Title:

The Application of Computational Models to Social Neuroscience: Promises and Pitfalls

Authors:

Charpentier, Caroline Juliette[1] and O'Doherty, John P.[1]

[1]*Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, USA*

Corresponding author:

Charpentier, Caroline Juliette

California Institute of Technology, MC 228-77
1200 E California blvd
Pasadena, CA 91125
USA

Email: ccharpen@caltech.edu

Telephone: +1-626-395-8457

Summary

Interactions with conspecifics are key to any social species. In order to navigate this social world, it is crucial for individuals to learn from and about others. Whether it is learning a new skill by observing a parent perform it, avoiding negative outcomes, or making complex collective decisions, understanding the mechanisms underlying such social cognitive processes has been of considerable interest to psychologists and neuroscientists, particularly to studies of learning and decision-making. Here, we review studies that have used computational modelling techniques, combined with neuroimaging, to shed light on how people learn and make decisions in social contexts. As opposed to previous methods used in social neuroscience studies, the computational approach allows one to directly examine where in the brain particular computations, as estimated by models of behavior, are implemented. Similar to studies of experiential learning, findings suggest that learning from others can be implemented using several strategies: vicarious reward learning, where one learns from observing the reward outcomes of another agent; action imitation, which relies on encoding a prediction error between the expected and actual actions of the other agent; and social inference, where one learns by inferring the goals and intentions of others. These strategies rely on distinct neural networks, which may be recruited adaptively depending on task demands, the environment and other social factors.

Keywords: computational modeling, fMRI, social learning, social decision-making

Introduction

Over the past decade, many cognitive neuroscience studies, particularly in the field of learning and decision-making, have used a combination of computational modelling of behavior together with neuroimaging. Internal variables, such as reward prediction errors or subjective values, often cannot be directly measured from the task design, but instead have to be extracted from a computational model estimated from participants' behavior. These variables or model parameters can in turn be regressed against a measure of brain activity during task performance, such as the fMRI (functional magnetic resonance imaging) BOLD (blood oxygen level-dependent) signal, giving insights into whether and where in the brain these variables are computed (Cohen et al., 2017; Corrado & Doya, 2007; J. P. O'Doherty, Hampton, & Kim, 2007; John P O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003).

More recently, this type of computational or model-based neuroimaging experiments have been conducted in the social domain, to better understand the signals computed by the brain during social interactions. In this review, we outline several approaches that have been taken to examine social neuroscience from a computational perspective. We focus on two aspects of the social cognition literature: (i) how people learn from observing others as well as the application to strategic interactions, and (ii) how people learn about other people's preferences and make collective decisions.

Learning FROM others

It is crucial for humans and other animals to learn about the world around them in order to make adaptive decisions, obtain rewards and avoid punishments. These 'objective' values of decision variables can be learned experientially, by trial and error. In social species however, there are many situations where one can learn by observing the behavior of another individual. Such observational learning can be clearly advantageous as it allows an individual to assess the

consequences of actions available in the environment without directly experiencing these potentially negative or threatening outcomes. Current theories suggest that three strategies are at play in this process (Dunne & O'Doherty, 2013): vicarious reinforcement-learning, action imitation, and inference about others' beliefs and intentions (**Figure 1**). It is worth noting that this distinction, at least between action imitation and inference over others, has been discussed at length in developmental and comparative psychology - also referred to as 'imitation versus emulation' distinction (Horner & Whiten, 2005; Nielsen, 2006; Thompson & Russell, 2004; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009).

Three observational learning strategies with distinct computational and neural signatures

In vicarious reinforcement-learning, an individual learns from observing someone else experiencing outcomes, rather than experiencing outcomes by themselves. Similar to experiential learning, associations between the actions taken and the outcomes experienced by another agent can be learned to inform the observer of the different action values. These associations can then act as a guide for the observer's actions. Computational mechanisms of such forms of observational learning involve computing a prediction error about the other agent's outcome, i.e. the difference between the other agent's predicted and actual outcome (Burke, Tobler, Baddeley, & Schultz, 2010; Cooper, Dunne, Furey, & O'Doherty, 2012; M. R. Hill, Boorman, & Fried, 2016; Suzuki et al., 2012). These observational reward prediction errors (oRPE) have been found to be encoded in the brain, in particular in the ventromedial prefrontal cortex (vmPFC; Burke, Tobler, Baddeley, & Schultz, 2010; Suzuki et al., 2012) and in the dorsal striatum (Cooper et al., 2012) in fMRI studies. A single-unit recording study in humans recently reported the encoding of observational RPEs at the single neuron level in the rostral anterior cingulate cortex (ACC; Hill, Boorman, & Fried, 2016). These neural signals are partly shared with the encoding of experiential RPEs in dopaminergic regions of the striatum and projections to the vmPFC (Behrens, Hunt, Woolrich, & Rushworth, 2008; Daw & Doya, 2006; Rangel,

Camerer, & Montague, 2008). The value of rewards obtained by others, as well as predictions about these rewards (i.e. the expected value) have also been found to be encoded in the brain, particularly in the ACC (Apps & Ramnani, 2014; Lockwood, Apps, Roiser, & Viding, 2015). Collectively, these findings suggest that vicarious reinforcement-learning is implemented in the brain and that this mechanism depends on neural circuits that at least partially overlap with those involved in experiential learning.

A second observational learning mechanism, action imitation, involves learning from observing another person's actions. In imitation learning, an observer learns to take a particular action based on the extent to which the other agent took that same action in the past and in the same context. This action imitation strategy can also be explained in a reinforcement-learning framework, whereby actions performed by the other agent are reinforced positively, while unchosen actions are reinforced negatively, leading to the computation of action values that can then be used by the observer. Action prediction errors – the difference between the action performed by another agent and the action that was expected of them by the observer – have been reported in the dorsolateral prefrontal cortex (dlPFC), dorsomedial prefrontal cortex (dmPFC), and bilateral inferior parietal lobule (Burke et al., 2010; Suzuki et al., 2012). While not being too computationally demanding, action imitation can be especially advantageous in situations where the other agent's outcomes are not available for the observer to see. At the level of neuronal implementation, it is possible that action imitation is implemented in part through mirror neurons, which have been found to fire when an individual performs an action but also observes another person performing the same action (Catmur, Walsh, & Heyes, 2009; Lametti & Watkins, 2016; Rizzolatti & Craighero, 2004; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). This imitation strategy involves some level of trust that the other person's actions are correct. In case of distrust, the same strategy can result in "reverse" action imitation, whereby the observer chooses the opposite action from that of the other agent. Other variables likely to modulate imitation learning

could then include factors such as how well the observer knows the agent, how reliable the agent's actions are, or whether the agent has a competitive interest.

Finally, a last strategy for observational learning involves a more complex inference process about other agents' intentions and hidden mental states. In such a strategy, an individual updates their beliefs about others' goals and intentions in a Bayesian manner, combining their prior beliefs with evidence they get from observing others' actions and/or outcomes. A mechanism that has been put forward to implement this strategy is inverse reinforcement-learning or inverse RL (Collette, Pauli, Bossaerts, & O'Doherty, 2017; Ng & Russell, 2000). Contrary to classical RL, in which an individual learns the value of an action from observing the rewards, in inverse RL an individual infers the reward distribution from observing the actions of another agent. In another study, a hierarchical Bayesian learning model best explained how people infer the intentions of others. In this model the observer learns about the volatility of the partner's intentions in order to optimize his/her own predictions about the validity of the partner's advice (Diaconescu et al., 2014). Interestingly, brain activity tracking these social inference computations was found in regions that are known to be part of the mentalizing and Theory of Mind network: dmPFC (Boorman, O'Doherty, Adolphs, & Rangel, 2013; Collette et al., 2017; Hampton, Bossaerts, & O'Doherty, 2008), temporoparietal junction (TPJ; Behrens, Hunt, Woolrich, & Rushworth, 2008; Boorman, O'Doherty, Adolphs, & Rangel, 2013) and posterior superior temporal sulcus (pSTS; Hampton, Bossaerts, & O'Doherty, 2008). These regions were originally identified with non-computational approaches in a range of social inference tasks (Fletcher et al., 1995; Frith & Frith, 2006; Koster-Hale & Saxe, 2013; Saxe & Kanwisher, 2003; Van Overwalle & Baetens, 2009). Overall, a neurocomputational approach to social inference during observational learning has helped providing a mechanistic account of Theory of Mind – the ability to attribute mental states and intentions to others. It also provides a global framework in which the observer can also take into account the possibility that the agent they are observing has different preferences, goals and intentions from their own, or a competing agenda.

An empirical question that remains to be examined in more detail is how much of these computations and circuits involved in observational learning strategies overlap with those of experiential reinforcement-learning, when an individual learns by directly experiencing outcomes. Extensive work points towards two major strategies underlying experiential learning: model-free or stimulus-driven learning, as well as model-based or goal-directed learning (Balleine & Dickinson, 1998; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Dickinson, 1985; John P. O'Doherty, Cockburn, & Pauli, 2017). A parallel could be drawn between the experiential and the social domains, whereby vicarious RL and action imitation RL could be considered model-free strategies, while social inference could constitute a model-based strategy, requiring the observer to build a model of world and to learn probability distributions of another agent's goals and transition between states (Dunne, D'Souza, & O'Doherty, 2016). However, neuroimaging results involving areas such as the TPJ, pSTS or dmPFC in social inference learning, which are not typically involved in model-based experiential learning, suggest that this parallel may be too simplistic. Even though the computations underlying social inference learning may fit the description of 'model-based' computations, the neural circuits recruited in the social domain seem to be distinct from those implementing model-based computations during experiential learning. More empirical evidence is needed to investigate how much overlap there is between the circuits for social and experiential learning and whether and how exactly the computations implemented by these circuits differ.

Application to strategic and competitive interactions

Many social interactions involve a strategic or competitive component (e.g. games), such that an individual has an incentive to exploit the knowledge they learn from their opponent, by relying on recursive beliefs about the opponent's intentions (e.g. "I think that he thinks that I think…"). These beliefs are acquired through learning from previous interactions with that opponent (for a detailed review, see Lee & Seo, 2016). People in such interactions can even have an incentive to

lie or purposefully deceive each other, and therefore to detect these deceptive strategies in their opponents. The study of strategy in social neuroscience has tended to utilize concepts and tasks from behavioral game theory (Camerer, 2003). Many studies have now developed computational models combined with neuroimaging to explain these strategic social interactions and how they are implemented in the brain.

For example, in a study using the trust game, in which an individual learns about another person's reputation and trust in a two-person economic exchange, activity in the dorsal striatum was found to predict reciprocity or 'intention to trust' in the game (King-Casas et al., 2005). In a recent study, participants had to learn about the trustworthiness of several partners and then decide to play with strangers who look more or less like the original partners (FeldmanHall et al., 2018). Amygdala tracked resemblance to untrustworthy partners, dmPFC tracked resemblance to trustworthy partners, and dorsal striatum (caudate) activation patterns supported the decision to trust new players. In another study (Hampton et al., 2008), pairs of participants played the inspector game, a variant of the competitive game 'matching pennies'. In this game one participant is an employer who can inspect or not inspect and the other an employee or can work or shirk. Both participants have different interests, such that the employee has an incentive to shirk if not being inspected, or to work if inspected; in contrast the employer's preference is to not inspect while the employee is working. Therefore, both participants have to try and predict what the other player's next action will be in order to choose the best action for themselves. The computational model that best explains participants' behavior consists of an algorithm that iteratively updates the probability of the opponent's action based on their previous actions, combined with a second-order mental state representation (i.e. an effect of the opponent's predictions on the participant's actions). Different components of this computational model were tracked by different neural substrates in the brain. The mPFC was found to incorporate second-order knowledge by tracking an individual's expectations given the degree of model-predicted influence from the opponent. The pSTS tracked a signal used to update the second-order

knowledge representation once the opponent's action is observed. A recent replication and extension of this work used theta-burst transcranial magnetic stimulation (TMS), to provide evidence for a causal influence of the rTPJ on mentalizing and integration of other people's beliefs during strategic social behavior (Hill et al., 2017).

Another strategic learning task which was used in combination with computational modelling and fMRI is a stag-hunt game, in which participants interact with a computerized agent using different levels of recursive inferences (sophistication). In the game, the participant and the computerized agent are hunters who can either individually hunt a rabbit for a small payoff, or collaborate to hunt a stag for a large payoff. A computational model of dynamic belief inference (Yoshida, Dolan, & Friston, 2008) was fit to the behavioral data. At the neural level, computations reflecting the uncertainty of the inference about the other agent's strategy were found in the dmPFC, while the estimated sophistication level (or depth of recursion) of the participant's strategy was encoded in the left dlPFC (Yoshida, Seymour, Friston, & Dolan, 2010). Involvement of the dmPFC in tracking an opponent's or partner's belief during strategic interaction was confirmed by electrophysiological recording of dmPFC neurons in non-human primates, who were found to engage in recursive learning and counter predictable exploitation by their opponent (Seo, Cai, Donahue, & Lee, 2014). Neurons in the dmPFC represented the animal's recent choice and reward history, as well as a switching signals that correlated with the animal's tendency to deviate from simple heuristic learning.

In a multi-strategy competitive learning task called the 'patent race game', a hybrid model integrating both RL and social belief inference best explained behavior (Zhu, Mathewson, & Hsu, 2012). Ventral striatum was found to track both an RL prediction error – the difference between expected and actual payoffs given the chosen strategy – and a belief-based prediction error – the difference between expected and actual payoffs taking into account all possible strategies weighted by the beliefs about future actions of opponents. Interestingly, the rostral

ACC exclusively encoded the belief prediction error, in a way that correlated with individual difference in the engagement of belief learning.

Finally, in a recent study (Hertz et al., 2017), the authors investigated the neural computations associated with the source of social influence during advice giving. The strategic aspect of the task is such that two advisers, one of which is the participant, compete for influence over a 'client'. Theory of mind regions were again associated with different components of behavior. Activity in the rTPJ was found to be involved in tracking whether the client chose them or not, which was argued to play a role in determining strategic influence over the client accordingly; while accuracy relative to the other adviser was found to be encoded in the mPFC.

These studies suggest that the same brain areas involved in learning from another agent by inferring their beliefs and intentions, mainly dmPFC, pSTS and rTPJ, can also perform these computations in strategic and competitive contexts. Several other sophisticated computational models of mentalizing and recursive belief inference have been put forward to explain strategic social interactions between people (Devaine, Hollard, & Daunizeau, 2014; Hula, Montague, & Dayan, 2015; Hula, Vilares, Lohrenz, Dayan, & Montague, 2018; Xiang, Ray, Lohrenz, Dayan, & Montague, 2012) and to predict sequential actions in complex environments (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017); however, the neural bases of these computations have yet to be fully examined using model-based fMRI.

Learning ABOUT others

So far we have described computational strategies that people use to learn **from** other people's actions, outcomes, beliefs and intentions, in order to perform a task correctly by themselves. However, in many situations, we also learn **about** others. This type of learning usually involves learning about subjective values and preferences of another person or a group, in a context where there is no right or wrong decision, but instead a desire to understand others and possibly 'fit in' with the group. Whether people know it or not, what they learn about others can influence their

own preferences and decisions and can help a group reach a consensus. In this part, we present literature that has shed light on the neural computations underlying these processes.

Learning about other people's attitudes and abilities

The neurocomputational mechanisms by which people learn about others have been examined in several recent studies. In Boorman et al. (2013), participants have to evaluate the expertise of other people as compared to that of algorithms in predicting the value of hypothetical assets. Model-based computations characterized subjects' behavior such that individuals credit people who agree with them more than equivalent algorithms when their predictions are correct, and penalize them less when they are incorrect. Beliefs about the expertise of other people and algorithms were represented and updated in the mentalizing network (mPFC, ACC, TPJ, precuneus), while behavioral differences between learning about people relative to algorithms were reflected in the lateral orbitofrontal cortex (OFC) and mPFC. In another study (Wittmann et al., 2016), participants played a reaction-time game in which they had to learn about other people's ability as well as estimate their own ability, both in cooperative and competitive social contexts. Given self and other performance history, a computational RL framework was used to model participants' estimates of self- and other-performance. At the neural level, effects were found in the dmPFC, which tracked two components of the model: others' estimated performance, as well as self-performance in 'compete' relative to 'cooperate' contexts.

In addition to learning about other people's expertise and ability in performing a task, individuals often learn about the preferences or subjective values of their peers. For example, in a social version of a temporal discounting task, participants learned about another person's subjective values and discounting rate (Garvert, Moutoussis, Kurth-Nelson, Behrens, & Dolan, 2015). Using fMRI repetition suppression, the authors showed that learning about another agent's subjective values induce plasticity in the mPFC. This plasticity is in turn explained by a striatal prediction error signal encoding the difference between self and other's values. The mPFC has

also found to be involved in social hierarchy learning (Kumaran, Banino, Blundell, Hassabis, & Dayan, 2016). In this task participants had to learn a hierarchy of nine people within a company, including either themselves or a friend. Learning behavior was better explained by a Bayesian inference scheme than by an RL model. Knowledge about one own's hierarchy, as opposed to that of a friend, was found to be selectively updated in the mPFC. Domain-general learning of other people's relative status within a hierarchy was mediated by learning signals in the amygdala and hippocampus.

A situation in which it is key to be able to learn about other people's preferences is when an individual has to make a decision on behalf of someone else. In Nicolle et al. (2012), the authors tested this using an intertemporal choice task in which participants sometimes choose for themselves and sometimes for someone else. Depending on which choice was relevant for the task, neural signals reflecting self-choice versus other-choice encoding were inter-changeable between the vmPFC and the dmPFC. The choice that needed to be executed was represented in the vmPFC, while the non-executed choice (i.e. the other person's preference when I choose for myself, or my choice when I choose for the other person) was reflected in the dmPFC. In another study, magnetoencephalography (MEG) was used during a learning task to identify how learning signals are attributed to oneself versus another agent. The representation of prediction errors in the brain showed separate signals depending on the identity of the agent being learned about, consistent with a 'neural self-other distinction' (Ereira, Dolan, & Kurth-Nelson, 2018).

Finally, a recent study developed a computational model of how people learn about other people's prudence, impatience, or laziness, in a task that involves observing another agent's cost-benefits decisions between a low-cost/low-reward option and a high-cost/high-reward option before the participant makes their own decision (Devaine & Daunizeau, 2017). There were three cost types associated with the three attitudes mentioned above: delay (impatience), effort (laziness) and risk (prudence). The computational model, based on Bayes-optimal information

processing principles, correctly predicted two biases that arise when individuals learn about others' attitudes. First, people overestimate the degree to which their preferences are similar to others (social projection bias) and second, they align their decisions with those of others (social influence bias). Another recent study provided more evidence and a computational account of the social projection bias, showing that people's own priors influence how they learn about the food preferences of others (Tarantola, Kumaran, Dayan, & De Martino, 2017). The neural mechanisms of such computations still remain to be established.

Social influence on individual preferences and choices

Some of the studies presented above already hint at the tendency that one's own attitudes are influenced by the attitudes of other people. In Garvert et al. (2015), the plasticity observed in the mPFC value representation following learning about another person's values predicted changes in participants' own preferences. In Devaine & Daunizeau (2017), the computational model suggests that the degree to which individual preferences align with the other agent's results from an interaction between the social-projection and the social-influence biases.

Social influence on risk preferences – the extent to which an individual makes a safe versus risky decision after observing the behavior of others – has been investigated in two recent studies (Chung, Christopoulos, King-Casas, Ball, & Chiu, 2015; Suzuki, Jensen, Bossaerts, & O'Doherty, 2016). In the former (Chung et al., 2015), the authors found that observing other people's gambling decisions increased the subjective utility of these gambles for the observer. Such 'other-conferred utility' was encoded in the vmPFC and the strength of this signal predicted the degree of social conformity. In Suzuki et al. (2016), behavioral contagion of risk preferences was better explained by a change in subjects' risk attitudes (curvature of the utility function) than by a change in their subjective evaluation of probabilities (probability-weighting). Neurally, risk was found to be represented in the caudate nucleus, while belief updating about others' risk

preference was encoded in the dlPFC. Across individuals, functional connectivity between these two regions was associated with the size of the contagion effect.

Collective decisions can also have an influence on individual choice (Charpentier, Moutsiana, Garrett, & Sharot, 2014). In this task, groups of five participants make collective decisions between pairs of food items, determined by the majority vote, then get to make decisions for themselves between these items. Activity in the OFC in response to the initial social influence (i.e. the result of the collective decision) was found to be mirrored at a later time when the individual chooses their own action. The strength of this mirroring predicted the extent to which participants altered their decisions to align with the group.

Not only can other people's preferences and decisions affect ours, but how confident other people are about their decisions should also matter in our own judgment. A recent study examined how other people's confidence is integrated in value computations (Campbell-Meiklejohn, Simonsen, Frith, & Daw, 2017). Such integration was found to rely on a posterior-anterior gradient of activity from the subgenual ACC to the vmPFC to the ventromedial Broadmann area 10 (BA 10). More posterior areas (ACC/vmPFC) encoded experiential values as well as values observed from others, while more anterior areas (BA 10) integrated values computed from other people's choices weighted by their confidence. This mechanism suggests that areas that are located in the most anterior part of the prefrontal cortex are able to perform more complex computations underlying social influence.

Finally, there is evidence that social conformity – the tendency of people to align their behavior with the group – may be computationally implemented as a reinforcement learning process. A line of studies suggests that an agent learns about the preferences or opinions of another agent or a group by computing the difference between their own judgment and the judgment of the group (similar to a prediction error) and integrates social information, possibly from several sources, together with individual information (Klucharev et al, 2009; Toelch et al, 2013; Huber et al,

2015). According to a recent meta-analysis of functional brain imaging studies of social conformity (Wu, Luo, & Feng, 2016), dmPFC responses to deviation between individual and group preferences constitute the main signal that predicts subsequent conformity to group opinions. In addition, the meta-analysis also points towards anterior insula activation and ventral striatum deactivation in response to these deviations, although these signals do not seem to be directly linked to preference changes.

Collective decision-making: neural computations involved in reaching a consensus

How people behave in a group, from collaborating or helping each other to reaching a consensus on a subjective question, is a key question of social cognition. Several studies have developed computational accounts of how these collective behaviors may arise.

In a first study (Suzuki, Adachi, Dunne, Bossaerts, & O'Doherty, 2015), groups of 4 or 6 participants repeatedly chose between pairs of items until they reach a consensus. This means that if they all choose the same item they get it as a reward, but if they disagree they have to make a choice again. Therefore, it is crucial that participants in this task incorporate their own preferences with the likely choices of other members of the group. The computational model predicted that the value assigned to one given item by an individual depended on the preference of that individual for the item, the group members' prior choices, as well as on the 'stickiness' of the round (i.e. how aggregated the preferences of other group members for that item are). Those three components had distinct neural representations: personal preferences for items in vmPFC, group members' prior choices in TPJ and pSTS, and stickiness in posterior parietal cortex. Participants' choices were predicted by an integration of these signals in the ACC. Another recent study investigated a similar mechanism, namely how individual and social information are integrated during group decisions, such as jury decisions for criminals (Park, Goïame, O'Connor, & Dreher, 2017). Participants appropriately integrated this information and adapted their judgments to groups of different sizes in a Bayesian manner. The best-fitting

Bayesian inference model also revealed that the strength of integration of social information with individual judgment depended on its credibility. Activity in the dorsal ACC reflected belief updates predicted by the model, while activity in the dlPFC and functional connectivity between the dlPFC and dorsal ACC were associated with the credibility of social information in larger groups.

Several other studies have developed interesting computational models of collective decision-making at the behavioral level. These models provide insights into how confidence escalate during the collective decision-making process (Mahmoodi, Bang, Ahmadabadi, & Bahrami, 2013), how people communicate their confidence to each other in the group (Bang et al., 2017) and how they integrate the opinion of group members who differ in their competence (Mahmoodi et al., 2015). Interestingly, the latter study revealed an equality bias, by which participants assign nearly equal weight to each other's opinion regardless of competence, a result replicated across three cultures. These studies have not used neuroimaging to investigate whether the computations predicted by the behavioral models are implemented in the brain. They could therefore have important implications for future neuroscientific research to help validating the behavioral models and their implementation at the neural level, as well as to improve our mechanistic understanding of these key social processes.

Finally, the study of collective behavior can also provide interesting evolutionary and societal perspectives. Mann & Helbing (2016) recently developed an evolutionary game-theoretic model of collective prediction to examine the role of incentives in maintaining useful diversity. They showed that an incentive scheme that rewards accurate minority predictions results in optimal diversity and collective intelligence, in comparison to market-based incentive systems, which produce herding effects, reduce information available and restrain collective intelligence. Such models could have important societal and policy-related implications.

Discussion

In this review, we explored studies using a combination of computational modelling of behavior with functional neuroimaging to examine learning and decision-making in social contexts. Overall, these studies help illustrate some of the core advantages of the computational approach relative to more traditional social psychology and neuroscience methods. They also point towards some potential pitfalls and issues associated with computational modelling, which we discuss below.

Methodological advantages of the computational approach

Traditionally, most social neuroscience studies have used task designs with multiple conditions (e.g. 2*2 factorial design), allowing to compare behavior and brain activity between two (or more) conditions, and infer underlying processes accordingly. A common example in social neuroscience could be comparing performance on a task where the participant interacts with another human participant versus with a computer. The inference is that brain responses to such a contrast reflects the specific involvement of that network in social processes. However, there are two main issues with such a "categorical" approach. First, if other factors, such as task difficulty, are not perfectly matched between the two conditions being compared, they could be driving differences in brain activity instead of the factor of interest. Second, many cognitive processes cannot be defined as simply as a binary contrast between conditions (e.g. quantifying the expected reward value of a stimulus, or the probability that an observed agent will perform a given action).

The computational approach, in contrast, allows a much more fine-tuned regression of variables of interest against brain activity. By deriving the behavioral computations associated with a specific mechanism and examining the neural correlates of these computations, this approach overall provides a more mechanistic account of brain function, and can offer answers as to how exactly a particular process is implemented in the brain. If two competing hypotheses about a

particular mechanism make different predictions as to what variables should be encoded in the brain, these predictions can be directly tested using neurocomputational methods. Finally, such methods are more flexible than traditional contrast approaches in the sense that multiple parametric variables can be added to the BOLD model at the same time, thus controlling for potential confounds and identifying the unique contribution of a variable to the BOLD signal.

Potential issues and pitfalls of neurocomputational methods

A general issue with any computational approach is overfitting (Vandekerckhove, Matzke, & Wagenmakers, 2015). If the behavioral or the BOLD models are defined with too many parameters or regressors than justified by the data, this can lead to findings that fail to replicate or generalize. To avoid this pitfall, it is important to proceed to a rigorous model comparison using methods that prevent overfitting, such as out of sample cross-validation, penalization of more complex models with Bayesian or Akaike Information Criteria (Akaike, 1974; Schwarz, 1978), and Bayesian Model Selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). Ultimately, it is also key to replicate both behavioral and neuroimaging findings in an independent sample.

Second, another potential issue is correlation between model-based regressors. Similar to the traditional contrasts approach, it is possible that a particular regressor of interest ends up being correlated with another variable, thus leading to misinterpretation of effects of interest. To prevent this, it is crucial to examine these potential correlations ahead of time, by collecting behavioral pilot data and defining the behavioral models and model-based regressors. Just as with more traditional approaches to neuroimaging, it may be necessary to structure the experimental design prospectively in such a way so as to minimize the correlation between the regressor of interest and confounding variables. If some correlations between regressors remain, they have to be controlled for by including both regressors in the BOLD model in order to obtain the unique contribution of the regressor of interest.

Finally, a major concern that seem to emerge from this field of research is that most studies, as illustrated by those described in this review, examine very specific questions with specific task designs and computational models. They report ad-hoc models that are applied uniquely to one particular situation or task, thus making generalization very difficult. Moving forward, we need a 'unified' theory that can be extended and generalized to all sorts of tasks and computational problems, at least within the realm of social inference.

Conclusions and benefits for behavioral and social sciences

The studies described in this review provide key insights into how a computational approach can inform the behavioral and neural mechanisms by which people learn from and about others. We suggest that Bayesian inference models, and their associated neural correlates in the mentalizing network, best explain people's social learning behavior, especially in complex tasks involving strategic or competitive interactions. Simpler computations derived from an RL framework may however perform very well in some contexts. **Figure 1** provides a summary of these strategies and associated neural computations, possibly paving the way for a more 'unified' theory of the classes of computational strategies involved in social learning.

Overall, we suggest that the benefits of a computational approach to social neuroscience outweighs its potential pitfalls, not only because of the more refined mechanistic accounts it can provide, but also given its ability to inform behavior. Indeed, a particular behavior can at times be equally well explained by the computations of two different variables, and model-based analysis of neuroimaging data can answer the question of which of these variables is preferably encoded in the brain, shedding light on the mechanism at play. Focusing on the example of observational learning, this neurocomputational approach has allowed disentangling specific computations associated with different learning strategies (e.g. vicarious reward learning versus action imitation), which could in turn have implications for situations or psychiatric conditions where social learning is impaired.

Finally, this review focused on social learning and decision-making, but this computational neuroimaging approach has the potential of being applied to other subfields of social neuroscience. Some already promising examples include studies of social feedback processing (Jones et al., 2011), altruism (Hutcherson, Bushong, & Rangel, 2015), moral behavior (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017) or social norm enforcement (Zhong, Chark, Hsu, & Chew, 2016).

Acknowledgments

This work was funded by the NIMH Caltech Conte Center for Social Decision Making.

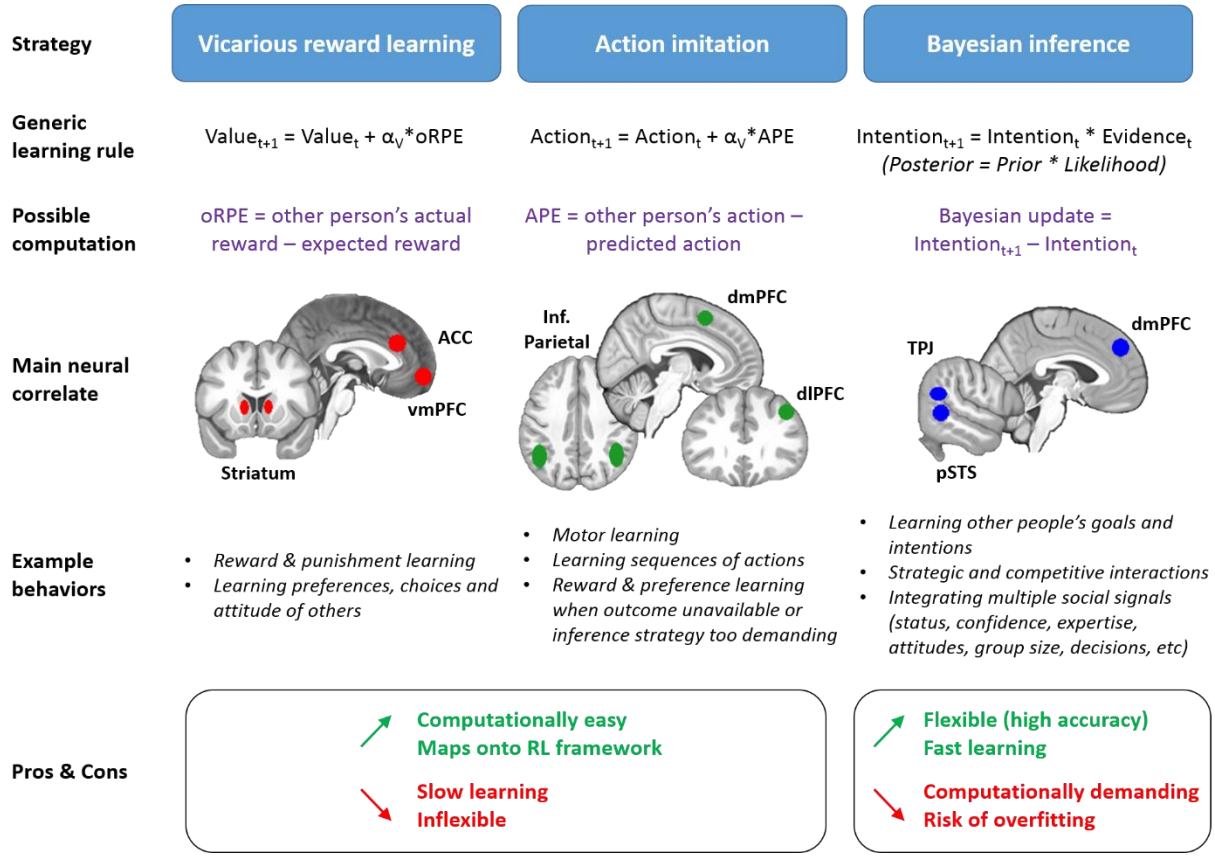

Figure 1. Summary of computational strategies underlying social learning. oRPE: observational reward prediction error; APE: action prediction error.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Apps, M. A. J., & Ramnani, N. (2014). The anterior cingulate gyrus signals the net value of others' rewards. *The Journal of Neuroscience*, *34*(18), 6190–6200. https://doi.org/10.1523/JNEUROSCI.2701-13.2014

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 64. https://doi.org/10.1038/s41562-017-0064

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419. https://doi.org/10.1016/S0028-3908(98)00033-1

Bang, D., Aitchison, L., Moran, R., Herce Castanon, S., Rafiee, B., Mahmoodi, A., … Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*(6), 117. https://doi.org/10.1038/s41562-017-0117

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, *456*(7219), 24524–9. https://doi.org/10.1038/nature07538

Boorman, E. D., O'Doherty, J. P., Adolphs, R., & Rangel, A. (2013). The behavioral and neural mechanisms underlying the tracking of expertise. *Neuron*, *80*(6), 1558–1571. https://doi.org/10.1016/j.neuron.2013.10.024

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, *107*(32), 14431–14436. https://doi.org/10.1073/pnas.1003111107

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

Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D., & Daw, N. D. (2017). Independent neural computation of value from other people's confidence. *The Journal of Neuroscience*, *37*(3), 673–684. https://doi.org/10.1523/JNEUROSCI.4490-15.2017

Catmur, C., Walsh, V., & Heyes, C. (2009). Associative sequence learning: The role of experience in the development of imitation and the mirror system. *Philosophical Transactions of the Royal Society: Biological Sciences*, *364*(1528), 2369–2380. https://doi.org/10.1098/rstb.2009.0048

Charpentier, C. J., Moutsiana, C., Garrett, N., & Sharot, T. (2014). The brain's temporal dynamics from a collective decision to individual action. *J. Neurosci.*, *34*(17), 5816–5823. https://doi.org/10.1523/JNEUROSCI.4107-13.2014

Chung, D., Christopoulos, G. I., King-Casas, B., Ball, S. B., & Chiu, P. H. (2015). Social signals of safety and risk confer utility and have asymmetric effects on observers' choices. *Nature Neuroscience*, *18*(6), 912–916. https://doi.org/10.1038/nn.4022

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., … Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, *20*(3), 304–313. https://doi.org/10.1038/nn.4499

Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. P. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife*, *6*, e29718. https://doi.org/10.7554/eLife.29718

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience*, *24*(1), 106–118. https://doi.org/10.1162/jocn_a_00114

Corrado, G., & Doya, K. (2007). Understanding Neural Coding through the Model-Based Analysis of Decision Making. *The Journal of Neuroscience*, *27*(31), 8178–8180. https://doi.org/10.1523/JNEUROSCI.1590-07.2007

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325. https://doi.org/10.1073/pnas.1408988111

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. https://doi.org/10.1038/nn.4557

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*, 199–204. https://doi.org/10.1016/j.conb.2006.03.006

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Devaine, M., & Daunizeau, J. (2017). Learning about and from others' prudence, impatience or laziness: The computational bases of attitude alignment. *PLoS Computational Biology*, *13*(3), e1005422. https://doi.org/10.1371/journal.pcbi.1005422

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, *10*(12), e1003992. https://doi.org/10.1371/journal.pcbi.1003992

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian learning. *PLoS Computational Biology*, *10*(9), e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society: Biological Sciences*, *308*(1135), 67–78.

Dunne, S., D'Souza, A., & O'Doherty, J. P. (2016). The involvement of model-based but not model-free learning signals during observational reward learning in the absence of choice. *Journal of Neurophysiology*, *115*(6), 3195–3203. https://doi.org/10.1152/jn.00046.2016

Dunne, S., & O'Doherty, J. P. (2013). Insights from the application of computational neuroimaging to social neuroscience. *Current Opinion in Neurobiology*, *23*(3), 387–392. https://doi.org/10.1016/j.conb.2013.02.007

Ereira, S., Dolan, R. J., & Kurth-Nelson, Z. (2018). Agent-specific learning signals for self-other distinction during mentalising. *PLoS Biology*, *16*(4), e2004752. https://doi.org/10.1371/journal.pbio.2004752

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A.

(2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128. https://doi.org/10.1016/0010-0277(95)00692-R

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/j.neuron.2006.05.001

Garvert, M. M., Moutoussis, M., Kurth-Nelson, Z., Behrens, T. E. J., & Dolan, R. J. (2015). Learning-induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron*, *85*(2), 418–428. https://doi.org/10.1016/j.neuron.2014.12.033

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, *105*(18), 6741–6746. https://doi.org/10.1073/pnas.0711099105

Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *bioRxiv*, 121947. https://doi.org/10.1101/121947

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience*, *20*, 1142–1149. https://doi.org/10.1038/nn.4602

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications*, *7*, 12722. https://doi.org/10.1038/ncomms12722

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition*, *8*(3), 164–181. https://doi.org/10.1007/s10071-004-0239-6

Hula, A., Montague, P. R., & Dayan, P. (2015). Monte Carlo planning method estimates planning horizons during interactive social exchange. *PLoS Computational Biology*, *11*(6), e1004254. https://doi.org/10.1371/journal.pcbi.1004254

Hula, A., Vilares, I., Lohrenz, T., Dayan, P., & Montague, P. R. (2018). A model of risk and mental state shifts during social interaction. *PLoS Computational Biology*, *14*(2), e1005935. https://doi.org/10.1371/journal.pcbi.1005935

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451–463. https://doi.org/10.1016/j.neuron.2015.06.031

Jones, R. M., Somerville, L. H., Li, J., Ruberry, E. J., Libby, V., Glover, G., … Casey, B. J. (2011). Behavioral and neural properties of social reinforcement learning. *The Journal of Neuroscience*, *31*(37), 13039–13045. https://doi.org/10.1523/JNEUROSCI.2972-11.2011

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: Reputation and trust in a two-person economic exchange. *Science*, *308*(5718), 78–83. https://doi.org/10.1126/science.1108062

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A neural prediction problem. *Neuron*,

79(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations underlying social hierarchy learning: Distinct neural mechanisms for updating and representing self-relevant information. *Neuron*, *92*(5), 1135–1147. https://doi.org/10.1016/j.neuron.2016.10.052

Lametti, D. R., & Watkins, K. E. (2016). Cognitive neuroscience: The neural basis of motor learning by observing. *Current Biology*, *26*(7), R288–R290. https://doi.org/10.1016/j.cub.2016.02.045

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences*, *39*(1), 40–48. https://doi.org/10.1016/j.tins.2015.11.002

Lockwood, P. L., Apps, M. A. J., Roiser, J. P., & Viding, E. (2015). Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *The Journal of Neuroscience*, *35*(40), 13720–13727. https://doi.org/10.1523/JNEUROSCI.1703-15.2015

Mahmoodi, A., Bang, D., Ahmadabadi, M. N., & Bahrami, B. (2013). Learning to make collective decisions: The impact of confidence escalation. *PLoS ONE*, *8*(12), e81195. https://doi.org/10.1371/journal.pone.0081195

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., … Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112*(12), 3835–3840. https://doi.org/10.1073/pnas.1421692112

Mann, R. P., & Helbing, D. (2016). Minorities report: Optimal incentives for collective intelligence. *Proceedings of the National Academy of Sciences*, *114*(20), 5077–5082. https://doi.org/10.1073/pnas.1618722114

Ng, A., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In J. P. De Sousa (Ed.), *Proceedings of the Seventeenth International Conference on Machine Learning, Vol. 67* (pp. 663–670). Morgan Kaufmann Publishers Inc.

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. E. J. (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron*, *75*(6), 1114–1121. https://doi.org/10.1016/j.neuron.2012.07.023

Nielsen, M. (2006). Copying actions and copying outcomes: Social learning through the second year. *Developmental Psychology*, *42*(3), 555–565. https://doi.org/10.1037/0012-1649.42.3.555

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

O'Doherty, J. P., Hampton, A. N., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53. https://doi.org/10.1196/annals.1390.022

Park, S. A., Goïame, S., O'Connor, D. A., & Dreher, J.-C. (2017). Integration of individual and

social information for decision-making in groups of different sizes. *PLoS Biology*, *15*(6), e2001958. https://doi.org/10.1371/journal.pbio.2001958

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. https://doi.org/10.1038/nrn2357

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*(1), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131–141. https://doi.org/10.1016/0926-6410(95)00038-0

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science*, *346*(6207), 340–343. https://doi.org/10.1126/science.1256254

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017.

Suzuki, S., Adachi, R., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2015). Neural mechanisms underlying human consensus decision-making. *Neuron*, *86*(2), 591–602. https://doi.org/10.1016/j.neuron.2015.03.019

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., … Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, *74*(6), 1125–1137. https://doi.org/10.1016/j.neuron.2012.04.030

Suzuki, S., Jensen, E. L. S., Bossaerts, P., & O'Doherty, J. P. (2016). Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, *113*(14), 3755–3760. https://doi.org/10.1073/pnas.1600092113

Tarantola, T., Kumaran, D., Dayan, P., & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature Communications*, *8*(1), 1–14. https://doi.org/10.1038/s41467-017-00826-8

Thompson, D. E., & Russell, J. (2004). The ghost condition: Imitation versus emulation in young children's observational learning. *Developmental Psychology*, *40*(5), 882–889. https://doi.org/10.1037/0012-1649.40.5.882

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*(3), 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model Comparison and the Principle of Parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology*. New York, NY: Oxford University Press.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society: Biological Sciences*, *364*(1528), 2417–2428. https://doi.org/10.1098/rstb.2009.0069

Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. F. S. (2016). Self-other mergence in the frontal cortex during cooperation and competition. *Neuron*, *91*(2), 482–493. https://doi.org/10.1016/j.neuron.2016.06.022

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *71*, 101–111. https://doi.org/10.1016/j.neubiorev.2016.08.038

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS Computational Biology*, *8*(12), e1002841. https://doi.org/10.1371/journal.pcbi.1002841

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12), e1000254. https://doi.org/10.1371/journal.pcbi.1000254

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *The Journal of Neuroscience*, *30*(32), 10744–10751. https://doi.org/10.1523/JNEUROSCI.5895-09.2010

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, *129*, 95–104. https://doi.org/10.1016/j.neuroimage.2016.01.040

Zhu, L., Mathewson, K. E., & Hsu, M. (2012). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proceedings of the National Academy of Sciences*, *109*(5), 1419–1424. https://doi.org/10.1073/pnas.1116783109