Chapter 2
The Neural Basis of Fairness

Peter Vavra, Jeroen van Baar, and Alan Sanfey

Introduction

Recent laboratory research in cognitive neuroscience has begun to explore paradigms that offer fruitful avenues to examine how processes involving a sense of fairness may be encoded in the human brain. Most of this research is embedded within the field of Decision Neuroscience (also known as Neuroeconomics), an interdisciplinary effort to better understand the fundamentals of human decision-making. Within this field, researchers endeavor to build accounts of decision-making that incorporate the psychological processes that influence decisions, that indicate how these processes are constrained by the underlying neurobiology of the brain, and at the same time developing formal models of these decisions, an approach extended from economics.

The emergence of this 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 decision tasks that allow for the exploration of fairness in a controlled laboratory environment. In contrast to standard behavioral studies, the combination of game theoretic models with online measurement of brain activity during decision-making allows for the discrimination and modeling of processes that are often hard to separate at the behavioral level. Within this 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 (land, political power, etc.). These tasks (see Box 2.1) are well suited to be used in combination with brain imaging methods, and produce a surprisingly rich pattern of decision-making, which allows for a wide range of questions to be answered about the

P. Vavra • J. van Baar • A. Sanfey (✉)
Donders Institute for Brain Cognition and Behavior, Radboud University Nijmegen, Nijmegen, The Netherlands
e-mail: p.vavra@donders.ru.nl; j.vanbaar@donders.ru.nl; a.sanfey@donders.ru.nl

M. Li, D.P. Tracer (eds.), *Interdisciplinary Perspectives on Fairness, Equity, and Justice*, DOI 10.1007/978-3-319-58993-0_2

motivations that underlie fairness behavior. In this chapter, we review the work to date in understanding fairness from a Decision Neuroscience perspective, with particular interest paid to the brain regions that appear to be prominently involved when we consider whether outcomes or procedures are fair or not, and what actions we are willing to take to redress the balance. Exploring these fundamental mechanisms can provide valuable insight into the associated psychological processes, and ultimately can help us better understand the complex but important concept of fairness.

Box 2.1 Experimental Tasks

Tasks used to investigate fairness-related decision-making have their root in behavioral economics and game theory. In these games, typically two players are facing a decision situation. We briefly describe here four games that have often been used in the context of fairness and equity.

The Ultimatum Game (UG; Güth, Schmittberger, & Schwarze, 1982) is a two-player game where the players each make a decision sequentially. The first player, termed the proposer, is endowed with a sum of money. The proposer has to decide how much of this sum to offer the second player. Then, the second player, the responder, decides whether to accept or reject the offer. If the responder accepts, the two players split the money accordingly. If the responder rejects, however, neither player receives any money. The proposer's decision is seen as reflecting strategic decision-making—how much is the responder probably willing to accept?—as well as reflecting some form of consideration of fairness—what do I consider a fair split of money? The responder's decision to reject a low offer is seen as a canonical example of altruistic punishment: the willingness to forgo a (monetary) payoff in favor of enforcing a social norm of fairness. Indeed, most people reject low offers and consider them unfair (Camerer, 2003).

The Dictator Game (DG) and Impunity Game are closely related to the UG. In the DG, the only difference with the UG is that the responder doesn't have the opportunity to reject the offer. Instead, the allocation of money is realized after the first player, now called a dictator, decides how much to transfer to the recipient. Given that there is no risk here of rejection for the dictator, the motivation to transfer money in this game is seen as genuine generosity. In the Impunity Game, the responder does have the option to accept or reject an offer. However, in contrast to the UG, when rejecting an offer, only the responder receives nothing; the proposer still receives the rest of the money. That is, in the Impunity Game, there is also no risk of losing money for oneself.

The above games are all related to equity or, more narrowly, equality norms. However, fairness and justice also relate to reciprocity. One simple, two-player game used for investigating reciprocity is the Trust Game. The first player, termed the investor, is endowed with a sum of money. They can decide how much of their money to transfer to the second player, the trustee. Importantly,

(continued)

Box 2.1 (continued)
whatever amount is transferred is multiplied by a fixed factor, e.g., four. For example, the investor could transfer $5. Then, the trustee would receive $20 and can, in turn, decide how much of this latter amount to transfer back. Importantly, the trustee can also decide to not transfer any money at all. Thus, the decision of the investor is seen as a sign of trusting the second player to return some amount of money. The second player's decision to return any amount is seen as a sign of reciprocating trust. Note that for the trustee, the decision is structurally identical to a dictator's decision in the DG, except for the history of how the trustee arrived in the position of having any endowment at all.

In their most straightforward form, these games are played as one single round with a completely anonymous partner. However, for the purpose of neuroimaging studies, it is important to have multiple observations, so many studies employ so-called single-shot multi-round games: as a participant, one plays the game multiple times, on each round paired with a new partner. Alternatively, studies focusing on learning processes often employ repeated paradigms. For example, by playing with the same set of partners, one can learn to trust or distrust trustees in the TG based on how often, and how much, money they return. These simple yet powerful tasks allow researchers to employ computational models to quantify key theoretical variables. By using simple variations of these tasks, e.g., by playing for a third person instead of oneself, it is also possible to disentangle the contributions of different motivations to the decisions. In sum, these tasks are exceptionally well suited for the study of fairness and justice, because they provide a unique balance between experimental control, rich psychological processes, and formal modeling.

By employing functional neuroimaging (see Box 2.2) to examine a range of tasks that are rooted in behavioral economics, a network of brain regions has been identified that support a decision as to whether to behave fairly or unfairly in a given situation, as well what underlies the response to the fair or unfair behavior of others. A major effort in this regard has been to elucidate the psychological and computational roles each of these brain regions might play in this process. In this chapter, we highlight the brain systems (Fig. 2.1) that have been most consistently identified in these processes, and review the respective roles they may play in fairness-related decision-making.

Trading-Off Self-Interest Versus the Greater Good

In most of the experiments used to study fairness-related processing in the brain, participants face a trade-off between self-interest and adhering to a fairness norm of some type. It should be noted that these two motivations are often explicitly pitted against each other by the researcher. While self-interest and fairness regularly

Box 2.2 Functional Neuroimaging Techniques

Since the early twentieth century, humans have been capable of measuring neural activity in the living brain, without damaging underlying tissue. The first of these noninvasive functional neuroimaging techniques was the electroencephalogram (EEG). Building on the pioneering work of Hans Berger in 1920s Germany, modern EEG methods are capable of measuring electric field changes at up to 256 electrode sites on the scalp. These electric field changes are assumed to be caused by synchronized neural firing in the cerebral cortex. Given that EEG picks up on electric field changes, its temporal resolution is very high and limited only by the sampling frequency of the EEG equipment. For this reason, EEG is extremely useful in measuring the brain's response to rapidly developing stimuli, like chunks of spoken language. The spatial resolution of EEG is, however, quite low, as electric fields generated by the brain are smeared by the layers of soft and hard tissue that lie between the brain and the electrode cap. Furthermore, EEG favors measuring superficial brain structures over deep ones, as the electric field changes caused by deep structures may not even reach the scalp.

A neuroimaging method which is complementary to EEG is functional magnetic resonance imaging (fMRI). fMRI picks up on magnetic field changes inside the brain, which are caused by changes in the relative flow of oxygenated and deoxygenated blood to and from active brain tissue. As such, the fMRI signal is termed the *blood oxygenation level-dependent* response, or BOLD. Due to numerous advances in fMRI techniques since its inception in the early 1990s, the spatial resolution of this method is in the order of millimeters, allowing, for example, for precise functional parcellation of brain structures. It is equally powerful in measuring deep brain regions as it is in measuring superficial ones, as signals from different brain regions do not interfere with one another. As a limitation however, the temporal resolution of fMRI is relatively low, as the response of the blood flow to brain activity is slow: it takes between 6 and 10 s after a brain region is active for the blood flow response to reach its peak, and the entire response can take up to 30 s. Still, by systematically varying time intervals between experimental stimuli, it is possible to connect the brain response to individual stimuli, even if they are very close together in time. One major downside of fMRI is its cost, both in purchasing the equipment (in the order of millions of dollars) and running it (hundreds of dollars per hour).

A method that is sometimes heralded as having both high spatial and high temporal resolution is the magnetoencephalogram, or MEG. MEG capitalizes on the fact that, by the laws of electromagnetic induction, neuronal electrical activity in the brain ever so slightly changes the magnetic field that exists around the head of a participant. By placing many tiny superconducting sensors in an array around the head, one can sense these magnetic field changes at a high temporal resolution. Compared to fMRI, an important advantage of

(continued)

Box 2.2 (continued)
MEG is that, like EEG, it directly measures neuronal activity, and not a secondary measure like blood flow. Localizing the source of the magnetic field change (the active brain region) in MEG at the centimeter scale is more feasible than in EEG, as the magnetic field can easily pass through skull and scalp. Unfortunately, the MEG signal is, like EEG, dominated by neuronal activity in superficial brain areas. Therefore, MEG is still seldom used to investigate the activity of brain structures that are crucial in understanding fairness processing, such as the striatum and the insula. In addition, using MEG requires advanced shielding of the experimental environment to external magnetic fields—even the movement of an elevator in the building generates a magnetic field change much larger than that caused by brain activity.

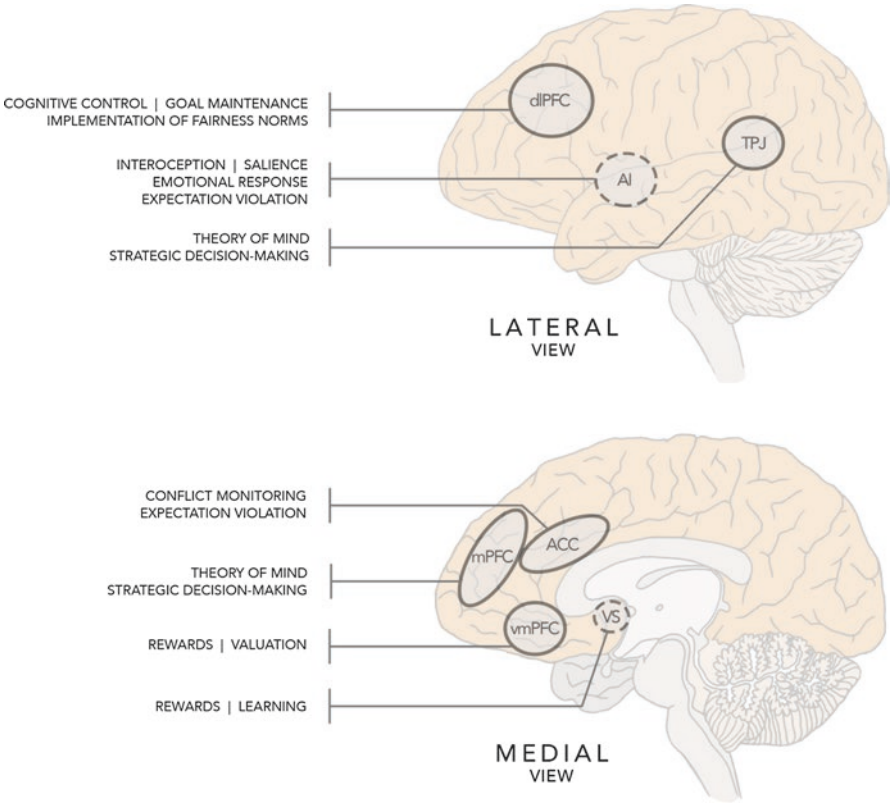

Fig. 2.1 Brain areas involved. Brain regions and their involvement in different processes during fairness-related decision-making, showing lateral (*top* panel) and medial (*bottom* panel) views of the human brain. *Solid lines* indicate surface structures, *dashed lines* indicate deep structures, *mPFC* medial prefrontal cortex, *TPJ* temporoparietal junction, *vmPCF* ventromedial prefrontal cortex, *ACC* anterior cingulate cortex, *dlPFC* dorsolateral prefrontal cortex, *VS* ventral striatum, *AI* anterior insula

motivate the same behavior in everyday situations (for example, when voting in support of wealth redistribution while being on the receiving end of such a measure), researchers are typically interested in isolating a single motivation. By pitting self-interest against fairness and observing subsequent behavior, they can deduce which motivation was the primary driver of the participants' decisions. This allows for careful study of the psychological and neural processes underlying fairness motivations. Note that in experimental practice, "unfair" decisions most often align with financially self-interested ones, while "fair" decisions usually serve the greater good (i.e., others' (financial) interests).

Turning to the brain, we first consider which neural systems instantiate self-interested behavior. The first structure that deserves mention in this respect is the ventral striatum, a collection of brain nuclei situated underneath the neocortex. It has long been known that the substructures of the ventral striatum play an important role in driving choice behavior. Ventral striatal structures are responsible for incentive salience (i.e., desire), pleasure, and learning. For example, dopamine neurons in the substantia nigra, which project to the ventral striatum, become more active when a rewarding stimulus (e.g., food) is presented to a participant. Interestingly, these neurons also fire when a cue is presented that is not rewarding in itself, but that has previously been associated with a primary reward through learning (Schultz, Dayan, & Montague, 1997). As such, the ventral striatum facilitates motivational learning, but is also involved in addiction (Everitt & Robbins, 2005).

Considering the ventral striatum's role in reward processing, it is no surprise that it also responds strongly to the rewarding stimulus of money in the context of economic games. It is perhaps less well-known that the ventral striatum can also be activated by social rewards, such as possessing a good reputation (Izuma, Saito, & Sadato, 2008). This finding speaks to the concept of a "common neural currency," that is, the integration of several sources of reward into a single neural signal. Another brain region that appears to carry a domain-general signal, tracking the subjective value of a stimulus to the participant, is the ventromedial prefrontal cortex (Bartra, McGuire, & Kable, 2013). This region likely plays an important role in integrating the subjective value of different choice options into a decision and then driving the acquisition of the chosen option (Ruff & Fehr, 2014).

For the remainder of this chapter, it is important to note that while fairness judgments involve many different parts of the brain, the reward system is also simultaneously processing financial self-interest. In order for an individual to behave fairly, therefore, they must balance out the impulse of self-interest with an inclination towards fairness. It is well-known that the prefrontal cortex is very important for executive control (Miller & Cohen, 2001; Seeley et al., 2007), and therefore the connections between the prefrontal cortex and the reward system are prime targets for the neurobiological study of fairness-related behavior. We will discuss these connections in more detail below.

On the fair, "greater good," side of the equation, it is useful to start by investigating what happens in the brain when someone observes both the fair and unfair behavior of another person.

Monitoring (Un)fairness: The Role of the Anterior Insula

Some of the earliest neuroscientific experiments concerning fairness implicated two brain regions whose involvement in fairness-probing tasks has been consistently replicated, these regions being the bilateral anterior insulae. The insula is a part of the cerebral cortex that is folded inward on the side of the brain, located between the frontal and temporal lobes. This region is broadly divided into a posterior part (towards the back of the brain) and an anterior part (towards the front).

From neuroimaging experiments using economic games, we know that the anterior insula becomes more active when processing the unfair behavior of their game partner. For example, in the Ultimatum Game, receiving a low, as compared to a high, offer is associated with increased anterior insula activity (see Feng, Luo, & Krueger, 2015 and Gabay, Radua, Kempton, & Mehta, 2014 for meta-analyses). Anterior insula activity has also been found to be correlated with the probability of subsequently rejecting an Ultimatum Game offer (Kirk, Downar, & Montague, 2011; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003), suggesting that participants with a more responsive anterior insula were less likely to accept unfair behavior. However, this was not the case in all studies (e.g., Civai, Crescentini, Rustichini, & Rumiati, 2012; see also Gabay et al., 2014).

When playing a UG on behalf of others, Civai et al. (2012) showed that the anterior insula was more active for more unequal allocations, whether they are advantageous or disadvantageous for the person one is playing for. Simply receiving unfair offers without being able to reject them, i.e., when playing the Dictator Game, also recruits the insula (Grecucci, Giorgetta, Bonini, & Sanfey, 2013). Similarly, in a Trust Game experiment, Delgado, Frank, and Phelps (2005) reported increased activity in the insular cortex of the investor when they learned that the trustee defected. The anterior insula, thus, appears to respond to observed unfair behavior on the part of a game partner, independent of whether one can act on this unfairness, e.g., by punishing the perpetrator, or not, and also independent of whether oneself is the target of the unfairness or not.

These findings on the anterior insula raise the question as to when a game partner's behavior is actually deemed unfair. One way to approach this question is through the lens of inequity aversion (Fehr & Schmidt, 1999; see Box 2.3). Inequity aversion theory posits that participants derive negative utility (i.e., diminished subjective value) from an unequal distribution of resources between individuals. In the Ultimatum Game, then, a player is thought to balance the conflicting goals of making money and minimizing inequity. This explains why responders in the UG sometimes reject low offers: although accepting a low offer would yield more financial payoff than rejecting, acceptance would also bring about an undesirable degree of inequity. By responding to unfairness, therefore, the anterior insula may play an instrumental role in the neural implementation of inequity aversion. In line with this interpretation, Hsu, Anen, and Quartz (2008) find insular cortex activity to be correlated with trial-by-trial inequity when deciding between different allocations for other people (third-party allocation).

Box 2.3 Computational Approaches

Computational models have greatly risen in use in recent years in order to better understand decision-making. The main appeal of these approaches is that they allow the formal specification of theories and the decomposition of the underlying psychological processes into useful subcomponents. Conceptually, there are three distinct classes of formal models typically employed: Utility models, Learning models, and Process models. We will briefly highlight an example of each in the context of fairness-related decision-making.

Utility models: This class of models specifies which features of a situation influence the evaluation of the available options by the decision maker. The inequity-aversion (Fehr & Schmidt, 1999) and Expectation (Battigalli et al., 2015) models are prominent examples here. They propose that the utility of accepting an offer in, for example, the Ultimatum Game comprise two parts: the value of the money, and the (dis)utility from deviating either from an equal split (Inequity-Aversion) or from expectations. When making the decision itself, the utility for accepting the offer is compared to the utility for rejecting it. By formalizing these utilities, it is possible to look for neural correlates and shed light on the specific contributions of brain regions to this decision-making process. For example, Chang & Sanfey (2013) compared these two models and found that anterior insula and anterior cingulate cortex showed neural activity consistent with the Expectation model specifically.

Learning models: A rapidly growing amount of work focuses on how we update our utilities based on prior experience. Reinforcement learning models (Sutton & Barto, 1998), for example, propose that we compare an experienced reward to our previous expectation of that reward, resulting in a prediction error which has been linked to phasic dopamine firing of midbrain neurons (Schultz et al., 1999; Niv & Schoenbaum, 2008). In the context of the Ultimatum Game, Xiang et al. (2013) used a Bayesian Observer model to extend the abovementioned expectation model. They demonstrated that people dynamically updated their expectations based on their experience, and that their norm prediction errors correlate with the subjective emotional experience. Recent reviews highlight such learning models, as many observed neural correlates may be related to incidental learning, and can help in disentangling the specific contributions of different brain regions (e.g., Apps, Rushworth, & Chang, 2016; Lee & Seo, 2016).

Process models: Whereas utility models formalize which features of a situation affect subjective utilities, a third group of formal models propose how a single decision is reached. A prime example of such algorithmic models is the drift-diffusion model (DDM; Ratcliff & Mckoon, 2008; Smith & Ratcliff, 2004). In a DDM, the computation of a utility is modeled as an accumulation of a noisy signal, with a choice being made when this value signal reaches a certain threshold, that is, after enough "evidence" has accumulated in favor of

(continued)

Box 2.3 (continued)
one of the options. Importantly, such models do predict not only the decision itself, but also the associated reaction times. For example, Hutcherson, Bushong, and Rangel (2015) modeled the decision to choose either a selfish or a generous offer in a modified Dictator Game as a noisy calculation of a relative value signal. Hutcherson and colleagues proposed that the decision process needs to compare the value for oneself and the value for the other player, and that these values are calculated independently. Among other regions, they found that activity in the striatum was related to the value of the options for the self, while right TPJ activity was related to value for the other. Finally, activity in the vmPFC showed overlap for self and other utilities, consistent with the idea that the vmPFC integrates multiple attributes into a final value (Basten, Biele, Heekeren, & Fiebach, 2010).

Further, people who weigh inequity more strongly in their decisions also show a larger insula response to inequity (Hsu et al., 2008).

Crucially, the inequity aversion account implies that fairness norms are static and always favor a precisely even distribution of money. An alternative interpretation is that the evaluation of a game partner's behavior is made in comparison to one's expectations of the partner's behavior (Battigalli, Dufwenberg, & Smith, 2015). After all, what we find "fair" in everyday life is strongly dependent on both mitigating and aggravating circumstances, as well as dependent on our moral expectations of the individual we are dealing with—one may expect fairer behavior from a nun than from a convicted conman. In line with this view, there is evidence that the response of the responder's anterior insula to Ultimatum Game offers is proportional to the difference between the offer in question and that which was a priori expected (Chang & Sanfey, 2013; Xiang, Lohrenz, & Montague, 2013). In line with this dynamic view of fairness norms underlain by the cognitive expectations we generate, Fareri, Chang, and Delgado (2012) showed that the insular and cingulate brain response to prediction error after seeing the outcome of the trust game correlated with the participant's individual learning rate. That is, participants with a higher learning rate (who respond more sensitively to deviations from expectation) show a greater brain response in cingulate and insular cortex when being disappointed by a trustee.

One important question in the practice of cognitive neuroscience is: what is the participant experiencing subjectively while completing the experimental task? Measurements of brain activity can offer a window into this experience. For one, we know that the insular cortex plays an important role in emotion processing, especially of anger and disgust (Damasio et al., 2000; Phillips et al., 1997), and in the visceral experience of negative feelings (Critchley et al., 2004; Singer, Critchley, & Preuschoff, 2009). Therefore, the increased anterior insula activity in the Ultimatum Game is often interpreted as an emotional response to unfair behavior (Sanfey et al., 2003). In line with this interpretation, several studies show the importance of

emotions in the UG. For example, Harlé, Chang, van't Wout, and Sanfey (2012) demonstrate that after watching a sad movie clip compared to a neutral one, people more often reject unfair UG offers. Importantly, the change in emotional state was accompanied by increased activity in the anterior insula, which was shown to mediate the relationship between the emotion condition and acceptance rate.

Even simply instructing participants to either up- or downregulate their emotional response can increase and decrease the rejection rate of unfair UG offers, respectively (Grecucci, Giorgetta, van't Wout, Bonini, & Sanfey, 2013). Importantly, the (posterior) insula activity decreased for downregulation and increased for upregulation of one's emotional arousal, in line with the changes in rejection rates. When playing the Dictator Game, that is, without having the opportunity to punish a low offer, insula activity is also affected by emotional reappraisal in the same pattern and is correlated with the subjective experience of anger (Grecucci, Giorgetta, Bonini, & Sanfey, 2013).

Nonetheless, the role of emotions and its link to the insula activity are less straightforward than these studies suggest. In a set of studies, Civai and colleagues compared playing the UG for oneself and playing it on behalf of a third party. When measuring emotional arousal using skin conductance response, Civai et al. (2010) found that participants had an increased emotional response to unfair offers only if playing for themselves, even though they would reject unfair offers as often as when playing for others. The anterior insula was associated with rejections in both contexts, and it was the mPFC which dissociated between the two situations (Corradi-Dell'Acqua, Civai, Rumiati, & Fink, 2013).

Therefore, to summarize, it has been known since the first neuroimaging experiments on fairness that the anterior insula responds to unfair behavior of oneself and others. This response is thought to reflect the difference between the observed behavior of others and one's prior expectations of this behavior. The result of this comparison, i.e., the deviation from expectations, may drive the emotional response as well as the decision to reject in some situations.

Conflict Monitoring and Cognitive Control in Economic Games

Aside from its role in emotion processing, the anterior insula is also thought to play a key role in the brain's salience network (Seeley et al., 2007). This ensemble of brain regions is thought to integrate sensory information with bodily cues from the autonomic nervous system, thereby enabling fast responding to the most homeostatically relevant events. This network additionally comprises, among other regions, the anterior cingulate cortex (ACC; Seeley et al., 2007). The ACC is hypothesized to monitor conflict in information processing, thereby triggering compensatory adjustments in cognitive control (Botvinick, Cohen, & Carter, 2004). In recent years, the role of the cognitive control system, and the role of the ACC in particular, in interactive decision-making has become clearer.

In the context of fairness and equity, the ACC has been related to several different psychological states. For example, in the Ultimatum Game, the ACC is more active when observing unfair as compared to fair offers (Feng et al., 2015; Gabay et al., 2014), and this activity is proportional to the deviation from fairness expectations (Chang & Sanfey, 2013), much like activity in the anterior insula (AI). Similarly, Haruno and Frith (2010) found that activity in ACC and AI tracked the difference between the payoffs of the participant and another person (i.e., inequity). Additionally, in the Trust Game, Baumgartner, Fischbacher, Feierabend, Lutz, and Fehr (2009) found increased activity in the ACC in trustees who were about to defect, breaking a promise they had previously made to their game partner, as compared to trustees who were about to keep their promise to reciprocate. A working hypothesis holds that ACC detects conflict between a norm (fairness, equity, etc.) and real or possible behavior (Chang, Smith, Dufwenberg, & Sanfey, 2011; Fehr & Krajbich, 2013).

Interestingly, however, contrary to the above findings, some research has found anterior cingulate cortex to be more active during reciprocation than during defection in Trust Games (Chang et al., 2011; Van Baar, Chang, & Sanfey, 2016). How to explain these seemingly contradictory findings? One should realize that many of these conflict detection operations can be carried out by ACC in the time it takes to acquire one snapshot of the brain with functional MRI. It may well be, for instance, that the increased ACC activity observed by Baumgartner et al. (2009) occurred in response to the participants' own decision to break their promise and defect, while the ACC activity observed by Chang et al. (2011) occurred in response to the participants merely considering defection. Strong ACC activity may have different effects on behavior when it occurs at different time points across the decision-making process.

Moreover, recent research points towards a subdivision of ACC into two regions with potentially distinct functions (e.g., Apps et al., 2016), as well as to multiple, but different, brain signals present in the same subregion of ACC (e.g., Kolling, Behrens, Wittmann, & Rushworth, 2016). Therefore, while intriguing thus far, more investigation of the location and time course of activity in ACC will be needed in order to clarify its role in fairness-related decision-making.

Other important nodes of the cognitive control network are dorsolateral prefrontal cortex (DLPFC) and supplementary motor area (SMA). Both have been found to be more active when trustees reciprocated in a Trust Game, thereby adhering to a fairness norm (Chang et al., 2011; Van Baar et al., 2016; van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2011). This evidence fits with the notion that cognitive control is required to overcome the temptation of making an unfair, though financially beneficial, decision. Fairness-based decisions can thus be likened to effortful actions: a prepotent (selfish) response needs to be overridden in order for an intentional (fair) action to occur. In line with this interpretation, it has been found that increased functional connectivity between the salience (AI and ACC) and central executive (DLPFC and posterior parietal cortex) networks is associated with increased reciprocity (Cáceda, James, Gutman, & Kilts, 2015).

There is currently a lively debate in experimental psychology as to whether prosocial behavior is prepotent and thereby intuitive, or alternatively requires overriding a prepotent selfish response and is thus deliberate. Using measurements of reaction time, Rand, Greene, and Nowak (2012) have made the case for intuitive cooperation—a typical prosocial behavior. They showed that faster responses in their task were on average more prosocial than slower responses. However, Krajbich, Bartling, Hare, and Fehr (2015) point out that it may in fact be strength-of-preference rather than selfishness that predicts longer reaction times. In response, Rand (2016) provided meta-analytic evidence that deliberation is associated with self-interested behavior in situations where prosocial behavior is not beneficial for the self, i.e., situations of "pure cooperation."

If we assume that "deliberation" maps onto DLPFC, there appears a contradiction between the aforementioned behavioral evidence and the available neuroscientific evidence about the role of DLPFC in social decision-making. Specifically, Knoch, Pascual-Leone, Meyer, Treyer, and Fehr (2006) temporarily disrupted neural function in the left and the right DLPFC using repetitive transcranial magnetic stimulation (TMS; see Box 2.4). They found that disrupting right (but not left) DLPFC reduced subjects' willingness to reject unfair offers in single-shot, anonymous Ultimatum Games. In other words, intact DLPFC function was associated with costly fair decisions on the part of the subjects, which suggests that deliberation can contribute to fair behavior. Interestingly, this stimulation method left the subjective unfairness ratings of the subjects unaffected. The researchers therefore concluded that the judgment of fairness was not supported by the right DLPFC, but rather the actions based on this judgment. Further, Baumgartner, Knoch, Hotz, Eisenegger, and Fehr (2011) showed that TMS stimulation decreased both activity in right DLPFC and functional connectivity between right DLPFC and ventromedial prefrontal cortex (valuation), and that this reduced connectivity could explain the change in offer acceptance rates. A working hypothesis is, therefore, that fairness judgments in the anterior insula are relayed to the DLPFC, which in turn inhibits the self-interested "greed" response in VMPFC to make costly fair behavior possible. This neuroscientific interpretation however is at odds with the intuitive cooperation findings of Rand and colleagues, and as such offers fruitful avenues for further research.

Much like with anterior insula, it is an open question how we should define the "fairness" that DLPFC appears to strive towards. Several different approaches have recently been proposed to help solve this question. First, Ruff, Ugazio, and Fehr (2013)) showed that increasing neural excitability in right lateral prefrontal cortex (LPFC) in Dictator Game Proposers, using anodal tDCS, led to decreased monetary transfers from the dictator to the receiver and thus, arguably, a decreased sense of fairness. When repeating this experiment with Ultimatum Game Proposers, however, upregulating LPFC with anodal tDCS now led to increased offer amounts from proposers to responders. As the only

Box 2.4 Brain Stimulation Techniques

To assess causal roles for brain regions in decision-making, two dominant noninvasive methods exist: transcranial magnetic stimulation (TMS) and transcranial current stimulation (tCS).

In TMS, researchers place a coil close to the skull. By running a brief, but strong, current though the coil, a transient magnetic field is created which in turn creates a secondary, induced, electric field inside the skull. This field can cause electrical currents in tissue and generate action potentials. When stimulating the primary motor cortex, for example, these impulses can lead to muscle contractions. For tCS, the researcher places two electrodes on the body, one electrode being the active electrode, i.e., positioned at the brain region one wants to stimulate, with the other being the reference electrode placed somewhere else. The reference electrode can be positioned either close by (e.g., only a one or two centimeters away) or far away, for example on a limb. By varying the size of the electrode itself, it is possible to vary the induced change in potential in the underlying tissue. The reference electrode is, thus, typically larger than the active electrode, causing less change to the tissue beneath it. One can use either direct current (tDCS) or alternating current (tACS) stimulation paradigms, with the former being more commonly used in the context of decision-making. The canonical interpretation for tDCS is that cathodal stimulation worsens performance (Stagg & Nitsche, 2011), while anodal improves it, but this dual-polarity effect is not always observed (e.g., Jacobson, Koslowsky, & Lavidor, 2012; Miniussi, Harris, & Ruzzoli, 2013).

There are two main types of paradigm for using noninvasive brain stimulation techniques: online or offline. Online paradigms use the stimulation at the time of the process itself. For example, by stimulating the dlPFC while playing the Ultimatum Game, it is possible to investigate how this stimulation alters the underlying neural processes (Knoch et al., 2008). In contrast, in offline paradigms one stimulates the brain region of interest first, up- or down-regulating its activity for several minutes, and only then is a task used to study the process of interest. Using TMS, this means that one uses a repetitive stimulation paradigm (rTMS) where pulses are created with typically 1 Hz frequency for several minutes. This is thought to cause a deactivation of the underlying brain region (Iyer, Schleper, & Wassermann, 2003). In decision-making research therefore, one first (de)activates a brain region of interest and then participants play, for example, the Ultimatum Game (e.g., Knoch et al., 2006; van't Wout, Kahn, Sanfey, & Aleman, 2005).

One strong limitation of all noninvasive brain stimulation techniques is the lack of spatial specificity. Although one might be interested in altering function in a single brain region (e.g., dlPFC), stimulation might affect even distant brain regions via neural connectivity. Indeed, it might be the connections themselves that are affected by the stimulation, such as dlPFC-vmPFC

(continued)

Box 2.4 (continued)
connectivity in the UG (Baumgartner et al., 2011). Thus, the conclusions that can be drawn from stimulation studies are greatly enhanced when conducted in conjunction with functional brain imaging. Alternatively, one can add multiple control conditions, using varied stimulation sites to show spatial specificity and a collection of tasks to assess cognitive specificity of the employed stimulation intervention. A more practical limitation is that only superficial brain regions can be targeted directly. Unfortunately, therefore it is difficult to stimulate for example the anterior insula, an especially important brain region for understanding fairness. Despite these limitations, noninvasive brain stimulation techniques such as rTMS and tDCS are valuable tools for the investigation of fairness-related decision-making. An opportunity for future studies is to combine stimulation techniques and formal modeling to arrive at a better understanding of the respective processes and computations.

difference between the DG and the UG for the proposers is the "sanction threat" of not getting any money at all, Ruff and colleagues concluded that the right lateral PFC processes voluntary and sanction-induced "fairness" differently. Sanfey, Stallen, and Chang (2014) added another interpretation of this finding: it is possible that increased activity in LPFC places participants' behavior more in line with what they believe other people would do in the same situation (their "descriptive social norm"). That is, participants may believe that other people would transfer relatively little money in the Dictator Game but a greater amount in the (potentially sanctioned) Ultimatum Game, and if this is the case, upregulating LPFC activity with tDCS could stimulate behavior to align these descriptive social norms. In either case, the findings by Ruff and colleagues suggest that the norm for "fair" or "correct" behavior is dependent on social interactions, sanction threats, and neural activity in lateral prefrontal cortex.

In an interesting addition to this line of reasoning, Bereczkei, Deak, Papp, Perlaki, and Orsi (2013); Bereczkei et al. (2015) reported that Iterative Trust Game players who scored high on a scale for Machiavellian (manipulative) personality traits showed increased activity in left DLPFC when responding to a cooperative move of their game partner. As the high-Machiavellian subjects responded to this cooperative move by sending back less money (thus profiting more), in this case DLPFC activity was associated with reduced fairness behavior. It may well be, therefore, that brain systems involved in cognitive control are simply producing goal-directed behavior, whatever one's goal is. If one values fairness, these areas may override greedy impulses to facilitate fair behavior; if one values maximizing personal gains, these areas may override a cooperative response in favor of the exploitation of others. Indeed, this interpretation is in line with the role of the DLPFC in goal maintenance and cognitive control independent of fairness-related decisions (Miller & Cohen, 2001).

Fairness as Reward

To this point, we have discussed the role of the brain's reward system in facilitating financially self-interested behavior. That is, however, not the complete story. Tricomi, Rangel, Camerer, and O'Doherty (2010) reported observations that neural activity in ventromedial prefrontal cortex and ventral striatum increased when money was transferred from another player to the participant—but only if that other player had begun the experiment with a large monetary endowment. If the participant was the one who was endowed with money, the opposite pattern was observed: monetary transfers from self to the other player were associated with increased ventral striatal and VMPFC activity. Thus, Tricomi and colleagues argue for evidence for a reward-based neural implementation of inequity aversion, by which the receipt of money is only rewarding if it reduces inequity between game partners, in either direction. Whether this inequity-sensitivity in the brain's reward system is a function of DLPFC-VMPFC connectivity is still unknown.

These findings relate to an earlier report by Harbaugh, Mayr, and Burghart (2007). Here, the transfer of money from a participant to a charity of their choice elicited neural activity in the ventral striatum, both when those transfers were voluntary (similar to real-world donation) and when they were mandatory (similar to real-world taxation). In addition, it seems that the ventral striatum also responds to reward receipt of others although this response is diminished by social distance to the other person (Mobbs et al., 2009). In sum, the role of the reward system in fairness-related decision-making is complex and deserves further inquiry.

The Link Between Theory of Mind and Fairness

Humans may have an intrinsic need for justice (Decety & Yoder, 2015), but can also act strategically in social interactions (Lee & Seo, 2016). One core ability underlying such strategic choices is that of theory of mind, i.e., the skill of maintaining a mental model of others' minds. Brain systems that facilitate theory of mind, such as medial prefrontal cortex (MPFC; Denny, Kober, Wager, & Ochsner, 2012; Van Overwalle & Baetens, 2009) and the temporoparietal junction (TPJ), have often been mentioned in the context of economic games, and their role in fairness-related decisions is potentially important.

The medial PFC is proposed to integrate emotional, deliberative, and social information (Amodio & Frith, 2006), especially when social interests are in conflict with self-interest (Koban, Pichon, & Vuilleumier, 2014). Indeed, in the UG, the mPFC plays a crucial role in rejecting offers. By comparing how people play for themselves versus play for others, Corradi-Dell'Acqua et al. (2013) showed that people reject unfair offers equally often, but recruit the mPFC more strongly when playing for themselves. Importantly, the insula shows a similar response in both situations. Civai, Miniussi, and Rumiati (2015) expand on this finding by

manipulating mPFC activity using tDCS, demonstrating a causal role: when playing for oneself, decreasing mPFC activity using cathodal stimulation leads to fewer unfair offers being rejected; however, when playing for a third party, the same stimulation does not affect the rejections of unfair offers, but instead leads to more fair offers being rejected. Together these findings suggest that the insular cortex evaluates the fairness of the allocation, while the mPFC integrates this with the direct impact for oneself.

Hutcherson et al. (2015) let participant play a DG as the proposers, and found that TPJ and vmPFC signals correlated with the value for the other. Since the vmPFC activity was also correlated with the value for oneself, they proposed that the TPJ represents the valuation for the other, while the vmPFC integrates this information with the amount for the self. This interpretation is in line with extensive work in nonsocial decision-making, where the vmPFC seems to integrate value-information of different choice options (Hare, Camerer, & Rangel, 2009; Kable & Glimcher, 2009).

In a recent study of third-party punishment, Feng et al. (2016) compared participants' willingness to punish when they were either alone or as part of a larger group of (potential) third-party players. They found that participants punished more when alone and that in the group condition, the dmPFC activity modulated the activity in vmPFC and AI.

In the Trust Game, multiple studies have found medial prefrontal cortex (MPFC) to be more active when trustees defected than when they reciprocated (Chang et al., 2011; Van Baar et al., 2016; van den Bos, van Dijk, Westenberg, Rombouts, & Crone, 2009; van den Bos et al., 2011). This may mean that trustees process the other's mental state when they decide to behave unfairly. In line with this interpretation, Van Baar et al. (2016) observed increased activity in posterior superior temporal sulcus (pSTS), another important region for theory of mind, when participants did not reciprocate trust. On the other hand, increased activity in TPJ was found by Chang et al. (2011) when participants reciprocated. We can, therefore, not simply state that the theory of mind network contributes either positively or negatively to fair behavior.

It may prove more fruitful to investigate not simply brain activity but rather functional and effective connectivity between brain regions. If the activity in two brain regions is strongly correlated, they may be influencing one another; if the strength of this correlation changes with task demands, the two regions are said to be "effectively" connected (Friston et al., 1997). When investigating the neural signals from the trustee through this lens, Van Baar et al. (2016) found that functional connectivity between TPJ (theory of mind) and VMPFC (valuation) is stronger in guilt-averse trustees than in inequity-averse subjects. That is, there were trustees who appeared to behave perfectly fairly, yet reached that fair behavior by reasoning only from the investor's expectations and not from their own norms about fair behavior. These participants showed strong functional connections between the theory of mind and valuation systems, whereas other participants, who made their decisions based on their own fairness norms, did not have these functional connections. In line with this "individual differences"

interpretation, Van den Bos et al. (2009) found that the right TPJ and precuneus were more responsive to defection in participants who were, in general, more prosocial. Just like the salience network, therefore, the theory of mind network may be flexibly activated during reciprocity decisions based on the personal preferences of the trustee. One should therefore be mindful of such personal differences in social preferences when studying the neural correlates of fairness.

Influencing Fairness Using Neuropharmacology

One final avenue for studying fairness-related behavior is via the use of pharmacological manipulations. By administering hormones like oxytocin and testosterone, or using procedures like acute tryptophan depletion, researchers are able to directly affect the nervous system and observe the behavioral outcomes.

The influence of several neuromodulators has been investigated in the context of the Ultimatum Game. In a series of studies, Crockett and colleagues studied how serotonin influences the behavior of the responder. Specifically, Crockett, Clark, Tabibnia, Lieberman, and Robbins (2008) showed that people with lower serotonin levels reject more unfair offers, independent of the stake size. Importantly, the manipulation of serotonin levels did not affect self-reported mood, nor what proportion of the stake participant considered a fair split. However, those participants for whom lower serotonin levels led to more rejections also became more impatient, as measured using a temporal discounting task in which participants have to choose between a lower reward sooner (impatient choice) and a larger reward later (patient choice) (Crockett, Clark, Lieberman, Tabibnia, & Robbins, 2010). In a follow-up study, Crockett, Clark, Hauser, and Robbins (2010) now increased serotonin levels with citalopram. Using the same paradigm with variable stake sizes, they found that increased levels of serotonin reduced rejection rates, without affecting fairness perceptions or self-reported mood. Based on additional tests, the authors proposed that serotonin might modulate how likely one is to cause harm to others. Finally, Crockett et al. (2013) combined these procedures with fMRI. The neuroimaging results showed that the activity in the dorsal striatum correlated with increased rejection rates under decreased levels of serotonin. These findings are indeed consistent with the interpretation that serotonin modulates the willingness to punish unfair behavior, without affecting the perception of fairness itself.

Other neuromodulators which have been proposed to play a role in social decision-making include testosterone and oxytocin. However, the relationship here with fairness and reciprocity is less clear. Increasing testosterone levels in women leads them to propose higher offers in the Ultimatum Game (Eisenegger, Naef, Snozzi, Heinrichs, & Fehr, 2010). However, this might be due to increased concerns for social status (Eisenegger, Haushofer, & Fehr, 2011), and not a concern for fairness in itself. The latter would imply that responders should reject unfair offers at a

greater rate as well. However, there does not seem to be an effect of testosterone on the responder's behavior in the UG (Cueva et al., 2016; Zethraeus et al., 2009). Additionally, oxytocin has been linked to trust. Early studies (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005) showed that oxytocin increases transfers by investors in a Trust Game, or whether they adapted their investments after receiving feedback that the trustee did not reciprocate (Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008). However, these findings have not been consistently replicated (Nave, Camerer, & McCullough, 2015).

Conclusion

As we have attempted to demonstrate, cognitive neuroscience can provide important biological constraints on the processes involved in decisions involving fairness, 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 studies is innovative, and combining insights from Psychology, Neuroscience, and Economics has the potential to greatly increase our knowledge about the psychological and neural basis of fairness. 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 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. Further, 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. Firstly, it ensures that models of fairness are formally described, as opposed to the ad-hoc models that have been typically proposed. 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 fairness and equity.

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. More comprehensive knowledge of people's attitudes to fairness could usefully be employed to inform how policy decisions are taken, for example in relation to tax compliance, environmental behavior, and legal judgments. 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 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 actions of fairness and unfairness can assist in designing policy to better achieve desired societal aims.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268–277. http://doi.org/10.1038/nrn1884.

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

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76*(1), 412–427. http://doi.org/10.1016/j.neuroimage.2013.02.063.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences of the United States of America, 107*(50), 21767–21772. http://doi.org/10.1073/pnas.0908104107.

Battigalli, P., Dufwenberg, M., & Smith, A. (2015). *Frustration & anger in games*. Working paper (pp. 1–44). http://doi.org/10.13140/RG.2.1.3418.4403.

Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., & Fehr, E. (2009). The neural circuitry of a broken promise. *Neuron, 64*(5), 756–770. http://doi.org/10.1016/j.neuron.2009.11.017.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*(4), 639–650. http://doi.org/10.1016/j.neuron.2008.04.009.

Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., & Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience, 14*(11), 1468–1474. http://doi.org/10.1038/nn.2933.

Bereczkei, T., Deak, A., Papp, P., Perlaki, G., & Orsi, G. (2013). Neural correlates of Machiavellian strategies in a social dilemma task. *Brain and Cognition, 82*(1), 108–116. http://doi.org/10.1016/j.bandc.2013.02.012.

Bereczkei, T., Papp, P., Kincses, P., Bodrogi, B., Perlaki, G., Orsi, G., & Deak, A. (2015). The neural basis of the Machiavellians' decision making in fair and unfair situations. *Brain and Cognition, 98*, 53–64. http://doi.org/10.1016/j.bandc.2015.05.006.

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

Cáceda, R., James, G. A., Gutman, D. A., & Kilts, C. D. (2015). Organization of intrinsic functional brain connectivity predicts decisions to reciprocate social behavior. *Behavioural Brain Research, 292*, 478–483. http://doi.org/10.1016/j.bbr.2015.07.008.

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

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

Chang, L. J., Smith, A., Dufwenberg, M., & Sanfey, A. G. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron, 70*(3), 560–572. http://doi.org/10.1016/j.neuron.2011.02.056.

Civai, C., Corradi-Dell'Acqua, C., Gamer, M., & Rumiati, R. I. (2010). Are irrational reactions to unfairness truly emotionally-driven? Dissociated behavioural and emotional responses in the Ultimatum Game task. *Cognition, 114*(1), 89-95. http://doi.org/10.1016/j.cognition.2009.09.001.

Civai, C., Crescentini, C., Rustichini, A., & Rumiati, R. I. (2012). Equality versus self-interest in the brain: Differential roles of anterior insula and medial prefrontal cortex. *NeuroImage, 62*(1), 102–112. http://doi.org/10.1016/j.neuroimage.2012.04.037.

Civai, C., Miniussi, C., & Rumiati, R. I. (2015). Medial prefrontal cortex reacts to unfairness if this damages the self: A tDCS study. *Social Cognitive and Affective Neuroscience, 10*(8), 1054–1060. http://doi.org/10.1093/scan/nsu154.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2013). Disentangling self- and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Social Cognitive and Affective Neuroscience, 8*(4), 424–431. http://doi.org/10.1093/scan/nss014.

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., Dolan, R. J., Öhman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience, 7*(2), 189–195. http://doi.org/10.1038/nn1176.

Crockett, M. J., Apergis-Schoute, A., Herrmann, B., Lieberman, M. D., Muller, U., Robbins, T. W., & Clark, L. (2013). Serotonin modulates striatal responses to fairness and retaliation in humans. *Journal of Neuroscience, 33*(8), 3505–3513. http://doi.org/10.1523/JNEUROSCI.2761-12.2013.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences of the United States of America, 107*(40), 17433–17438. http://doi.org/10.1073/pnas.1009396107.

Crockett, M. J., Clark, L., Lieberman, M. D., Tabibnia, G., & Robbins, T. W. (2010). Impulsive choice and altruistic punishment are correlated and increase in tandem with serotonin depletion. *Emotion, 10*(6), 855–862. http://doi.org/10.1037/a0019861.

Crockett, M. J., Clark, L., Tabibnia, G., Lieberman, M. D., & Robbins, T. W. (2008). Serotonin modulates behavioral reactions to unfairness. *Science, 320*(5884), 1739. http://doi.org/10.1126/science.1155577.

Cueva, C., Roberts, R. E., Spencer, T. J., Rani, N., Tempest, M., Tobler, P. N., … Rustichini, A. (2016). Testosterone administration does not affect men's rejections of low ultimatum game offers or aggressive mood. *Hormones*. http://doi.org/10.1016/j.surfcoat.2016.08.074.

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience, 3*(10), 1049–1056. http://doi.org/10.1038/79871.

Decety, J., & Yoder, K. J. (2015). Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social Neuroscience, 919*(January), 1–14. http://doi.org/10.1080/17470919.2015.1029593.

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

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742–1752. http://doi.org/10.1162/jocn_a_00233.

Eisenegger, C., Haushofer, J., & Fehr, E. (2011). The role of testosterone in social interaction. *Trends in Cognitive Sciences, 15*(6), 263–271. http://doi.org/10.1016/j.tics.2011.04.008.

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature, 463*(7279), 356–359. http://doi.org/10.1038/nature08711.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience, 8*(11), 1481–1490. http://doi.org/10.1038/nn1579.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience, 6*(October), 148. http://doi.org/10.3389/fnins.2012.00148.

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y. J., & Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: A functional magnetic resonance imaging effective connectivity study. *Human Brain Mapping, 37*(2), 663–677. http://doi.org/10.1002/hbm.23057.

Feng, C., Luo, Y. J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human Brain Mapping, 36*(2), 591–602. http://doi.org/10.1002/hbm.22649.

Fehr, E., & Krajbich, I. (2013). Social preferences and the brain. In *Neuroeconomics: Decision making and the brain* (2nd ed.). Amsterdam: Elsevier. http://doi.org/10.1016/B978-0-12-416008-8.00011-5.

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

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

Gabay, A. S., Radua, J., Kempton, M. J., & Mehta, M. A. (2014). The ultimatum game and the brain: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews, 47*, 549–558. http://doi.org/10.1016/j.neubiorev.2014.10.014.

Grecucci, A., Giorgetta, C., Bonini, N., & Sanfey, A. G. (2013). Reappraising social emotions: The role of inferior frontal gyrus, temporo-parietal junction and insula in interpersonal emotion regulation. *Frontiers in Human Neuroscience, 7*(September), 523. http://doi.org/10.3389/fnhum.2013.00523.

Grecucci, A., Giorgetta, C., Van't Wout, M., Bonini, N., & Sanfey, A. G. (2013). Reappraising the ultimatum: An fMRI study of emotion regulation and decision making. *Cerebral Cortex, 23*(2), 399–410. http://doi.org/10.1093/cercor/bhs028.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367–388. http://doi.org/10.1016/0167-2681(82)90011-7.

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. http://doi.org/10.1126/science.1140738.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*, 646–648. http://doi.org/10.1126/science.1168450.

Harlé, K. M., Chang, L. J., van't Wout, M., & Sanfey, A. G. (2012). The neural mechanisms of affect infusion in social economic decision-making: A mediating role of the anterior insula. *NeuroImage, 61*(1), 32–40. http://doi.org/10.1016/j.neuroimage.2012.02.027.

Haruno, M., & Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature Neuroscience, 13*(2), 160–161. http://doi.org/10.1038/nn.2468.

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. http://doi.org/10.1126/science.1153651.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron, 87*(2), 451–462. http://doi.org/10.1016/j.neuron.2015.06.031.

Iyer, M. B., Schleper, N., & Wassermann, E. M. (2003). Priming stimulation enhances the depressant effect of low-frequency repetitive transcranial magnetic stimulation. *Journal of Neuroscience, 23*(34), 10867–10872.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*(2), 284–294. http://doi.org/10.1016/j.neuron.2008.03.020.

Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). tDCS polarity effects in motor and cognitive domains: A meta-analytical review. *Experimental Brain Research, 216*, 1–10.

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: Consensus and controversy. *Neuron, 63*(6), 733–745. http://doi.org/10.1016/j.neuron.2009.09.003.

Kirk, U., Downar, J., & Montague, P. R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimatum game. *Frontiers in Neuroscience, 5*(April), 1–11. http://doi.org/10.3389/fnins.2011.00049.

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., & Fehr, E. (2008). Studying the neurobiology of social interaction with transcranial direct current stimulation—The example of punishing unfairness. *Cerebral Cortex, 18*(9), 1987–1990. http://doi.org/10.1093/cercor/bhm237.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science, 314*(5800), 829–832. http://doi.org/10.1126/science.1129156.

Koban, L., Pichon, S., & Vuilleumier, P. (2014). Responses of medial and ventrolateral prefrontal cortex to interpersonal conflict for resources. *Social Cognitive and Affective Neuroscience, 9*(5), 561–569. http://doi.org/10.1093/scan/nst020.

Kolling, N., Behrens, T. E. J., Wittmann, M. K., & Rushworth, M. F. S. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology, 37*, 36–43. http://doi.org/10.1016/j.conb.2015.12.007.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673–677. http://doi.org/10.1038/nature03701.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications, 6*(May), 7455. http://doi.org/10.1038/ncomms8455.

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

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202. http://doi.org/10.1146/annurev.neuro.24.1.167.

Miniussi, C., Harris, J. A., & Ruzzoli, M. (2013). Modelling non-invasive brain stimulation in cognitive neuroscience. *Neuroscience Biobehavioural Review, 37*, 1702–1712.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A.J., Schweizer, S., Frith, C.D., & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science, 324*(5929), 900-900. http://doi.org/10.1126/science.1170539.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science, 10*(6), 772–789. http://doi.org/10.1177/1745691615600138.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12*(7), 265–272. http://doi.org/10.1016/j.tics.2008.03.006.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., … David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389*(October), 495–498. http://doi.org/10.1038/39051.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*(9), 1192–1206. http://doi.org/10.1177/0956797616654455.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427–430. http://doi.org/10.1038/nature11467.

Ratcliff, R., & Mckoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549–562. http://doi.org/10.1038/nrn3776.

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science, 342*(6157), 482–484. http://doi.org/10.1126/science.1241399.

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 (New York, N.Y.), 300*(5626), 1755–1758. http://doi.org/10.1126/science.1082976.

Sanfey, A. G., Stallen, M., & Chang, L. J. (2014). Norms and expectations in social decision-making. *Trends in Cognitive Sciences, 18*(4), 172–174. http://doi.org/10.1016/j.tics.2014.01.011.

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

Schultz, W. (1999). The reward signal of midbrain dopamine neurons. *Physiology, 14*(6), 249-255.

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., … Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *Journal of Neuroscience, 27*(9), 2349–2356. http://doi.org/10.1523/JNEUROSCI.5587-06.2007.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences, 13*(8), 334–340. http://doi.org/10.1016/j.tics.2009.05.001.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences, 27*(3), 161–168. http://doi.org/10.1016/j.tins.2004.01.006.

Stagg, C. J., & Nitsche, M. A. (2011). Physiological basis of transcranial direct current stimulation. *The Neuroscientist, 17*(1), 37–53. http://doi.org/10.1177/1073858410386614.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature, 463*(7284), 1089–1091. http://doi.org/10.1038/nature08785.

Van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2016). *Separating guilt aversion and inequity aversion in Trust Game reciprocity*. Poster presented at the Annual Meeting of the Society for Neuroeconomics, Berlin, Germany.

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A. R. B., & Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Social Cognitive and Affective Neuroscience, 4*(3), 294–304. http://doi.org/10.1093/scan/nsp009.

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A. R. B., & Crone, E. A. (2011). Changing brains, changing perspectives: The neurocognitive development of reciprocity. *Psychological Science, 22*(1), 60–70. http://doi.org/10.1177/0956797610391102.

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

van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport, 16*(16), 1849–1852. http://doi.org/10.1097/01.wnr.0000183907.08149.14.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience, 33*(3), 1099–1108. http://doi.org/10.1523/JNEUROSCI.1642-12.2013.

Zethraeus, N., Kocoska-Maras, L., Ellingsen, T., von Schoultz, B., Hirschberg, A. L., & Johannesson, M. (2009). A randomized trial of the effect of estrogen and testosterone on economic behavior. *Proceedings of the National Academy of Sciences of the United States of America, 106*(16), 6535–6538. http://doi.org/http://dx.doi.org/10.1073/pnas.0812757106.