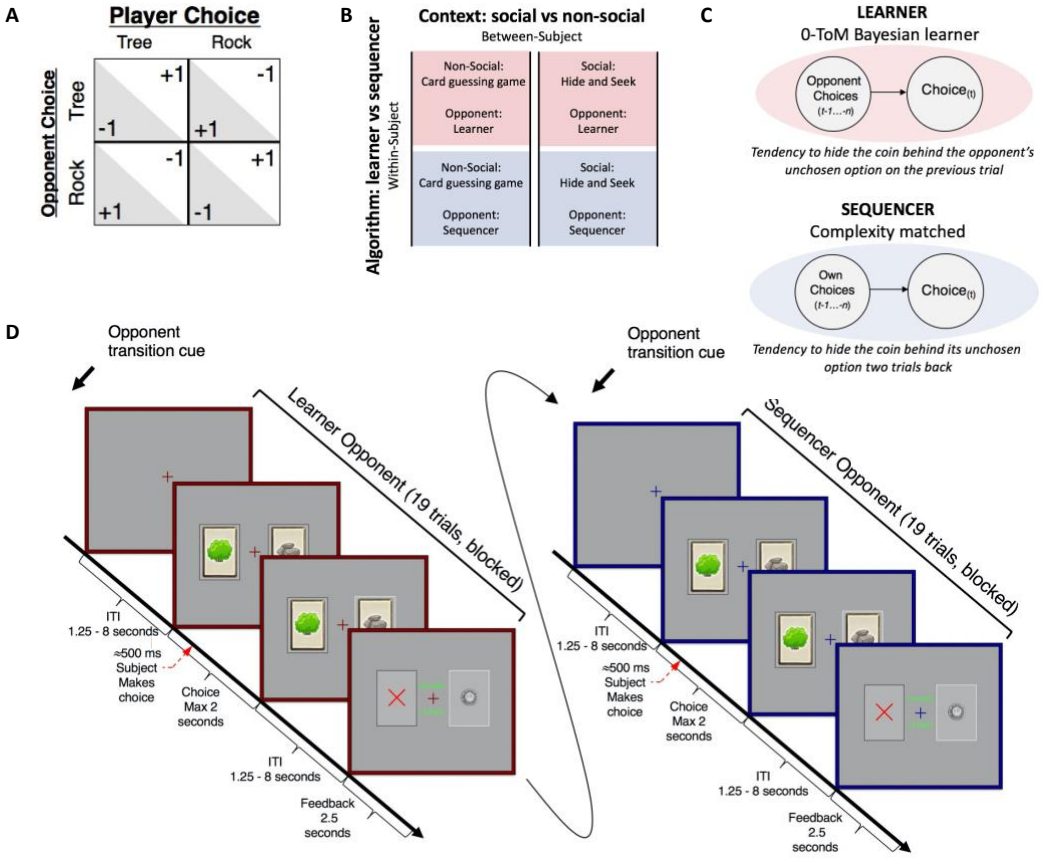

Figure 1. Experimental design. (A) The payoff matrix of the game. (B) 2x2 factorial design: Subjects were randomly sorted into two groups, one playing in social framing and the other playing against a deck of cards. Each subject played against two distinct opponents (or decks) represented by two computer algorithms: sequencer and learner. (C) The learner algorithm employed a 0-ToM strategy (Devaine et al., 2014a). The sequencer algorithm produced a noisy sequence of choices without reacting to the subject's choices. (D) fMRI task design. The subject played against each algorithm for a block of 19 trials before switching to the other algorithm, with 228 trials in total. The start of a new block was indicated by a transition screen, and the identity of the opponent was indicated with the color of the screen frame. See also Methods S1.

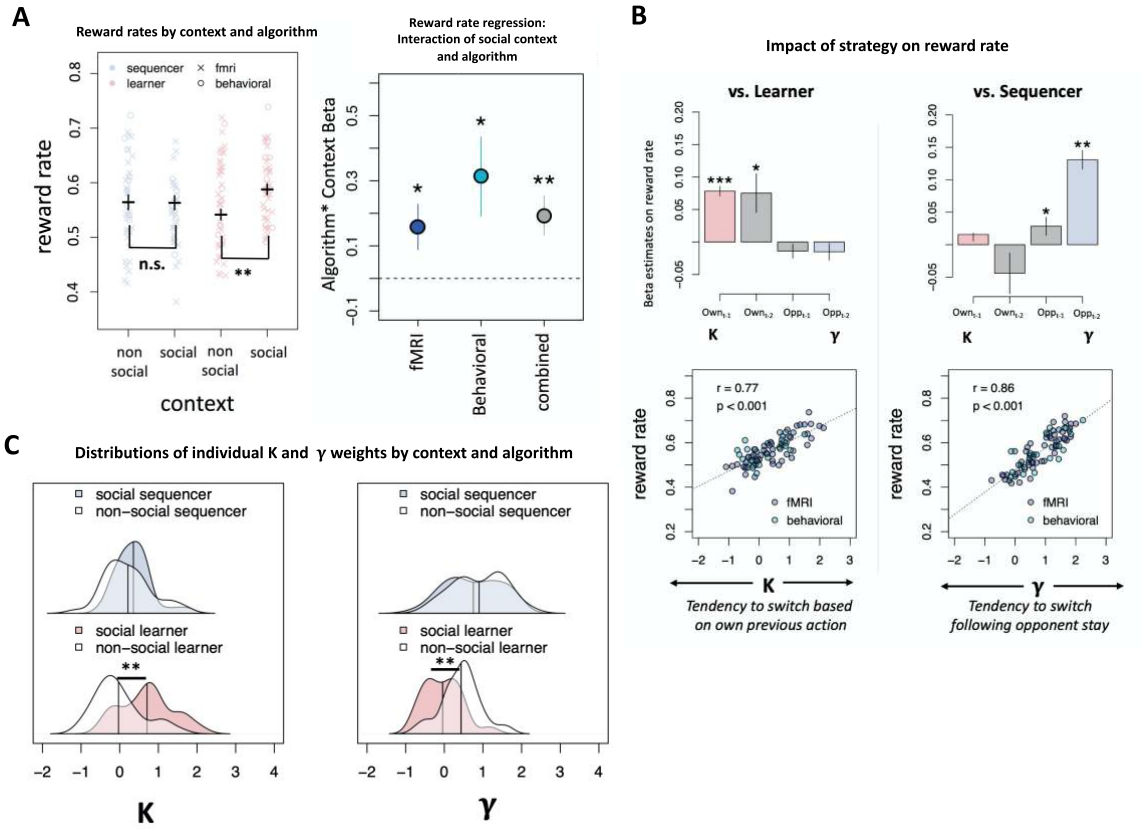

Figure 2: **Behavioral results.** (A) Left panel: Reward rate (cross: mean across subjects, dots: single subjects) split by algorithm type and social context. Right panel: Fixed effects coefficients for algorithm x context interaction in mixed-effects logistic regression of reward rates. (B) Top panel: Fixed-effects coefficients from a model regressing individual reward rates on individual choice model coefficients (impact of own and opponent's previous trial choice on the current trial choice). Note that for ease of presentation, we have inverted the coefficient signs so that successful strategy implementation results in a positive correlation with reward rate. Bottom panel: correlations (Pearson) between κ and behavioral success (reward rate) against the learner opponent and between γ and behavioral success against the sequencer opponent. (C) Distributions of individual coefficients (fixed + random effects) split by opponent type (sequencer vs learner) and context (social vs non-social). See also Figure S1.

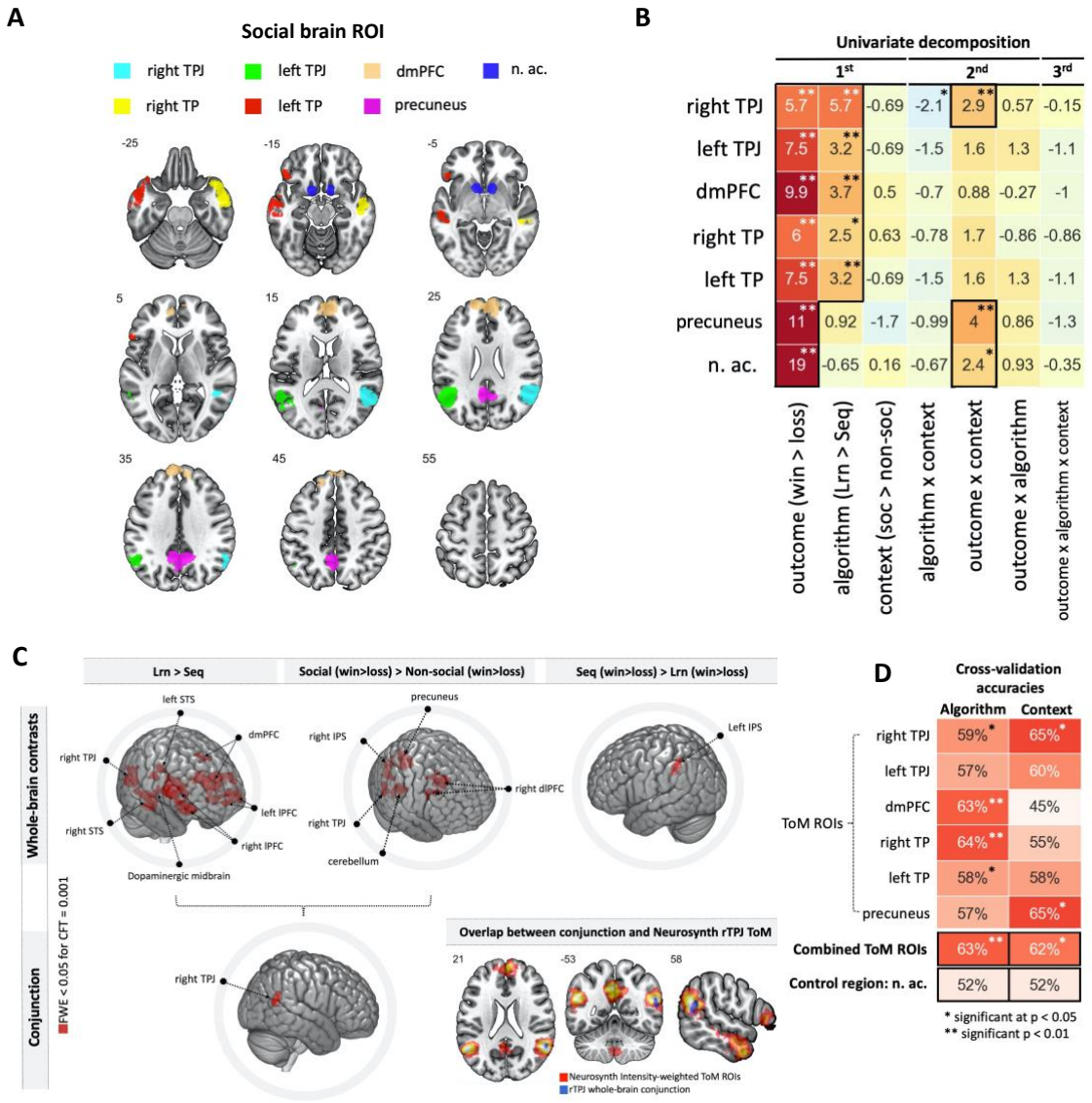

Figure 3: **Model-free fMRI results.** (A) ROIs for the social brain network identified by Neurosynth meta-analysis using the term "theory of mind": the right and left temporoparietal junctions (TPJ), right and left temporal poles (TP), dorsomedial prefrontal cortex (dmPFC), precuneus, and nucleus accumbens (n. acc). The ROIs are shown in axial slices from z = -25 to z = 55. (B) Map of t-values for coefficients of a mixed-effect model with the dependent variable BOLD beta estimate during feedback (simple contrast against the baseline neural activity) and independent dummy variables representing the 8 conditions (wins and losses against learner and sequencer opponents in social and non-social context) and their interactions. * denotes $p < 0.05$, ** denotes $p < 0.01$. The black frame denotes p(FDR-corrected) < 0.05. (C) Top left panel: Maximum-intensity-projection (MIP) of areas significantly activated for contrast of feedback in learner > sequencer conditions, overlaid on template brain. Top middle panel: MIP for the interaction between reward feedback (win > loss) and social context (social > non-social). Top right panel: MIP for the interaction between reward feedback (win > loss) and algorithm type (sequencer > learner). Bottom left panel: MIP of cluster in the TPJ produced by conjunction of the contrasts in the top left and top middle panels. All MIPs shown with threshold p(FWE) < 0.05, permutation-based FWE-corrected at cluster-level with cluster-forming threshold (CFT) of p = 0.001. Bottom right panel: Overlap between the TPJ conjunction cluster (presented in the bottom left panel) and the mask generated from Neurosynth "theory of mind" meta-analysis (presented in panel A). (D) Cross-validation prediction accuracies for multi-variate pattern decoding analyses, split by condition labels (algorithm type and social context) and ToM regions. See also Tables S2-3.

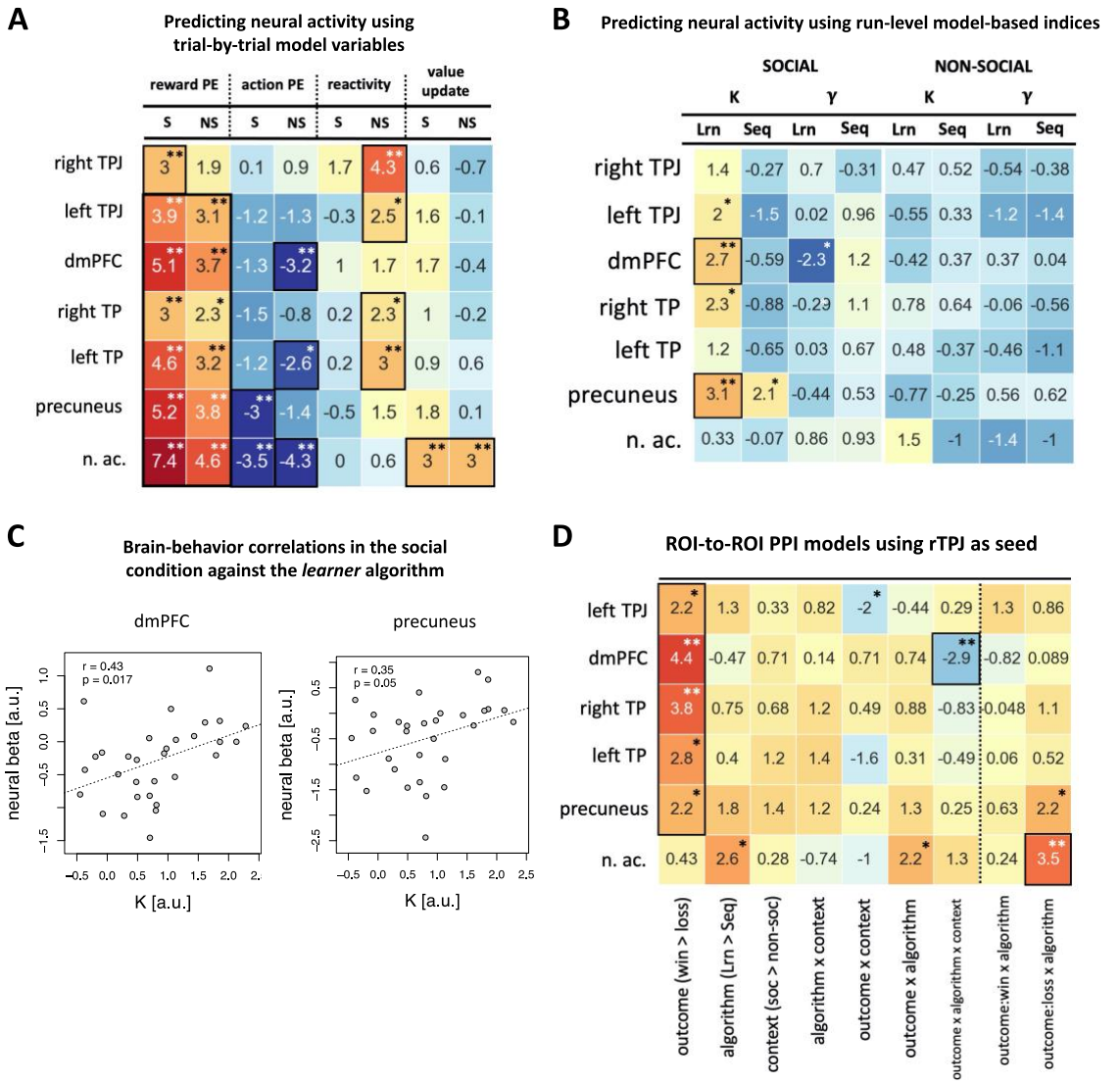

Figure 4: Model-based fMRI results and PPI analysis. (A) Model-based analysis of learning and valuation signals. T-statistics for group-level tests of individual neural betas reflecting coefficients of parametric modulators included in the model-based analysis (see Methods), split by social (S) and non-social (NS) contexts. These modulators include: the signed reward prediction error (reward PE), action prediction error (action PE), reactivity measure (calculated via predicting the algorithm's choice using the subject's last action), and strength of choice value update. * denotes $p < 0.05$, ** denotes $p < 0.01$. The black frame denotes p(FDR-corrected) < 0.05. (B) Strength of strategy use relates to activity in dmPFC and precuneus. T-statistics from mixed-effect models regressing behavioral indices of strategy use (κ and γ) on beta estimates for the outcome stage (using run-level data and treating individual subjects as random effects), split by algorithm type (learner and sequencer) and context (social and non-social). * denotes $p < 0.05$, ** denotes $p < 0.01$. The black frame denotes p(FDR-corrected) < 0.05. (C) Across-subject illustration of the key results in panel C, showing the correlation between the individual neural beta (BOLD response at the feedback stage) and the individual κ coefficient. Each dot is one subject, r is Pearson correlation. (D) The rTPJ connectivity depends flexibly on algorithm and context. Map of t-values for functional connectivity PPI analyses using the rTPJ as the seed, demonstrating main effect of conditions (win vs loss, learner vs sequencer, social vs non-social) as well as their interactions. * denotes $p < 0.05$, ** denotes $p < 0.01$. The black frame denotes p(FDR-corrected) < 0.05. See also Table S3.

STAR Methods

RESOURCE AVAILABILITY

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants

We recruited 20 subjects for a behavioral study and 66 healthy volunteers (28 male, 38 female; age 18-25) for a functional magnetic resonance imaging (fMRI) experiment. All subjects were right-handed, non-smokers, medication-free, and with normal eyesight. They provided written consent and the ethics committee of the Canton of Zurich approved the study.

Before running the analysis, we had to exclude a small number of subjects based on clearly defined criteria. First, we excluded all data from one subject in the non-social condition who ended the experiment with a negative score against both opponent types. Second, we excluded one additional subject in the social condition who was erroneously given the wrong instruction set, which compromised the social context induction as evidenced by post-experimental manipulation checks. In addition, we excluded the data from one subject in the social condition due to a partial loss of fMRI data. Finally, we excluded one subject in the social condition and one participant in the non-social condition whose sudden head movements inside the scanner exceeded 3mm. Thus, we performed the final analyses on 29 participants in the non-social condition and 31 in the social condition (60 subjects total).

METHOD DETAILS

Experimental Task

Social context manipulation. Before the experiment, we randomly assigned participants to the social or non-social condition. Subjects in the *non-social condition* were told that they will play a card-guessing game featuring two distinct decks, a blue deck and a red deck. These decks contained either Rock or Tree cards (**Figure 1**). The goal for the participant was to predict the next card drawn from the deck. Participants were explicitly told that this was not a game of chance and that the two decks may be sorted according to different criteria. They were also instructed that the screen would indicate in each block which deck they were currently drawing from (red or blue).

In contrast, subjects in *the social condition* were told that they will play the game of hide and seek against two human opponents. The goal of the game was to find the coin, which could be hidden (by the opponent who is making the choice simultaneously) behind either a Tree or a Rock card. Participants read that they would be facing two different human opponents and that the screen would indicate which one they are facing at any point in time (red or blue). To reinforce the plausibility of our social context manipulation, the first experimenter was called on the phone while he instructed the participant to open the door of the lab for the simulated human opponent, and the second experimenter interacted with this opponent (played by a confederate sitting in the adjacent room) in a clearly audible fashion.

All other elements of the task (such as the interface, response method, duration, number of blocks and runs, see the task description below) were identical across both conditions.

While our analyses would have higher statistical power if the social/non-social comparison would have been run in a within-subject design, our pilot tests showed that the subjects can become suspicious if they are presented with the same opponents but are told that in one condition they deal with a computer and in the other one they are

facing a human. Our main goal was to convince subjects that they are playing in a specific setting (i.e., social or non-social), so they do not try to guess the true underlying identity of the opponent. To minimize these suspicions and put the subjects in the "social" or "non-social" mindset for the whole duration of the experiment, we thus chose a between-subject design for the social/non-social comparison.

At the end of the experiment, participants filled out a post-experimental questionnaire to measure their beliefs about the social or non-social nature of their opponent. We asked the subjects to which degree they agreed with the following statements: "I felt I was playing against a computer" (q1) and "I felt I was playing against a human opponent" (q2). These statements were rated on an 8-point Likert scale and collected for both the learner and sequencer algorithm. To capture the contrast between these beliefs, a social belief index was constructed as follows: $SBI_j = q2_j - q1_j$, where j denotes the subject. Analyses of this SBI showed that participants in the social setting felt that they were playing against real opponents (mean SBI 5.35, t-test, $t(30) = 4.68$, $p = 0.00002$), and the participants in non-social setting believed that they played against a computer (mean SBI -5.17, $t(28) = -3.95$, $p = 0.0005$), with a strong significant difference between the conditions ($t(56) = 6.05$, $p < 10^{-6}$). At the same time, there was no overall effect of algorithm type on social belief ($t(59) = -0.96$, $p = 0.33$), underlining that participants were mainly affected in their perceptions by the social context manipulation. Moreover, to test whether or not participants in the social condition believed that the *learner* opponent was more human, we added the interaction term algorithm*context to a regression of the social belief on context and algorithm type. This interaction was not significant ($t(116) = 0.16$, $p = 0.87$), suggesting that perception of the social context was not different across both algorithms.

Task. In both conditions, the subjects played the same simple card-prediction game with two choice options (Tree and Rock). Participants won a point if they played the same action as the opponent in the social context or predicted the card that was revealed in the non-social context, and they lost a point if they played the option that was different from the opponent's choice or predicted the card that was different from the one that was revealed (i.e., subjects were required to match the opponent's choice). The task had 228 trials, organized in blocks of 19 trials, separated by explicit transition screens, indicating that the participant was about to face the other opponent (for the social context) or predict draws from the other deck of cards (for the non-social context). We color-coded the edge of the screen with the opponent/deck type (blue and red) to keep this information salient throughout the game.

*Algorithms generating the opponent choice***.** All participants played against two distinct computer algorithms, which we label as the *learner* and *sequencer*. The subjects were unaware of these labels and about the nature of the algorithmic computations.

The first algorithm, *learner*, kept track of the player's play history and used it to make choices. Specifically, the algorithm played as a so-called "0-ToM" learning agent tracking the player's bias and the associated uncertainty in a Bayesian fashion (Devaine et al., 2014a, 2014b). Simply put, this algorithm hid the coin where it predicted the player was least likely to look, based on the history of its opponent's choices (see **Supplementary Materials (Methods S1)** on the details of the computation). Specifically, the algorithm estimated the relative frequency of the participant's choices and played the less frequent option.

The second algorithm, the *sequencer*, did not react to the player's choices but produced a sequence of decisions that was probabilistically determined based on its own

previous choices. Simply put, the algorithm played a sequence that switched every two trials (e.g. "tree-tree-rock-rock-tree-tree…") with some degree of randomness (see **Supplementary Materials (Methods S1)** on the details).

To match the *sequencer* to *learner* in terms of statistical/computational complexity, we started by simulating the *learner* behavior and decomposed the predictability of its choices given the history of the other player's actions across last 8 trials using Volterra Kernels (Daunizeau et al., 2014). We then generated sequences of choices with similar statistical predictability, but now based on Volterra decompositions of the past choices of the sequence itself rather than of the other player's actions. Specifically, we generated sequences of choices for which the autocorrelation structure was such that choices were anti-correlated with choices two trials back (to make the *sequencer* robust to a simple win-stay-lose-shift or Q-learning strategy). Out of these candidate sequences, we then selected those that satisfied the following two criteria: (1) the use of the optimal strategy (switching every two trials) yielded the same performance (winning on 80% of trials) as the optimal strategy against the learner algorithm (switching own action every trial), and (2) the use of a simple win-stay-lose-shift heuristic or Q-learning yielded suboptimal performance around or below chance level (see **Supplementary Materials** and **Figures S6-7** for details).

Participant payoff

Every subject started with an initial endowment of 50 Swiss Francs (around 50 dollars at the time of testing) in addition to a 30 Swiss Francs show-up fee and minimum pay. Every win trial added one Swiss Franc to their endowment, and every loss trial removed one Swiss Franc. Furthermore, the three best-performing participants (in each condition, so

six in total) received an additional 50 Swiss Francs as a competitive incentive to perform well in the task.

fMRI data-acquisition and pre-processing

We optimized the fMRI sequence for measuring BOLD signals in the valuation and social-brain networks given our scanner setup (Philips Achieva 3T whole-body scanner with an 8-channel Philips sensitivity-encoded (SENSE) head coil). For this purpose, we employed a sequence with a slice angle of 45° to maximally reduce dropout in the vmPFC. We set the imaging parameters as follows: 2624 ms repetition time (TR); 40 slices; 2.5 mm^2 voxel size, 2.5 mm slice thickness; 0.65 mm gap; 90° flip angle. To equilibrate the magnetic field before measurements, we administered 5 dummy image excitations prior to the image acquisition stage. In addition, we acquired a T1-weighted whole brain structural image with 1x1x1 mm^3 cubic voxel size for each subject. We performed the preprocessing routine using SPM12 (Wellcome Trust Centre for Neuroimaging) following the standardized procedure of our research group (Hill et al., 2017). Specifically, we performed slice-timing correction on the middle slice, realigned the images accounting for head movement, co-registered the T1 image with the functional image and performed spatial normalization to the T1 MNI template employing the "new-segment" procedure in SPM12. Finally, we smoothed the functional images with an 6mm FWHM Gaussian kernel.

Peripheral measures

To control for the effects of heart rate and breathing on fMRI recordings, we acquired cardiac and respiratory signals using electrocardiogram and breathing belt. We transformed the physiological time series following the RETROICOR procedure (Glover

et al., 2000), which uses Fourier expansions of various orders for the phase and cardiac pulsation (3rd order), respiration (4th order) and cardio-respiratory interaction (1st order) (Harvey et al., 2008). We carried out these transformations using the TAPAS toolbox (Kasper et al., 2009).

QUANTIFICATION AND STATISTICAL ANALYSIS

Choice modeling

We assumed that the subject's choices were determined by the recent history of both players' actions. As the optimal play against both opponent algorithms involved using the history of choices for up to 2 trials in the past, we used a simple choice model that decomposes the influence of different past actions (of the player and the opponent) on choice by computing the weights of past actions of both players on the subject's current decision.

More specifically, the choice of subject s on trial t (a_t^s) was modelled as a softmax function of a weighted mixture of the subjects' own recent actions a_{t-1}^s and a_{t-2}^s, as well as the actions of the opponent o (a_{t-1}^o and a_{t-2}^o), with each past action being assigned its own independent weight:

$$P(a_t^s) = \frac{1}{1 + e^{-(\beta + \kappa a_{t-1}^s + \lambda a_{t-2}^s + \delta a_{t-1}^o + \gamma a_{t-2}^o)}}. \quad (1)$$

We fit this model as a mixed-effect regression model with binomial link as implemented in R's lme4 package. We used each combination of subject, opponent, and run as a grouping variable for random effects. For each group (subject-run-opponent combination) we estimated posterior weights of subject's own choice at $t-1$ (labelled κ, reflecting the tendency to switch on each trial) and opponent's choice at $t-2$ (labelled γ, reflecting the tendency to switch every two trials).

The model included both fixed effects (coefficients showing the effect across subjects) and random effects (posterior estimates of coefficients showing the subject's individual propensity to use own or opponent's previous choices). We focused on these coefficients since by design, they were crucial to the subject's performance against each type of algorithm. We used these estimates for the main analyses. In addition, we used this model to predict the subject's chosen value $P(a_t^s)$ at each trial t for the fMRI analyses, thereby validating with neural data that subjects employed the strategies incorporated in our model (see below).

As a further validation of our model, we formally compared our model to several standard models developed to model either sequence learning or learning during strategic interactions (Konovalov et al., 2018). Since these models are specialized for one specific kind of behavior (either learning the sequence or strategic interactions), it is not surprising that they provided a worse fit to our data than our chosen model (**Figure S1**). We fit all the models to each subject individually using the variational Bayes approach with the VBA toolbox (Daunizeau et al., 2014; https://mbb-team.github.io/VBA-toolbox/). Using this method, the base model we selected outperformed all other standard models. For the main analyses, we decided to fit the same model using a mixed effects regression for the sake of simplicity, easier replicability, and possibility of hierarchical estimation.

fMRI design matrix

We used SPM12 (Wellcome Trust Centre for Neuroimaging) general linear models (GLMs) for all first-level fMRI analysis. All analyses were performed without orthogonalizing regressors (Mumford et al., 2015).

GLM analysis. For the main analyses (**Figures 3B-D, 4A-D**), we set up an fMRI design matrix that included series of delta functions, placed at the temporal onset of the corresponding events, and convolved with the canonical hemodynamic response function in SPM12 (no derivatives were added). We added the following events as separate regressors: Choice with value of the chosen option as a parametric modulator, win outcome against learner, loss outcome against learner, win outcome against sequencer, loss outcome against sequencer. Stick functions were used to model the outcome period, while epochs of the actual reaction-time duration were used to model the decision period (Grinband et al., 2008). In addition to movement parameters, we added heart rate and respiration as nuisance regressors, using the procedure described in the peripheral measure section.

We used this model to extract from a set of predefined ROIs (see below) the regression weights (betas) corresponding to the impact of these events, and their comparison for main effects and interactions, upon the BOLD response.

PPI analysis. We used the CONN toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012) for the functional connectivity analyses, employing the default band-pass filtering of 0.01–0.1 Hz to the preprocessed images recommended by the toolbox. As the seed region, we used the Neurosynth-defined mask for the right temporoparietal junction (rTPJ) and the set of other social-brain regions and nucleus accumbens defined above as the target regions. The period of interest was set to the feedback. Our comprehensive analysis included all regressors also employed in the main GLM used in the ROI activity analyses: outcome (win vs loss), algorithm type (learner vs sequencer), and context (social vs non-social). We corrected the statistical results for multiple comparisons across ROIs by means of the FDR correction implemented in the toolbox.

Neuroimaging analysis strategy and correction for multiple comparisons

In all analyses (whole-brain, ROIs, PPIs), we systematically analyzed all seven independent contrasts provided by our factorial design. These contrasts were: outcome (win > loss), algorithm (learner > sequencer), context (social > non-social), outcome *algorithm, outcome *context, algorithm*context and outcome *algorithm*context.

For whole-brain analyses, we employed a cluster-level correction with threshold of FWE $p < 0.05$ using an initial cluster-forming threshold of $p = 0.001$ uncorrected. We performed all whole-brain analyses using non-parametric tests (5000 permutations, no t-map smoothing) as implemented in the package SnPM (open source code available at http://warwick.ac.uk/snpm). This method is more robust to deviations from test assumption (Nichols and Holmes, 2002) and has been shown to optimally correct for type-1 error rates (Eklund et al., 2016).

All ROI-based analyses employed functional masks defined a priori using the platform Neurosynth (neurosynth.org). We searched for the term "theory of mind" and extracted the corresponding activation masks based on all 181 studies on the 26.10.2018. Then we broke down this mask into sub-masks for each functional region of interest: the right TPJ (voxels = 1519), left TPJ (voxels = 1301), precuneus (voxels = 1358), left temporal pole (voxels = 1741), right temporal pole (voxels = 1665), and dorsomedial prefrontal cortex (voxels = 1961). Additionally, we used a Neurosynth mask of the left and right nucleus accumbens (voxels = 386 and 323).

In all ROI-based analyses we implemented the FDR-based multiple comparison correction of p-values (Benjamini and Hochberg, 1995).

Decoding and univariate testing of ROIs

For all ROI-based analyses, we opted for the conservative approach of performing ROI-level inference. This means we extracted all voxels within the masks to measure differences in BOLD between conditions.

We extracted the average beta over all voxels associated with wins and losses weighted by their meta-analytic intensity value from our first-level model using the standard procedure implemented in the Marsbar package (Brett et al., 2002). Intensity weighting has the advantage of giving more weight to voxels which are more associated with the terms 'Theory of Mind' as defined by the meta-analysis. We repeated the procedure for each run separately to account for unspecific temporal effects.

Model-free univariate analyses were performed using R and the package lme4 (Bates et al., 2018). We regressed beta coefficients extracted from the ROIs on outcome (win = 1), algorithm type (learner = 1), and context (social = 1) to obtain main effects, and added three possible double interactions and one triple interaction of these variables to obtain the interaction effects, using mixed-effects models in which subjects and runs were entered as nested random-effect intercepts.

For model-based analyses, we ran a GLM including 4 parametric modulators of BOLD signal at the time of feedback. These modulators included (1) trial-by-trial standard ("signed") reward prediction error (RPE), calculated as a difference between the obtained outcome (reward or no reward) and the value of the chosen option calculated using the choice model; (2) trial-by-trial action prediction error, calculated as the absolute difference between the action selected by the algorithm (coded as 0 or 1) and the probability of the subject's choice of these actions predicted by the model (as shown in equation (1)); (3) a measure of opponent reactivity, which we calculated from fitting the same model to the opponent's choices treating each block of 19 trials as grouping variable in the mixed effects regression and using δ in our notation (weight on

the subject's action in the previous trial; see Figure S4 for the histogram of this measure across the two algorithm types); (4) the trial-by-trial strength of the choice value update at the time of feedback, calculated as the absolute difference between the pre-choice and the post-choice value of the chosen option. We included all 4 regressors in the GLM to control for shared variance. Most regressors were very weakly correlated ($r < 0.1$), with one exception being reward prediction error and value update ($r = 0.48$ across all pooled data). This allowed us to identify the unique BOLD signal variance impact for each modulator. We then extracted neural betas for each of these variables for each ROI and each subject, and performed t-tests against 0 to obtain t-values displayed in Figure 4A.

For the out-of-sample decoding analysis, we trained a regularized linear support vector machine (SVM) classifier with cross-validation on the run-level ROI data used for the previous analysis (see above) to predict the algorithm type and the context at the subject-level. We repeated this analysis for each ROI in isolation but also trained the classifier on all ROIs of the ToM network simultaneously. All decoding analyses were performed in Python 3.5 using Scikit learn (Pedregosa et al., 2011). As for our decoding model, we used a linear support vector machine (SVM) with elastic-net regularization with stochastic gradient descent optimization. Because SVMs are not scale-insensitive, all classification analysis was performed on standardized data ($\mu=0$, $\sigma=1$). To obtain prediction accuracies and p-values while accounting for the non-independence of cross-validation folds, we used the permutation test score function implemented with 10000 permutations (Ojala and Garriga, 2009).

All decoding models with a single ROIs were comprised of six features – betas for wins and losses against baseline in three separate runs used to predict 60 labels for context, and 120 labels for algorithm type. To prevent subject-level information leakage

into the algorithm type label prediction, we stratified our cross-validation scheme using the subject ID to form 60 different groups.

References

Adolphs, R. (2009). The Social Brain: Neural Basis of Social Knowledge. Annual Review of Psychology.

Atzil, S., Gao, W., Fradkin, I., and Barrett, L.F. (2018). Growing a social brain. Nature Human Behaviour *2*, 624–636.

Baron-Cohen, S. (2000). Theory of Mind and Autism : A Review. International Review of Research in Mental Retardation *23*, 169–184.

Baron-Cohen, S., Leslie, A.M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition *21*, 37–46.

Baron-Cohen, S., Ring, H.A., Wheelwright, S., Bullmore, E.T., Brammer, M.J., Simmons, A., and Williams, S.C. (1999). Social intelligence in the normal and autistic brain: an fMRI study. European Journal of Neuroscience *11*, 1891–1898.

Barraclough, D.J., Conroy, M.L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nature Neuroscience *7*, 404.

Bartra, O., McGuire, J.T., and Kable, J.W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., and et al (2018). Package "lme4." R Foundation for Statistical Computing, Vienna, Austria.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., and Rushworth, M.F.S. (2008). Associative learning of social value. Nature *456*, 245–249.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) *57*, 289–300.

Bhatt, M., and Camerer, C.F. (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. Games and Economic Behavior *52*, 424–459.

Bhatt, M. a, Lohrenz, T., Camerer, C.F., and Montague, P.R. (2010). Neural signatures of strategic types in a two-person bargaining game. Proceedings of the National Academy of Sciences of the United States of America *107*, 19720–19725.

Blakemore, S.-J., Fonlupt, P., Pachot-Clouard, M., Darmon, C., Boyer, P., Meltzoff, A.N., Segebarth, C., and Decety, J. (2001). How the brain perceives causality: an event-related fMRI study. Neuroreport *12*, 3741–3746.

Blakemore, S.-J., Boyer, P., Pachot-Clouard, M., Meltzoff, A., Segebarth, C., and Decety, J. (2003). The Detection of Contingency and Animacy from Simple Animations in the Human Brain. Cereb Cortex *13*, 837–844.

Boorman, E.D., Rushworth, M.F., and Behrens, T.E. (2013). Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. The Journal of Neuroscience *33*, 2242–2253.

Boylan, R.T., and El-Gamal, M.A. (1993). Fictitious play: A statistical study of multiple economic experiments. Games and Economic Behavior *5*, 205–222.

Brett, M., Anton, J.-L., Valabregue, R., and Poline, J.-B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. Neuroimage *16*, S497.

Burke, C.J., Tobler, P.N., Baddeley, M., and Schultz, W. (2010). Neural mechanisms of observational learning. Proceedings of the National Academy of Sciences *107*, 14431–14436.

Buschman, T.J., and Miller, E.K. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. Science *315*, 1860–1862.

Cabeza, R., Ciaramelli, E., and Moscovitch, M. (2012). Cognitive contributions of the ventral parietal cortex: An integrative theoretical account. Trends in Cognitive Sciences.

Calvo-Merino, B., Glaser, D.E., Grèzes, J., Passingham, R.E., and Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral Cortex *15*, 1243–1249.

Camerer, C. (2003). Behavioral game theory: Experiments in strategic interaction (Princeton University Press).

Camerer, C., Ho, T., and Chong, J. (2004). A cognitive hierarchy model of games. The Quarterly Journal of Economics *119*, 861–898.

Carrington, S.J., and Bailey, A.J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. Human Brain Mapping *30*, 2313–2335.

Carter, R.M., and Huettel, S. a. (2013). A nexus model of the temporal-parietal junction. Trends in Cognitive Sciences *17*, 328–336.

Carter, R.M., Bowling, D.L., Reeck, C., and Huettel, S.A. (2012). A Distinct Role of the Temporal-Parietal Junction in Predicting Socially Guided Decisions. Science *337*, 109–111.

Cavanna, A.E., and Trimble, M.R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. Brain *129*, 564–583.

Chaminade, T., Hodgins, J., and Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. Social Cognitive and Affective Neuroscience *2*, 206–216.

Charpentier, C.J., Iigaya, K., and O'Doherty, J.P. (2020). A Neuro-computational Account of Arbitration between Choice Imitation and Goal Emulation during Human Observational Learning. Neuron *106*, 687-699.e7.

Clegg, B.A., DiGirolamo, G.J., and Keele, S.W. (1998). Sequence learning. Trends in Cognitive Sciences *2*, 275–281.

Clithero, J.A., and Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. Social Cognitive and Affective Neuroscience *9*, 1289–1302.

Collette, S., Pauli, W.M., Bossaerts, P., and O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. ELife *6*.

Cooper, J.C., Dunne, S., Furey, T., and O'Doherty, J.P. (2012). Human Dorsal Striatum Encodes Prediction Errors during Observational Learning of Instrumental Actions. Journal of Cognitive Neuroscience *24*, 106–118.

Coricelli, G. (2005). Two-levels of mental states attribution: from automaticity to voluntariness. Neuropsychologia *43*, 294–300.

Coricelli, G., and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America *106*, 9163–9168.

Cross, E.S., Ramsey, R., Liepelt, R., Prinz, W., and Hamilton, A.F. de C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. Philosophical Transactions of the Royal Society B: Biological Sciences *371*, 20150075.

Cullen, H., Kanai, R., Bahrami, B., and Rees, G. (2014). Individual differences in anthropomorphic attributions and human brain structure. Soc Cogn Affect Neurosci *9*, 1276–1280.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. PLoS Comput Biol *10*, e1003441.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Raymond, J. (2012). Prediction Errors. *69*, 1204–1215.

De Martino, B., O'Doherty, J.P., Ray, D., Bossaerts, P., and Camerer, C. (2013). In the mind of the market: Theory of mind biases value computation during financial bubbles. Neuron *79*, 1222–1231.

De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., and Love, B.C. (2017). Social information is integrated into value and confidence judgments according to its reliability. Journal of Neuroscience *37*, 6066–6074.

Decety, J., and Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. Neuroscientist.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. From Monkey Brain to Human Brain 133–157.

Dehaene, S., and Cohen, L. (2007). Cultural recycling of cortical maps. Neuron *56*, 384–398.

Devaine, M., Hollard, G., and Daunizeau, J. (2014a). The Social Bayesian Brain: Does Mentalizing Make a Difference When We Learn? PLoS Computational Biology *10*, e1003992.

Devaine, M., Hollard, G., and Daunizeau, J. (2014b). Theory of mind: did evolution fool us? PloS One *9*, e87619.

Diaconescu, A.O., Mathys, C., Weber, L.A.E., Kasper, L., Mauer, J., and Stephan, K.E. (2017). Hierarchical prediction errors in midbrain and septum during social learning. Social Cognitive and Affective Neuroscience.

Donaldson, P.H., Rinehart, N.J., and Enticott, P.G. (2015). Noninvasive stimulation of the temporoparietal junction: a systematic review. Neuroscience & Biobehavioral Reviews *55*, 547–572.

Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science *293*, 2470–2473.

Eklund, A., Nichols, T.E., and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences 201602413.

Epstein, R.A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. Trends in Cognitive Sciences *12*, 388–396.

FeldmanHall, O., Mobbs, D., and Dalgleish, T. (2013). Deconstructing the brain's moral network: dissociable functionality between the temporoparietal junction and ventro-medial prefrontal cortex. Social Cognitive and Affective Neuroscience *9*, 297–306.

FeldmanHall, O., Dunsmoor, J.E., Kroes, M.C.W., Lackovic, S., and Phelps, E.A. (2017). Associative Learning of Social Value in Dynamic Groups. Psychol Sci *28*, 1160–1170.

Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S., and Frith, C.D. (1995a). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition *57*, 109–128.

Fletcher, P.C., Frith, C.D., Baker, S.C., Shallice, T., Frackowiak, R.S., and Dolan, R.J. (1995b). The mind's eye—precuneus activation in memory-related imagery. Neuroimage *2*, 195–200.

Forgeot d'Arc, B., Devaine, M., and Daunizeau, J. (2020). Social behavioural adaptation in Autism. PLOS Computational Biology *16*, e1007700.

Frith, C.D. (2007). The social brain? Philosophical Transactions of the Royal Society B: Biological Sciences *362*, 671–678.

Frith, C., and Frith, U. (2005). Theory of mind. Current Biology *15*, R644–R645.

Frith, U., and Frith, C. (2001). The biological basis of social interaction. Current Directions in Psychological Science *10*, 151–155.

Frith, U., and Happé, F. (1994). Autism: beyond "theory of mind." Cognition.

Gallagher, H.L., and Frith, C.D. (2003). Functional imaging of 'theory of mind.' Trends in Cognitive Sciences *7*, 77–83.

Garvert, M.M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T.E.J., and Dolan, R.J. (2015). Learning-Induced Plasticity in Medial Prefrontal Cortex Predicts Preference Malleability. Neuron *85*, 418–428.

Geng, J.J., and Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: Contextual updating? Neuroscience and Biobehavioral Reviews *37*, 2608–2620.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., and Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. NeuroImage *43*, 509–520.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., and Blake, R. (2000). Brain areas involved in perception of biological motion. Journal of Cognitive Neuroscience *12*, 711–720.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences of the United States of America *105*, 6741–6746.

Hebscher, M., Levine, B., and Gilboa, A. (2018). The precuneus and hippocampus contribute to individual differences in the unfolding of spatial representations during episodic autobiographical memory. Neuropsychologia *110*, 123–133.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences *6*, 242–247.

Heyes, C. (2014). Submentalizing: I Am Not Really Reading Your Mind. Perspectives on Psychological Science.

Hill, C.A., Suzuki, S., Polania, R., Moisa, M., O'doherty, J.P., and Ruff, C.C. (2017). A causal account of the brain network computations underlying strategic social behavior. Nature Neuroscience.

Hill, M.R., Boorman, E.D., and Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. Nature Communications *7*, 12722.

Hooker, C.I., Verosky, S.C., Germine, L.T., Knight, R.T., and D'Esposito, M. (2010). Neural activity during social signal perception correlates with self-reported empathy. Brain Res. *1308*, 100–113.

Jamali, M., Grannan, B.L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., and Williams, Z.M. (2021). Single-neuronal predictions of others' beliefs in humans. Nature.

Janowski, V., Camerer, C., and Rangel, A. (2013). Empathic choice involves vmPFC value signals that are modulated by social processing implemented in IPL. Social Cognitive and Affective Neuroscience *8*, 201–208.

Kan, I.P., Barsalou, L.W., Olseth Solomon, K., Minor, J.K., and Thompson-Schill, S.L. (2003). Role of mental imagery in a property verification task: fMRI evidence for perceptual representations of conceptual knowledge. Cognitive Neuropsychology *20*, 525–540.

Kana, R.K., Libero, L.E., Hu, C.P., Deshpande, H.D., and Colburn, J.S. (2014). Functional brain networks and white matter underlying theory-of-mind in autism. Social Cognitive and Affective Neuroscience *9*, 98–105.

Kana, R.K., Maximo, J.O., Williams, D.L., Keller, T.A., Schipul, S.E., Cherkassky, V.L., Minshew, N.J., and Just, M.A. (2015). Aberrant functioning of the theory-of-mind network in children and adolescents with autism. Molecular Autism *6*, 1–12.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience *17*, 4302–4311.

Katsuki, F., and Constantinidis, C. (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. The Neuroscientist *20*, 509–521.

Knutson, B., and Gibbs, S.E.B. (2007). Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology.

Knyazev, G., Merkulova, E., Savostyanov, A., Bocharov, A., and Saprigyn, A. (2019). Personality and EEG correlates of reactive social behavior. Neuropsychologia *124*, 98–107.

Kolling, N., Wittmann, M.K., Behrens, T.E.J., Boorman, E.D., Mars, R.B., and Rushworth, M.F.S. (2016). Value, search, persistence and model updating in anterior cingulate cortex. Nature Neuroscience *19*, 1280–1285.

Kolodny, T., Mevorach, C., and Shalev, L. (2017). Isolating response inhibition in the brain: parietal versus frontal contribution. Cortex *88*, 173–185.

Konovalov, A., and Krajbich, I. (2018). Neurocomputational Dynamics of Sequence Learning. Neuron.

Konovalov, A., Hu, J., and Ruff, C.C. (2018). Neurocomputational approaches to social behavior. Current Opinion in Psychology *24*, 41–47.

Koster-Hale, J., and Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. Neuron.

Kovach, C.K., Daw, N.D., Rudrauf, D., Tranel, D., O'Doherty, J.P., and Adolphs, R. (2012). Anterior Prefrontal Cortex Contributes to Action Selection through Tracking of Recent Reward Trends. Journal of Neuroscience *32*, 8434–8442.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. PLOS ONE *3*, e2597.

Krall, S.C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P.T., Eickhoff, S.B., Fink, G.R., and Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. Brain Structure and Function *220*, 587–604.

Lee, S.M., and McCarthy, G. (2016). Functional heterogeneity and convergence in the right temporoparietal junction. Cerebral Cortex *26*, 1108–1116.

Lee, D., Conroy, M.L., McGreevy, B.P., and Barraclough, D.J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. Cognitive Brain Research *22*, 45–58.

Lockwood, P.L., Apps, M.A.J., Valton, V., Viding, E., and Roiser, J.P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. PNAS *113*, 9763–9768.

Lockwood, P.L., Apps, M.A.J., and Chang, S.W.C. (2020). Is There a 'Social' Brain? Implementations and Algorithms. Trends in Cognitive Sciences.

Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., Birbaumer, N., and Grodd, W. (1999). Activation of cortical and cerebellar motor areas during executed and imagined hand movements: an fMRI study. Journal of Cognitive Neuroscience *11*, 491–501.

Mar, R.A., Kelley, W.M., Heatherton, T.F., and Macrae, C.N. (2007). Detecting agency from the biological motion of veridical vs animated agents. Social Cognitive and Affective Neuroscience *2*, 199–205.

Marsh, L.E., and Hamilton, A.F. de C. (2011). Dissociation of mirroring and mentalising systems in autism. NeuroImage *56*, 1511–1519.

Matsuzaka, Y., Akiyama, T., Tanji, J., and Mushiake, H. (2012). Neuronal activity in the primate dorsomedial prefrontal cortex contributes to strategic selection of response tactics. Proceedings of the National Academy of Sciences *109*, 4633–4638.

McClure, S.M., Berns, G.S., and Montague, P.R. (2003). Temporal prediction errors in a passive learning task activate human striatum. Neuron *38*, 339–346.

McDonald, K.R., Pearson, J.M., and Huettel, S.A. (2020). Dorsolateral and dorsomedial prefrontal cortex track distinct properties of dynamic social behavior. Social Cognitive and Affective Neuroscience *15*, 383–393.

Mengotti, P., Dombert, P.L., Fink, G.R., and Vossel, S. (2017). Disruption of the right temporoparietal junction impairs probabilistic belief updating. Journal of Neuroscience *37*, 5419–5428.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. Cerebral Cortex.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C., and Fehr, E. (2012). Linking Brain Structure and Activation in Temporoparietal Junction to Explain the Neurobiology of Human Altruism. Neuron *75*, 73–79.

Mumford, J.A., Poline, J.-B., and Poldrack, R.A. (2015). Orthogonalization of Regressors in fMRI Models. PLOS ONE *10*, e0126255.

Murdaugh, D.L., Nadendla, K.D., and Kana, R.K. (2014). Differential role of temporoparietal junction and medial prefrontal cortex in causal inference in autism: an independent component analysis. Neuroscience Letters *568*, 50–55.

Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping.

Ogawa, A., and Kameda, T. (2019). Dissociable roles of left and right temporoparietal junction in strategic competitive interaction. Soc Cogn Affect Neurosci.

Ojala, M., and Garriga, G.C. (2009). Permutation tests for studying classifier performance. In Proceedings - IEEE International Conference on Data Mining, ICDM, p.

Ong, W.S., Madlon-Kay, S., and Platt, M.L. (2021). Neuronal correlates of strategic cooperation in monkeys. Nat Neurosci *24*, 116–128.

O'Reilly, J.X., Schuffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., and Rushworth, M.F.S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proceedings of the National Academy of Sciences *110*, E3660–E3669.

Pedregosa, F., Weiss, R., and Brucher, M. (2011). Nscep. *12*, 2825–2830.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage *45*, S199–S209.

Ruff, C.C., and Fehr, E. (2014). The neurobiology of rewards and values in social decision making. Nature Reviews Neuroscience *15*, 549–562.

Ruff, C.C., Bestmann, S., Blankenburg, F., Bjoertomt, O., Josephs, O., Weiskopf, N., Deichmann, R., and Driver, J. (2008). Distinct causal influences of parietal versus frontal areas on human visual cortex: evidence from concurrent TMS–fMRI. Cerebral Cortex *18*, 817–827.

Rushworth, M., Walton, M.E., Kennerley, S.W., and Bannerman, D. (2004). Action sets and decisions in the medial frontal cortex. Trends in Cognitive Sciences *8*, 410–417.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal Cortex and Reward-Guided Learning and Decision-Making. Neuron *70*, 1054–1069.

Santiesteban, I., Banissy, M.J., Catmur, C., and Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. Current Biology *22*, 2274–2277.

Santiesteban, I., Banissy, M.J., Catmur, C., and Bird, G. (2015). Functional lateralization of temporoparietal junction–imitation inhibition, visual perspective-taking and theory of mind. European Journal of Neuroscience *42*, 2527–2533.

Saxe, R. (2006a). Uniquely human social cognition. Current Opinion in Neurobiology *16*, 235–239.

Saxe, R. (2006b). Why and how to study Theory of Mind with fMRI. Brain Research *1079*, 57–65.

Saxe, R.R. (2010). The right temporo-parietal junction: a specific brain region for thinking about thoughts. Handbook of Theory of Mind 1–35.

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." Neuroimage *19*, 1835–1842.

Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., and Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. Trends in Cognitive Sciences *19*, 65–72.

Schultz, J., Friston, K.J., Doherty, J.O., Wolpert, D.M., and Frith, C.D. (2005). Activation in Posterior Superior Temporal Sulcus Parallels Parameter Inducing the Percept of Animacy. *45*, 625–635.

Seo, H., and Cai, X. (2014). Neural correlates of strategic reasoning during competitive games. *346*, 340–344.

Smith, D.V., Clithero, J.A., Boltuck, S.E., and Huettel, S.A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. Social Cognitive and Affective Neuroscience *9*, 2017–2025.

Spiliopoulos, L. (2013). Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching. Games and Economic Behavior *81*, 69–85.

Stanley, D.A. (2016). Getting to know you: general and specific neural computations for learning about people. Social Cognitive and Affective Neuroscience *11*, 525–536.

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

Taren, A.A., Venkatraman, V., and Huettel, S.A. (2011). A parallel functional topography between medial and lateral prefrontal cortex: evidence and implications for cognitive control. Journal of Neuroscience *31*, 5026–5031.

Toneatto, T., Blitz-Miller, T., Calderwood, K., Dragonetti, R., and Tsanos, A. (1997). Cognitive Distortions in Heavy Gambling. Journal of Gambling Studies.

Tso, I.F., Rutherford, S., Fang, Y., Angstadt, M., and Taylor, S.F. (2018). The "social brain" is highly sensitive to the mere presence of social information: An automated meta-analysis and an independent study. PloS One *13*, e0196503.

Tusche, A., Böckler, A., Kanske, P., Trautwein, F.-M., and Singer, T. (2016). Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. J. Neurosci. *36*, 4719–4732.

Utevsky, A. V., Smith, D. V., and Huettel, S.A. (2014). Precuneus Is a Functional Core of the Default-Mode Network. The Journal of Neuroscience.

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. Hum Brain Mapp *30*, 829–858.

Vanderwal, T., Hunyadi, E., Grupe, D.W., Connors, C.M., and Schultz, R.T. (2008). Self, mother and abstract other: an fMRI study of reflective social processing. Neuroimage *41*, 1437–1446.

van Veluw, S.J., and Chance, S.A. (2014). Differentiating between self and others: an ALE meta-analysis of fMRI studies of self-recognition and theory of mind. Brain Imaging and Behavior *8*, 24–38.

Venkatraman, V., Rosati, A.G., Taren, A.A., and Huettel, S.A. (2009). Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. Journal of Neuroscience *29*, 13158–13164.

Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connectivity *2*, 125–141.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nature Methods *8*, 665–670.

Young, L., Camprodon, J.A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010a). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. Proceedings of the National Academy of Sciences *107*, 6753–6758.

Young, L., Dodell-Feder, D., and Saxe, R. (2010b). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. Neuropsychologia.

Zhu, L., Mathewson, K.E., and Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proceedings of the National Academy of Sciences *109*, 1419–1424.