The Neuroscientist
http://nro.sagepub.com/

# The Cooperative Brain

Mirre Stallen and Alan G. Sanfey

*Neuroscientist* 2013 19: 292 originally published online 8 January 2013
DOI: 10.1177/1073858412469728

The Neuroscientist
19(3) 292–303

DOI: 10.1177/1073858412469728
nro.sagepub.com
SAGE

# The Cooperative Brain

**Mirre Stallen[1,2] and Alan G. Sanfey[2,3]**

## Abstract

Cooperation is essential for the functioning of human societies. To better understand how cooperation both succeeds and fails, recent research in cognitive neuroscience has begun to explore novel paradigms to examine how cooperative mechanisms may be encoded in the brain. By combining functional neuroimaging techniques with simple but realistic tasks adapted from experimental economics, this approach allows for the discrimination and modeling of processes that are important in cooperative behavior. Here, we review evidence demonstrating that many of the processes underlying cooperation overlap with rather fundamental brain mechanisms, such as, for example, those involved in reward, punishment and learning. In addition, we review how social expectations induced by an interactive context and the experience of social emotions may influence cooperation and its associated underlying neural circuitry, and we describe factors that appear important for generating cooperation, such as the provision of incentives. These findings illustrate how cognitive neuroscience can contribute to the development of more accurate, brain-based, models of cooperative decision making.

## Keywords

neuroscience, cooperation, game theory, fMRI, decision making

*Picture yourself in a rural village surrounded by meadows open to herdsmen to graze their cows. It is in each herder's interest to put every new cow he acquires onto the land, even if this means that the pasture will be damaged by overgrazing in the long run. After all, more cows means more income for the herdsman and the disadvantage of less food per cow is spread among all the other herdsmen. But therein is the problem. If all herdsmen act this way, the meadows will be depleted to the detriment of all.*

The above scenario, as described by Hardin (1968), has become famous as a parable outlining the inherent conflict between self-interest and cooperation. Cooperative actions by individuals help the collective, but a selfish individual can benefit even more by not cooperating and instead pursuing his or own private interests. Modern examples of "the tragedy of the commons" can be found in countless domains, from relatively trivial instances such as littering, vandalism, and illegal downloading of media, to more consequential situations such as the use and overuse of environmental resources. For instance, oceans are not owned by individuals and provide a common resource of fish. However, as the amount of fishing increases year after year, the fish population loses its ability to restore itself, resulting in an overall diminution of the fish stock. So while the individuals who overfish benefit from greater supply, the collective suffers from a depleted resource. Similarly, the release of carbon dioxide has resulted in high concentrations of greenhouse gases that are harmful for everyone in the long term. Therefore, to prevent depletion of common resources, such as oceans or clean air, and to ensure the availability of public goods, such as medical care or inexpensive music on the Internet, cooperation is required.

A growing number of studies in both the field and in the laboratory demonstrate that people are imperfect cooperators—they tend to cooperate only if others do so, and there is a substantial minority of people who never cooperate, instead "free-riding" at the cost of others (Fischbacher and others 2001; Fig. 1). This suboptimal pattern of behavior causes unstable cooperation levels and often results in the disappearance of positive collective action over time. Thus, people must often be persuaded to sacrifice self-interest for the collective benefit. But how is cooperation induced? What processes are

[1] Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, Netherlands
[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands
[3] Behavioral Science Institute, Radboud University Nijmegen, Nijmegen, Netherlands

**Corresponding Author:**
Mirre Stallen, Erasmus University Rotterdam, Kapittelweg 29, 6525 EN Nijmegen, Netherlands
Email: m.stallen@donders.ru.nl

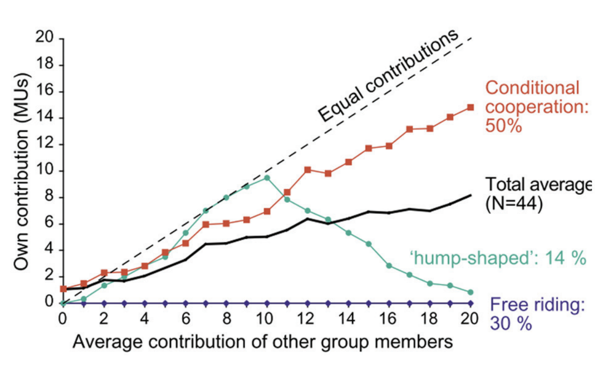

**Figure 1.** People's willingness to contribute to a public good is typically conditional on the average contribution of others. Here, 44 individuals interacted anonymously with each other (Fischbacher and others, 2001), with participants raising their contributions to the public good only if the average contribution of the others increased (conditional cooperation: 50%, red). However, about one third of the participants never contributed anything (free-riding: 30%, purple), or only contributed if the average contribution of others was low (hump-shaped: 14%, green). The remaining 6% of the participants in the experiment exhibited irregular contribution behavior. Reprinted from Fehr and Fischbacher (2004), Trends in Cognitive Sciences, 8(4), 185–90, Copyright 2004, with permission from Elsevier.

important in encouraging cooperation within a group? In this review, we will outline the basic processes that underlie cooperation, such as reward and learning mechanisms, and discuss what is known about the neural bases of these processes. In addition, we review how expectations induced by either the social context or social emotions may influence cooperation, also noting the associated underlying neural circuitry, and we describe factors important for generating cooperation, such as the provision of incentives. We conclude by discussing how a broad perspective in understanding the mechanisms of cooperation may help in developing more effective ways of promoting cooperative behavior. Understanding at a fundamental level how cooperation both succeeds and fails can provide valuable clues as to how interventions could be structured to maximize cooperative interactions in important social policy contexts.

Recent laboratory research in cognitive neuroscience has begun to explore novel paradigms that offer fruitful avenues to examine how cooperative processes may be encoded in the brain. Most of this research is embedded within the field of neuroeconomics/decision neuroscience, an interdisciplinary effort to better understand the fundamentals of human decision making. Within this field, researchers are building models of decision making that incorporate both the psychological processes that influence decisions, how these processes are constrained by the underlying neurobiology, and also developing formal models of these decisions, an approach developed from economics.

## Game Theory

An early effort to formally model how cooperation and non-cooperation can occur emerged from game theory (von Neumann and Morgenstern 1947), a collection of rigorous models attempting to understand and explain situations in which decision makers must interact with one another, such as bidding in auctions and salary negotiations. Consequently, these models were applied to large-scale social scenarios, in particular strategic decision making during the Cold War era. For example, theorists were influential in applying formal game theoretic principles to the Vietnam War (Schelling 1960). However, a fundamental flaw of this approach, as was painfully evident from efforts to formulate policy based on these theoretical principles, is that actual observed decision behavior typically deviates, often quite substantially, from the predictions of the model. Ample research has shown that players typically do not play according to the purely self-interested strategies predicted by classical game theory (Camerer 2003). In reality, decision makers are influenced by a wide range of psychological factors, which can enhance, though sometimes reduce, cooperative behavior. For example, people are typically both less selfish and more willing to consider factors such as reciprocity and equity than the classical model predicts. They care about status and social hierarchies, often seek vengeance, but are also affected by factors such as empathy and guilt. So, to develop policy principles that can accurately predict the development, and ideally, the enhancement of cooperation, the formal models require elaboration with detailed information regarding the psychological principles that guide decisions in social interactions.

The emergence of a neuroeconomic approach to examining interactive decision making offers real promise for the development of such models. This nascent research field combines psychological insight and brain imaging with realistic social tasks that allow for the exploration of cooperation in a controlled laboratory environment. In contrast to standard behavioral studies, the combination of game theoretic models with the online measurement of brain activity during decision making allows for the discrimination and modeling of processes that are hard to separate at the behavioral level. Within this neuroeconomic approach, tasks have been designed that ask people to decide about monetary divisions in an interactive setting, with money used both as a reward in itself and also as a proxy for other "rights" that affect cooperation

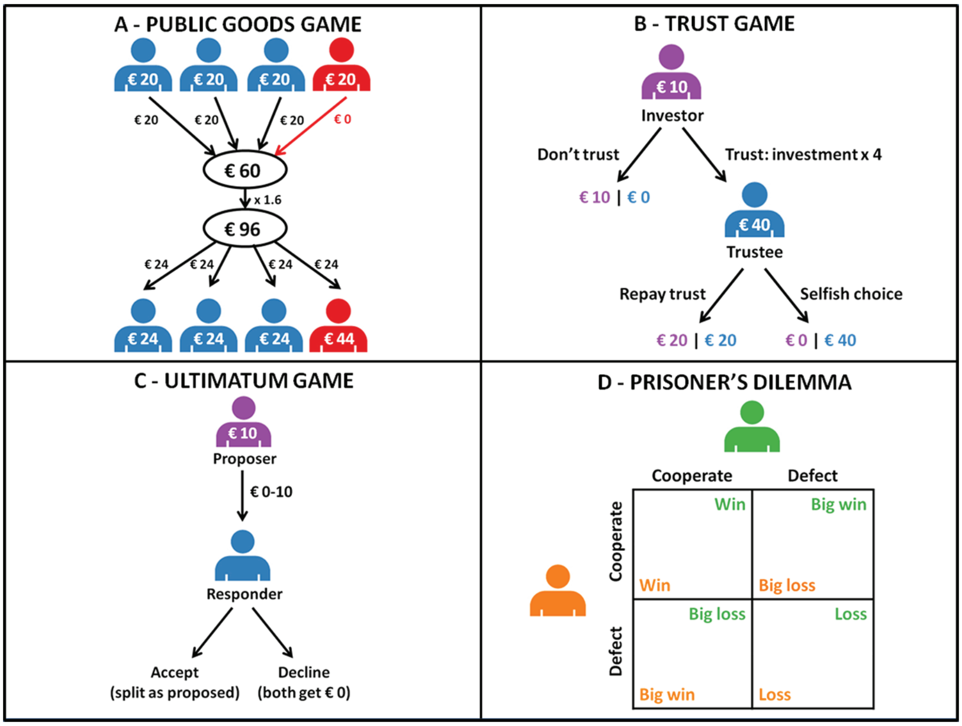

**Figure 2.** Outline of game theoretic tasks that are most commonly used to examine motivations involved in cooperative behavior. (A) Public Goods Game. Four players are provided with equal monetary endowments. Each individual decides how much of this endowment to contribute to the public pot. The experimenter multiplies all contributions by a factor of 1.6 and the result is equally divided among all players. Note that the player on the right is a "free-rider." This player enjoys the benefit of the group donations, while not contributing anything. (B) Trust Game. One player, the Investor, decides how much of his endowment to invest with a partner, the Trustee. The experimenter multiplies the transferred investment by a factor of 4. The Trustee has the option to return some of the final investment amount to the Investor but does not need to. (C) Ultimatum Game. One player, the Proposer, specifies how to divide a sum of money. The other player, the Responder, then has the option of accepting or rejecting this offer. If the offer is accepted, the sum is divided as proposed. If it is rejected, neither player receives anything. (D) Prisoner's Dilemma Game. Two players simultaneously choose to either cooperate or defect with the other. The largest payoff to a player occurs when they defect and their partner cooperates, while conversely, the worst outcome occurs when they cooperate and their partner defects. Mutual cooperation yields a modest payoff to both players, with mutual defection providing a still lower amount to each.

(land, political power, etc.). These tasks are well suited to be used in combination with brain imaging methods and produce a surprisingly rich pattern of decision making, allowing a wide range of questions to be answered about motivations to engage in cooperative behavior.

The games used in these experiments are generally simple and offer compelling social scenarios (Fig. 2). The Public Goods Game (PGG; Fehr and Gächter 2000) is the most commonly used game to study cooperation. In this game, four participants at a time are provided with a monetary endowment, and each individual then decides how much of this endowment they wish to keep for themselves and how much they want to contribute to a public pot. The experimenter multiplies the total contributions in the pot by a factor (typically 1.6), and this "public good" is then distributed equally among all players, irrespective of their contribution. Additionally, each participant retains the part of their endowment that was not shared. After all participants have indicated their decisions, outcomes are revealed and a new round starts. In a similar fashion to societal public goods such as clean air or medical care, the defining characteristic of a public good in the PGG is that all participants consume an equal share of the good, even those who did not bear the cost of providing the good. So, while the group as a whole is best off if all participants contribute equally, each individual has a competing incentive to free-ride, that is, to contribute nothing to the good, and the PGG nicely captures this conflict between self-interest and group cooperation in a controlled laboratory setting.

Other economic games are also useful in examining different aspects of cooperation. In the Trust Game (TG; Berg and others 1995), a player (the Investor) decides how much of an endowment to invest with a partner (the Trustee). Once transferred, the experimenter multiplies this money by a factor of 4. Then, the Trustee has the opportunity to return some of this increased pot of money back to the Investor, but, importantly, need not return anything. If the Trustee honors trust and returns money, both players usually end up with a higher monetary payoff than originally endowed. However, if the Trustee abuses trust and keeps the entire amount, the Investor takes a loss. As the Investor and Trustee interact only once during the one-shot version of the game, game theory predicts that a rational Trustee will never honor the trust given by the Investor. The Investor, realizing this, should never place trust in the first place, and so will invest zero in the transaction. Despite these grim theoretical predictions, a majority of investors do in fact send some amount of money to the Trustee, and, again contrary to predictions, this trust is generally reciprocated.

The well-studied Prisoner's Dilemma Game (PDG; Poundstone 1992) is similar to TG, except that both players now simultaneously choose whether or not to trust each other, without knowledge of their partner's choice. In PDG, each player chooses to either cooperate or not with their opponent, with their payoff dependent on the interaction of the two choices. The largest payoff to the player occurs when he or she defects and their partner cooperates, with the worst outcome when the decisions are reversed. Mutual cooperation yields a modest payoff to both players, whereas mutual defection provides a lesser amount to each. The classical game theoretic prediction for the PDG is mutual defection, which, interestingly, is a worse outcome for both players than mutual cooperation, but again, in most iterations of the game players exhibit much more trust than expected, with mutual cooperation occurring about 50% of the time. These latter two games model two-person situations in which players must decide to what degree they can increase their payoff by relying on a cooperative partner.

Finally, the Ultimatum Game (UG; Güth and others 1982) is often used to examine responses to fairness. Here, two players must divide a sum of money, with the Proposer specifying the division. The Responder then has the option of accepting or rejecting this offer. If the offer is accepted, the sum is divided as proposed. If it is rejected, neither player receives anything. The UG therefore models decisions about resource allocation on the part of the proposer and responses to fairness and inequity by the responder. If people are motivated purely by self-interest, the responder should accept any offer, and, knowing this, the proposer will offer the smallest nonzero amount. However, once again, this game theoretic prediction is at odds with observed behavior across a wide range of societies, with rejections of unequal offers standardly observed. Thus, people's choices in the UG do not conform to a model in which decisions are driven by financial self-interest.

While using these tasks, researchers have also employed a variety of neuroscientific methods to investigate the respective underlying brain systems, including functional neuroimaging, the study of brain-damaged neurological patients, transcranial magnetic stimulation, pharmacologic manipulations, genetic association studies, and studies of psychiatric patients, as well as lesion and single-cell recording studies in non-human primates. Here, we will focus on how brain imaging studies, in particular those using functional magnetic resonance imaging (fMRI), can yield insights into the processes underlying cooperation. The following sections will provide an overview of how external incentives may motivate cooperation, and examine the basic neural mechanisms underlying cooperation, in particular the role of reward and learning. Thereafter we review some more recent work which demonstrates that social context and social emotions have important roles to play in determining how and when we cooperate, and we outline the possible neural pathways via which these factors affect cooperation.

## Incentives

To encourage cooperation in social dilemma situations, and of course to reduce the likelihood of free-riding, authorities frequently reward cooperators (e.g., by providing awards or tax benefits) or punish non-cooperators (e.g., by levying fines or supplementary taxes). Indeed, experimental evidence has shown that incentives are quite effective in promoting increases in cooperation. For instance, a large amount of studies have demonstrated that stable cooperation levels are rarely attained in a PGG. In the standard version of this game, there is usually substantial cooperation across the initial rounds of the game, but over time cooperation drops, and by the final few rounds cooperation is typically at a very low level (Fehr and Fischbacher 2004). This is generally attributed to the diminishing "shadow of the future" in these later rounds, that is, the lowered likelihood of negative future consequences for non-cooperation. However, the addition of a punishment or reward mechanism to the standard game increases cooperation considerably (Balliet and others 2011). In PGGs with incentive options, participants are provided with the opportunity to either punish or reward the other players in each round. Whether a punishment or a reward is administered depends on the specific experimental manipulation, but

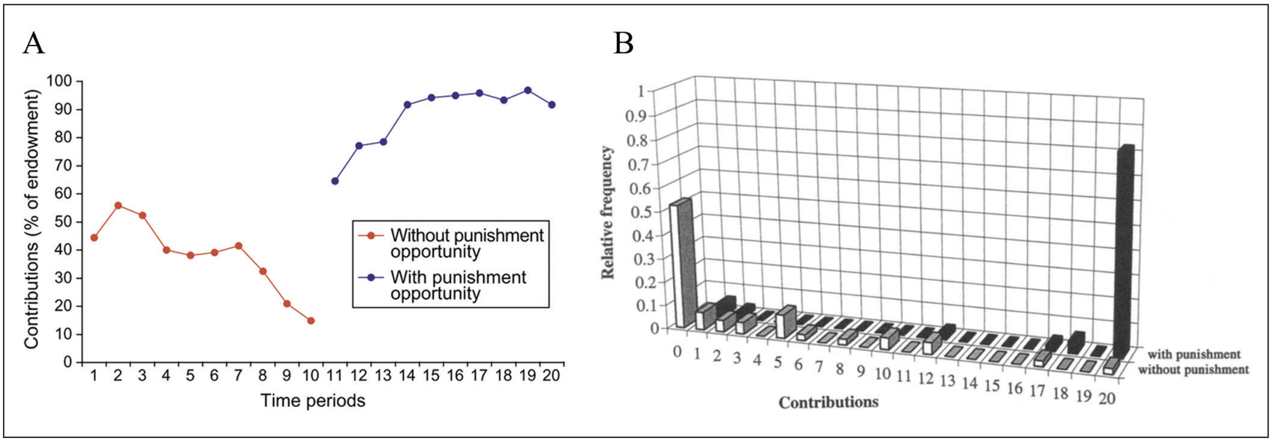

**Figure 3.** Fehr and Gächter (2000) demonstrated that the addition of a punishment option to the Public Goods Game considerably increased cooperation levels. (A) In the absence of a punishment opportunity (periods 1–10) cooperation levels dropped. However, after the introduction of a punishment mechanism there was an immediate increase in cooperation (period 11). During the remaining rounds, cooperation increased, until there was almost 100% cooperation in the final period. (B) Distribution of the average contributions in the final period of a Public Goods Game played by the same participants, with and without a punishment option. In the punishment condition, 82.5% of the participants cooperate, contributing their entire endowment. In the no-punishment condition, 53% of the participants free-ride in the final period. Figure 3A is reprinted from Fehr and Fischbacher (2004), Trends in Cognitive Sciences, 8(4), 185–90, Copyright 2004, with permission from Elsevier. Figure 3B is reprinted from Fehr and Gächter (2000), The American Economic Review, 90(4), 980–94, Copyright 2000, with permission from the American Economic Association.

the relevant award is usually made immediately after being informed about the group's contributions on that round. In these experiments, punishments and rewards are typically dispensed anonymously, and, importantly, are costly to the participant, as well as having a real effect on the target player. So, every monetary unit spent to punish (reward) decreases (increases) the income of the targeted player by 2 to 4 units. When a reward or a punishment can potentially be meted out, cooperation generally does not decrease, and may even increase over time. Moreover, full cooperation levels are commonly observed even in the concluding rounds of the game (Fehr and Gächter 2000; Fig. 3). Although there are fewer studies on rewards than punishments, both types of incentives appear to be equally effective in inducing cooperation (Balliet and others 2011).

Surprisingly, though there is now a substantial body of research demonstrating clear effects of incentive mechanisms on cooperative behavior, relatively little is known to date about how the brain may encode this type of cooperation. One study has examined the neural systems involved in incentive mechanisms though in a slightly different context, that of fairness norms (Spitzer and others 2007). Participants played a variant of the UG known as the Dictator Game with punishment and non-punishment conditions. As expected, participants transferred more units to their partner in the punishment condition than in the non-punishment condition, indicating that participants complied more with the norm of fairness under the threat of punishment. Brain areas in which activations were observed in the punishment condition were the dorsolateral prefrontal cortex (DLPFC), orbitolateralfrontal cortex (OLFC), and caudate (see Fig. 4 for an overview of all brain areas referred to in this review). The authors suggest these findings reflect the involvement of processes that evaluate social threat and implement cognitive control in cooperative decision making under potential punishment. These findings provide an initial glimpse into the neural systems that may be involved in changes in social decision making under incentive conditions. However, whether this mechanism underlies the effect of external incentives in general, including the effect of rewards, remains an open question, and future research could usefully examine the neural mechanisms via which incentives modify cooperative behavior.

## Reward

As described in the preceding section, there is limited knowledge about how the brain encodes external motivations of cooperation; however, there is a growing literature on how cooperation is internally motivated, and the associated neural mechanisms. One of the most consistent findings across studies on the neural mechanisms of

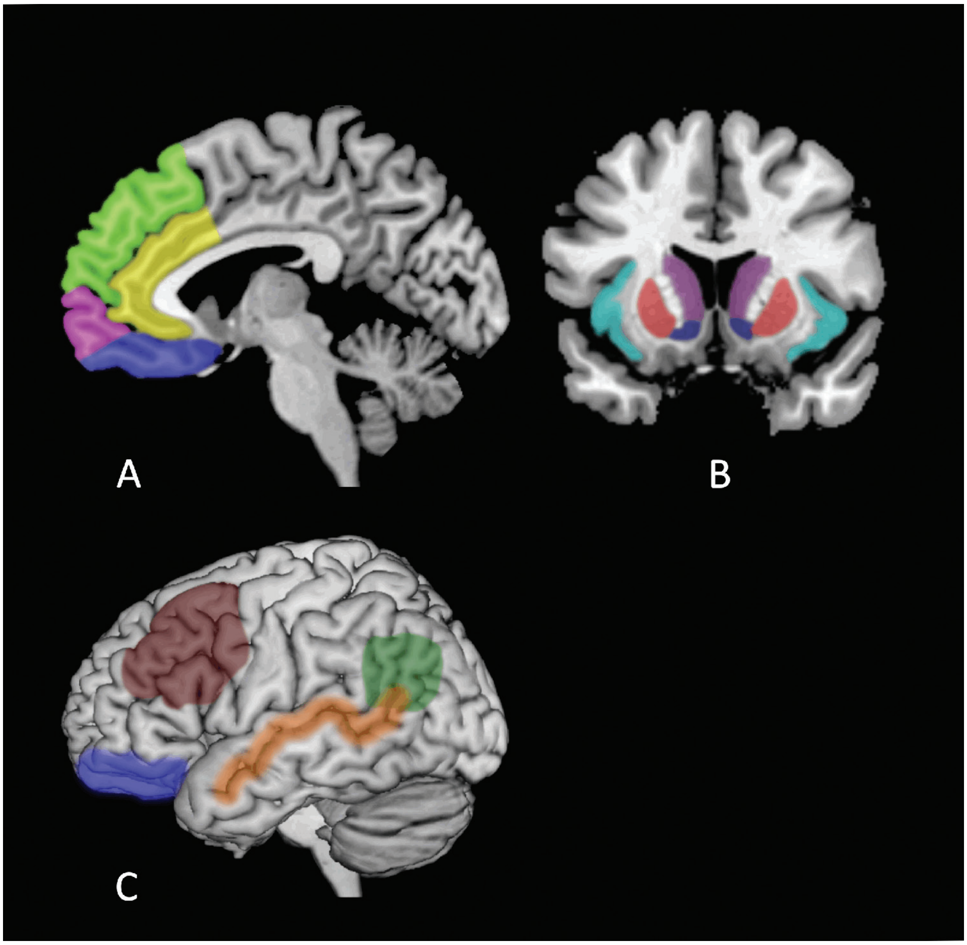

**Figure 4.** Overview of brain areas associated with cooperation. Colors indicate schematic locations. (A) Green, dorsal medial prefrontal cortex (DMPFC); yellow, anterior cingulate cortex (ACC); pink, ventromedial prefrontal cortex (VMPFC); blue, orbital frontal cortex (OFC). MNI coordinate: $x = -5$. (B) Turquoise, insula; red, putamen; purple, caudate; dark blue, nucleus accumbens (NACC). The striatum comprises the putamen, caudate, and NACC. MNI coordinate: $y = 13$. (C) Blue, OFC; brown, dorsolateral prefrontal cortex (DLPFC); orange, superior temporal sulcus (STS); dark green, temporal parietal junction (TPJ).

cooperative behavior is that cooperation is highly associated with activation in brain areas known to be involved in reward-based learning (Decety and others 2004; Rilling and others 2002; Rilling and others 2004a). For instance, these studies have shown that reciprocated cooperation in the PDG and TG is associated with activity in the ventral striatum and ventromedial prefrontal cortex (VMPFC), brain regions that have been consistently found to be activated by both social and monetary rewards (McClure and others 2004). Relatedly, viewing the faces of individuals who had previously cooperated in a PDG, as compared to faces with whom the player had no history, elicited enhanced neural activity in the reward-related areas such as the striatum, nucleus accumbens, and orbitofrontal cortex (Singer and others 2004). Though there are obvious dangers in making inferences about the cognitive processes reflected by activation in specific brain regions (Poldrack 2006), these findings suggest that by labeling mutual cooperation as rewarding in and of itself, that is, independent of whatever monetary gain was obtained by the cooperative action, people are motivated to resist the temptation to selfishly accept but not reciprocate favors. For example, a PDG study (Rilling and others 2004a) demonstrated increased ventral striatum and ventromedial prefrontal activity for mutual cooperation decisions, even when controlling for the amount of the money earned by the decision itself. Indeed, when contrasting play with either a human or a computer partner in the PDG, it was found that activation in these regions was strongest when participants interacted with another human, even when the two types of partners played identical strategies. These studies provide further support for the hypothesis that cooperation with other people is inherently rewarding, with this interpretation also in line with theories from both evolutionary psychology and developmental science that argue that cooperation is rewarding per se, and that, although the material payoff from cooperation may be delivered at a later remove, the psychological reward seems to be immediate.

## Learning

When interacting with others who frequently reciprocate one's cooperative behavior, people typically continue a pattern of cooperation, whereas interacting with non-cooperators decreases cooperation markedly (Fehr and Fischbacher 2004). This behavior is consistent with the notion that one of the best predictors of an individual's trustworthiness is their behavior in previous interactions (Axelrod and Hamilton 1981; King-Casas and others 2005), that is, we are of course more likely to invest trust in someone who has shown to be cooperative than to trust someone who has previously betrayed us. Indeed, experiments have demonstrated that repeated interactions with a partner influence participants' subjective ratings of this partner's character in both PDG (Singer and others 2004) and TGs (Delgado and others 2005), and, importantly, subsequently moderate the participants' investment behavior toward this partner (Chang and others 2010). These findings indicate that people learn about the cooperative nature of another player based on the history of that partner's behavior and that this social learning provides the basis for future cooperation (or non-cooperation).

Brain imaging work on the learning of cooperative behavior suggest that cooperation is facilitated by a mechanism similar to reward-dependent learning based on the computation of basic reward prediction errors (Rilling and others 2004a). Reward prediction errors are signals reflecting the discrepancy between the predicted probability of a reward and its actual outcome (Fiorillo and others 2003). Changes in neural activity related to reward prediction errors are thought to be critical for reinforcement learning (Schultz and others 1997), and may motivate behavioral change, such that, over time, behavior that is more rewarding than predicted will be adopted more easily, whereas behavior that is less rewarding than predicted will be reduced (Tricomi and others 2004). Indeed, neuroimaging data from the PDG showed that mutual cooperation in this game was associated with a positive blood oxygen level–dependent (BOLD) response in reward-processing areas, whereas non-reciprocated cooperation was associated with a negative BOLD response, suggesting that reciprocated cooperation involves a positive reward prediction error (i.e., an outcome is more rewarding than expected), and non-reciprocated cooperation involves a negative reward prediction error (i.e., an outcome is less rewarding than expected; Rilling and others 2004a). Additionally, similar reinforcement signals have been found to positively reinforce cooperation in the TG (King-Casas and others 2005). In this study, an iterated version of the TG was used in which homologous regions of two participants' brains were scanned simultaneously. Results showed that the head of the caudate nucleus received information about the fairness of the decision of the other, and encoded the intention to reciprocate the other's trusting decision. In line with the idea that people build a model of their partner's behavior to predict the other's next move, this temporal transfer of the neural signal correlated with future increases in trust, and activity in the caudate decreased over time as feedback from one's partner became more reliable. This shift of activity in the caudate suggests that this area keeps track of the reputation of one's partner by a mechanism that resembles reinforcement learning, showing that people learn about the cooperativeness of their partner over time (King-Casas and others 2005). Taken together, these findings indicate that the modulation of the BOLD response in dopaminergic regions during (non-)cooperative interactions reflects the learning of who will and will not reciprocate our trust, thereby helping us to decide with whom to cooperate and whom to avoid.

## Social Context

In everyday situations, individuals' willingness to cooperate is not only based on actual interactions with others but may also be influenced by additional information that is gathered from the specific social context. For instance, research has demonstrated that trustworthiness judgments are influenced by factors other than direct experience with a partner. In a TG, more trust was placed in partners who were described as having a praiseworthy character than in partners with a neutral or bad moral character (Delgado and others 2005). Also, participants typically invest more money when partners look trustworthy, with this trustworthiness assessed by an independent group of raters, indicating that participants believed that certain facial cues were predictive of the reciprocation of trust (Chang and others 2010; van't Wout and Sanfey 2008; Fig. 5). These findings suggest that initial impressions may function as a risk signal, which in turn influences the amount of money an individual expects to

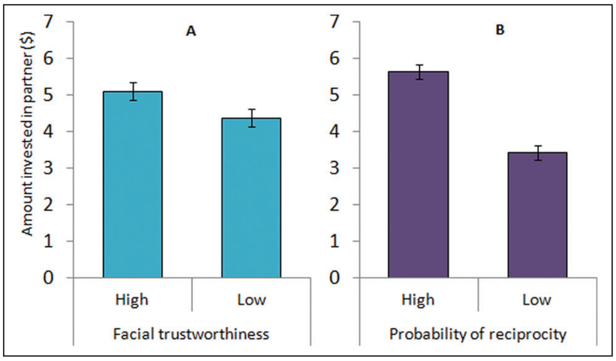

**Figure 5.** Chang and others (2010) studied the impact of both facial trustworthiness and previous direct experience on the amount of trust people place in a partner. In this experiment, participants played a repeated Trust Game in the role of Investor, while both the facial trustworthiness (high or low) and the probability of reciprocation (high: 80% or low: 20%) of the partner was systematically manipulated. Error bars indicate standard error of the mean. (A) In the first trial of the Trust Game, participants invested more money partners who looked more trustworthy as compared with those who looked less trustworthy. (B) Participants overall gave more money to partners who reciprocated 80% of the time as compared with participants who reciprocated 20% of the time.

be sent back. Importantly, however, implicit judgments, such as those derived from personality or facial information, actually interact with experienced trustworthiness. When repeatedly interacting with the same partners in a TG, both facial expressions and actual game experience influenced participants' behavior, so that partners who were initially judged as most trustworthy, and who actually turned out to be most trustworthy, were entrusted with most money (Chang and others 2010).

A possible mechanism via which implicit and explicit social signals may influence cooperative decision making is via the development of context-specific expectations. That is, based on initial impressions or previous experience, people may develop expectations about the trustworthiness of their partner, and in turn use these context-specific expectations as a behavioral reference point (Chang and others 2010). For instance, prior information about partners' personality traits influences learning about the trustworthiness of another, with this reflected in reduced caudate activation when learning about the game behavior of previously described morally good and bad partners (Delgado and others 2005). Similarly, expectations may reflect a social norm about what other people would do in a given situation. For instance, rejection rates of unfair offers in the UG increase when participants are provided with information about how other people respond (Bohnet and Zeckhauser 2004), but decrease when participants believe an unfair offer is "typical" (Sanfey 2009). Additional support for the role of expectations in everyday decision making comes from a study in which a computational model of expectations is developed and used to identify the brain networks involved in the tracking of social expectation violations (Chang and Sanfey 2011). Here, participants played a UG in the role of the responder while undergoing fMRI. Prior to scanning, expectations were elicited by asking participants to report their beliefs about the kind of offers they expected to encounter. Results demonstrated that the anterior cingulate cortex (ACC) was integral in tracking the violations of these expectations. The ACC has previously been associated with other processes related to the detection of expectation violations such as the signaling of social norm deviation (Klucharev and others 2009), the weighting of social prediction errors (Behrens and others 2008), and responses to unfairness in the UG (Sanfey and others 2003), suggesting that this area plays a critical role in the calculation of conflict between individual preferences and social norms. These findings are in accordance with the proposal that the ACC is involved in the processing of both negative affect and cognitive control (Shackman and others 2011) and indicates that people generate a specific neural signal when others violate their expectations, which in turn may serve as an emotional indicator motivating people to enforce a social norm.

## Social Emotions

Both psychological and neuroscience data have extensively demonstrated that emotions play an important role in decision making (Loewenstein and Lerner 2003; Phelps 2009). Surprisingly, however, there is little experimental research examining what specific emotions are recruited in cooperative decisions. One potential emotion underlying the decision to cooperate is the anticipation of guilt. That is, a motivation for cooperating is that we would expect to feel guilty if we did not reciprocate generous behavior. Initial evidence for this guilt hypothesis comes from studies examining the social behavior of patients with damage to the VMPFC, who are impaired in decision making, learning, and planning (e.g., Damasio, 2003; Koenigs and others 2007; Koenigs and Tranel, 2007). Based on these studies, one hypothesis is that damage to the VMPFC impairs concern for other people and that the abnormal behavior following damage to this region may, at least in part, result from an inability to experience guilt. To address this hypothesis, Krajbich and others (2009) used a formal economic model incorporating measures of envy and guilt to analyze the behavior of VMPFC patients. They found that VMPFC patients were less trustworthy and transferred less money in UG, TG, and Dictator game, than healthy control participants.

Moreover, the model showed that VMPFC patients are relatively insensitive to guilt, thereby demonstrating that the expression of guilt, and perhaps more generally the elicitation of imagined outcomes, plays an important role in cooperative decision making, and that the VMPFC may be central to the implementation of these processes.

To test whether the anticipation of guilt can indeed motivate cooperative behavior, researchers used a formal model of guilt aversion in conjunction with brain imaging data to identify the brain mechanisms that mediate cooperation while participants deciding whether or not to reciprocate trust in the TG (Chang and others 2011). In this model, the construct of guilt was formalized as the deviation between the player's belief about what their partner expects them to do and the amount of money this player actually returned, and posited that cooperation depends on the avoidance of the expected negative affective state associated with guilt. Results showed increased activity in the VMPFC, dorsolateral prefrontal cortex (DLFPC) and nucleus accumbens (NACC) when Trustees chose to abuse trust and maximize their gains. These findings are in line with previously mentioned patient work showing that the VMPFC plays an important role in the experience of guilt and further suggests that the insensitivity to guilt in these patients may result from the inability to form accurate expectations about the social behavior of others. When Trustees chose to match the Investor's expectations, and thus tried to minimize their anticipated guilt, Trustees exhibited increased activity in the insula, supplementary motor area (SMA), ACC, DLPFC, and parietal areas, including the temporal parietal junction (TPJ; Chang and others 2011). The insula, SMA, and ACC have been associated with a number of negative states, such as guilt, anger, and disgust, as well as social and physical pain (Calder and others 2000; Damasio and others 2000; Eisenberger and others 2003; Shin and others 2000; Singer and others 2004). Therefore, these results demonstrate that not fulfilling the expectations of another may result in the experience of negative affect, which in turn can motivate cooperation. Consistent with this interpretation, participants who reported they would have experienced more guilt had they returned less than they believed their partner expected them to return, showed increased activity in the insula and SMA when they matched expectations. Thus, people who are more guilt sensitive show increased activity in the network associated with negative affective states, providing further support for the argument that the anticipation of guilt may be used as a guide to cooperative decision making. The DLPFC in turn may function to override the affective feelings originating in the insula, as this area is known for its role in executive processing, and has been shown to play a key role in the implementation of fairness related behaviors (Knoch and others 2006; van't Wout and others 2005).

## Social Ties

In addition to the specific emotion of guilt, a variety of affective states elicited by emotional bonds, or social ties, may also influence cooperation (van Winden and others 2008). For example, people cooperate more with others they like, they feel close to, or with whom they have something in common (Bohnet and Zeckhauser 2004; Komorita and Parks 1995). Moreover, group membership has a strong influence on cooperation, and people are more likely to cooperate with in-group than out-group members (Goette and others 2006). One potential mechanism underlying the influence of social ties on cooperation that may underlie the above behavioral findings is the generation of empathy (De Dreu 2012; Goette and others 2006). That is, social ties may foster greater empathy between individuals, which in turn may enhance cooperative behavior. This notion is supported by brain imaging studies showing that empathic neural responses in the insula and ACC are modulated by the behavior of others (Singer and others 2004). For instance, when watching another individual getting a painful shock, empathy-related activation in pain areas, including the insula and ACC, was notably absent when this individual had previously shown non-cooperative behavior in a PDG (Tania Singer and others 2006). Moreover, these empathic pain-related activations in the insula were stronger when participants witnessed an in-group member receiving shocks as compared with an out-group member (Hein and others 2010). Additional evidence for the view that increased empathic concern moderates the influence of social bonds on cooperation comes from recent pharmacological studies using oxytocin, a hormone implicated in many aspects of human social cognition, including trust, in-group favoritism and empathy (for a review, see Bartz and others 2010). Intranasal administration of oxytocin increases cooperative behavior in particular with in-group members (De Dreu and others 2010), suggesting that oxytocin may amplify trust and empathy toward relevant others and, in turn, motivate cooperation.

Closely related to the capacity to empathize is the ability to understand the mental states of others, traditionally referred to as theory of mind. The neural circuitry of theory of mind has been well-studied and areas implicated in this process include the dorsal medial prefrontal cortex (DMPFC), as well as regions within the parietal and temporal lobes, such as the TPJ, and posterior part of the superior temporal sulcus (Gallagher and Frith 2003; McCabe and others 2001; Rilling and others 2004b). Indeed, cooperative decisions reliably engage brain systems implicated in theory of mind processes, suggesting that this ability to perspective-take plays an important role in cooperation. For instance, a study examining the

TG showed that DMPFC activity is high during the initial stages of building trust with another, with this activity decreasing once trust is firmly established (Krueger and others 2007). This suggests that, in conjunction with brain systems involved in reward-based learning, this region may encode the degree to which another player is trustworthy or not, this providing vital information in the decision to cooperate. Similarly, the receipt of partner feedback in both PDG and UG has been found to reliably engage brain systems implicated in theory of mind such as the DMPFC and posterior part of the superior temporal sulcus, with each of these areas activated more strongly when playing with a human than a computer partner (Rilling and others 2004b). To investigate the role of the DMPFC in cooperation in more detail, Yoshida and others (2010) colleagues applied a computational model of dynamic belief inference to neuroimaging data of a stag-hunt game. In the stag-hunt game, each of two players has to decide whether to hunt for a valuable stag or a less valuable rabbit, without knowing the choice of the other. To hunt the stag, both players must cooperate, while each player can acquire the rabbit by himself or herself. Combining this game with a computational model allowed the assessment of both the neural correlates of the sophistication of players' strategic thinking as well as the degree of uncertainty regarding their opponents' level of sophistication, where here sophistication was defined by the degree of belief inference (first order, second order, etc.). Different regions in the prefrontal cortex were involved in the implementation of these two separate components of belief, with activity in the DMPFC greater when players were more uncertain about their opponents' level of inference, suggesting that this area has a specific role in encoding the uncertainty of belief inference (Yoshida and others 2010). In contrast, the players' sophistication itself was associated with activation in the DLPFC, an area important for executive processing (Smith and Jonides 1999).

This result suggests that the DLPFC is involved in the strategic processes required for the implementation of social goals governing mutual cooperation. This study is a good template for demonstrating how using both formal mathematical models of high-level cognitive behavior in conjunction with brain imaging measures can add much useful knowledge to our understanding of the processes involved in cooperative social decision making. These innovative approaches only recently adopted within decision neuroscience offer much promise for more detailed understanding of how humans cooperate and how this process fails on occasion.

## Conclusion

As we have attempted to demonstrate, neuroscience can provide important biological constraints on the processes involved in decisions involving cooperation, and indeed the research reviewed here is revealing that many of the processes underlying these complex social decisions may overlap with rather fundamental brain mechanisms, such as those involved in reward, punishment, and learning.

Though still occupying a small subfield, the cross-disciplinary nature of these neuroeconomic studies are innovative, and combining insights from psychology, neuroscience, and economics has the potential to greatly increase our knowledge about the psychological and neural basis of cooperation. Participants in these studies are generally directly embedded in meaningful social interactions, and their decisions carry real weight in that their compensation is typically based on their cooperative decisions. Importantly, observed decisions in these tasks often do not conform to the predictions of classical game theory, and therefore more precise characterizations of both behavioral and brain mechanisms are important in adapting these models to better fit how decisions are actually made in an interactive environment. Furthermore, the recent use of formal modeling approaches in conjunction with psychological theory and fMRI offers a unique avenue for the study of social dynamics, with the advantages of this approach being twofold. First, it ensures that models of cooperative behavior are formally described, as opposed to the rather ad hoc models that are typically constructed. And secondly, by assessing whether these models are neurally plausible, it provides a more rigorous test of the likelihood that these models are good representations of how people are actually making decisions about whether or not to cooperate.

Finally, as we mentioned earlier, there is the potential for this work to ultimately have a significant practical impact in terms of understanding how interactive decision making works. While this is useful general knowledge to disseminate to the public, a more important potential gain is related to public policy. Results gleaned from laboratory studies in experimental economics have been found to generalize to behavior in the field (Carpenter and Seki 2011; Karlan 2005), suggesting that these tasks can be usefully employed to inform as to how real-world decisions regarding cooperative behavior are taken. More comprehensive knowledge of the neural processes underlying cooperation could in turn generate useful hypotheses as to how policy interventions could be structured, for example in relation to tax compliance, medical decision making, investment behavior, and social norms. Typically, these policy decisions are based on the standard economic models of behavior that often do not accurately capture how individuals actually decide. The development of more accurate, brain-based, models of cooperative decision making has the potential to greatly help with these policy formulations as they relate to our interactive choices. Knowing what signals commonly trigger both cooperation and non-cooperation can assist in designing policy to better achieve desired societal aims.

## Acknowledgments

We thank Ale Smidts for valuable comments and Veerle van Son for help with the figures.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## References

Axelrod R, Hamilton WD. 1981. The evolution of cooperation. Science 211:1390–6.

Balliet D, Mulder LB, Van Lange PAM. 2011. Reward, punishment, and cooperation: a meta-analysis. Psychol Bull 137:594–615.

Bartz JA, Zaki J, Bolger N, Hollander E, Ludwig NN, Kolevzon A, and others. 2010. Oxytocin selectively improves empathic accuracy. Psychol Sci 21:1426–8.

Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS. 2008. Associative learning of social value. Nature 456: 245–9.

Berg J, Dickhaut J, McCabe K. 1995. Trust, reciprocity, and social history. Games Econ Behav 10:122–42.

Bohnet I, Zeckhauser R. 2004. Social comparisons in ultimatum bargaining. Scand J Econ 106:495–510.

Calder AJ, Keane J, Manes F, Antoun N, Young AW. 2000. Impaired recognition and experience of disgust following brain injury. Nat Neurosci 3:1077–8.

Camerer CF. 2003. Behavioral game theory. Princeton, NJ: Princeton University Press.

Carpenter J, Seki E. 2011. Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. Econ Inquiry 49:612–30.

Chang LJ, Doll BB, van't Wout M, Frank MJ, Sanfey AG. 2010. Seeing is believing: trustworthiness as a dynamic belief. Cogn Psychol 61:87–105.

Chang LJ, Sanfey AG. 2011. Great expectations: neural computations underlying the use of social norms in decision-making. Soc Cogn Affect Neurosci. doi:10.1093/scan/nsr094

Chang LJ, Smith A, Dufwenberg M, Sanfey AG. 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70:560–72.

Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LLB, Parvizi J, and others. 2000. Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci 3:1049–56.

Damasio AR. 2003. Looking for Spinoza: joy, sorrow, and the feeling brain. New York: Harcourt

Decety J, Jackson PL, Sommerville JA, Chaminade T, Meltzoff AN. 2004. The neural bases of cooperation and competition: an fMRI investigation. Neuroimage 23:744–51.

Delgado MR, Frank RH, Phelps EA. 2005. Perceptions of moral character modulate the neural systems of reward during the trust game. Nat Neurosci 8:1611–8.

De Dreu CKW. 2012. Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda. Horm Behav 61:419–28.

De Dreu CKW, Greer LL, Handgraaf MJJ, Shalvi S, Van Kleef GA, Baas M, and others. 2010. The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328:1408–11.

Eisenberger NI, Lieberman MD, Williams KD. 2003. Does rejection hurt? An fMRI study of social exclusion. Science 302:290–2.

Fehr E, Fischbacher U. 2004. Social norms and human cooperation. Trends in cognitive sciences 8:185-90.

Fehr E, Gächter S. 2000. Cooperation and punishments in public goods experiments. Am Econ Rev 90:980–94.

Fiorillo CD, Tobler PN, Schultz W. 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299:1898–902.

Fischbacher U, Gächter S, Fehr E. 2001. Are people conditionally cooperative ? Evidence from a public goods experiment. Econ Lett 71:397–404.

Gallagher HL, Frith CD. 2003. Functional imaging of "theory of mind." Trends Cogn Sci 7:77–83.

Goette L, Huffman D, Meier S. 2006. The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. Am Econ Rev 96:212–6.

Güth W, Schmittberger R, Schwarze B. 1982. An experimental analysis of ultimatum game bargaining. J Econ Behav Organ 3:367–88.

Hardin G. 1968. The tragedy of the commons. Science 162: 1243–48.

Hein G, Silani G, Preuschoff K, Batson CD, Singer T. 2010. Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. Neuron 68:149–60.

Karlan DS. 2005. Using experimental economics to measure social capital and predict financial decisions. Am Econ Rev 95:1688–99.

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. 2005. Getting to know you: reputation and trust in a two-person economic exchange. Science 308: 78–83.

Klucharev V, Hytönen K, Rijpkema M, Smidts A, Fernández G. 2009. Reinforcement learning signal predicts social conformity. Neuron 61:140–51.

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E. 2006. Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science 314:829–32.

Koenigs M, Tranel D. 2007. Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. J Neurosci 27:951–6.

Koenigs M, Young L, Adolphs R, Tranel D, Cushman F, Hauser M, and others. 2007. Damage to the prefrontal cortex increases utilitarian moral judgements. Nature 446:908–11.

Komorita SS, Parks CD. 1995. Interpersonal relations: mixed-motive interaction. Annu Rev Psychol 46:183–207.

Krajbich I, Adolphs R, Tranel D, Denburg NL, Camerer CF. 2009. Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. J Neurosci 29:2188–92.

Krueger F, McCabe K, Moll J, Kriegeskorte N, Zahn R, Strenziok M, and others. 2007. Neural correlates of trust. Proc Natl Acad Sci U S A 104:20084–9.

Loewenstein G, Lerner JS. 2003. The role of affect in decision-making. In: Davidson RJ, Scherer KR, Goldsmith HH, editors. Handbook of affective sciences. New York: Oxford University Press. p 619–42.

McCabe K, Houser D, Ryan L, Smith V, Trouard T. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. Proc Natl Acad Sci U S A 98:11832–5.

McClure SM, York MK, Montague PR. 2004. The neural substrates of reward processing in humans: the modern role of FMRI. Neuroscientist 10:260–8.

Phelps EA. 2009. The study of emotion in neuroeconomics. In: Glimcher PW, Camerer CF, Fehr E, Poldrack RA, editors. Neuroeconomics: decision making and the brain.London: Academic Press. p 233–47.

Poldrack RA. 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci 10:59–63.

Poundstone W. 1992. Prisoner's dilemma. New York: Anchor Books/Doubleday.

Rilling JK, Gutman D, Zeh T, Pagnoni G, Berns GS, Kilts C. 2002. A neural basis for social cooperation. Neuron 35:395–405.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. 2004a. Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. Neuroreport 15:2539–43.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. 2004b. The neural correlates of theory of mind within interpersonal interactions. Neuroimage 22:1694–703.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. 2003. The neural basis of economic decision-making in the Ultimatum Game. Science 300:1755–8.

Sanfey AG. 2009. Expectations and social decision-making: biasing effects of prior knowledge on ultimatum responses. Mind Soc 8:93–107.

Schelling TC. 1960. The strategy of conflict. Cambridge, MA: Harvard University Press.

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science 275:1593–9.

Shackman AJ, Salomons TV, Slagter HA, Fox AS, Winter JJ, Davidson RJ. 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. Nat Rev Neurosci 12:154–67.

Shin LM, Dougherty DD, Orr SP, Pitman RK, Lasko M, Macklin ML, and others. 2000. Activation of anterior paralimbic structures during guilt-related script-driven imagery. Biol Psychiatry 48:43–50.

Singer T, Kiebel SJ, Winston JS, Dolan RJ, Frith CD. 2004. Brain responses to the acquired moral status of faces. Neuron 41:653–62.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD. 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303:1157–62.

Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD. 2006. Empathic neural responses are modulated by the perceived fairness of others. Nature 439:466–9.

Smith EE, Jonides J. 1999. Frontal lobes. Science 283:1657–61.

Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E. 2007. The neural signature of social norm compliance. Neuron 56:185–96.

Tricomi EM, Delgado MR, Fiez JA. 2004. Modulation of caudate activity by action contingency.Neuron 41:281–92.

van Winden F, Stallen M, Ridderinkhof KR. 2008. On the nature, modeling, and neural bases of social ties. Adv Health Econ Health Serv Res 20:125–59.

van't Wout M, Kahn R, Sanfey AG, Aleman A. 2005. Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. Neuroreport 16:1849–52.

van't Wout M, Sanfey AG 2008. Friend or foe: the effect of implicit trustworthiness judgments in social decision-making. Cognition 108:796–803.

von Neumann J, Morgenstern O. 1947. Theory of games and economic behavior. Princeton, NJ: Princeton University Press.

Yoshida W, Seymour B, Friston KJ, Dolan RJ. 2010. Neural mechanisms of belief inference during cooperative games. J Neurosci 30:10744–51.