17

Computational Models in Social Neuroscience

Jin Hyun Cheong[1]*, Eshin Jolly*[1]*, Sunhae Sul*[2]*, and Luke J. Chang*[1]

[1] *Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA*
[2] *Department of Psychology, Pusan National University, Busan, Republic of Korea*

Introduction

The burgeoning field of social neuroscience aims to characterize the neural and psychological processes that are involved in successfully navigating the social world. Building on a rich history of integrating the disciplines of social and cognitive psychology, social neuroscience attempts to use methods from cognitive neuroscience to better understand how social cognitions are instantiated in the brain (Ochsner & Lieberman, 2001; Sarter, Berntson, & Cacioppo, 1996). Early pioneering efforts attempted to map brain regions to social processes by using social psychological experimental paradigms while participants underwent functional neuroimaging. However, this work has had limited success integrating social processes into a broader processing stream that includes cognitive, perceptual, and motoric processes. Making these types of inferences requires formulating models that include multiple types of computations.

The application of computational models has the potential to dramatically improve inferences in social neuroscience. A survey of members from two social neuroscience scientific societies provided strong consensus that computational modeling is among the most promising approaches for progressing the field (Stanley & Adolphs, 2013). First, computational models provide a way to parsimoniously represent how different social, affective, and cognitive processes might interact to produce behavior, simultaneously bridging different levels of analysis from molecular and cellular processes to systems and psychological level processes (Fox, Chang, Gorgolewski, & Yarkoni, 2014). Second, computational models force explicit formalization of a scientific hypothesis to make quantifiable predictions that can be tested and falsified. This is critical to evaluating how well a model can account for a phenomenon while also comparing the predictions of competing models. Importantly, models can be compared not only by their ability to better account for behavior, but also by their capacity to explain neural processes, which can substantially improve the model development and testing cycle. Finally, models can be combined, providing a principled approach to developing a cumulative scientific understanding of how the brain operates.

In this chapter, we review an emerging literature that uses computational models to study the psychological and brain processes involved in social learning and inferring others' mental states. We focus specifically on reviewing literature that uses a reinforcement learning framework to understand how people learn from interacting with others.

Computational Models of Brain and Behavior, First Edition. Edited by Dr Ahmed A. Moustafa.

Reinforcement Learning Models

Reinforcement learning is a class of models that are commonly used to formalize trial-to-trial learning in response to feedback. They formalize how an agent interacts with the environment by learning the mapping between actions in situations (states) and outcomes (reward) with the goal of maximizing reward over time. The idea stems from work by behavioral psychologist Edward Thorndike (1911), who posited that actions followed by positive outcomes are likely to be repeated and those followed by negative outcomes are likely to be avoided.

One of the most simple and popular reinforcement learning models is the Rescorla–Wagner model (1972), which updates an agent's estimated value of states or actions based on the discrepancy between predicted and observed outcomes. For example, the value of a certain state at time t can be written as $V(t)$ with the subsequent value at $V(t+1)$ updated by:

$$V(t+1)=V(t)+\alpha\cdot\left[r-V(t)\right] \quad \text{(Eq. 17.1)}$$

where $\alpha=[0,1]$ and refers to the learning rate or the rate at which an agent adapts to new information. The discrepancy between the observed reward r and the expected reward $V(t)$ is described as the prediction error. When the learning rate is 0, the agent does not update newly observed values and when the learning rate is 1, the agent completely discards previous experiences and replaces them with the new information.

A classic finding in computational neuroscience is that dopamine neurons located in the ventral tegmental area (VTA) of the macaque brain appear to fire in response to reward prediction errors (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). The frequency of firing increases when a larger reward is received than expected, and decreases when a smaller reward is received than expected. Importantly, as the animal learns how much reward to expect, these same neurons start to predict the amount of expected reward by firing in response to a reward cue that precedes the actual reward outcome. Outcomes that are correctly predicted result in no firing of dopamine neurons upon receipt of the reward. In humans, functional magnetic resonance imaging (fMRI) studies have found that prediction error signals correlate with activity in the ventral striatum (VS; McClure, Berns, & Montague, 2003; O'Doherty et al., 2004; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003) consistent with the notion that the VS receives synaptic signaling from midbrain dopamine neurons (Ferenczi et al., 2016). In the context of social cognition, the reinforcement learning framework has been successfully used by many researchers to study various types of social learning processes.

Social Learning

Computational models of social learning describe (1) how we learn from the social world, and (2) how we learn about the social world. A growing body of evidence suggests substantial overlap between nonsocial (individual) and social learning. For instance, areas in the prefrontal cortex (PFC) and basal ganglia centering around the VS are involved in comparing and updating both nonsocial and social values (Behrens, Hunt, & Rushworth, 2009; Hampton et al., 2008; Sul et al., 2015; Vickery, Chun, & Lee, 2011). Studies using economic utility models have found similar overlap between individual and social decision-making (Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Ruff & Fehr, 2014). During the social learning process, these overlapping regions interact with other brain regions such as the anterior cingulate cortex (ACC), dorsomedial prefrontal cortex (DMPFC), and temporoparietal junction (TPJ) typically associated with social processing (Behrens et al., 2008; Corbetta, Maurizio, Gaurav, & Shulman 2008; Lee & Seo, 2016; Saxe & Wexler, 2005). In this section, we review two lines of computational neuroscience research on social learning: learning from others and learning about others.

Observational/Vicarious Learning

In psychology, social learning is defined as learning that occurs without direct reinforcement, such as learning through observation, instruction, or vicarious rewards and punishments (Bandura, 1977). This type of learning by observing others' reactions is adaptive in learning to avoid harmful or fearful stimuli (Olsson & Phelps, 2007) as well as fundamental for human society for its critical role in socialization and cultural transmission. Neural mechanisms of learning from observation and social feedback have been extensively studied, mainly in the framework of reinforcement learning. For example, Behrens and colleagues (2008) designed a reward-based associative learning task in which participants learned not only from their own experience (nonsocial value) but also from a confederate's advice (social value) and used a Bayesian reinforcement learning model to estimate learning parameters such as trial-by-trial reward prediction error, volatility, and outcome probability for nonsocial and social values. They found that neural activity in the DMPFC, middle temporal gyrus, and TPJ/superior temporal sulcus (STS) correlated with reward prediction error for the confederate's advice, while the VS, ventromedial prefrontal cortex (VMPFC), and ACC were associated with the prediction errors for participants' own experience of reward. VMPFC reflected reward probabilities for both participants' experience and confederate's advice, suggesting that this region integrates nonsocial and social information into a common currency when making decisions.

In another study on observational learning, Burke and colleagues (2010) estimated vicarious (observed) prediction errors extending a standard reinforcement learning model. They found that the activity in the dorsolateral prefrontal cortex (DLPFC) reflected the difference between the actual and predicted choice of others (action prediction error), and VMPFC was related to the difference between the actual and predicted outcome earned by others (outcome prediction error). Similar to this finding, Apps, Lesage, and Ramnani (2015) investigated the relationship between instructors' expectations and students' actions, finding that instructors' ACC, insula, and VMPFC tracked prediction errors and the expected values resulting from the students' actions.

Social Norm Learning and Conformity

Another important source of social learning is normative information from group behavior. People are highly motivated to conform to social norms and appear to value fitting in over being correct (Asch, 1951). Social norms are created and adhered to because people want to act effectively in a social environment and/or preserve social relationships (Cialdini & Trost, 1998; Deutsch & Gerard, 1955). Descriptive norms refer to shared expectations of context appropriate behavior held by most people and can be learned by observing how people generally act in certain situations (Cialdini, Kallgren, & Reno, 1991; Sherif, 1936). Conformity can arise from the rewarding value of sharing approval and affiliation with others (Campbell-Meiklejohn, Bach, Roepstorff, Dolan, & Frith, 2010) or from the fear of punishment for inappropriate behavior (Fehr & Fischbacher, 2004).

Campbell-Meiklejohnandcolleagues(2010) investigated how social agreement or disagreement from others influences one's valuation of music. They did this by asking participants to rate their enjoyment of different songs before and after receiving feedback of agreement or disagreement from other "music experts." This allowed the researchers to not only track how much participants valued each song, but also to estimate a social influenceability parameter B_{inf} for each participant using a linear regression that estimated how song ratings change after learning about expert opinions. Participants more sensitive to social influence (high B_{inf}) reacted to disagreements from experts with greater activity in the insula, dorsal ACC, lateral PFC, and right TPJ. Moreover, the VS for participants with high social influenceability was maximally activated when their opinions conformed with others. In contrast, the VS activity decreased in the same situation for participants with antisocial influenceability (negative B_{inf}), who decreased their liking of songs when experts agreed with their opinion. Overall, the authors were able to disentangle how participants' own value representations changed as a function of their agreement with others' opinions.

In addition to modifications of value, violations of social norms may lead to error signals as defined in reinforcement learning models, that motivate adjustment of actions (Chang & Koban, 2013; Montague & Lohrenz, 2007; Sanfey, Stallen, & Chang, 2014). Studies in cognitive control have implicated the ACC in conflict monitoring (Botvinick, Cohen, & Carter, 2004) and that cognitive and affective conflicts share similar systems in the dorsal ACC (Ochsner, Hughes, Robertson, Cooper, & Gabrieli, 2009).

To test how reinforcement learning signals support social conformity, Klucharev and colleagues (2009) had participants rate faces on attractiveness and subsequently view the social norm framed as the "average European rating." Discrepancies between individual ratings and the normative ratings were associated with activations in the rostral ACC and deactivations in the nucleus accumbens (NACC), and the magnitude of activity in these regions predicted whether participants changed their behavior to conform to the norm on a subsequent rating of the face.

How violations of social norms impact social interactions were also investigated by Chang and Sanfey (2013). This study examined bargaining behavior using the Ultimatum Game (UG; Fig. 17.1A), in which Player *A* proposes a split of an endowment to Player *B*, who then decides whether to accept the proposed split or reject the offer, in which case both players receive nothing (Güth, Schmittberger, & Schwarze, 1982). Receiving an offer that violated the players' expectations about the descriptive norm was associated with increased activity in the left anterior insula, ACC, and pre-supplementary motor area regions, consistent with an error-monitoring process (Fig. 17.1B).

In a related experiment, Xiang et al. (2013) provide even stronger evidence for how the brain tracks norm violations in the UG. In this study, the experimenters manipulated participants' expectations by exposing different groups of participants to three different distributions of offers (high, medium, low). In the test phase, all groups of participants were given offers from the medium distribution. Participants decided whether to accept or reject the offers and reported their affective responses by selecting from a set of emoticons. The authors combined an Ideal Bayesian Observer model to track how beliefs about the social norm are updated after each offer with a social preference utility function (Chang & Sanfey, 2013; Fehr & Schmidt, 1999) to provide trial-to-trial estimates of the prediction error and variance prediction error for a given offer conditional on prior beliefs.

Behaviorally, identical offers were rejected more frequently when participants expected offers from a high distribution compared to a low distribution, indicating that the social norm manipulation successfully changed attitudes towards the offers. Results from the fMRI analyses found norm prediction errors positively correlated with activity in the VS and medial orbitofrontal cortex (mOFC), while variance prediction errors were tracked by activities in the anterior insula and ACC (Fig. 17.1C). These findings suggest that both norm prediction errors and affective prediction errors comprise a form of reinforcement learning and share similar neural circuitry comprising the VS, mOFC, anterior insula, and ACC (O'Doherty et al., 2003; Preuschoff,

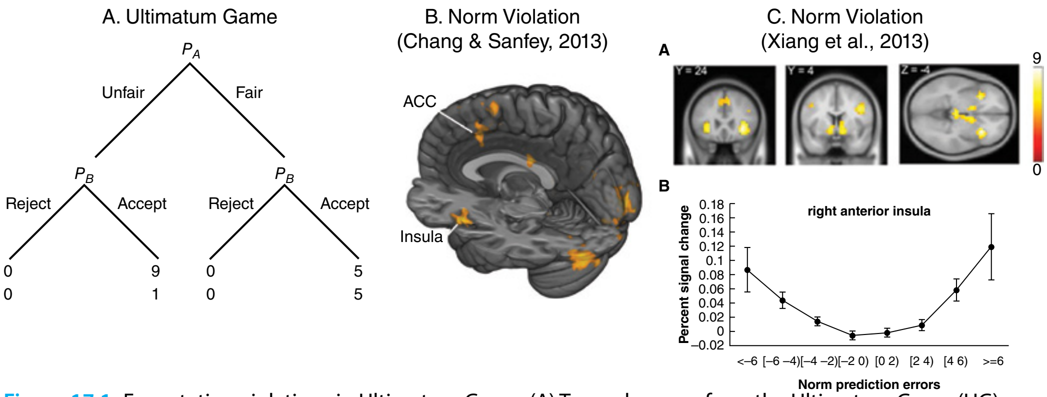

Figure 17.1 Expectation violations in Ultimatum Game. (A) Two subgames from the Ultimatum Game (UG). (B) ACC and left anterior insula track norm violations based on anger model in UG reported in Chang and Sanfey (2013). From Chang and Smith (2015). (C_A) Activity in bilateral anterior insula and ventral striatum correlates with variance prediction errors. (C_B) BOLD response of the right anterior insula tracks norm prediction errors. From Xiang et al. (2013).

Quartz, & Bossaerts, 2008; Schoenbaum, Roesch, Stalnaker, & Takahashi, 2009).

These experiments provide examples of how social norms are learned and why people conform to them. Norm compliance can lead to increases in subjective utility and norm violations can result in an error signal that contributes to updating beliefs and behavior via computations from the VS and the ACC.

Learning About Others

Another area of research involves learning about others. For successful social functioning, one needs to accurately infer personal characteristics (Mende-Siedlecki, Cai, & Todorov, 2013; Stanley, 2015), beliefs, and intentions of others (Delgado, Frank, & Phelps, 2005; Fareri, Chang, & Delgado, 2015; Hampton et al., 2008; King-Casas et al., 2005, 2008). Computational approaches provide a useful way to understand how the human brain dynamically updates beliefs about others in a continuously changing social environment.

Psychological and neural mechanisms of this type of social learning are often studied using repeated games. For instance, a repeated Trust Game (TG; Fig. 17.2) can allow two players (investor and trustee) to build a trusting relationship with each other. Delgado and colleagues (2005) investigated neural mechanisms of updating trustworthiness information during this iterated TG and found that the caudate nucleus was involved in differentiating positive and negative information about others. However, this neural signal was modulated by participants' initial impression of their partner's moral character, leading them to ignore information inconsistent with their initial beliefs. This social confirmation bias effect was subsequently replicated in other studies using reinforcement learning models (Fareri, Chang, & Delgado, 2012; Fouragnan et al., 2013). A hyperscanning fMRI study by King-Casas

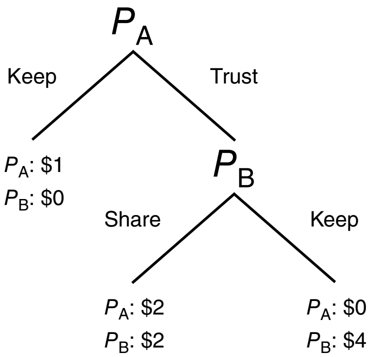

Figure 17.2 Example of the Trust Game. Player A is endowed with \$1 and may choose to keep the endowment or trust Player B, in which case the endowment is multiplied by a factor of 4. Player B can then decide to share the multiplied endowment with Player A or keep the \$4.

and colleagues (2005) found that the dorsal striatum was associated with building reputation about a partner's reciprocity, reflected by trustees' perceptions of an investor's "intention to trust." The peak of the "intention to trust" signal was temporally shifted from late to earlier occurrence, resembling reward prediction error signals commonly observed in standard reinforcement learning paradigms.

Chang et al. (2010) used reinforcement learning models to describe how initial trustworthiness information changes with experience in a repeated TG. They formalized trust as a belief about the probability of a relationship partner reciprocating and found that these beliefs dynamically change with experience. Extending this approach, Fareri, Chang, and Delgado (2015) examined how a prior relationship with a partner might affect behavior in the game. They proposed a social value model in which participants learn the probability of their partner reciprocating using a reinforcement learning model. This probability scales the amount of reward they expect to receive if their partner reciprocates. Importantly, this reward is composed of their financial incentive and also social incentive that is proportional to their partner's perceived trustworthiness. Formally, the probability of trusting partner c on trial t is:

$$Trust_c(t) = \frac{e^{\frac{\Big(P_c(t-1)+\alpha\cdot(\gamma_c(t)-P_c(t-1)))\cdot(1.5+\Big(\theta\cdot\frac{T_c}{max(T_c)}\Big)\Big)}{\beta}}}{e^{\frac{\Big(P_c(t-1)+\alpha\cdot(\gamma_c(t)-P_c(t-1)))\cdot(1.5+\Big(\theta\cdot\frac{T_c}{max(T_c)}\Big)\Big)}{\beta}} + e^{\frac{1}{\beta}}} \qquad \text{(Eq. 17.2)}$$

where $\alpha = [0,1]$ is the learning rate on how quickly participants update their beliefs, $\beta = [0,1]$ is the degree of stochasticity in the decision, γ is 1 if the partner reciprocated or 0 if they defected, $P_c(t-1)$ is the participant's belief about the likelihood of partner c reciprocating on the previous round, T_c is the participant's subjective trustworthiness rating for partner c, and $\theta = [0, 5]$ reflects the scaling of the social value term.

The authors found that this social value model provided a better account of participants' decisions than a standard reinforcement learning model (Eq. 17.1) in which the initial starting values were biased by participants' initial trustworthiness ratings. Moreover, the authors used the predictions of the social value model to identify regions of the brain that were associated with trial-to-trial learning and also to find regions that correlated with the magnitude of the social value signal. Replicating previous work (Fareri et al., 2012; Fouragnan et al., 2013), they found activity in the VS significantly correlated with trial-to-trial prediction errors. Importantly, they found that activity in the DMPFC and VS significantly correlated with the model-derived social value metric, providing further validation that the model was accurately capturing this psychological construct (Fig. 17.3).

In addition to trust, learning about others' personal attributes in impression formation (Stanley, 2015) and inferring mental states during strategic interactions (Hampton et al., 2008) also provide insight into the neural basis of social learning. Stanley (2015) used a Bayesian learning model to compare neural mechanisms of learning about generosity of target figures (social learning) and learning about winning probability of a lottery (nonsocial learning). The fMRI analyses showed that the DMPFC, DLPFC, and right lateral parietal cortex were related to prediction error signals for both social and nonsocial learning, whereas activations in the precuneus comprised prediction error signals more specific to social learning.

Consistent with this finding, Sul et al. (2015) compared neural mechanisms of reward-based learning for self and other and found that VS, precuneus, and posterior STS (pSTS; extending to inferior parietal lobule and temporal parietal junction) reflected general reward prediction errors both for self and other, whereas the medial prefrontal cortex (MPFC) reflected the chosen value. In this study, the degree of spatial segregation of the value computation signals between the VMPFC and DMPFC for self and other reflected individual differences in prosociality, such that the more prosocial participants were, the greater the overlap between self and other.

In summary, computational research on social learning involves learning from and about others. Different variants of reinforcement learning models are commonly used to formally describe how individuals update and integrate social information into their own experiences and adapt to a constantly changing social environment. Prediction error signals for nonsocial and social learning seem to share the same neural substrates including ventral and dorsal striatum, and ACC. The subregions of MPFC seem to compute a common currency for both social and nonsocial decision utility, reflecting personal characteristics. Though DMPFC, TPJ, pSTS, and precuneus are regarded as "social" regions in many studies, the distinction between social

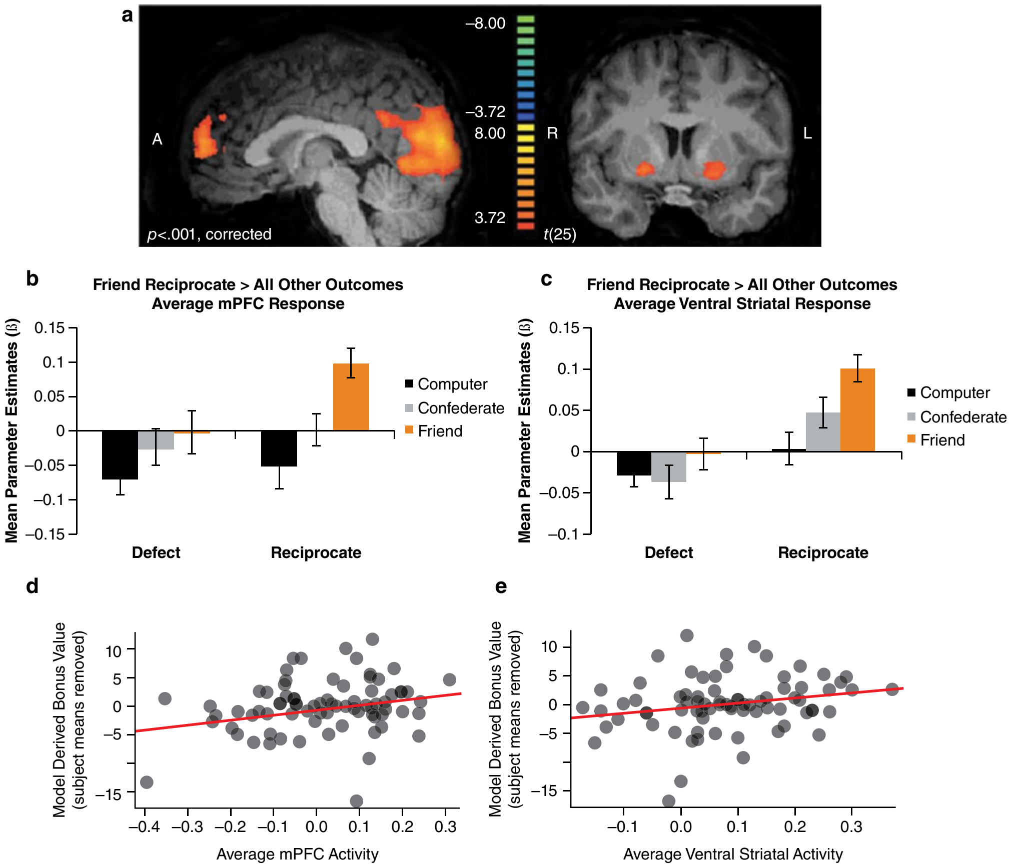

Figure 17.3 Neural representations of social value. (*a*) Bilateral VS and MPFC are more active when friends reciprocate compared to all other outcomes. (*b, c*) Mean parameter estimates of average MPFC (*b*) and VS (*c*) activity from the contrast in (*a*). (*d, e*) Average activations in MPFC (*d*) and VS (*e*) show significant predictive relationship with the model-derived bonus values. From Fareri et al. (2015).

learning and nonsocial learning in these regions deserves further research.

Mentalizing and Strategic Reasoning

Decisions made in social contexts also involve considering the decisions made by conspecifics. However, predicting others' choices presents a unique challenge: while behavioral outcomes themselves are directly observable, intentions are not and therefore must be inferred. This process of inferential reasoning has most often been called "mentalizing" or employing a "theory-of-mind," whereby an individual forms a theory (prediction) about the unobservable causes for an observed behavior, and uses this theory to guide predictions about future behavior (Frith & Frith, 2012; Lee & Seo, 2016; Premack & Woodruff, 1978).

Early work in social cognitive neuroscience was primarily concerned with identifying which brain regions subserved this type of inferential reasoning (Adolphs, 2001; Lieberman, 2007). Utilizing paradigmatic approaches from social and moral psychology, a reliable network of brain regions was quickly identified encompassing the STS, posterior cingulate cortex (PCC), and two key nodes, namely the TPJ/pSTS and the MPFC. The TPJ responds preferentially when individuals make inferences about the beliefs of others (Saxe & Wexler, 2005; Saxe & Kanwisher, 2003), such as reckoning the intentions of morally questionable actions (Young, Cushman, Hauser, & Saxe, 2007), trustworthiness of a partner (Behrens et al., 2008), and making empathic and agentic evaluations (Decety & Lamm, 2007). The MPFC responds preferentially when distinguishing between the thoughts and feelings of the self and others (Amodio & Frith, 2006; Jenkins & Mitchell, 2011), forming impressions about others (Mende-Siedlecki et al., 2013; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009), and judging others' preferences (Koster-Hale & Saxe, 2013; Mitchell, Macrae, & Mahzarin, 2006).

More recently, studies have employed computational modeling to go beyond identifying *where* in the brain mentalizing-related processing occurs to answer *how* this type of processing occurs. By using game theoretic approaches from behavioral economics, investigators can model how individuals engage strategic reasoning in both competitive and cooperative contexts. Game theory provides a set of solutions to such contexts that advise the best strategy an agent should follow given complete information. However, individuals often diverge from this strategy and act according to their subjective beliefs about the strategies of others, constrained by their own cognitive limitations (Gigerenzer & Selten, 2002; Lee & Seo, 2016). For example, consider the responder in the UG who may either accept or reject a proportion of the endowment suggested by the proposer. Both players receive the proposed split if the responder accepts, but both players receive nothing if the responder rejects the offer. If players were motivated purely by financial interests, the proposer would propose the lowest possible nonzero offer and the responder would accept any nonzero offer (referred to as the subgame perfect equilibrium). Instead, responders in an UG are highly sensitive to both outcomes (payoffs from actual offers made) and intentions (knowledge about what offers a decider could have made) and incorporate both when accepting decisions in the UG (Falk, Fehr, & Fischbacher, 2003).

As such, recent efforts to develop models of mentalizing have been based on boundedly rational theories of cognitive hierarchy rather than equilibrium analyses (Camerer, Ho, & Chong, 2015; Stahl & Wilson, 1995). Such models continue to assume that individuals choose a utility-maximizing strategy, but relax the assumption that individuals are consistently correct regarding their predictions about others' actions (Camerer et al., 2015). In this way, variance occurs across individuals' strategies based on the depth of reasoning they employ (e.g., I believe you will choose *X*; I believe that you believe I will choose *X*; I believe that you believe that I believe . . . etc.). Such models have found recent success in explaining the activity commonly observed in mentalizing brain regions. For example, Coricelli and Nagel (2009) had individuals play a Keynesian Beauty Contest Game whereby individuals were paid commensurate to choosing a number between 0 and 100 that was *M* times the average of guesses made by all others playing the game. Through backwards induction, the Nash equilibrium strategy dictates choosing 0 (e.g., for $M = 2/3$), yet most individuals chose values predicted by a step-by-step reasoning of an iterated best reply model ($50 \times M^k$) where k is the depth of reasoning that an individual employed (participants typically employed strategies with k between 1 and 3). Furthermore, individuals with higher k values (i.e., greater depth of reasoning) demonstrated greater activity in ventral and dorsal regions of the MPFC, what the authors referred to as "strategic IQ."

In a similar study, Bhatt, Lohrenz, and Camerer (2010) had individuals engage in a bargaining game where buyers provided information to sellers regarding the valuation of an item in an attempt to influence price setting and negotiate a sale. Because sales were only enacted if sellers set prices below the true value of the item and buyers were free to be as truthful as they wanted, the authors were able to establish different depths of strategies that individuals employed during the game. Consistent with a second-level depth of strategic reasoning (i.e., $k = 2$), 20% of players employed a deceptive strategy that earned them more money. These individuals exhibited stronger activity in DLPFC and TPJ during bargaining bluffs relative to others, suggesting a role for these regions in tracking the degree of influence that one individual has on others during a strategic interaction.

Hampton et al. (2008), more directly modeled social influence using an fMRI paradigm in which individuals played a competitive game known as the inspection game. In this game, participants played the role of either an employer or an employee. The employer chose whether to inspect the employee and the employee decided whether to work. To maximize payoffs, the employer had to inspect when the employee was not working and the employee had to work when the employer inspected, but not otherwise. The authors fit three different models to individuals' decisions: (1) a simple reinforcement learning model in which future actions were chosen based on previously successful actions, (2) a fictitious play (or elementary mentalizing) model in which future actions were chosen based on best responses to a competitor's previous actions, and (3) an influence model in which future actions were chosen based on a prediction of a competitor's belief regarding one's own action (i.e., incorporating the influence one has on their competitor). The experimenters found that participants track both their opponent's actions and the influence of their own strategy on their opponent's strategy confirming that mentalizing is a key component of social decisions. The MPFC incorporated influence information and reflected each individual's expectation, while pSTS and VS were involved in updating new information by capturing prediction error-like signals, namely the difference between expected and actual influence. Activations in the MPFC appeared to reflect participants belief about their level of influence over their partner as activity in this region correlated with individual variability in the degree to which the influence model provided a better account of participant's behavior compared to the fictitious play model. Moreover, activity in pSTS and VS covaried with MPFC activity, providing evidence that these regions communicate to support mentalizing computations.

Yoshida, Dolan, and Friston (2008; Yoshida, Seymour, Friston, & Dolan, 2010) built on this work to develop a more sophisticated "belief inference mode." In this model, individuals try to infer the strategy of another agent by watching how game states change as a consequence of others' decisions. This process allows an agent to infer the depth of reasoning k that another agent is utilizing, and respond by utilizing a strategy of $k + 1$. In other words, their model assumes that an individual chooses a strategy by first inferring the strategy in use by the other agent, and then responds by picking a strategy that uses a "deeper" level of reasoning.

To test their model, they utilized a stag hunt game in which individuals worked either competitively or cooperatively with a computer agent to "hunt" either a low-value and easy-to-catch reward (rabbit) or a high-value but difficult-to-catch reward (stag). In order to estimate participants' inferences, the computer agent operated at different levels of recursive inference which changed randomly throughout the game. By modeling cooperation rates and participants decisions, Yoshida and colleagues (2010) were able to infer the depth of strategic recursion individuals were employing and found that individuals respond to strategy changes employed by the computer agent. MPFC tracked individuals' uncertainty regarding computer strategies, and DLPFC, superior parietal lobule, and frontal eye fields tracked the recursive depth of individuals' own strategies, consistent with findings from Bhatt and colleagues (2010). The novelty in this study lies in explicitly modeling the depth of recursion that individuals utilize during strategic reasoning (and thereby their beliefs about the computer agent's depth of strategic reasoning) and the dynamic generation of beliefs over repeated play.

The ability to infer the intentions of others also seems to improve with age. Sul et al. (under revision) investigated how participants aged between 9 and 23 years old responded to multiple rounds of an UG in which information about the alternative split that the proposer could have offered was revealed. In this modified UG, participants could make decisions based on an egalitarian strategy (e.g., Was the split 50/50?) represented using an inequity aversion model (Fehr & Schmidt, 1999), or alternatively, participants could infer the intentions motivating the other player's decision (e.g., Why did he/she choose this offer rather than the alternative?), which was modeled using a reciprocity model (Dufwenberg & Kirchsteiger, 2004; Rabin, 1993). Younger participants used the simpler rule-based egalitarian strategy, but adolescents shifted to using a more sophisticated intention-based reciprocity strategy around 17 years of age. Importantly, the degree to which the intention-based reciprocity strategy was preferred to the egalitarian strategy was mediated by cortical thinning in the DMPFC and the posterior temporal lobes, suggesting that the development of these regions is integral in making social inferences (Coricelli & Nagel, 2009; Güroğlu, van den Bos, & Crone, 2009; Lee & Seo, 2016).

While the explicit modeling of mentalizing processes is a relatively new research effort, findings from these groups demonstrate how computational models can be utilized to explicitly test theories about both behavioral and neural mechanisms. In particular, these results and several others (Carter, Bowling, Reeck, & Huettel, 2012; Seo, Cai, Donahue, & Lee, 2014; Suzuki et al., 2012) demonstrate how social information is utilized by brain regions involved in mentalizing and strategic reasoning, and that this process specifically involves estimating the degree of influence one's own decisions have on others' beliefs and dynamically updating these estimations in order to choose optimal actions (Lee & Seo, 2016).

Conclusion

In this chapter, we reviewed a number of studies that have employed computational modeling to help us understand the neural and psychological processes underlying

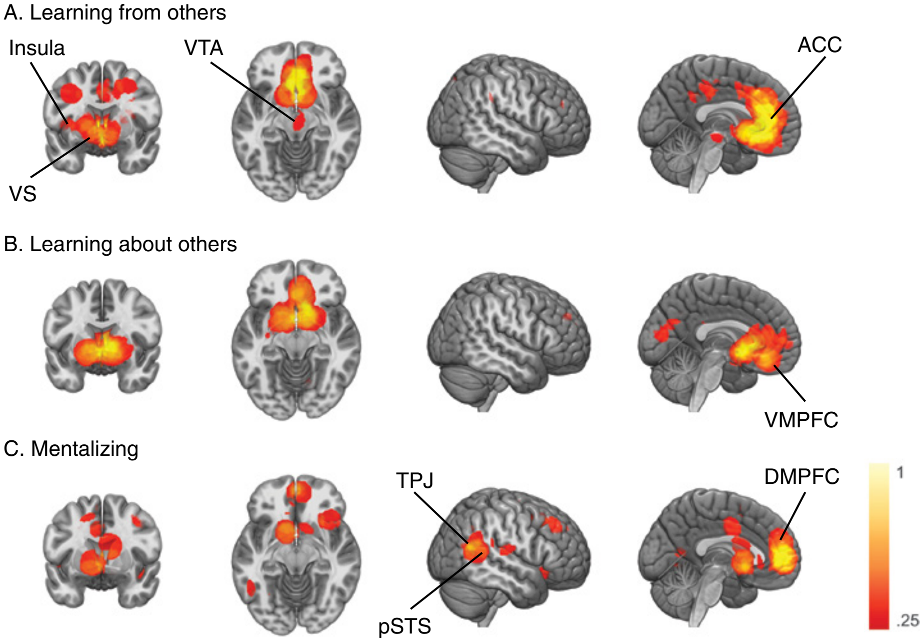

Figure 17.4 Meta-analysis representation of commonly activated regions for social learning and mentalizing. (A) Regions commonly activated for observational/vicarious learning and social conformity by norm prediction error signals tracked in VS, VTA, ACC and VMPFC. (B, C) Inferring other people's intentions or characteristics recruit DMPFC, TPJ, and pSTS also incorporating valuations represented in the VMPFC that are updated via prediction error signals from VS and ACC. Brighter regions indicate greater common activation as percentage overlap across studies. See Table 17.1 for list of studies included in the analysis.

social cognition in the context of learning and decision making. The bulk of the work to date has leveraged modeling frameworks from economic utility theory and reinforcement learning. Overall, several consistent findings have begun to emerge. The DMPFC, PCC, and TPJ appear to be reliably computing processes related to inferring others' mental states (Figure 17.4B, C). The VMPFC and VS are involved in representing monetary or social value, while the VTA and VS as well as the ACC and insula are involved in calculating different types of prediction errors (Figure 17.4A). The insula and dorsal ACC appear to be involved in errors involving stronger negative affective responses from norm violations, while the VTA and VS are more reliably involved in learning probabilities (Table 17.1 includes all studies included in this analysis).

The use of computational modeling in the various studies discussed in this chapter has permitted researchers to formally test specific theories that describe the functional processing in these brain regions. In each of these cases, this approach involved outlining a mathematical account of a possible strategy utilized by participants (e.g., recursive reasoning) or learning process performed by a brain region (e.g., reinforcement learning).

Table 17.1 List of studies included in Fig. 17.4.

Study	Learning from others	Learning about others	Mentalizing
Apps et al. (2015)	×		
Behrens et al. (2008)	×		
Bhatt et al. (2010)			×
Boorman et al. (2013)		×	
Burke et al. (2010)	×		
Campbell-Meiklejohn et al. (2010)	×		
Cooper et al. (2012)	×		
Coricelli & Nagel (2009)			×
Delgado et al. (2005)		×	
Fareri et al. (2012)		×	
Faereri et al. (2015)		×	
Fouragnan et al. (2013)		×	
Hampton et al. (2008)			×
Hill et al. (2016)	×		
Jones et al. (2011)		×	
King-Casas et el. (2015)		×	
Klucharev et al. (2009)	×		
Lin et al. (2011)		×	
Stanley (2015)		×	
Sul et al. (2015)		×	
Sul et al. (Under Review)			×
Xiang et al. (2013)		×	
Yoshida et al. (2010)			×
total number of studies	7	11	5

Coordinates were dilated into a 15-mm-radius sphere and overlapped to generate Fig. 17.4.

The work discussed here demonstrates the power of the computational approach to draw inferences beyond those afforded by simple social psychological paradigms, which often lack mechanistic explanations.

Overall, we believe that the application of computational techniques to the study of the social and affective brain is an exciting endeavor with immense potential for growth and innovation. We encourage more researchers from both computational and social disciplines to consider collaboratively developing new approaches to contribute to this enterprise.

References

Adolphs, R. (2001). The neurobiology of social cognition. *Current Opinion in Neurobiology, 11*(2), 231–239.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews. Neuroscience, 7*(4), 268–277.

Apps, M. A. J., Lesage, E., & Ramnani, N. (2015). Vicarious reinforcement learning signals when instructing others. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 35*(7), 2904–2913.

Asch, M. J. (1951). Nondirective teaching in psychology: An experimental study. *Psychological Monographs: General and Applied, 65*(4), i–24.

Bandura, A. (1977). *Social learning theory*. PToronto, Canada: Prentice-Hall of Canada.

Behrens, T. E. J., Hunt, L. T., & Rushworth, M. F. S. (2009). The computation of social behavior. *Science, 324*(5931), 1160–1164.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*(7219), 245–249.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & 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*(46), 19720–19725.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences, 8*(12), 539–546.

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.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America, 107*(32), 14431–14436.

Camerer, C. F., Ho, T., & Chong, J. K. (2015). A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences, 3*, 157–162.

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., & Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Current Biology, 20*(13), 1165–1170.

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science, 337*(6090), 109–111.

Chang, L. J., Doll, B. B., van 't Wout, M., Frank, M. J., & Sanfey, A. G. (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105.

Chang, L. J., & Koban, L. (2013). Modeling emotion and learning of norms in social interactions. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(18), 7615–7617.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*(3), 277–284.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In M.P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 201–234). NewYork: Academic Press.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 151–192). New York: McGraw-Hill.

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.

Corbetta, M., Maurizio, C., Gaurav, P., & Shulman, G. L. (2008). The reorienting system of the human brain: From environment to theory of mind. *Neuron, 58*(3), 306–324.

Coricelli, G., & 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*(23), 9163–9168.

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist, 13*(6), 580–593.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgement. *Journal of Abnormal Psychology, 51*(3), 629–636.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268–298.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry, 41*(1), 20–26.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*, 148.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 35*(21), 8170–8180.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185–190.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817–868.

Ferenczi, E. A., Zalocusky, K. A., Liston, C., Grosenick, L., Warden, M. R., Amatya, D., . . . & Deisseroth, K. (2016). Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior. *Science, 351*(6268), aac9698.

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational priors magnify striatal responses to violations of trust. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(8), 3602–3611.

Fox, A. S., Chang, L. J., Gorgolewski, K. J., & Yarkoni, T. (2014, January 1). Bridging psychology and genetics using large-scale spatial analysis of neuroimaging and neurogenetic data. *bioRxiv*. https://doi.org/10.1101/012310

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology, 63*, 287–313.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Güroğlu, B., van den Bos, W., & Crone, E. A. (2009). Fairness considerations: Increasing understanding of intentionality during adolescence. *Journal of Experimental Child Psychology, 104*(4), 398–409.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization, 3*(4), 367–388.

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 of the United States of America, 105*(18), 6741–6746.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience, 30*(2), 583–590.

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications, 7*, 1–12.

Jenkins, A. C., & Mitchell, J. P. (2011). Medial prefrontal cortex subserves diverse forms of self-reflection. *Social Neuroscience, 6*(3), 211–218.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science, 321*(5890), 806–810.

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.

Klucharev, V., Vasily, K., Kaisa, H., Mark, R., Ale, S., & Guillén, F. (2009). Reinforcement learning signal predicts social conformity. *Neuron, 61*(1), 140–151.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836–848.

Lee, D., & Seo, H. (2016). Neural basis of strategic decision making. *Trends in Neurosciences, 39*(1), 40–48.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology, 58*, 259–289.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38*(2), 339–346.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623–631.

Mitchell, J. P., Neil Macrae, C., & Mahzarin, B. (2006). Dissociable neural systems underlying impression formation. *Neuron, 50*, 655–663.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 16*(5), 1936–1947.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: Norm violations and their enforcement. *Neuron, 56*(1), 14–18.

Ochsner, K. N., Hughes, B., Robertson, E. R., Cooper, J. C., & Gabrieli, J. D. E. (2009). Neural systems supporting the control of affective and cognitive conflicts. *Journal of Cognitive Neuroscience, 21*(9), 1842–1855.

Ochsner, K. N., & Lieberman, M. (2001). The emergence of social cognitive neuroscience. *American Psychologist, 56*(9), 717–734.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452–454.

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.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 28*(11), 2745–2752.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*(5), 1281–1302.

Rescorla, R. A., Wagner, A. R., & others (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2*, 64–99.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews, Neuroscience, 15*(8), 549–562.

Sanfey, A. G., Stallen, M., & Chang, L. J. (2014). Norms and expectations in social decision-making. *Trends in Cognitive Sciences, 18*(4), 172–174.

Sarter, M., Berntson, G. G., & Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. *American Psychologist, 51*(1), 13–21.

Saxe, R., & Kanwisher, N. (2003/8). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage, 19*(4), 1835–1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508–514.

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews, Neuroscience, 10*(12), 885–892.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Seo, H., Cai, X., Donahue, C. H., & Lee, D. (2014). Neural correlates of strategic reasoning during competitive games. *Science, 346*(6207), 340–343.

Sherif, M. (1936). *The psychology of social norms* (Vol. xii). Oxford: Harper.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*(1), 218–254.

Stanley, D. A. (2015). Getting to know you: General and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience, 11*(4), 525–536.

Stanley, D. A., & Adolphs, R. (2013). Toward a neural basis for social behavior. *Neuron, 80*(3), 816–826.

Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., & Kim, H. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences of the United States of America, 112*(25), 7851–7856.

Sul, S., Güroglu, B., Crone, E. A., & Chang L. J. (under revision). Medial prefrontal cortical thinning Mediates shifts in other-regarding preferences during adolescence.

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.

Thorndike, E. L. (1911). *Animal intelligence; experimental studies* (p. 324). New York: Macmillan.

Vickery, T. J., Chun, M. M., & Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron, 72*(1), 166–177.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 33*(3), 1099–108a.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology, 4*(12), e1000254.

Yoshida, W., Seymour, B., Friston, K. J., & Dolan, R. J. (2010). Neural mechanisms of belief inference during cooperative games. *Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(32), 10744–10751.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences of the United States of America, 104*(20), 8235–8240.

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage, 129*, 95–104.

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 of the United States of America, 109*(5), 1419–1424.