CORTEX 127 (2020) 221–230

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue "Understanding Others": Review Paper

Breaking human social decision making into multiple components and then putting them together again

Shinsuke Suzuki [a,b,*] *and John P. O'Doherty* [c,d]

[a] *Brain, Mind and Markets Laboratory, Department of Finance, Faculty of Business and Economics, The University of Melbourne, Parkville, Australia*

[b] *Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai, Japan*

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

[d] *Computation and Neural Systems, California Institute of Technology, Pasadena, USA*

ARTICLE INFO

Article history:
Received 4 July 2019
Reviewed 17 November 2019
Revised 23 January 2020
Accepted 28 February 2020
Published online 9 March 2020

Keywords:
Social cognition
Decision-making
Model-based fMRI
Reinforcement learning
Computational neuroscience

ABSTRACT

Most of our waking time as human beings is spent interacting with other individuals. In order to make good decisions in this social milieu, it is often necessary to make inferences about the internal states, traits and intentions of others. Recently, some progress has been made toward uncovering the neural computations underlying human social decision-making by combining functional magnetic resonance neuroimaging (fMRI) with computational modeling of behavior. Modeling of behavioral data allows us to identify the key computations necessary for social decision-making and to determine how these computations are integrated. Furthermore, by correlating these variables against neuroimaging data, it has become possible to elucidate where in the brain various computations are implemented. Here we review the current state of knowledge in the domain of social computational neuroscience. Findings to date have emphasized that social decisions are driven by multiple computations conducted in parallel, and implemented in distinct brain regions. We suggest that further progress is going to depend on identifying how and where such variables get integrated in order to yield a coherent behavioral output.

1. Introduction

A fundamental question in neuroscience is how do we make decisions. A popular framework developed in economics, psychology and machine-learning called value-based decision-making (Rangel, Camerer, & Montague, 2008), posits that (i) our brain assigns a scalar quantity, *subjective value*, to each available option, then (ii) selects the option with the highest value, and finally (iii) updates values of the options based on experienced outcomes (i.e., learning). Application of formal models of value-based decision-making (e.g., reinforcement learning) to behavioral and neuroimaging data has uncovered neural mechanisms underlying human decision-making (Daw

* *Corresponding author*. Brain, Mind and Markets Laboratory, Department of Finance, Faculty of Business and Economics, The University of Melbourne, 198 Berkeley St, Carlton, VIC 3053, Australia
E-mail address: shinsuke.szk@gmail.com (S. Suzuki).

https://doi.org/10.1016/j.cortex.2020.02.014
0010-9452/

& Doya, 2006; Glimcher & Rustichini, 2004; O'Doherty, Hampton, & Kim, 2007; Schultz, Dayan, & Montague, 1997): for instance, subjective value signals are encoded in the medial prefrontal cortex (Chib, Rangel, Shimojo, & O'Doherty, 2009; Kable & Glimcher, 2007; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Suzuki, Cross, & O'Doherty, 2017) while learning signals (i.e., reward prediction errors) are encoded in the striatum and dopaminergic midbrain (McClure, Berns, & Montague, 2003; O'Doherty et al., 2004; Rutledge, Dean, Caplin, & Glimcher, 2010).

In the last decade, researchers have employed formal theoretical approaches from economics, game theory and machine-learning to investigate the neural underpinnings of human social behavior with reference to the value-based decision-making framework (Behrens, Hunt, & Rushworth, 2009; Dunne & O'Doherty, 2013; Hackel & Amodio, 2018; Lee, 2008; Ruff & Fehr, 2014). Social decision-making is very complex, because it often requires inference about hidden states such as another's intentions, state of mind, traits and/or predispositions. Indeed, accumulating evidence suggests that multiple forms of computation performed in distinct brain regions might underlie social decision-making (Behrens et al., 2009; Charpentier & O'Doherty, 2018; Dunne & O'Doherty, 2013; Joiner, Piva, Turrin, & Chang, 2017; Konovalov, Hu, & Ruff, 2018; Lee & Seo, 2016; Ruff & Fehr, 2014; Wittmann, Lockwood, & Rushworth, 2018). In other words, to compute values for available decision options in a social situation, one might need to integrate multiple computations about one's own individual preferences, one's preferences about the outcomes that others can receive, socially-specific inferences about others, domain-general inferences about the environment and so on. Yet, little is known about how these multiple computations necessary for social decision-making are integrated in the human brain.

Here, in this review, we discuss these issues, while maintaining a focus on studies that extend the value-based decision-making framework to social behavior. We first outline a simple extension of this framework to the social domain: decision-making for others, in which consideration of another individuals' welfare works as a modulatory factor. Next we consider learning through observing others, in which anothers' choice and its consequence work as a source of learning. Finally, we consider an expansion of this framework to the domain of strategic decision-making/learning.

2. Value-based decision-making for others

In our daily life, we make value-based decisions not only for our own interest but also for the benefit of other individuals. For instance when making charitable donations, decisions are made about how resources are allocated between oneself and deserving others. In such situations, an individual computes the value of each option by considering the reward outcomes for both oneself and others according to various social preferences, such as the warm–glow, inequity–aversion and envy–aversion effects (Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017; Fehr & Camerer, 2007; Fehr & Shmidt, 1999; Fukuda et al., 2019; Harbaugh, Mayr, & Burghart, 2007; Hula, Vilares, Lohrenz, Dayan, & Montague, 2018; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Takahashi et al., 2009). These findings suggest that multiple forms of social and non-social information are represented in the brain to guide choices. One study (Hutcherson, Bushong, & Rangel, 2015) examined simple decisions about different allocations of monetary reward between oneself and an anonymous partner. They found that choice behavior and reaction-times can be well-captured by a computational model called the Multi-attribute Drift-Diffusion Model (Ratcliff & McKoon, 2008). Furthermore, monetary reward for oneself and for the anonymous partner were found to be encoded in the ventral striatum and temporoparietal junction (TPJ) respectively.

Another study (Hsu, Anen, & Quartz, 2008) investigated decision-making between different donation plans to two groups of children living in an orphanage in northern Uganda. Importantly, the plans differed in terms of efficiency (i.e., the overall amounts of money donated to the two groups) and inequity (i.e., the difference in the amounts donated to the two groups). Information about efficiency was found to be represented in a region of dorsal striatum (the putamen) while information about inequity was encoded in insula.

Moreover, we sometimes make decisions on behalf of others (e.g., leadership decisions) (Edelson, Polania, Ruff, Fehr, & Hare, 2018; Jung, Sul, & Kim, 2013; Nicolle et al., 2012; Ogawa, Ueshima, Inukai, & Kameda, 2018). In one study (Edelson et al., 2018), participants had to decide whether to take the lead (i.e., by making a choice on behalf of the group) or not (i.e., following the majority's choice). Participants in that experiment tended to avoid assuming leadership, especially when the choice was difficult; and patterns of connectivity among brain regions encoding task-relevant variables (e.g., choice difficulty, probability of leading, and so on) were found to predict individual differences in leadership decisions and self-reported leadership scores.

Another important component of social value-based decision-making involves learning on behalf of others (i.e., learning about the consequence of one's own action for other individuals). Researchers have identified neural underpinnings of learning for others to attain monetary reward (Christopoulos & King-Casas, 2015; Lockwood, Apps, Valton, Viding, & Roiser, 2016), to avoid painful electric shock (Lockwood, Klein-Flügge, Abdurahman, & Crockett, 2019), and to reduce exposure to unpleasantly loud noise (Sul et al., 2015). Some of these studies (Lockwood et al., 2016, 2019; Sul et al., 2015) have found that the ventral striatum tracks learning signals, reward prediction errors, when learning for others as well as oneself (but see Christopoulos & King-Casas (2015) for counter evidence). On the other hand, prediction error signals specific to learning for others have been found in the thalamus/caudate (Lockwood et al., 2019) and vmPFC expanding to subgenual anterior cingulate cortex (sgACC) (Christopoulos & King-Casas, 2015; Lockwood et al., 2016).

It is worth noting that there is likely to be considerable variation in social preference across individuals (Fehr & Camerer, 2007; Lange, Bruin, Otten, & Joireman, 1997). Some people appear to care a lot about others' payoffs, while others seem to care much more about their own payoff. Such individual differences modulate neural responses to fairness (Haruno & Frith, 2009), other-regarding values (Sul et al., 2015)

and prediction errors about others' rewards (Christopoulos & King-Casas, 2015).

While we have so far emphasized contributions of anatomically distinct brain regions to self-regarding and other-regarding computations, recent studies suggest the existence of more flexible representations in the ventral–dorsal axis of medial prefrontal cortex (Nicolle et al., 2012; Sul et al., 2015). Comparing decision-making for oneself and that on behalf of others, one study (Nicolle et al., 2012) demonstrated that vmPFC tracked "executed value" utilized for the current choice. That is, the brain region signaled self- and other-referential values, when making a choice for self and for others, respectively. These authors further found that dorsomedial prefrontal cortex (dmPFC) encoded "modeled value", which is not used for the current choice but could be internally simulated (i.e., self- and other-values, when making a choice for others and for self, respectively). Another study (Sul et al., 2015) revealed that social preference modulated a spatial gradient from vmPFC predominantly representing self-value to dmPFC encoding other-value. These studies together challenge the view that distinct and non-overlapping neural mechanisms are utilized for social and non-social inferences.

Most of the studies discussed above have investigated social interactions with anonymous others. However, several studies suggest that neural responses to social decision-making can be modulated as a function of the relationship between self and other such as the degree of social distance (Strombach et al., 2015), friendship (Fareri, Chang, & Delgado, 2015) and group membership (Hackel, Zaki, & Bavel, 2017; Hein, Engelmann, et al., 2016). For example, in the context of decision-making for others, willingness to pay for another individual's benefit declines with an increase in social distance. Social-distance-dependent choices were found to be associated with neural activity in TPJ and vmPFC and in the functional coupling between those areas (Strombach et al., 2015).

3. Value-based decision-making through observing others

To make appropriate decisions, one needs to learn the values of available actions (options) and of other features of the environment. Such learning can be accomplished not only by one's own experience but also by observing another individual's experience. This type of observational learning exists in multiple species, and has been directly examined in a range of species from rodents to humans (Burke, Tobler, Baddeley, & Schultz, 2010; Cooper, Dunne, Furey, & O'Doherty, 2011; Hill, Boorman, & Fried, 2016; Zentall, 2012). This form of learning is likely to be beneficial for survival as it enables individuals to efficiently acquire knowledge about the world without direct experience.

Recent human neuroimaging studies suggest that, for observational learning, human observers utilize two sources of information in order to acquire knowledge from an observee: the rewards obtained by the observee in relation to particular choices, and the actions performed by the observee (Burke et al., 2010; Suzuki et al., 2012). More precisely, prediction errors about the reward outcomes received by the observee have been found in vmPFC and dorsal striatum, and these may be used to update the value of the option chosen by the observee, in a similar manner to that which occurs during conventional reinforcement learning through direct experience. Indeed, a meta-analysis found that the vmPFC tracks reward prediction errors about both experienced and observed outcomes (Morelli, Sacchet, & Zaki, 2015). On the other hand, prediction errors about the observee's actions (i.e., the discrepancy between the observee's actual choice and the observer's prediction of the choice) have been found to be encoded in the dorsolateral prefrontal cortex (dlPFC) as well as other brain structures such as dorsomedial prefrontal cortex (dmPFC) and inferior parietal lobule (IPL) (Burke et al., 2010; Suzuki et al., 2012). Moreover, another line of studies focusing on fear conditioning, highlighted the pivotal role of amygdala in both learning from experienced and observed outcomes (Olsson, Nearing, & Phelps, 2007; Olsson & Phelps, 2007).

When making decisions and learning through direct experience, there exist two key strategies (Balleine & O'Doherty, 2010): one is a goal-directed strategy that tracks causal relations between actions and outcomes, and the other is a habitual strategy in which actions are automatically elicited by environmental states. Interestingly, a similar dichotomy between goal-directed and habitual strategies has been suggested to apply when learning through observation (Liljeholm, Molloy, & O'Doherty, 2012). However, the underlying neural mechanisms of these two strategies still remain elusive (Dunne, D'Souza, & O'Doherty, 2016; Liljeholm et al., 2012).

Learning about another individual's reliability or trustworthiness and competence through observing her behavior is also useful, especially when making a decision about whether or not to take into account her advice (Behrens, Hunt, Woolrich, & Rushworth, 2008; Boorman, O'Doherty, Adolphs, & Rangel, 2013). One study (Behrens et al., 2008) examined a case, in which in order to choose optimally, participants had to combine learning about option values through direct reward feedback and learning about the adviser's reliability through observation. They found that these two types of learning are formed in parallel in the brain: while the ventral striatum tracked learning about value from reward feedback, the right posterior superior temporal sulcus (pSTS) and TPJ tracked learning about the adviser's reliability. Furthermore, neural signatures of uncertainty in the two types of learning were found in distinct sub-regions of dACC, consistent with a theoretical proposition that uncertainty in the estimation of a prediction modulates the speed of learning about that prediction (Behrens, Woolrich, Walton, & Rushworth, 2007). Another study (Wittmann et al., 2016) used a mini-game that required participants to estimate their own and the other players' ability in cooperative and competitive contexts. The results of that study showed that the abilities of the participants themselves on the task and those of the other players were estimated based on past performance, and were represented in the vmPFC and dmPFC respectively. In addition to reliability and ability, researchers have assessed social learning about other types of traits (Delgado, Frank, & Phelps, 2005; Hackel, Doll, & Amodio, 2015; Stanley, 2016). For example, Hackel et al. (2015) devised a task that allows one to

dissociate learning about others' generosity from learning about the reward obtained from others. Utilizing this task, they revealed that, while both types of learning recruited ventral striatum, learning about others' generosity specifically employed a network of brain regions associated with social cognition: the TPJ and precuneus.

More broadly, neural signatures of learning signals (i.e., prediction errors) have been reported in many other social situations. For example, in the case of learning about ownership, prediction errors about others' and self ownership are represented in distinct sub-regions of dACC along the antero-posterior axis: the anterior part tracks prediction errors about whether objects belong to others, while the posterior part tracks prediction errors about individuals' own ownership (Lockwood et al., 2018). Furthermore, another study examined a teacher–student interaction in which the teacher informed the student whether the student's choice was rewarding or not (Apps, Rushworth, & Chang, 2016). Those authors demonstrated a role for anterior dACC in the teacher's brain in signaling prediction errors about the student's rewards, suggesting that teachers vicariously kept track of their students' learning progress. Importantly, these prediction error signals in ACC cannot be attributed to the domain-general process of error/conflict detection (Botvinick, Cohen, & Carter, 2004), because the signals were observed only in a particular condition (Lockwood et al., 2018), and were significant after controlling for effects of error trials and surprise (i.e., unsigned prediction error) (Apps, Lesage, & Ramnani, 2015).

Apart from value-based decision-making, information provided by others is useful for forming and updating one's belief about oneself. One study (Will, Rutledge, Moutoussis, & Dolan, 2017) examined how appraisals from others shape an individual's own self-esteem. Fluctuation in self-esteem was found to be driven by a prediction error corresponding to the discrepancy between expected and received social feedback. This was in turn represented in the ventral striatum and sgACC.

4. Value-based decision-making in strategic interactions

In the real world, we are often engaged in bilateral or reciprocal interactions, in which an individual needs to take into account predictions about another agent's intentions in order to make an advantageous decision, while the other agent also strives to predict the individual's intentions at the same time (Camerer, 2003). Researchers have begun to uncover computational processes for such strategic decision-making/learning and its neural underpinnings (Fareri et al., 2015; Hampton, Bossaerts, & O'Doherty, 2008; Haruno & Kawato, 2009; Hill et al., 2017; Lee & Seo, 2016; Suzuki, Adachi, Dunne, Bossaerts, & O'Doherty, 2015; Xiang, Ray, Lohrenz, Dayan, & Montague, 2012; Yoshida, Seymour, Friston, & Dolan, 2010; Zhu, Mathewson, & Hsu, 2012).

One important form of strategic decision-making arises when it is necessary to coordinate or form a consensus within a group. One study (Suzuki et al., 2015) devised a novel experimental task in which in the main condition participants had to make a unanimous consensus with other human participants. Behavioral modeling together with analysis of neuroimaging data demonstrated that one's decisions were guided by three separate factors: knowledge about one's own preference, information about the prior choices made by the majority of group-members, as well as an inference about how much each option is doggedly stuck to by the other group-members. These three different variables were found to be represented in the vmPFC, TPJ and IPL respectively (Fig. 1A, B). Note that although TPJ and IPL activations sometimes overlap, but this was not the case in this study (the peak MNI coordinates were: [60 −46 10] for TPJ and [30 −52 34] for IPL). Importantly, the experimental task used in this study had a control condition in which participants interacted with computer agents programmed to mimic actual human behavior. Comparison of the main and the control conditions revealed a significant difference in the neural representation of group-members' prior choices in TPJ, but not for the other variables. This suggests that information about the group-members' prior choices is processed in a social-specific manner in TPJ, while information about one's own preference and inference about the stickiness of an option (i.e., how much each option is stuck to by the other human group-members or computer agents) are processed in a domain-general manner in vmPFC and IPL respectively.

Another study (Hampton et al., 2008) examined a competitive interaction between two individuals, by using an experimental task originally developed in economics. The task, called the *inspection game*, models repeated interactions between an employee and her employer, in which the employee decides to work or to shirk (i.e., not to work) while the employer decides to inspect or not to inspect her employee. Each player gets a higher payoff if they can outsmart the opponent. For example, an employee obtains a higher reward by shirking without being inspected by the employer. The authors identified two computations related to forming a prediction about the opponent's next move. One is learning from the opponent's past choices, driven by prediction error, which was found to be encoded in the ventral striatum. The other is higher-order reasoning about how one's own current move will influence the opponent's next choice, which was found to be encoded in pSTS/TPJ. It is also worth noting that a recent study combining computational modeling and neuroimaging with non-invasive brain stimulation largely replicated the original findings of the Hampton et al. study, while establishing that the TPJ signal is causally relevant for computing a higher-order inference about the influence of one's own action on the opponent's next choice (Hill et al., 2017). In that study, theta burst stimulation over the TPJ which temporarily disrupts the functions of that region was found to result in a reduced tendency to engage the higher order inference compared to a sham control condition.

Complex strategic interactions often involve deep recursive reasoning about another's mental state: inference about your inference about my inference about your inference and so on (Camerer, 2003). A cognitive hierarchy theory, originally developed in game theory, posits that an individual conducts recursive inferences to a one-step higher level of depth than one's opponent in order to gain an advantage over the opponent (Camerer et al., 2004). In other words, a reasonable strategy is to estimate the opponent's depth-of-reasoning and adjust one's own behavior so as to be tailored to that

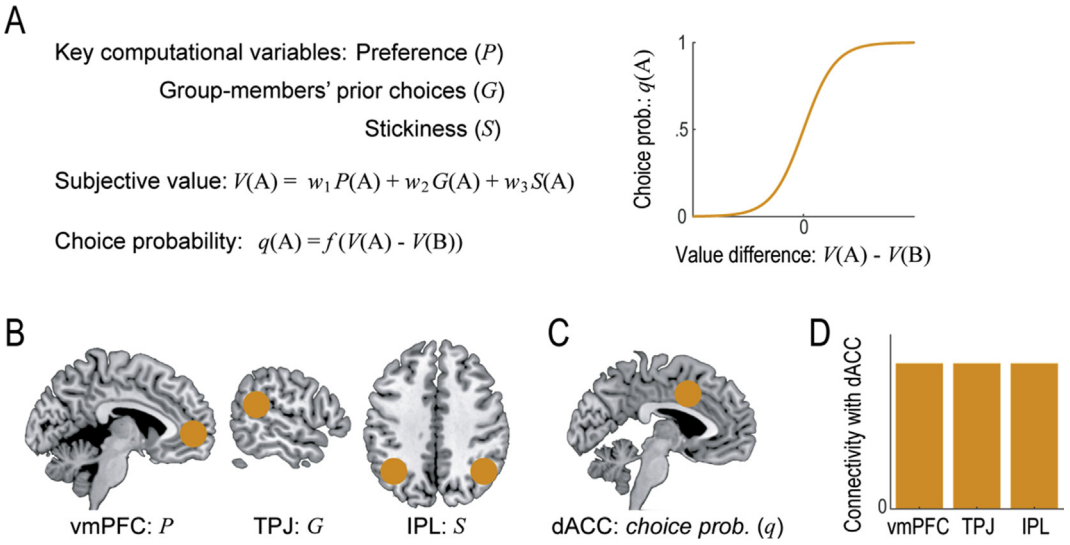

Fig. 1 – Neural mechanism underlying human consensus decision-making (Suzuki et al., 2015). (A) Illustration of the computational model. Subjective value of each option is computed through integrating one's own preference, group-members' prior choices and one's inference about how much each option is stuck to by the other group-members, and then finally converted to the choice probability. (B) Neural correlates of the three key computational variables. vmPFC: ventromedial prefrontal cortex; TPJ: temporoparietal junction; and IPL: inferior parietal lobule. (C) Neural correlates of the integrated choice probability. dACC, dorsal anterior cingulate cortex. (D) Functional connectivity between dACC and each of the three regions individually tracking the key computational variables.

estimation (by being one step more sophisticated). Although in general this type of inference is itself computationally costly, several formal models based on Bayesian inference have been proposed (Ray, King-Casas, Montague, & Dayan, 2008; Yoshida, Dolan, & Friston, 2008). Correlating these models with functional magnetic resonance imaging (fMRI) signals, Yoshida et al. showed that uncertainty about the estimation of the opponent's depth-of-reasoning was represented in dmPFC, while dlPFC tracked the depth of one's own strategy (Yoshida et al., 2010). Furthermore, Xiang et al. demonstrated differential brain regions encoded reward prediction error signals with respect to individual difference in participants' depth-of-reasoning (Xiang et al., 2012).

Note that the issues concerning strategic strategic decision-making are not mutually exclusive from the issues discussed in the previous sections. For example, inference about the opponent's depth-of-reasoning (Yoshida et al., 2010) can be interpreted as a form of learning about her traits (see the section of Value-based decision-making through observing others). Furthermore, decision-making in a strategic game called the Ultimatum game has been found to be affected by one's preference for fairness (see the section of Value-based decision-making for others) (Chang & Sanfey, 2013; Falk, Fehr, & Fischbacher, 2003; Sanfey et al., 2003).

5. Integration of multiple computations in social value-based decision-making

We have considered evidence supporting the possibility that multiple computational strategies are involved in parallel during many different forms of social value-based decision-making. However, the different computations need to be integrated somehow in order to generate a coherent behavioral output. So far we have reviewed neuroimaging studies that have identified a number of different forms of computation during social-decision-making, represented in discrete brain regions. In most of the computational models mentioned above, the form of the integration of the variables is to compute a subjective value or choice probability for a given option. For example, in models of decision-making for others (e.g., Fukuda et al., 2019; Hutcherson et al., 2015), a subjective value for each option is computed by integrating information about the amount of reward delivered to oneself and others (Fig. 2A). In observational learning models (e.g., Burke et al., 2010; Suzuki et al., 2012), leaning signals from others' choices and reward outcomes are integrated to compute the value of each option (Fig. 2B). In strategic decision-making models (e.g., Hampton et al., 2008; Hill et al., 2017; Suzuki et al., 2015), value computation requires the integration of multiple types of inference (Figs. 1 and 2C). Given these model structures (Figs. 1 and 2), we suggest that a brain region engaged in the integration process must (1) encode the integrated subjective value or choice probability signals assigned by the computational model (Fig. 1C) and (2) manifest functional connectivity with regions encoding each of the individual key computational variables (Fig. 1D). In other words, if a brain region is implicated in social information integration, the region must satisfy the above two criteria.

The first criterion has been examined in many studies. When correlating fMRI signals with model-derived overall subjective value or choice probability, converging evidence suggests that key computational variables necessary for decision-making are integrated in the vmPFC including the rostral ACC (rACC) and/or the dmPFC including the dACC (e.g.,

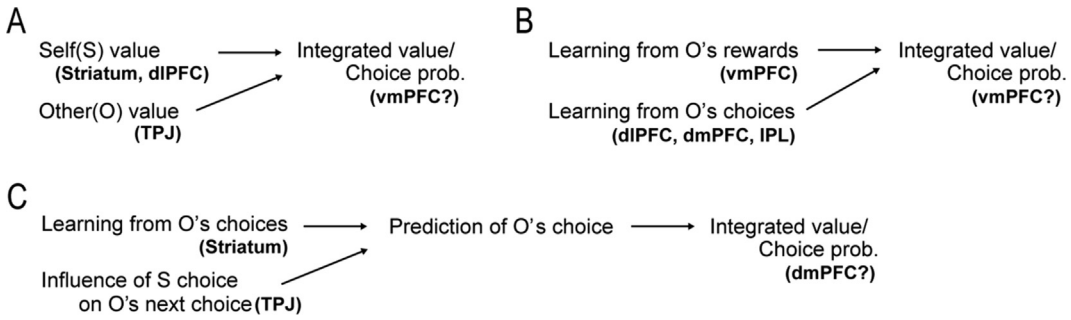

Fig. 2 – Schematic illustrations of example computational models for social value-based decision-making and their neural correlates. (A) Decision-making for others (Fukuda et al., 2019; Hutcherson et al., 2015). Overall value and choice probability of each option is computed by integrating information about self and other reward values. S: self; O: other; dlPFC: dorsolateral prefrontal cortex; TPJ: temporoparietal junction; and vmPFC: ventromedial prefrontal cortex. (B) Decision-making through observing others (Burke et al., 2010; Suzuki et al., 2012). Value and choice probability of each option is computed by integrating two types of learning from others' rewards and choices. dmPFC: dorsomedial prefrontal cortex; and IPL: inferior parietal lobule. (C) Decision-making in strategic interactions (Hampton et al., 2008; Hill et al., 2017). Value of each option is computed by the prediction of the other's choice that integrates learning from the other's past choices and higher-order inference about influence of self-choice on the other's next choice.

Behrens et al., 2008; Hsu et al., 2008; Hutcherson et al., 2015; Suzuki et al., 2012).

On the other hand, to our knowledge, only a few studies (Fukuda et al., 2019; Hampton et al., 2008; Hill et al., 2017; Suzuki et al., 2015) have examined both of the two criteria. For example, as mentioned in the previous section, Suzuki et al. (Suzuki et al., 2015) first identified three key computational variables and their neural correlates for consensus formation in a group: vmPFC encoding one's own preference for each of the available options, TPJ encoding group-members' prior choices and IPL encoding one's inference about how much each option was stuck to by the other group-members (Fig. 1A, B). Next, they found two brain regions, rACC and posterior dACC, that satisfied the first criterion: that is, fMRI signals in these two regions were correlated with modeled choice probability derived by integration of the three key computational variables (Fig. 1C). Finally, to examine the second criterion, a functional connectivity analysis was employed [Psycho-Physiological Interaction analysis (Friston et al., 1997)] which demonstrated that the posterior dACC, but not the rACC, had increased connectivity at the time of decision with each of the three regions, the vmPFC, the TPJ, and the IPL, that individually tracked the three key computational variables (Fig. 1D). Taken together, only the posterior dACC was foiund to satisfy both of the two criteria, suggesting that the three key computational variables involved in consensus decision-making are integrated in that region.

In the context of competitive interactions, the results from Hampton et al. suggest that the integration process occurs in the dmPFC (Hampton et al., 2008). The dmPFC was found to represent an overall subjective value and to have functional connectivity with the other regions, the ventral striatum and the pSTS/TPJ, responsible for individual computations underlying the decision-making (i.e., learning from the opponent's past choices and higher-order reasoning about how one's own choice will influence the opponent's next choice). This account of integration is further supported by a recent study showing that disruption of the TPJ alters its functional connectivity with the dmPFC (Hill et al., 2017).

In the context of decision-making for others, one study (Fukuda et al., 2019) tested for areas meeting the two criteria, by combining two types of connectivity analysis: Psycho-Physiological Interaction analysis (PPI) (Friston et al., 1997), and Dynamical Causal Modeling, DCM (Friston, Harrison, & Penny, 2003). Note that PPI analyses are based on a regression model and thus cannot examine the directionality of the connectivity, while DCM is based on a model of causal interactions of brain regions and can thus enable inference about the directionality of the effects. In Fukuda et al. (2019), these two connectivity analysis approaches consistently showed that integration of information about self- and others' rewards occurred in the vmPFC, including the adjacent rACC.

Furthermore, some other studies have aimed to address the issue of integration by using connectivity analyses, although they did not directly test the above two criteria. For example, van den Bos, Talwar, & McClure (2013) suggest that multiple computations necessary for bidding behavior in an auction are integrated in the vmPFC and striatum. Smith et al. proposed that the value of social stimuli (i.e., attractiveness of others' faces) are computed in the vmPFC via interactions with other regions such as the TPJ and middle temporal gyrus (Smith, Clithero, Boltuck, & Huettel, 2014). Finally, based on a meta-analytic connectivity analysis evaluating co-activation patterns across various tasks, the authors of one study (Alcalá-López et al., 2017) concluded that diverse neural circuits for from low-level sensory to high-level associative processes mediate human social cognitive capacities.

These findings could, we believe, provide a possible account for HOW social information is integrated with simple non-social decision-making processes. In studies on simple decision-making, it has been suggested that values of available options and goals in the vmPFC (including rACC) are utilized as inputs for computing values for actions in the posterior dmPFC (posterior dACC), and then finally transformed into a motor command in the motor cortex (Hare,

Schultz, Camerer, O'Doherty, & Rangel, 2011). The findings obtained in Suzuki et al., Hampton et al. and Fukuda et al. could suggest that, in the contexts of strategic decision-making, social information modulates basic decision-making processes at the stage of action value computation in the posterior dmPFC, while social information operates on the upstream stage (i.e., the computation of option and goal values in the vmPFC) in the context of decisions for others. This account further motivates another fascinating question: how is anterior dmPFC (anterior dACC), located between vmPFC and posterior dmPFC along the rostral–caudal axis of medial prefrontal cortex, involved in the social decision-making process? This question is of particular importance as the anterior dmPFC has been proposed to play a central role in social cognition (Amodio & Frith, 2006; Apps et al., 2016). Given the sparsity of studies to date that have addressed both of the above two criteria for how integration might happen across different computational strategies, however, future studies will need to address how and where discrete social computations are integrated across a wide array of different task domains and computational variables.

6. Conclusions

It is widely believed that decisions in social contexts are made through integrating multiple types of inference about one's own rewards, others' rewards, others' mental-states and so on. In the last decade, the notion has been supported by a computational modeling approach combined with neuroimaging (Behrens et al., 2009; Charpentier & O'Doherty, 2018; Dunne & O'Doherty, 2013; Joiner et al., 2017; Konovalov et al., 2018; Lee & Seo, 2016; Wittmann et al., 2018). By constructing formal models that can account for behavioral data during social decision making, it has become possible to identify key variables thereby, providing significant insights into the specific computations that underlie human social behavior. Furthermore, by correlating these variables against neuroimaging data, it has become possible to identify where in the brain these computations are implemented.

On the other hand, a more challenging and less explored issue is how these computations are integrated in the brain to guide social behavior. In a broader sense, this issue is related to a long-lasting question in neuroscience, known as "The Binding Problem" (Roskies, 1999). While in this review we have introduced some studies addressing this issue, more evidence is needed for a more comprehensive understanding of the information integration process. For example, to better understand the information integration process in the brain, it would also be essential to examine the nature of causal interactions (i.e., the direction of information flow) among multiple brain regions, which cannot be tested by correlation-based connectivity analysis methods alone such as psychophysiological interaction analyses (Friston et al., 1997). In future studies, we suggest that it will be increasingly important to emphasize the use of approaches more suited to address causal integration, such as Dynamic Causal Modeling or/and non-invasive brain stimulation together with computational modeling (Hein, Morishima, et al., 2016; Hill et al., 2017).

Important issues we have not addressed in this review are the perceptual aspects of social decision-making. Perception of social stimuli, especially others' faces, plays an important role in real-world decisions such as mate choice (Fletcher, Simpson, Thomas, & Giles, 1999), electoral behavior (Todorov, Mandisodza, Goren, & Hall, 2005) and sentencing judgments (Blair, Judd, & Chapleau, 2004). While several brain regions have been found to represent others' attractiveness (O'Doherty et al., 2003), emotion (Wegrzyn et al., 2015) and trustworthiness (Todorov, Baron, & Oosterhof, 2008; Winston, Strange, O'Doherty, & Dolan, 2002) as perceived from their faces, much less is known about how these perceptions are constructed from low-level visual inputs. Another line of studies has examined neural mechanisms underlying perception of animacy or biological motion (Giese & Poggio, 2003; Schultz & Bülthoff, 2019). Such studies have implicated a brain network including pSTS in detection of animacy of abstract stimuli (e.g., moving dots), but the underlying computations still remain elusive. Future studies could fruitfully provide neurocomputational accounts that bridge low-level sensory inputs and the higher-order perception of social stimuli (Chang & Tsao, 2017; Lin, Keles, & Adolphs, 2019; Oosterhof & Todorov, 2008).

To conclude, in this review, we discuss recent advances in the study of human social value-based decision-making. Despite the consensus that multiple types of computations underlie social decision-making, our understanding of how these computations are integrated to guide behavior is still in its infancy. Further elucidation of the integration process by combining neuroimaging, brain stimulation, computational modeling and connectivity analyses would be a critical step towards a more comprehensive understanding of human social decision-making.

Declaration of Competing Interest

The authors declare no competing financial interest.

Acknowledgments

This work was supported by the JSPS KAKENHI Grants JP17H05933 and JP17H06022 (S.S.) and the NIMH Caltech Conte Center for the Neurobiology of Social Decision Making (J.P.O).

REFERENCES

Alcalá-López, D., Smallwood, J., Jefferies, E., Overwalle, F., Vogeley, K., Mars, R. B., et al. (2017). Computing the social brain connectome across systems and states. *Cerebral Cortex, 28*(7), 2207–2232. https://doi.org/10.1093/cercor/bhx121.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), nrn1884. https://doi.org/10.1038/nrn1884.

Apps, M., Lesage, E., & Ramnani, N. (2015). Vicarious reinforcement learning signals when Instructing others. *The Journal of Neuroscience, 35*(7), 2904–2913. https://doi.org/10.1523/jneurosci.3669-14.2015.

Apps, M., Rushworth, M., & Chang, S. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron, 90*(4), 692–707. https://doi.org/10.1016/j.neuron.2016.04.018.

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent Homologies in action control: Corticostriatal Determinants of goal-directed and habitual action. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 35*(1), 48–69. https://doi.org/10.1038/npp.2009.131.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *Science, 324*(5931), 1160–1164. https://doi.org/10.1126/science.1169694.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature, 456*(7219), 245. https://doi.org/10.1038/nature07538.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221. https://doi.org/10.1038/nn1954.

Blair, I., Judd, C., & Chapleau, K. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10), 674–679. https://doi.org/10.1111/j.0956-7976.2004.00739.x.

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.

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. https://doi.org/10.1016/j.tics.2004.10.003.

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. https://doi.org/10.1073/pnas.1003111107.

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

Camerer, C., Ho, H. T., & Chong, K. J. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics, 119*(3), 861–898. https://doi.org/10.1162/0033553041502225.

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. https://doi.org/10.1093/scan/nsr094.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell, 169*(6). https://doi.org/10.1016/j.cell.2017.05.011, 1013–1028.e14.

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience, 13*(6), 637–647. https://doi.org/10.1080/17470919.2018.1518834.

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *The Journal of Neuroscience, 29*(39), 12315–12320. https://doi.org/10.1523/jneurosci.2575-09.2009.

Christopoulos, G. I., & King-Casas, B. (2015). With you or against you: Social orientation dependent learning signals guide actions made for others. *NeuroImage, 104*, 326–335. https://doi.org/10.1016/j.neuroimage.2014.09.011.

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2011). 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.

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. 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*(2), 199–204. https://doi.org/10.1016/j.conb.2006.03.006.

Delgado, M., Frank, R., & Phelps, E. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience, 8*(11), 1611–1618. https://doi.org/10.1038/nn1575.

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.

Edelson, M. G., Polania, R., Ruff, C. C., Fehr, E., & Hare, T. A. (2018). Computational and neurobiological foundations of leadership decisions. *Science, 361*(6401). https://doi.org/10.1126/science.aat0036. eaat0036.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry, 41*(1), 20–26. https://doi.org/10.1093/ei/41.1.20.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *The Journal of Neuroscience, 35*(21), 8170–8180. https://doi.org/10.1523/jneurosci.4775-14.2015.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences, 11*(10), 419–427. https://doi.org/10.1016/j.tics.2007.09.002.

Fehr, E., & Shmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868. https://doi.org/10.1162/003355399556151.

Fletcher, G. J., Simpson, J. A., Thomas, G., & Giles, L. (1999). Ideals in intimate relationships. *Journal of Personality and Social Psychology, 76*(1), 72–89. https://doi.org/10.1037/0022-3514.76.1.72.

Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., & Dolan, R. (1997). Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage, 6*(3), 218–229. https://doi.org/10.1006/nimg.1997.0291.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage, 19*(4), 1273–1302. https://doi.org/10.1016/s1053-8119(03)00202-7.

Fukuda, H., Ma, N., Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., et al. (2019). Computing social value conversion in the human brain. *Journal of Neuroscience*, 3117–3118. https://doi.org/10.1523/jneurosci.3117-18.2019.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179–192. https://doi.org/10.1038/nrn1057.

Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: The consilience of brain and decision. *Science, 306*(5695), 447–452. https://doi.org/10.1126/science.1102566.

Hackel, L. M., & Amodio, D. M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology, 24*, 92–97. https://doi.org/10.1016/j.copsyc.2018.09.001 (Psychol. Sci. 15 2004).

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience, 18*(9), 1233–1235. https://doi.org/10.1038/nn.4080.

Hackel, L. M., Zaki, J., & Bavel, J. J. (2017). Social identity shapes social valuation: Evidence from prosocial behavior and vicarious reward. *Social Cognitive and Affective Neuroscience, 12*(8), nsx045. https://doi.org/10.1093/scan/nsx045.

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. https://doi.org/10.1073/pnas.0711099105.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316*(5831), 1622–1625. https://doi.org/10.1126/science.1140738.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences of the United States of America, 108*(44), 18120–18125. https://doi.org/10.1073/pnas.1109322108.

Haruno, M., & Frith, C. D. (2009). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience, 13*(2), 160–161. https://doi.org/10.1038/nn.2468.

Haruno, M., & Kawato, M. (2009). Activity in the superior temporal sulcus highlights learning competence in an interaction game. *The Journal of Neuroscience, 29*(14), 4542–4547. https://doi.org/10.1523/jneurosci.2707-08.2009.

Hein, G., Engelmann, J. B., Vollberg, M. C., & Tobler, P. N. (2016a). How learning shapes the empathic brain. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 80–85. https://doi.org/10.1073/pnas.1514539112.

Hein, G., Morishima, Y., Leiberg, S., Sul, S., & Fehr, E. (2016b). The brain's functional network architecture reveals human motives. *Science, 351*(6277), 1074–1078. https://doi.org/10.1126/science.aac7992.

Hill, M. R., Boorman, E. D., & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nature Communications, 7*(1), 12722. https://doi.org/10.1038/ncomms12722.

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*(8), 1142–1149. https://doi.org/10.1038/nn.4602.

Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science, 320*(5879), 1092–1095. https://doi.org/10.1126/science.1153651.

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), 451–462. https://doi.org/10.1016/j.neuron.2015.06.031.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. (2017). Social learning through prediction error in the brain. *Npj – Science of Learning, 2*(1), 8. https://doi.org/10.1038/s41539-017-0009-2.

Jung, D., Sul, S., & Kim, H. (2013). Dissociable neural processes underlying risky decisions for self versus other. *Frontiers in Neuroscience, 7*, 15. https://doi.org/10.3389/fnins.2013.00015.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience, 10*(12), nn2007. https://doi.org/10.1038/nn2007.

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2018.04.009.

Lange, P. A., Bruin, E. M., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology, 73*(4), 733. https://doi.org/10.1037/0022-3514.73.4.733.

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation System in the human brain: Evidence from functional neuroimaging. *Neuron, 64*(3), 431–439. https://doi.org/10.1016/j.neuron.2009.09.040.

Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience, 11*(4), 404–409. https://doi.org/10.1038/nn2065.

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.

Liljeholm, M., Molloy, C. J., & O'Doherty, J. P. (2012). Dissociable brain systems mediate vicarious learning of stimulus–response and action–outcome contingencies. *The Journal of Neuroscience, 32*(29), 9878–9886. https://doi.org/10.1523/jneurosci.0548-12.2012.

Lin, C., Keles, U., & Adolphs, R. (2019). *Comprehensive trait attributions show that face impressions are organized in four dimensions*. https://doi.org/10.31234/osf.io/87nex.

Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., & Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proceedings of the National Academy of Sciences of the United States of America, 113*(35), 9763–9768. https://doi.org/10.1073/pnas.1603198113.

Lockwood, P. L., Klein-Flügge, M., Abdurahman, A., & Crockett, M. J. (2019). Neural signatures of model-free learning when avoiding harm to self and other. *BioRxiv*, 718106. https://doi.org/10.1101/718106.

Lockwood, P. L., Wittmann, M. K., Apps, M. A., Klein-Flügge, M. C., Crockett, M. J., Humphreys, G. W., et al. (2018). Neural mechanisms for learning self and other ownership. *Nature Communications, 9*(1), 4747. https://doi.org/10.1038/s41467-018-07231-9.

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. https://doi.org/10.1016/s0896-6273(03)00154-5.

Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage, 112*, 244–253. https://doi.org/10.1016/j.neuroimage.2014.12.056 (Ann. N. Y. Acad. Sci. 1191 2010).

Nicolle, A., Klein-Flügge, M. C., Hunt, L. T., Vlaev, I., Dolan, R. J., & Behrens, T. (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.

Ogawa, A., Ueshima, A., Inukai, K., & Kameda, T. (2018). Deciding for others as a neutral party recruits risk-neutral perspective-taking: Model-based behavioral and fMRI experiments. *Scientific Reports, 8*(1), 12857. https://doi.org/10.1038/s41598-018-31308-6.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience, 2*(1), 3–11. https://doi.org/10.1093/scan/nsm005.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102. https://doi.org/10.1038/nn1968.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105.

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. https://doi.org/10.1126/science.1094285.

O'Doherty, J., Hampton, A., & 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.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D., & Dolan, R. (2003). Beauty in a smile: The role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41*(2), 147–155. https://doi.org/10.1016/s0028-3932(02)00145-8.

Rangel, A., Camerer, C., & Montague, R. P. (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.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420.

Ray, D., King-Casas, B., Montague, R. P., & Dayan, P. (2008). *Bayesian model of behaviour in economic games*.

Roskies, A. L. (1999). The binding problem. *Neuron, 24*(1), 7–9. https://doi.org/10.1016/s0896-6273(00)80817-x.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8). https://doi.org/10.1038/nrn3776. nrn3776.

Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *The Journal of Neuroscience, 30*(40), 13525–13536. https://doi.org/10.1523/jneurosci.1747-10.2010.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755–1758. https://doi.org/10.1126/science.1082976.

Schultz, J., & Bülthoff, H. H. (2019). Perceiving animacy purely from visual motion cues involves intraparietal sulcus. *NeuroImage, 197*, 120–132. https://doi.org/10.1016/j.neuroimage.2019.04.058.

Schultz, W., Dayan, P., & Montague, R. P. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593.

Smith, D. V., Clithero, J. A., Boltuck, S. E., & Huettel, S. A. (2014). Functional connectivity with ventromedial prefrontal cortex reflects subjective value for social rewards. *Social Cognitive and Affective Neuroscience, 9*(12), 2017–2025. https://doi.org/10.1093/scan/nsu005.

Stanley, D. A. (2016). Getting to know you: General and specific neural computations for learning about people. *Social Cognitive and Affective Neuroscience, 11*(4), 525–536. https://doi.org/10.1093/scan/nsv145.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., et al. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences of the United States of America, 112*(5), 1619–1624. https://doi.org/10.1073/pnas.1414715112.

Sul, S., Tobler, P. N., Hein, G., Leiberg, S., Jung, D., Fehr, E., et al. (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. https://doi.org/10.1073/pnas.1423895112.

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., Cross, L., & O'Doherty, J. P. (2017). Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nature Neuroscience, 20*(12), 1780–1786. https://doi.org/10.1038/s41593-017-0008-x.

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

Takahashi, H., Kato, M., Matsuura, M., Mobbs, D., Suhara, T., & Okubo, Y. (2009). When your gain is my pain and your pain is my gain: Neural correlates of envy and schadenfreude. *Science, 323*(5916), 937–939. https://doi.org/10.1126/science.1165604.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience, 3*(2), 119–127. https://doi.org/10.1093/scan/nsn009.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589.

van den Bos, W., Talwar, A., & McClure, S. M. (2013). Neural correlates of reinforcement learning and social preferences in competitive bidding. *The Journal of Neuroscience, 33*(5), 2137–2146. https://doi.org/10.1523/jneurosci.3095-12.2013.

Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S., et al. (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex: A Journal Devoted To the Study of the Nervous System and Behavior, 69*, 131–140. https://doi.org/10.1016/j.cortex.2015.05.003.

Will, G.-J., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *ELife, 6*, e28098. https://doi.org/10.7554/elife.28098.

Winston, J. S., Strange, O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5*(3), 277–283. https://doi.org/10.1038/nn816.

Wittmann, M. K., Kolling, N., Faber, N. S., Scholl, J., Nelissen, N., & Rushworth, M. (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.

Wittmann, M. K., Lockwood, P. L., & Rushworth, M. F. (2018). Neural mechanisms of social cognition in primates. *Annual Review of Neuroscience, 41*(1), 99–118. https://doi.org/10.1146/annurev-neuro-080317-061450.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, R. P. (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.

Zentall, T. R. (2012). Perspectives on observational learning in animals. *Journal of Comparative Psychology, 126*(2), 114. https://doi.org/10.1037/a0025381.

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. https://doi.org/10.1073/pnas.1116783109.