Neuroscience of Moral Decision Making

Yang Hu[a,b,*], **Xiaoxue Gao**[a,b,*], **Hongbo Yu**[c], **Zhewen He**[b,d], **and Xiaolin Zhou**[a,b,e], [a] School of Psychology and Cognitive Science, East China Normal University, Shanghai, People's Republic of China; [b] School of Psychological and Cognitive Sciences, Peking University, Beijing, People's Republic of China; [c] Department of Psychology, University of California, Santa Barbara, CA, United States; [d] Division of Biosciences, University College London, London, United Kingdom; and [e] PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, People's Republic of China

© 2022 Elsevier Ltd. All rights reserved.

Introduction

As moral agents, we human beings are equipped with the capability to make judgments about the moral appropriateness of the other's behaviors (i.e., moral judgment) (Baron, 2014; Malle et al., 2014; Wojciszke et al., 2015) and henceforth to form the beliefs about the moral character of the other (i.e., moral inference) (Everett et al., 2016; Kleiman-Weiner et al., 2017; Knobe, 2010). Moreover, we are able to handle situations in which we need to trade off our own profits against the other's welfare (i.e., moral decision making) (Andreoni, 1990; Batson et al., 1981; Charness and Rabin, 2002; Fehr and Schmidt, 1999; Gneezy, 2005; Rand et al., 2012; Shalvi et al., 2015).

Given the complexity and multifaceted nature of morality, questions such as "what is morality" and "how the moral domain is defined" are very often raised by philosophers, social scientists (especially psychologists) and the general public. Rather than coming up with a comprehensive and commonly agreed definition of morality (or moral domain), researchers often define morality from different angles (Bartels et al., 2014; Crockett, 2013; Haidt, 2007). Two of those are commonly adopted. The first approach highlights the compliance with norms, which describes expectancies, beliefs, and rules of how people should (not) act in all or certain cases (e.g., fairness, honesty, justice) (Graham et al., 2011, 2013). The other addresses the concern for the other's welfare, typically accompanied by the cost of personal interests (e.g., relieving the other's suffering) (Cushman, 2015). Accordingly, immorality refers to the behaviors violating norms and causing harm to others' welfare (typically accompanied by the increase of personal profits), which intrinsically deserve blame and punishment and can elicit outrage as well as other negative affect (Schein and Gray, 2018). Notably, these two definitions are not mutually exclusive and can be seen as different sides of the same coin, because most moral rules concern others' interests (despite some exceptions that do not explicitly involve others' welfare, such as purity).

How does our brain enable (im)morality? This is one of the most crucial research questions at the intersection of several fields, such as ethics, social psychology, and cognitive neuroscience. With the rapid development of non-invasive brain imaging techniques (e.g., functional magnetic resonance imaging, fMRI) in the past two decades, there has been a large body of literature investigating the neurobiological basis of human morality (Dolan, 1999; Eres et al., 2017; Fede and Kiehl, 2019; Garrigan et al., 2016; Moll et al., 2005). In brief, most of these studies have focused on uncovering the neural underpinnings of moral judgment (Greene, 2015; Killen and Smetana, 2008; Liao, 2016). While these studies have made indelible contributions to our knowledge of the moral brain, the design properties and analytical approaches of these studies potentially preclude us from a deeper understanding of how real-life moral decisions are made at the neurobiological level. One of the properties is that stimuli commonly adopted in these studies consist of vignettes inspired by moral philosophy describing certain hypothetical moral-laden scenarios. Decisions made in these studies do not lead to any real consequences. However, recent evidence has clearly shown that people's (im)moral behaviors (FeldmanHall et al.,

[*] These authors equally contributed to this work.

https://doi.org/10.1016/B978-0-12-819641-0.00151-1

Encyclopedia of Behavioral Neuroscience, Second Edition, 2021, 481–495

2012) and the underlying neural activation patterns (FeldmanHall et al., 2012; Gospic et al., 2013) are different in hypothetical versus real contexts. Moreover, most of these previous studies were not designed to provide a mechanistic account for the moral behaviors, namely the computation our brains perform to transform the input information (e.g., components of moral dilemmas) into behavioral outputs (e.g., moral judgments and decisions; but see Crockett, 2016; Yu et al., 2019).

In this article, we provide an overview of the latest progress in the field of moral neuroscience, with a specific highlight on human fMRI studies investigating the neural substrates of moral decision-making. To distinguish the current article from previous review articles (Forbes and Grafman, 2010; Garrigan et al., 2016; Greene, 2015; Moll et al., 2008; Moll et al., 2005) and meta-analyses. (Eres et al., 2017; Fede and Kiehl, 2019) that are mainly based on traditional moral neuroscience studies, we will mainly consider studies adopting interactive games which are usually based on incentivized economic paradigms. In these tasks, individuals are required to trade off their own profits against others' welfares (or certain moral principles) or to interact with real persons, and their decisions will bring real consequences. Notably, a fair amount of them carries out the approach of computational modeling, which can specify the latent variables involved in the neurocomputational process during the decision period in certain morally relevant contexts (Charpentier and O'Doherty, 2018; Crockett, 2016; Cushman and Gershman, 2019; Hackel and Amodio, 2018; Konovalov et al., 2018) under the general framework of value-based decision making.

Moral Decision-Making in the Brain: A Multi-Stage Framework

We make moral decisions in everyday life. For example, how would you decide when facing the conflict between receiving illicit money and sticking to the bottom line of being an honest person? To uphold the inner conscience by forgoing personal gains, or to succumb to material interests at the cost of moral value? A recent theory in neuroeconomics has offered a computational account of how people make such moral decisions. Essentially, it assumes that these decisions are made by computing a subjective value for all the potential actions (or options) available on a commeasurable scale and then executing the one with the highest value (Levy and Glimcher, 2012; Padoa-Schioppa, 2011). Such computational process can be decomposed into three stages which recruit several neural networks (Platt and Plassmann, 2014; Rangel et al., 2008; Ruff and Fehr, 2014) (see **Fig. 1**). Stage 1 focuses on the

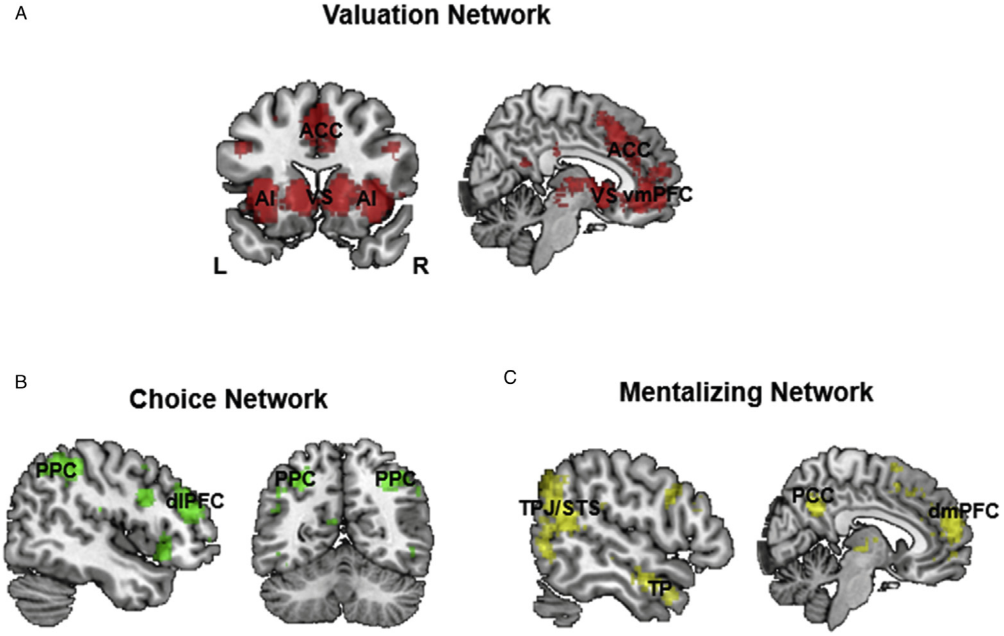

Fig. 1 Neural Networks Engaged in Value-Based Decision-Making. (A) **The valuation network.** This network includes regions encoding positive and negative value of stimuli (i.e., VS and AI), and regions involved in value integration and comparison (i.e., vmPFC and ACC). These regions mainly function in Stage 1 (value-based choice); some of the regions are also involved in Stage 2 (outcome evaluation; e.g., VS and AI) and Stage 3 (calculation of prediction error for optimization of later choices; e.g., VS). (B) **The choice network.** This network includes regions that transform decision values to choice behaviors (i.e., dlPFC, PPC), which mainly function in Stage 1. (C) **The mentalizing network.** This network mainly contains regions involved in inferring the other's intention (e.g., dmPFC, TP, TPJ/STS, PCC), which influences the evaluation of contexts before making choices (Stage 1). Notably, these networks are also found to be critically involved in moral decision-making. *Abbreviations*: L = left, R = right; ACC = anterior cingulate cortex, AI = anterior insula, dlPFC = dorsolateral prefrontal cortex, dmPFC = dorsomedial prefrontal cortex, PCC = posterior cingulate cortex, PPC = posterior parietal cortex, STS = superior temporal sulcus, TP = temporal pole, TPJ = temporo-parietal junction, vmPFC = ventromedial prefrontal cortex, VS = ventral striatum. All brain maps are defined based on the uniformity test meta-analytical maps generated from the Neurosynth database (https://neurosynth.org/analyses/terms/), using "value", "choice", and "mentalizing" as the key terms respectively.

process of value-based choice. Specifically, the decision-maker is supposed to represent multiple attributes with regard to each option (stimulus); this representation is often supported by the reward circuitry (e.g., ventral striatum [VS], including Nucleus Accumbens [NAcc]) (Fareri and Delgado, 2014; Haber and Knutson, 2009) and the network encoding negative information (e.g., anterior insula [AI]) (Namkung et al., 2017). In some complex scenarios, the decision-maker also needs to evaluate the other's intention via recruitment of the mentalizing network (e.g., temporoparietal junction [TPJ] and dorsomedial prefrontal cortex [dmPFC]) (Schaafsma et al., 2014; Schurz et al., 2014). Then the integrated subjective value (SV) of each option is computed and compared against each other on a common scale, potentially via the ventral medial prefrontal cortex (vmPFC) and the anterior cingulate cortex (ACC) (Bartra et al., 2013; Kolling et al., 2016; Levy and Glimcher, 2012). This is followed by a stimulus-action value transformation enabling choice selection (Rangel and Hare, 2010). At the neural level, the frontoparietal network, typically consisting of the dorsolateral prefrontal cortex (dlPFC) and the posterior parietal cortex (PPC) (Domenech et al., 2017), converts such value signal into a choice and finally a motor command, which recruits the motor areas to implement the action (Hare et al., 2011; Rangel and Hare, 2010). Stage 2 mainly involves the evaluation of the outcome brought by certain acts (e.g., a reward/punishment feedback) via the value encoding system (e.g., VS, amygdala, AI). In a dynamic environment, the decision-maker is required to compute the disparity between the expected and the actual outcomes, forming the so-called "prediction error (PE)", which is typically reflected in (but not limited to) the VS (Schultz and Wolfram, 2015). In Stage 3, the decision-maker is assumed to use such PE signals to optimize future decisions (Sutton and Barto, 2018). Notably, Stages 2 and 3 take place in more complex decision contexts, such as learning (e.g., inferring others' moral characters based on the observation of their behaviors).

In the remaining part of this section, we will introduce recent advancements in the neuroscience of moral decision-making within this framework. Because most of the studies mentioned below adopted a task that requires the tradeoff between personal profits and various moral costs, the current article focuses on the neurocomputational mechanisms underlying value-based choice in moral contexts (Stage 1), mainly covering the topics such as *harm* (e.g., harming others for personal gains), *help* (e.g., helping others in need or donating to a charity at the cost of personal gains) *(un)fairness* (e.g., preferring selfish or generous resource distribution) *(dis)honest* (e.g., lying for personal gains), and *betrayal* (e.g., breaking a promise for personal gains). After that, we introduce the existing neural evidence about how decision-makers respond to given the outcome of the other's behavior (Stage 2). We will also briefly discuss the emerging evidence regarding the morality-relevant learning process (Stage 3).

Value-Based Choice

Harm

Harm is considered as one of the core components (or even the only core component) of morality (Graham et al., 2013; Haidt, 2008; Schein and Gray, 2018). Supporting this claim, previous studies have shown that people take advantage of the cue of the other's suffering to distinguish immorality from unconventional behaviors (Hauser et al., 2010; Turiel et al., 1987) and they universally regard harm avoidance as a critical moral principle (Gert, 2004). Even some non-human primates have been shown to exhibit an aversion to profiting themselves at the cost of harming conspecific partners (Masserman et al., 1964).

To investigate harm-based moral decision-making in laboratory settings, researchers have designed behavioral assays in which participants trade off personal monetary profits against physically harming others (Crockett et al., 2014; FeldmanHall et al., 2012). Leveraging such a behavioral paradigm, Crockett and colleagues examined the computational mechanisms and individual differences underlying harm-based moral decision-making (Crockett et al., 2014, 2015). Participants in these studies were requested to voluntarily decide between two options consisting of different amounts of monetary reward and different numbers of painful electric shocks. Critically, the investigators manipulated the recipient of these painful shocks (self vs. other) while always keeping the participants as the beneficiary. Combining computational modeling with choice behaviors, they were able to quantify the latent parameter, namely the harm aversion (defined as the reluctance to cause harm for personal gains), which determines the computational process underlying such decision-making. Model-based analyses across several studies surprisingly revealed that participants displayed a higher level of harm aversion for others than for themselves. A follow-up fMRI study further uncovered the neural mechanisms underlying such "hyperaltruistic" behaviors (Crockett et al., 2017). Specifically, reduced money-sensitive signals of profiting from harming others (vs. oneself) in the lateral prefrontal cortex (lPFC) and dorsal striatum was positively correlated with the individual differences in hyperaltruism (i.e., the differential harm aversion for others than for oneself), indicating that morality originates from a devaluation of ill-gotten profits.

Another line of research stems from studies on proactive aggression, i.e., behaviors deliberately aiming to achieve personal gains (or goals) by planned attacks that cause physical or psychological harm on others (Anderson and Bushman, 2002; Wrangham, 2018). To our knowledge, three studies have so far explored the neural basis of proactive aggression through non-invasive brain stimulation techniques, with a focus on the role of dlPFC, but with mixed results. In an earlier transcranial magnetic stimulation (TMS) study (Perach-Barzilay et al., 2013), participants first received an inhibitory continuous theta-burst magnetic stimulation (cTBS) on dlPFC of either hemisphere. They then participated in a point subtraction task where the aggressive behaviors could be specifically measured by times of button press to cause monetary loss of a fictitious gender-matched partner. Importantly, proactive aggression was defined as aggressive behaviors only when these behaviors were not preceded by the partner's provocative act in previous trials. Results showed that compared with the ones after inhibition of the right dlPFC through cTBS, proactive aggressive responses increased after inhibition of the left dlPFC, suggesting a hemispheric asymmetry in dlPFC modulating proactive aggression. A later study (Dambacher et al., 2015) revealed that after the right dlPFC activity was enhanced by anodal transcranial direct current stimulation (tDCS), male participants displayed less proactive aggression (i.e., the intensity of noise administered to punish the partner), which was measured by the no-provocation trials in the Taylor reaction-time aggression paradigm (Taylor, 1967). However, a recent tDCS study showed that enhancing the right dlPFC activity merely reduced the intention to commit aggressive behaviors in hypothetical scenarios rather than influencing the real proactive aggressive behaviors (Choy et al., 2018). Given the mixed results, more studies are needed to clarify the specific role of dlPFC in regulating proactive aggression.

Help

Helping behaviors reflect the care for the other's welfare at the cost of the helper's own interest, touching upon the key value of morality. One of the most representative helping behaviors is charity donation, which has been extensively investigated in the arena of moral neuroscience. In a pioneering fMRI study, participants decided whether to accept or oppose a proposal of donating to real charitable organizations with or without a personal cost. Results showed that the subgenual part of the ACC was more active when participants accepted the donation proposal than when they received a monetary reward for themselves. Moreover, the donation-related activity in the VS was positively correlated with inter-individual differences in the frequency of accepting the costly donations (Moll et al., 2006). Using a similar design, investigators also revealed a crucial role of reward-related circuitry (especially the VS) in charitable donations. For instance, an increased reward-related signal in the VS persisted even when such donations were mandatory (Harbaugh et al., 2007). A later study further revealed that such VS signal during the charitable decision period could be specifically enhanced by the presence of observers, accompanied by an increase in donation rates (Izuma et al., 2010).

To increase the external validity of the paradigm, Hare et al. (2010) modified the task such that the participants were required to indicate the exact amount of money they would like to donate to a charity, rather than making a binary choice. They found that the monetary amount of voluntary donation was encoded in the vmPFC. Interestingly, the functional connectivity between the vmPFC and the social brain network, including the right TPJ extending to pSTS and bilateral AI, was stronger when participants made donation choices than when they made purchasing decisions, suggesting a specific neural network underpinning the valuation of donation. The contribution of these regions to charitable decisions was also confirmed in a recent study using multivariate decoding techniques (Tusche et al., 2016).

Given these findings, researchers further explored whether and how neurocomputational mechanisms underlying charitable decisions differ from those of immoral choices, namely profiting oneself at the cost of moral values. To answer this question, Qu et al. (2019) established a novel task in which participants decided whether to accept or reject an offer involving either a monetary cost to oneself and an amount of money donating to a charity (a positive moral value) or a personal gain and an amount of money sending to a morally-negative social cause (a negative moral value). Surprisingly, they observed two separate valuation networks functioning for each decision context, with the bilateral caudate engaged in value computation for the charitable decisions and the AI along with the (left) dlPFC for the immoral choices. A separate TMS study uncovered the causal role of the right TPJ in resolving such context-dependent tradeoff between personal profits and moral values (Obeso et al., 2018). Combining multivariate analyses with clinical populations, a recent fMRI showed that the representation of moral contexts (i.e., weighing personal gains/losses against positive/negative moral values) in the right TPJ was selectively impaired in individuals with autism spectrum disorders, further identifying a specific role of the right TPJ in representing moral contexts (Hu et al., 2021).

Another task widely used for measuring altruism, a generalized form of helping, is the dictator game (DG) (Kahneman et al., 1986a,b). In this task, participants are endowed with an amount of money and could voluntarily decide how much to distribute it between themselves and a matched partner (this task is also adopted to investigate fairness, see the next section). One recent fMRI study adopted the modified version that the participants needed to decide between two pre-determined monetary distributions between themselves and another person in which the gains for each party vary independently (Hutcherson et al., 2015). They found that the VS encoded the personal profits, while the right TPJ encoded the other's gains. In addition, the vmPFC was involved in representing gains for both parties but was more sensitive to personal profits. In another study, the right TPJ was found to encode the egoism bias, measured by the difference between one's own reward and the social-distance-dependent other-regarding utilities, especially during generous choices (Strombach et al., 2015). This area's connectivity with the vmPFC was also enhanced when participants made generous choices.

Apart from charity donation and resource distribution, reducing the other's suffering at a cost to oneself also denotes the nature of helping behavior and is commonly seen in our daily life. To explore the neural substrates of such decisions, Feldmanhall et al. (2012) established a paradigm where participants in the MRI scanner decided how much money they would forgo to reduce the intensity of painful electrical shocks inflicted on an anonymous stranger. They found that the socio-affective neurocircuitry, consisting of bilateral TPJ, ACC, and amygdala, was strongly activated during the decision period, especially when the participants' decisions brought real consequences. In a follow-up fMRI study adopting the same behavioral paradigm, the authors provided direct neural evidence supporting the empathy-altruism hypothesis, stating that the trait empathic concern motivated costly helping behaviors by modulating the decision-relevant activity in the neural network crucial for social attachment and caregiving, including the ventral tegmental area and the subgenual part of ACC (FeldmanHall et al., 2015). A recent fMRI study extended these findings by using computational modeling. Model-based analyses suggest that individuals were likely to integrate their own monetary costs with non-linearly transformed recipients' benefits. At the neural level, it identified a functional dissociation of adjacent but different sub-regions within the AI for different processes underlying altruistic behaviors. While the dorsal AI was involved in the valuation of benefactors' costs, the ventral/middle AI, as empathy-related regions, reflected individual variations in valuating recipients' benefits (Hu et al., 2021). Another study investigated the costly helping behaviors in a virtual reality environment where participants decided whether to risk their own lives to save someone trapped in a building on fire in the virtual scenario, revealing an association between the anatomical structure of right AI and the decision to help (Patil et al., 2018).

Notably, all studies mentioned above assumed that helping behaviors would surely reduce the other's suffering, which was not always true in real life. To address this issue, a study combining both fMRI and tDCS techniques developed a new paradigm in which participants were asked to consider the probability of being punished (i.e., receiving a 1s noise administration) for both themselves (self-risk) and a partner in need (other-need) while deciding whether to help. At the neural level, the right dlPFC was shown to causally influence both the effect of self-risk and that of other-need on helping behaviors, whereas the right inferior parietal lobule (IPL) selectively modulated the other-need effect (Hu et al., 2018).

(Un)Fairness

When distributing resources between oneself and the other, individuals commonly prefer fairness as they dislike the difference between themselves and the other (Fehr and Schmidt, 1999). This inequity aversion can emerge not only when individuals receive less (i.e., disadvantageous inequity) than others, but also when individuals receive more (i.e., advantageous inequity) than others (Fehr and Schmidt, 1999). The distinction between these two types of inequity aversion has been demonstrated in different disciplines, providing clues for potentially differential neurocognitive mechanisms underpinning these two types of inequity aversion. For example, behavioral studies showed that individuals' responses to advantageous inequity are usually not as strong as the ones to disadvantageous inequity (Bechtel et al., 2018; Fehr and Schmidt, 1999; Loewenstein et al., 1989). While disadvantageous inequity aversion emerges at early stages of evolution and human development, advantageous inequity aversion has only been observed in chimpanzees (Brosnan and de Waal, 2014) and humans over eight years old, who are equipped with relatively mature social and cognitive control abilities (McAuliffe et al., 2017).

In a seminal study (Tricomi et al., 2010), participants evaluated monetary transfers from the experimenter to him/herself or to another person. The researchers found that the process of inequity aversion was associated with activity in the reward system that computes abstract subjective value (Bartra et al., 2013), including the ventral striatum and ventromedial prefrontal cortex. The activity of these areas was more responsive to transfers to others than to oneself in the "high-pay" participants, whereas the activity of such areas in the "low-pay" participants showed the opposite pattern, suggesting that the brain's reward circuitry is sensitive to both advantageous and disadvantageous inequality. A few studies (Güroğlu et al., 2014; Yu et al., 2014) investigated whether the processes of advantageous and disadvantageous inequity involve shared or distinct neural mechanisms using the dictator game, where participants distribute resources as a dictator (i.e., the dictator game, DG) (Kahneman et al., 1986a,b). For example, one study (Güroğlu et al., 2014) focused on participants' decisions to share in advantageous and disadvantageous contexts, in which participants were asked to choose between an equal split of money (e.g., 1 coin for self and 1 coin for other) and an advantageous split (e.g., 2 coins for self and 0 coin for other) or a disadvantageous split (e.g., 1 coin for self and 2 coins for other). The brain activity for choosing the advantageous options in contrast to the one for choosing the equal option was regarded as the neural correlates of advantageous inequity; similar analysis was also conducted for disadvantageous inequity. However, it is difficult to discern whether the brain activations revealed in these comparisons are driven by the "inequity," by the actual payoff, or by both. Gao et al. (2018) addressed this question by manipulating the social context in which the resource allocation occurs. By using computational models to characterize individual weights on inequity aversion (Fehr and Schmidt, 1999), they found that after causing pain to the co-player (i.e., the guilt context), participants cared more about the advantageous inequity and became more tolerant of the disadvantageous inequity, as compared to other conditions. fMRI results showed that the two types of inequity were associated with differential neurocognitive substrates, subserved by different brain regions, and in particular by the spatial gradient in insular activity. While the context-dependent processing of advantageous inequity was associated with the social- and mentalizing-related processes, involving the left anterior insula, right dlPFC, and dmPFC, the context-dependent processing of disadvantageous inequity was primarily associated with emotion- and conflict-related processes, involving the left posterior insula, right amygdala, and dACC. These results were consistent with a previous study investigating the structural basis of inequity aversion; the latter showed that the gray matter volume in TPJ, a region implicated in the mentalizing process, predicted altruism in advantageous situations but not in disadvantageous situations (Morishima et al., 2012). In sum, the evidence so far suggests that the process of inequity aversion involves not only the valuation system, but also the social cognition and emotional processing systems in the brain (Oberliessen and Kalenscher, 2019).

People sometimes also need to decide how to distribute resources between individuals who are strangers to them. Such distributions are often not limited to fairness concern but extend to other types of distributive justice. In an early study (Hsu et al., 2008), participants were asked to decide between two options concerning whether different amounts of meals would be taken away from either of two groups of orphans. Notably, these options were designed based on two justice principles, namely the equity rule (i.e., the smallest meal variances), and the efficiency rule (i.e., the largest total meals). Results showed that efficiency was encoded in the putamen, whereas inequity was represented in the insula. In a later study using a similar task (Kameda et al., 2016), participants decided between two monetary distributions among three other persons. Apart from the options with the feature of either equity or efficiency, this study additionally considered another justice principle, namely the max-min rule (i.e., maximizing the minimum of the payoffs). Results identified a critical role of the right TPJ in representing the minimum payoffs.

In everyday life, the violation of fairness norms can be accompanied by sanction threats, and norm compliance under such threat may have a distinct neural basis in comparison with voluntary norm compliance, i.e., compliance without sanction threat (Ruff et al., 2013). For instance, using the DG with and without sanction, Spitzer et al. (2007) showed that participants allocated more to the co-player (i.e., closer to fairness norm) when a monetary sanction threat was introduced. In parallel, neural activations in the lateral prefrontal cortex (lPFC) and lateral orbitofrontal cortex (lOFC) were enhanced in the sanction condition relative to the sanction-free condition. Ruff et al. (2013) combined this task with tDCS and demonstrated that modulations of the right lPFC function had opposite effects on voluntary norm compliance and norm compliance under sanction threat (see also Strang et al., 2014). However, studies using the trust game (TG) showed opposite results as the trustee returns less money to the investor when the investor imposes a punishment threat on the trustee, and decreased activations were observed in the lOFC and vmPFC when punishment threat was present (Li et al., 2009). A further study (Zhang et al., 2016) indicated that the discrepancy in previous evidence might arise from the intention behind the threat. In this study, participants divided an amount of money between themselves and a co-player. The co-player (intentionally) or a computer program (unintentionally) decided to retain or waive the right to punish the participant upon selfish distribution. As compared to the unintentional condition, participants allocated more when the co-player intentionally waived the power of punishment, but less when the co-player retained such power. The right lOFC showed greater activation when the co-player waived than when the computer waived or when the co-player retained the power. The functional connectivity between the right lOFC and the brain network associated with intention/mentalizing processing (e.g., dmPFC and TPJ) was associated with the allocation difference induced by intention. The role of lOFC in intention-based fairness norm compliance was further confirmed by the brain stimulation evidence, showing inhibition or activation of the right lOFC decreased or increased, respectively, the participants' reliance on the co-player's intention during monetary allocation (Yin et al., 2017; Zhang et al., 2016).

(Dis)Honesty

Honesty serves as the cornerstone of morality (Graham et al., 2011, 2013). However, individuals often break this moral rule and lie for various reasons, especially for their own profits (Bazerman and Gino, 2012). As one of the most popular topics in the field of social neuroscience, the neural basis of (dis)honesty (or deception) has been widely investigated using neuroimaging techniques, especially fMRI (Abe, 2009, 2011). Earlier studies usually adopted the paradigm in which participants were instructed to deceive in certain conditions. However, such instructed lie was considered to be divergent from the essence of (dis)honesty, as it failed to be distinguished from behaviors in need of executive control (Christ et al., 2008) which was underpinned by a separate neural network (Lisofsky et al., 2014; Yin et al., 2016).

Addressing this concern, a distinct stream of studies utilized more ecologically valid paradigms in which individuals could voluntarily decide when to tell a lie or to be honest, usually with the goal of profiting oneself. In a seminal study of the latter type (Greene and Paxton, 2009), participants in the MRI scanner were provided with opportunities to deceive the experimenters to earn a higher monetary payoff for themselves (i.e., reporting the side of a coin and winning a certain amount of money if they reported correctly; the coin task). Critically, they were aware that their (dis)honest behaviors were not recorded, which reduced to a large extent the social desirability effects (i.e., participants behave honestly because this would make them more socially acceptable) that could blur their true preferences. As a main result, dishonest individuals showed increased signals in the control-related network (especially bilateral dlPFC) during the decision period when they had the chance to lie, whereas no such activations were observed in the honest group. The causal role of dlPFC in modulating (dis)honest choices was confirmed in a later tDCS study (Maréchal et al., 2017) using a dice-rolling task where participants could misreport the outcome of a fair dice to be better off. Strikingly, individuals became more honest after the right dlPFC activity was enhanced by the anodal stimulation. However, this effect disappeared if dishonest behavior benefited another person, indicating a unique function of dlPFC in resolving the conflict between personal interests and honesty. A similar conclusion could also be drawn from a brain lesion study (Zhu et al., 2014) which recruited patients with lesions in the dlPFC along with healthy and lesion controls to participate in a message game. Specifically, two roles were included in this task, a sender and a receiver. The sender was presented with two options, comprising different monetary payoffs for both oneself and the receiver, and could send a false message deceiving the uninformed receiver to make more profits for oneself. Model-based analyses showed that lesions in dlPFC specifically reduced the effect of honesty concern on the parameter pitting personal profit against the other's welfare. Combining the same paradigm with fMRI, Volz et al. (2015) investigated the neural basis of a more complicated voluntary dishonesty, i.e., the behavior that conveys literally the truth but is intentionally expected to be perceived as a lie. Such sophisticated dishonesty via truth-telling, relative to the plain lies, elicited a stronger activity in the left part of TPJ, superior temporal gyrus (STG), and insular cortex.

Another question of interest is why dishonest behaviors vary hugely from person to person. One behavioral study employing the dice-rolling task, for example, found that around 40% of participants were completely honest whereas 20% of them always lied, with the rest falling in between (Fischbacher and Föllmi-Heusi, 2013). Is there any neural substrate sensitive to such individual differences in dishonesty? Yin and Weber (2018) directly examined this question in a fMRI study adopting a novel paradigm in which participants could benefit themselves by reporting incorrect responses but might also be punished (i.e., losing personal profits if being caught) with a certain probability. Individual differences in dishonest decisions were negatively correlated with the dishonesty-sensitive activity in the lateral prefrontal areas (e.g., dlPFC) and the left caudate. Moreover, the functional connectivity between these regions and the right AI, an area relevant to dishonest decisions, negatively correlated with the frequency of dishonest decisions. With an economic paradigm involving a conflict between honest costs and personal profits, another fMRI study showed consistent findings, namely the signal encoding the cost of truth-telling in the left dlPFC (along with dmPFC) positively predicted the individual differences in actual honest decisions (Dogan et al., 2016). The inter-individual difference in dishonesty could even be predicted by the reward signals in the bilateral VS in a separate task (Abe and Greene, 2014).

Despite diverse paradigms, a common aspect of the studies above is that they all focused on self-serving dishonesty (lies). However, there are also other forms of dishonesty which benefit other individuals even at the cost of the deceiver (Erat and Gneezy, 2012). For instance, doctors sometimes would hide the actual outcome of a certain disease to reduce the patient's anxiety, which in turn may facilitate the patient's recovery. Does such other-serving dishonesty share the same neural representation as self-serving dishonesty? If not, how do the neural mechanisms differ between the two forms of dishonesty? To our knowledge, Abe et al. (2014) first addressed these questions in a fMRI study in which participants were asked to decide whether they would lie in hypothetical life scenarios associated with either harmful or helpful outcomes. They found a stronger activity in the mentalizing network, especially the dorsomedial prefrontal cortex (dmPFC) and the right TPJ, when participants made harmful dishonest (vs. honest) decisions, whereas no such effect was detected in the helpful decisions. In a later study with a modified message game (Yin et al., 2017), researchers showed that compared with the self-serving dishonesty, the other-serving dishonesty (here refers to the lie benefiting a charity) showed reduced activity in the right AI. Moreover, the activity in AI specific to the other-serving dishonesty also positively correlated with the relevant behavioral index that measures the relative financial costs due to the other-serving honesty. However, similar results were not observed in another study adopting the coin task (Pornpattananangkul et al., 2018). Instead, the vmPFC along with its functional network with the dlPFC was commonly activated in both forms of dishonesty, whereas the striatum-middle MPFC coupling sensitive to individual differences distinguished the two forms of dishonesty. From a different angle, Garrett et al. (2016) revealed that only the intensity of self-serving dishonesty increased with time (i.e., escalation), accompanied by a time-dependent reduced amygdala activity. More intriguingly, such escalation of self-serving dishonesty could be explained by the adapted amygdala activity, indicating a critical role of the amygdala in supporting the gradual enhancement of the self-serving dishonesty.

Betrayal

Betrayal is widely seen in every aspect of our real life, ranging from the unfaithfulness in a marriage (e.g., as a husband), the disloyalty to a sports team (e.g., as a soccer fan), to the infidelity to a country (e.g., as an official). These behaviors of betrayal are commonly regarded as moral violations and pervasively unacceptable (Feldman et al., 2000), as they ubiquitously disobey the moral principles of maintaining an interpersonal relationship (Turiel, 1998) and cause intentional harm to other's well-being, particularly those that one has a trusting bond with (Rachman, 2010).

In the social/moral neuroscience literature, betrayal is usually operationalized as the return behavior in the trust game (Berg et al., 1995). The standard version of this game includes two roles, i.e., an investor and a trustee. The investor is initially endowed with a certain amount of money and then decides whether (and how much) to invest an anonymous trustee. The investment amount would be multiplied by some factor (often 3 or 4) and be sent to the trustee, who decides whether to return a certain proportion (e.g., 50%) of the received investment to the investor or to keep it to him/herself. Combining this paradigm with hyper-scanning fMRI, King-Casas et al. (2005) in a pioneering study investigated the neural processing of the investor-trustee dyad during the dynamic interaction. They showed that the signals in the caudate of the trustee could track the return behavior (i.e., the amount transferring back to the investor) depending on the intention of the investor (i.e., the amount giving to the trustee). More complex analyses further revealed that the peak activity in caudate shifted its temporal occurrence as the trustee formed the impression of the investor's reputation in time, indicating the involvement of the reinforcement learning in the social context. A more direct test of the betray-brain causal relationship came from a lesion study, which showed a decreased repayment when patients with a lesion in the vmPFC acted as trustees, relative to the lesion and healthy control groups (Moretto et al., 2013).

Later studies examined additional factors that potentially influence the neural activation related to the trustee's return behaviors, with a focus on the role of guilt, a negative emotional state elicited by the violation of social norms or personal standards (Haidt, 2003). For example, Chang et al. (2011) showed increased activity in the left NAcc when the trustee returned less than what the investor had expected. Interestingly, such betrayal-like NAcc signal was modulated by the trustee's degree of guilt sensitivity. Using both fMRI and tDCS in combination with computational models, investigators also revealed the crucial role of the right dlPFC in gating the trustees' level of guilt aversion (Nihonsugi et al., 2015). Taking a novel approach of inter-subject representational similarity analysis, a model-based fMRI study published recently further differentiated trustees employing different moral strategies according to the association between model-based parameters and decision-related neural patterns involving the contributions of dlPFC, vmPFC, ACC, and AI (van Baar et al., 2019). Other factors were also demonstrated to affect the betrayal of trustees as well as decision-related neural signals, including the threat of investors (Li et al., 2009), the benefit of betrayal (van Den Bos et al., 2009), and the developmental characteristics of the trustee (van den Bos et al., 2011).

Under some circumstances, betrayal involves breaking an explicit promise, which is often considered as a stronger violation of moral values given the key role of promise in facilitating cooperation (Ellingsen and Johannesson, 2004; Kerr and Kaufman-Gilliland, 1994) and enhancing trustworthiness (Blue et al., 2020; Charness and Dufwenberg, 2006; Ismayilov and Potters, 2015). Baumgartner et al. (2009) explored the neural foundation of promise-based betrayal by using a modified trust game in a fMRI study. Here, participants in the role of trustee, in half of the trials, were additionally asked to make a promise at the beginning whether they plan to send back half of the money to the paired investor for the next three trials. Categorizing participants into two groups based on the average return rates, this study found that the amygdala signal in the untrustworthy participants was stronger during the decision period in the promise (vs. no promise) condition than that in the trustworthy group. Moreover, the promise-specific neural activity in the frontoinsular cortex during the promise and anticipation period was negatively correlated with the return rate regardless of groups. A follow-up study further revealed that the resting-state activity of the left AI reflected by the electroencephalography (EEG) signals positively predicted inter-individual difference in the degree of betrayal (measured by the difference between the average rate of promise and the rate of return) (Baumgartner et al., 2013).

Outcome

In everyday life, individuals evaluate the outcomes of the other's moral decision and make corresponding behavioral responses, such as acting kindly to the other's helpful behavior and unkindly to the other's harmful behavior. This kind of reciprocal behaviors happens not only when interactions involve the individuals directly (direct reciprocity), but also when these acts have been directed not to us but to others (indirect reciprocity). Both direct and indirect reciprocity are vital for human cooperation, adaption, and survival (Nowak and Sigmund, 2005).

Direct Reciprocity

Positive Reciprocity

Previous neuroimaging studies have mainly focused on reciprocal behaviors in the contexts of trust (see the section **Betrayal** above) and favor-receiving. When receiving favors, individuals commonly feel grateful and are motivated to reciprocate the benefactor. Such motivation in gratitude has been emphasized as a core feature of this emotion (McCullough et al., 2001). Two studies have investigated the neural bases underlying gratitude-induced reciprocity in the favor-receiving context. In one study (Yu et al., 2017), participants played a multi-round interactive game where they received pain stimulation. In each round, the participant interacted with an anonymous co-player who either intentionally or unintentionally (i.e., determined by a computer program) bore part of the participant's pain; the participant could transfer monetary points to the co-player with the knowledge that the co-player was unaware of this transfer. Relative to unintentional help, intentional help led to higher reciprocity (money allocation) and higher activation in value-related structures such as the vmPFC. Moreover, the vmPFC activation was predictive of the individual differences in gratitude ratings and subsequent reciprocal behaviors. A follow-up study (Yu et al., 2018) further demonstrated that neural signals representing cognitive antecedents of gratitude (e.g., benefactor-cost and self-benefit), were passed to the vmPFC via effective connectivity, suggesting an integrative role of the vmPFC in generating gratitude. Moreover, participants who were most willing to translate their grateful feelings into actual reciprocation showed stronger responses in the gyral part of ACC to the benefactor's help.

Negative Reciprocity

A widely used behavioral task in the research of negative reciprocity is the ultimatum game (UG). In a typical UG, participants act as a responder and decide whether to accept a fair or unfair division of money suggested by a proposer (Sanfey et al., 2003). If the division is accepted, the money would be split as proposed; but if the division is rejected, neither one would receive anything. Participants commonly accepted offers when the divisions comply with the fairness norm (fair offers). Although participants could have obtained a certain amount of money by accepting the unfair offers, they rejected more offers (i.e., receiving nothing) as the extent of the proposer's norm violation increase (i.e., the offers become less fair), indicating the negative reciprocity and negative cost enforcement. In one line of research, neuroimaging studies using this task have consistently demonstrated the involvements of brain areas related to the initial evaluation of norm compliance/violation (Aoki et al., 2015; Feng et al., 2015; Gabay et al., 2014). Specifically, responders gave higher happiness ratings to more equal offers (Tabibnia et al., 2008); this observation was consistent with the greater responses in the vmPFC to fair (vs. unfair) offers, suggesting that the vmPFC contributed to the processing of the social rewards of fairness norm compliance (Baumgartner et al., 2011; Dawes et al., 2012; Tabibnia et al., 2008; Xiang et al., 2013). In contrast, compared with fair offers, unfair offers would activate the anterior insula, an area implicated in detecting norm violation (Cheng et al., 2017; Civai, 2013; Civai et al., 2012; Guo et al., 2013; Strobel et al., 2011; Xiang et al., 2013) or signaling emotional processing via representations of aversive internal states (Chang and Sanfey, 2011; Corradi-Dell'Acqua et al., 2012; Guo et al., 2013; Sanfey et al., 2003), and the amygdala, which was linked to signal negative emotional response to norm violation (Gospic et al., 2013; Haruno and Frith, 2010; Yu et al., 2014).

Another line of research revealed greater activations in brain regions related to the integration of social norms and economic self-interest in favor of flexible decision-making in the unfair condition as compared to the fair condition (Aoki et al., 2014; Feng et al., 2015; Gabay et al., 2014). Specifically, the unfairness-evoked aversive responses and the self-interest that would be obtained by acceptance contradict each other, resulting in a motivational conflict that was suggested to be monitored by the dACC (Fehr and Camerer, 2007; Sanfey et al., 2003). Neural evidence suggested two ways to resolve this conflict: first, the unfairness-evoked aversive responses may be suppressed, probably implemented by brain regions associated with emotion regulation such as the vlPFC and dmPFC, resulting in an increase in acceptance rates (Civai et al., 2012; Grecucci et al., 2012; Tabibnia et al., 2008). Second, the conflict may be resolved by inhibiting selfish motives to promote norm compliance; this would rely on the cognitive control functions in the right dlPFC (Knoch et al., 2006; Ruff et al., 2013; Zhu et al., 2014). In addition, it was shown that, as compared to the gain frame used in the traditional UG, participants were more likely to reject unfair offers in the loss frame, where the proposers proposed unfair offers to share the loss (Zhou and Wu, 2011). Neuroimaging data indicated that loss reduced the responsiveness of the dopamine system (ventral striatum) to fairness while enhancing the motivation to reject the offer. This process was complemented by increased responses of dlPFC to insultingly unfair offers (Guo et al., 2013; Wu et al., 2014).

Notably, the reciprocal behaviors in UG are based not only on the preference for fair outcomes (i.e., egalitarianism) but also on reciprocal considerations regarding the others' intentions (i.e., intention-based reciprocity) (Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk et al., 2003; Rabin, 1993; Zheng et al., 2014). For example, the same unfair offers are more likely to be accepted if the proposer demonstrates good intentions by choosing the inequitable division over an even more unfair division (Falk et al., 2003). This increase in acceptance rates is associated with activity in the anterior medial prefrontal cortex and the TPJ, implying that higher demands in moral mentalizing are required in social decision-making when the decision to reject could not be readily justified (Güroğlu et al., 2010). Moreover, a gradual shift in other-regarding preferences was observed from simple rule-based egalitarianism to complex intention-based reciprocity from early childhood to young adulthood (Sul et al., 2017). The preference shift was associated with cortical thinning of the dmPFC and posterior temporal cortex, which were involved in social inference as indicated by the meta-analytic reverse-inference analysis.

Moreover, Yu et al. (2015) investigated the neural substrates underlying the processing of both intention and consequence of the other's harm using an interactive game. In the task, participants interacted with anonymous co-players, who decided to deliver pain stimulation either to him/herself or to the paired participant to earn a monetary reward. In some cases, the decision was reversed by the computer. Unbeknownst to the co-player, the participant was then allowed to punish the co-player by reducing his/her monetary reward after seeing the co-player's intention. Behaviorally, the punishment was lower in the accidental condition (unintended harm relative to intended harm) but higher in the failed-attempt condition (intended no-harm relative to unintended no-harm). Neurally, the left amygdala was activated in conditions with blameworthy intention (i.e., intentional harm and failed attempt). The accidental (relative to intentional) harm activated the right TPJ and IFG, while the failed attempt (relative to genuine no-harm) activated the anterior insula and posterior IFG. Effective connectivity analysis revealed that in the unintentional conditions (i.e., accidental and failed attempt) the IFG received input from the TPJ and AI and sent regulatory signals to the amygdala. These findings demonstrate that the processing of intention may gate the emotional responses to transgression and regulate subsequent reactive punishment.

Indirect Reciprocity

Humans never only care about the consequences of other's (im)moral acts on themselves and pay it back accordingly. Rather, they are also concerned with the helpful or harmful acts that occur to others (e.g., "I won't scratch your back because you did not scratch his") or simply pass the behavior to someone else (e.g., "You scratch my back and I'll scratch others'"), which are formally called indirect reciprocity (Nowak and Sigmund, 2005). Third-party punishment (or reputation-based reciprocity) and pay-it-forward are regarded as the most representative types of indirect reciprocity.

Third-Party Punishment

Third-party punishment usually refers to the phenomenon that unaffected third-party bystanders are willing to sacrifice their own resources to punish the norm transgressor, who breaks a moral norm (e.g., fairness) or the law and harms the interest of another person (Fehr and Fischbacher, 2004). In other words, the third-party bystander makes the (punishment) decision based on the reputation of the first party. In the past decades, a fair amount of human neuroscience studies have interrogated the neural mechanisms of third-party punishment, mainly using fMRI and TMS techniques. One of the earliest studies has adopted the judicial-arbitration paradigm, in which participants (third-party deciders) in the scanner determined the punishment for criminals in a series of hypothetical scenarios varying in the perpetrators' responsibility and severity of the crime (Buckholtz et al., 2008). Results showed that the right dlPFC was selectively activated when participants decided the punishment levels in the scenarios in which the perpetrator was supposed to take full responsibility for the crime. In a follow-up TMS study, investigators confirmed the above findings by showing that only the punishment level, but not the blameworthiness judgment, toward the full-responsibility crime was reduced after the inhibition of the right dlPFC, indicating the causal role of dlPFC in regulating third-party punishment (Buckholtz et al., 2015). Consistently, a later fMRI study employing multivariate pattern analyses (MVPA) further demonstrated that the signals of right dlPFC could specifically distinguish the punishment level but not other components of the context (e.g., the consequence or the intention of the perpetrator's behavior) (Ginther et al., 2016). Another line of studies adopting the unfairness-based economic decision paradigms focused more on the potential factors that modulate the third-party punishment and its neural correlates, ranging from genetic/trait basis (Krueger et al., 2014; Strobel et al., 2011) to social levels (Baumgartner et al., 2012; Feng et al., 2016). Some recent studies also compared how the third-party punishment differs from the third-party compensation, another option to restore justice, at the neural level (Civai et al., 2019; David et al., 2017; Hu et al., 2016; Hu et al., 2015; Stallen et al., 2018).

Building upon these empirical findings, a recent review proposed a tentative neural-cognitive model describing the cognitive processes and neural substrates underlying the third-party punishment (Krueger and Hoffman, 2016). Briefly, the model assumed that the third party generates an aversive affective experience in response to the transgression of moral norms (e.g., unfairness), which is underpinned by regions including AI and amygdala. The mentalizing network, especially the TPJ, additionally represents the inferred intention of the norm violation. The vmPFC is in charge of integrating all the information into a single value signal, which is then converted to a punishment decision via the central-executive network (especially dlPFC and PCC).

Paying-It-Forward

Comparatively, much less neuroscience research has been done to unveil the neural mechanisms of paying-it-forward (PIF), known as a behavioral response directing to someone else rather than the person who brings benefits (or causes harm). To address this question, Hu et al. (2018) adopted a modified PIF paradigm where participants first received an offer from either a human partner or a computer, and then decided on one of the two options concerning extra amounts of money they would share with a third-party. The study showed that participants integrated the social impact of previous treatments (i.e., human vs. computer) into value signals computed by the vmPFC and the right TPJ, which guided subsequent decision-making. Moreover, Watanabe et al. (2014) directly compared the neural correlates of the two types of indirect reciprocity, showing that the reputation-based third-party behavior selectively recruited the precuneus, whereas the PIF behavior specifically activated the AI. These findings cohere with a previous study revealing that the resting-state brain activity in the left ventral AI (as well as other regions) was correlated with the PIF response (Wu et al., 2015). Together, these findings suggest that the AI is not only engaged in signaling social norm violation during UG but also recruited in guiding subsequent adaptive behaviors (e.g., PIF response).

Learning

In real life, we not only make moral choices in one shot, but often need to form and update our beliefs about the moral trait of others, thereby guiding how we should get along with them in the future (Siegel et al., 2018). Although a substantial amount of evidence has revealed the neurocomputational mechanisms underlying how people learn through feedbacks under the general framework of reinforcement learning (O'Doherty et al., 2017), the neural underpinnings through which we infer the moral character of other people are still poorly understood. To investigate this issue, Hackel et al. (2015) performed a fMRI study in which participants were asked to learn how generous an anonymous partner was via trial-and-error learning based on the proportion of resources shared by the partner. As a control condition, participants also needed to learn which slot machine earned themselves more. Model-based analyses revealed that participants relied more on generosity information than on reward value during the task. Trial-wise prediction error (PE) of both types of information was commonly encoded in the right VS. However, the generosity prediction error recruited an additional network in association with the formation of social impression, including the ventral lateral prefrontal cortex (vlPFC), IPL, PCC extending to precuneus, as well as the right TPJ. Another study with a similar learning paradigm also found a signal of generosity PE in the PCC/precuneus (Stanley, 2016). Furthermore, our ability to infer others' moral character (i.e., trustworthiness) could be generalized to new partners who resemble the previous ones in appearance, supported by the neural patterns of the amygdala and caudate selectively encoding the transfer of learned moral value (FeldmanHall et al., 2018).

Learning about others' moral traits is not the full picture. Sometimes we also need to learn for the sake of others' welfares. How does our brain represent PE driving such prosocial learning? Does it recruit the same neurocircuitry as the standard reinforcement learning with the goal of maximizing one's own profit? To answer these questions, Sul et al. (2015) adopted a modified two-armed bandit probabilistic learning task in which participants in the MRI scanner needed to learn which one of the two options had a higher (fixed) probability leading to a reward. Critically, participants learned to profit themselves in some cases but benefit a paired partner in other trials. The authors found a spatial gradient in the mPFC for the value signals of the chosen option, that is, the ventral parts of mPFC were more sensitive to the chosen value when learning for oneself, whereas the dorsal parts predominantly encoded the chosen value when learning for the partner. Splitting all participants into two groups based on the preference of social value orientation, this study further revealed that the prosocial individuals differentiated themselves from the selfish ones by exhibiting a stronger mPFC-striatum functional coupling when learning for others (vs. oneself). In a later fMRI study with a similar task (Lockwood et al., 2016), investigators showed that the PE in both types of learning was commonly encoded in the bilateral VS, whereas the PE signal in the subgenual part of ACC only existed in prosocial learning. Intriguingly, such PE signal biased toward prosocial learning was positively correlated with individual differences in empathy.

Open Questions and Future Directions

Several issues should be kept in mind for future studies to explore. To begin with, a large neural network has been shown to engage in moral decision-making given the findings mentioned above. Most of the brain regions, especially a wide range of prefrontal areas (e.g., vmPFC, dmPFC, dlPFC), AI, TPJ, PCC, and some subcortical clusters (e.g., VS, amygdala), are detected in previous meta-analytical studies on moral cognition, typically using moral judgment tasks (Bzdok et al., 2012; Eres et al., 2017; Fede and Kiehl, 2019). This network is also well-known for its contribution to non-moral decisions (Ruff and Fehr, 2014). However, activations of anatomically overlapping areas do not necessarily reflect the same psychological process. First, one region might contain distinct sub-areas or neural populations that serve different functions, which is difficult to distinguish using fMRI due to its inadequate spatial resolution. For instance, recent neurophysiological studies on monkeys have identified that the ACC sulcus (ACCs) tracks the reward for oneself, whereas the ACC gyrus (ACCg) is more selective in encoding the partner's reward (Lockwood et al., 2020). Moreover, the activation pattern of multiple neurons (voxels) within the same region might differ. As a prime example, dACC has been identified with distinct and non-transferable activation patterns between physical and social pain (Woo et al., 2014), challenging the well-established conclusion that this region commonly encodes both affective processes (Eisenberger et al., 2003). Thus, it remains to be clarified whether and how these regions (and the relevant functional connectivities) work differently during decision making than during other cognitive processes, and how they function distinctively during the decision-making period in moral and non-moral contexts.

A related issue concerns the diversity of (im)morality and the corresponding neural substrates. For instance, it is still unclear how we maintain morality merely at the cost of others and how the neural basis of such kind of moral decisions differs from those involving the trade-off between personal interest and the other's interest. Another interesting topic would be to investigate the neurocomputational mechanisms underlying moral decision-making in a collective setting. Moreover, our moral principles are not limited to the ones we discussed above; according to the pluralism of (im)morality, other dimensions such as purity (degradation) and respect (subversion) are also fundamental to morality. Despite numerous studies at the behavioral level, no study, as far as we know, has yet explored the neural substrates arbitrating decisions in these moral contexts.

The third issue is related to methodological approaches that should be taken to provide additional information from different viewpoints, thereby characterizing a panoramic view of the moral brain. Obviously, the current literature predominantly considers which parts of the brain (and the inter-regional connections) are associated with a specific form of moral decision using fMRI, supplemented by the causality methods such as brain lesion and non-invasive brain stimulation (e.g., TMS, tDCS). There have been several studies adopting the EEG technique (e.g., event-related potential, ERP) to explore the temporal features of moral decision making at the neural level, especially those related to (un)fairness (Alexopoulos et al., 2012; Boksem and De Cremer, 2010; Cui et al., 2019; Ma et al., 2015; Mothes et al., 2016; Qu et al., 2013; Sun et al., 2015; Wu et al., 2011; Wu et al., 2011; Yu et al., 2015). However, most of these studies concern with the evaluation of the resource distribution, rather than the decision *per se*. Future studies could incorporate more advanced EEG analyses and other techniques with a high temporal resolution (e.g., magnetoencephalography, MEG) and/or a high spatial resolution (e.g., intracranial EEG) to uncover the temporal-frequency characteristics underlying the process of moral decision making.

From a more realistic perspective, how neuroscience could inform our real-life moral behaviors is another promising but challenging direction guiding the future development of moral neuroscience. For instance, many real-life choices cannot be made individually but rather collectively (Black, 1948; Hwang and Lin, 2012), with no exception to the moral domain (e.g., a decision made by members of the board regarding whether to misreport the sewage discharge to profit the enterprise). Recently, there is a growing trend interrogating the neural mechanisms of such group-based decisions with multi-brain measures, especially via hyper-scanning functional near-infrared spectroscopy (fNIRS) (Ferrari and Quaresima, 2012; Quaresima and Ferrari, 2019). This technique, in combination with more ecologically valid paradigms, enables future investigation on how neural information would flow between and be coordinated among several persons while they are making a collective decision in the moral context.

In sum, this article provides an overview of the recent advances in neuroscientific studies of moral decision making, highlighting human fMRI research which adopted economically incentivized paradigms with real consequences. Although these findings have extended and deepened our knowledge about the neural mechanisms of moral decision making, our current understanding of the moral brain is still far from complete, thereby urgently requiring insights from future investigations addressing the issues above and other related issues with novel methodological approaches.

Acknowledgments

This work was supported by National Natural Science Foundation of China (31630034, 31900798, 71942001) and China Postdoctoral Science Foundation (2019M660007, 2019M650008).

References

Abe, N., 2009. The neurobiology of deception: evidence from neuroimaging and loss-of-function studies. Curr. Opin. Neurol. 22 (6), 594–600.
Abe, N., 2011. How the brain shapes deception: an integrated review of the literature. Neuroscientist 17 (5), 560–574.
Abe, N., Fujii, T., Ito, A., Ueno, A., Koseki, Y., Hashimoto, R., et al., 2014. The neural basis of dishonest decisions that serve to harm or help the target. Brain Cognit. 90, 41–49.
Abe, N., Greene, J.D., 2014. Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. J. Neurosci. 34 (32), 10564–10572.
Alexopoulos, J., Pfabigan, D.M., Lamm, C., Bauer, H., Fischmeister, F.P.S., 2012. Do we care about the powerless third? An ERP study of the three-person ultimatum game. Front. Hum. Neurosci. 6.
Anderson, C.A., Bushman, B.J., 2002. Human aggression. Annu. Rev. Psychol. 53.
Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. Econ. J. 100 (401), 464–477.
Aoki, R., Matsumoto, M., Yomogida, Y., Izuma, K., Murayama, K., Sugiura, A., et al., 2014. Social equality in the number of choice options is represented in the ventromedial prefrontal cortex. J. Neurosci. 34 (18), 6413–6421.
Aoki, R., Yomogida, Y., Matsumoto, K., 2015. The neural bases for valuing social equality. Neurosci. Res. 90, 33–40.
Baron, J., 2014. Moral judgment. In: Zamir, E., Teichman, D. (Eds.), The Oxford Handbook of Behavioral Economics and the Law. Oxford Handbooks, New York, pp. 61–89.
Bartels, D., Bauman, C., Cushman, F., Pizarro, D., McGraw, A.P., 2014. Moral judgment and decision making. In: Keren, G., Wu, G. (Eds.), The Wiley Blackwell Handbook of Judgment and Decision Making. Wiley, Chichester, UK.
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, 412–427.
Batson, C.D., Duncan, B.D., Ackerman, P., Buckley, T., Birch, K., 1981. Is empathic emotion a source of altruistic motivation? J. Pers. Soc. Psychol. 40 (2), 290.
Baumgartner, T., Fischbacher, U., Feierabend, A., Lutz, K., Fehr, E., 2009. The neural circuitry of a broken promise. Neuron 64 (5), 756–770.
Baumgartner, T., Gianotti, L.R., Knoch, D., 2013. Who is honest and why: baseline activation in anterior insula predicts inter-individual differences in deceptive behavior. Biol. Psychol. 94 (1), 192–197.
Baumgartner, T., Götte, L., Gügler, R., Fehr, E., 2012. The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. Hum. Brain Mapp. 33 (6), 1452–1469.
Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., Fehr, E., 2011. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. Nat. Neurosci. 14 (11), 1468–1474.
Bazerman, M.H., Gino, F., 2012. Behavioral ethics: toward a deeper understanding of moral judgment and dishonesty. Annu. Rev. Law Soc. Sci. 8, 85–104.
Bechtel, M.M., Liesch, R., Scheve, K.F., 2018. Inequality and redistribution behavior in a give-or-take game. Proc. Natl. Acad. Sci. U. S. A. 115 (14), 3611–3616.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Game. Econ. Behav. 10 (1), 122–142.
Black, D., 1948. On the rationale of group decision-making. J. Polit. Econ. 56 (1), 23–34.
Blue, P.R., Hu, J., Peng, L., Yu, H., Liu, H., Zhou, X., 2020. Whose promises are worth more? How social status affects trust in promises. Eur. J. Soc. Psychol. 50, 189–206.
Boksem, M.A., De Cremer, D., 2010. Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. Soc. Neurosci. 5 (1), 118–128.
Brosnan, S.F., de Waal, F.B., 2014. Evolution of responses to (un) fairness. Science 346 (6207), 1251776.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R., 2008. The neural correlates of third-party punishment. Neuron 60 (5), 930–940.

Buckholtz, J.W., Martin, J.W., Treadway, M.T., Jan, K., Zald, D.H., Jones, O., Marois, R., 2015. From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. Neuron 87 (6), 1369–1380.

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A.R., Langner, R., Eickhoff, S.B., 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct. Funct. 217 (4), 783–796.

Chang, L.J., Sanfey, A.G., 2011. Great expectations: neural computations underlying the use of social norms in decision-making. Soc. Cognit. Affect Neurosci. 8 (3), 277–284.

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.

Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 817–869.

Charpentier, C.J., O'Doherty, J.P., 2018. The application of computational models to social neuroscience: promises and pitfalls. Soc. Neurosci. 13 (6), 637–647.

Cheng, X., Zheng, L., Li, L., Zheng, Y., Guo, X., Yang, G., 2017. Anterior insula signals inequalities in a modified Ultimatum Game. Neuroscience 348, 126–134.

Choy, O., Raine, A., Hamilton, R.H., 2018. Stimulation of the prefrontal cortex reduces intentions to commit aggression: a randomized, double-blind, placebo-controlled, stratified, parallel-group trial. J. Neurosci. 38 (29), 6505–6512.

Christ, S.E., Van Essen, D.C., Watson, J.M., Brubaker, L.E., McDermott, K.B., 2008. The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. Cerebr. Cortex 19 (7), 1557–1566.

Civai, C., 2013. Rejecting unfairness: emotion-driven reaction or cognitive heuristic? Front. Hum. Neurosci. 7, 126.

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.

Civai, C., Huijsmans, I., Sanfey, A., 2019. Neurocognitive mechanisms of reactions to second- and third-party justice violations. Sci. Rep. 9 (9271), 1–11.

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R.I., Fink, G.R., 2012. Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. Soc. Cognit. Affect Neurosci. 8 (4), 424–431.

Crockett, M.J., 2013. Models of morality. Trends Cognit. Sci. 17 (8), 363–366.

Crockett, M.J., 2016. How formal models can illuminate mechanisms of moral judgment and decision making. Curr. Dir. Psychol. Sci. 25 (2), 85–90.

Crockett, M.J., Kurth-Nelson, Z., Siegel, J.Z., Dayan, P., Dolan, R.J., 2014. Harm to others outweighs harm to self in moral decision making. Proc. Natl. Acad. Sci. U. S. A. 111 (48), 17320–17325.

Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Dayan, P., Dolan, R.J., 2017. Moral transgressions corrupt neural representations of value. Nat. Neurosci. 20 (6), 879–885.

Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Ousdal, O.T., Story, G., Frieband, C., et al., 2015. Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. Curr. Biol. 25 (14), 1852–1859.

Cui, F., Wang, C., Cao, Q., Jiao, C., 2019. Social hierarchies in third-party punishment: a behavioral and ERP study. Biol. Psychol. 107722.

Cushman, F., 2015. From moral concern to moral constraint. Curr. Opin. Behav. Sci. 3, 58–62.

Cushman, F., Gershman, S., 2019. Editors' introduction: computational approaches to social cognition. Topics Cogn. Sci. 11 (2), 281–298.

Dambacher, F., Schuhmann, T., Lobbestael, J., Arntz, A., Brugman, S., Sack, A.T., 2015. Reducing proactive aggression through non-invasive brain stimulation. Soc. Cognit. Affect Neurosci. 10 (10), 1303–1309.

David, B., Hu, Y., Krüger, F., Weber, B., 2017. Other-regarding attention focus modulates third-party altruistic choice: an fMRI study. Sci. Rep. 7, 43024. https://doi.org/10.1038/srep43024.

Dawes, C.T., Loewen, P.J., Schreiber, D., Simmons, A.N., Flagan, T., McElreath, R., et al., 2012. Neural basis of egalitarian behavior. Proc. Natl. Acad. Sci. U. S. A. 109 (17), 6479–6483.

Dogan, A., Morishima, Y., Heise, F., Tanner, C., Gibson, R., Wagner, A.F., Tobler, P.N., 2016. Prefrontal connections express individual differences in intrinsic resistance to trading off honesty values against economic benefits. Sci. Rep. 6, 33263.

Dolan, R.J., 1999. On the neurology of morals. Nat. Neurosci. 2 (11), 927.

Domenech, P., Redouté, J., Koechlin, E., Dreher, J.-C., 2017. The neuro-computational architecture of value-based selection in the human brain. Cerebr. Cortex 28 (2), 585–601.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Game. Econ. Behav. 47 (2), 268–298.

Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. Econ. J. 114 (495), 397–420.

Eisenberger, N.I., Lieberman, M.D., Williams, K.D., 2003. Does rejection hurt? An fMRI study of social exclusion. Science 302 (5643), 290–292.

Erat, S., Gneezy, U., 2012. White lies. Manag. Sci. 58 (4), 723–733.

Eres, R., Louis, W.R., Molenberghs, P., 2017. Common and distinct neural networks involved in fMRI studies investigating morality: an ALE meta-analysis. Soc. Neurosci. 1–15.

Everett, J.A., Pizarro, D.A., Crockett, M., 2016. Inference of trustworthiness from intuitive moral judgments. J. Exp. Psychol. Gen. 145 (6), 772.

Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Econ. Inq. 41 (1), 20–26.

Fareri, D.S., Delgado, M.R., 2014. The importance of social rewards and social networks in the human brain. Neuroscientist 20 (4), 387–402.

Fede, S.J., Kiehl, K.A., 2019. Meta-analysis of the moral brain: patterns of neural engagement assessed using multilevel kernel density analysis. Brain Imag. Behav. 1–14.

Fehr, E., Camerer, C.F., 2007. Social neuroeconomics: the neural circuitry of social preferences. Trends Cognit. Sci. 11 (10), 419–427.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25 (2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 817–868.

Feldman, S.S., Cauffman, E., Jensen, L.A., Arnett, J.J., 2000. The (un) acceptability of betrayal: a study of college students' evaluations of sexual betrayal by a romantic partner and betrayal of a friend's confidence. J. Youth Adolesc. 29 (4), 499–523.

FeldmanHall, O., Dalgleish, T., Evans, D., Mobbs, D., 2015. Empathic concern drives costly altruism. Neuroimage 105, 347–356.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., Mobbs, D., 2012. Differential neural circuitry and self-interest in real vs hypothetical moral decisions. Soc. Cognit. Affect Neurosci. 7 (7), 743–751.

FeldmanHall, O., Dunsmoor, J.E., Tompary, A., Hunter, L.E., Todorov, A., Phelps, E.A., 2018. Stimulus generalization as a mechanism for learning to trust. Proc. Natl. Acad. Sci. U. S. A. 115 (7), E1690–E1697.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., Dalgleish, T., 2012. What we say and what we do: the relationship between real and hypothetical moral choices. Cognition 123 (3), 434–441.

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. Hum. Brain Mapp. 37 (2), 663–677.

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. Hum. Brain Mapp. 36 (2), 591–602.

Ferrari, M., Quaresima, V., 2012. A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. Neuroimage 63 (2), 921–935.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise—an experimental study on cheating. J. Eur. Econ. Assoc. 11 (3), 525–547.

Forbes, C.E., Grafman, J., 2010. The role of the human prefrontal cortex in social cognition and moral judgment. Annu. Rev. Neurosci. 33, 299–324.

Gabay, A.S., Radua, J., Kempton, M.J., Mehta, M.A., 2014. The Ultimatum Game and the brain: a meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. 47, 549–558.

Gao, X., Yu, H., Sáez, I., Blue, P.R., Zhu, L., Hsu, M., Zhou, X., 2018. Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. Proc. Natl. Acad. Sci. U. S. A. 115 (33), E7680–E7689.

Garrett, N., Lazzaro, S.C., Ariely, D., Sharot, T., 2016. The brain adapts to dishonesty. Nat. Neurosci. 19 (12), 1727.

Garrigan, B., Adlam, A.L., Langdon, P.E., 2016. The neural correlates of moral decision-making: a systematic review and meta-analysis of moral evaluations and response decision judgements. Brain Cognit. 108, 88–97.

Gert, B., 2004. Common Morality: Deciding what to Do. Oxford University Press.

Ginther, M.R., Bonnie, R.J., Hoffman, M.B., Shen, F.X., Simons, K.W., Jones, O.D., Marois, R., 2016. Parsing the behavioral and brain mechanisms of third-party punishment. J. Neurosci. 36 (36), 9420–9434.

Gneezy, U., 2005. Deception: the role of consequences. Am. Econ. Rev. 95 (1), 384–394.

Gospic, K., Sundberg, M., Maeder, J., Fransson, P., Petrovic, P., Isacsson, G., et al., 2013. Altruism costs—the cheap signal from amygdala. Soc. Cognit. Affect Neurosci. 9 (9), 1325–1332.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H., 2013. Moral foundations theory: the pragmatic validity of moral pluralism. Adv. Exp. Soc. Psychol. 47, 55–130.

Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Koleva, S., Ditto, P.H., 2011. Mapping the moral domain. J. Pers. Soc. Psychol. 101 (2), 366.

Grecucci, A., Giorgetta, C., Van't Wout, M., Bonini, N., Sanfey, A.G., 2012. Reappraising the ultimatum: an fMRI study of emotion regulation and decision making. Cerebr. Cortex 23 (2), 399–410.

Greene, J.D., 2015. The cognitive neuroscience of moral judgment and decision-making. In: Gazzaniga, M.S., Wheatley, T. (Eds.), The moral brain: A multidisciplinary perspective, pp. 197–220. Boston (Review).

Greene, J.D., Paxton, J.M., 2009. Patterns of neural activity associated with honest and dishonest moral decisions. Proc. Natl. Acad. Sci. U. S. A. 106 (30), 12506–12511.

Guo, X., Zheng, L., Zhu, L., Li, J., Wang, Q., Dienes, Z., Yang, Z., 2013. Increased neural responses to unfairness in a loss context. Neuroimage 77, 246–253.

Güroğlu, B., van den Bos, W., Rombouts, S.A., Crone, E.A., 2010. Unfair? It depends: neural correlates of fairness in social context. Soc. Cognit. Affect Neurosci. 5 (4), 414–423.

Güroğlu, B., Will, G.-J., Crone, E.A., 2014. Neural correlates of advantageous and disadvantageous inequity in sharing decisions. PLoS One 9 (9), e107996.

Haber, S.N., Knutson, B., 2009. The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology 35 (1), 4–26.

Hackel, L.M., Amodio, D.M., 2018. Computational neuroscience approaches to social cognition. Curr. Opin. Psychol. 24, 92–97.

Hackel, L.M., Doll, B.B., Amodio, D.M., 2015. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. Nat. Neurosci. 18 (9), 1233–1235.

Haidt, J., 2003. The moral emotions. In: Schrerer, K.R., Goldsmith, H.H. (Eds.), Handbook of Affective Sciences, vol. 11. Oxford University Press, Oxford, pp. 852–870.

Haidt, J., 2007. The new synthesis in moral psychology. Science 316 (5827), 998–1002.

Haidt, J., 2008. Morality. Perspect. Psychol. Sci. 3 (1), 65–72.

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.

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

Hare, T.A., Schultz, W., Camerer, C.F., O'Doherty, J.P., Rangel, A., 2011. Transformation of stimulus value signals into motor commands during simple choice. Proc. Natl. Acad. Sci. U. S. A. 108 (44), 18120–18125.

Haruno, M., Frith, C.D., 2010. Activity in the amygdala elicited by unfair divisions predicts social value orientation. Nat. Neurosci. 13 (2), 160–161.

Hauser, M., Lee, J., Huebner, B., 2010. The moral-conventional distinction in mature moral competence. J. Cognit. Cult. 10 (1–2), 1–26.

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.

Hu, J., Hu, Y., Li, Y., Zhou, X., 2021. Computational and neurobiological substrates of cost-benefit integration in altruistic helping decision. J. Neurosci. 41 (15), 3545–3561.

Hu, J., Li, Y., Yin, Y., Blue, P.R., Yu, H., Zhou, X., 2018. How do self-interest and other-need interact in the brain to determine altruistic behavior? Neuroimage 157, 598–611.

Hu, Y., He, L., Zhang, L., Wolk, T., Dreher, J.C., Weber, B., 2018. Spreading inequality: neural computations underlying paying-it-forward reciprocity. Soc. Cognit. Affect Neurosci. 13 (6), 578–589. https://doi.org/10.1093/scan/nsy040.

Hu, Y., Pereira, A.M., Gao, X., Campos, B.M., Derrington, E., Corgnet, B., et al., 2021. Right temporoparietal junction underlies avoidance of moral transgression in autism spectrum disorder. J. Neurosci. 41 (8), 1699–1715. https://doi.org/10.1523/JNEUROSCI.1237-20.2020.

Hu, Y., Scheele, D., Becker, B., Voos, G., David, B., Hurlemann, R., Weber, B., 2016. The effect of oxytocin on third-party altruistic decisions in unfair situations: an fMRI Study. Sci. Rep. 6, 20236. https://doi.org/10.1038/srep20236.

Hu, Y., Strang, S., Weber, B., 2015. Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. Front. Behav. Neurosci. 9, 24. https://doi.org/10.3389/fnbeh.2015.00024.

Hutcherson, C., Bushong, B., Rangel, A., 2015. A neurocomputational model of altruistic choice and its implications. Neuron 87 (2), 451–462.

Hwang, C.-L., Lin, M.-J., 2012. Group Decision Making under Multiple Criteria: Methods and Applications, vol. 281. Springer Science & Business Media.

Ismayilov, H., Potters, J.J.J., 2015. Promises as Commitments.

Izuma, K., Saito, D.N., Sadato, N., 2010. Processing of the incentive for social approval in the ventral striatum during charitable donation. J. Cognit. Neurosci. 22 (4), 621–631.

Kahneman, D., Knetsch, J.L., Thaler, R., 1986a. Fairness as a constraint on profit seeking: entitlements in the market. Am. Econ. Rev. 728–741.

Kahneman, D., Knetsch, J.L., Thaler, R.H., 1986b. Fairness and the assumptions of economics. J. Bus. S285–S300. https://doi.org/10.1086/296367.

Kameda, T., Inukai, K., Higuchi, S., Ogawa, A., Kim, H., Matsuda, T., Sakagami, M., 2016. Rawlsian maximin rule operates as a common cognitive anchor in distributive justice and risky decisions. Proc. Natl. Acad. Sci. U. S. A. 113 (42), 11817–11822.

Kerr, N.L., Kaufman-Gilliland, C.M., 1994. Communication, commitment, and cooperation in social dilemma. J. Pers. Soc. Psychol. 66 (3), 513.

Killen, M., Smetana, J., 2008. Moral judgment and moral neuroscience: intersections, definitions, and issues. Child Dev. Perspect. 2 (1), 1–6.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., Montague, P.R., 2005. Getting to know you: reputation and trust in a two-person economic exchange. Science 308 (5718), 78–83.

Kleiman-Weiner, M., Saxe, R., Tenenbaum, J.B., 2017. Learning a commonsense moral theory. Cognition 167, 107–123.

Knobe, J., 2010. Person as scientist, person as moralist. Behav. Brain Sci. 33 (4), 315–329.

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.

Kolling, N., Wittmann, M.K., Behrens, T.E., Boorman, E.D., Mars, R.B., Rushworth, M.F., 2016. Value, search, persistence and model updating in anterior cingulate cortex. Nat. Neurosci. 19 (10), 1280–1285.

Konovalov, A., Hu, J., Ruff, C.C., 2018. Neurocomputational approaches to social behavior. Curr. Opin. Psychol. 24, 41–47.

Krueger, F., Hoffman, M., 2016. The emerging neuroscience of third-party punishment. Trends Neurosci. 39 (8), 499–501.

Krueger, F., Hoffman, M., Walter, H., Grafman, J., 2014. An fMRI investigation of the effects of belief in free will on third-party punishment. Soc. Cognit. Affect Neurosci. nst092.

Levy, D.J., Glimcher, P.W., 2012. The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. 22 (6), 1027–1038.

Li, J., Xiao, E., Houser, D., Montague, P.R., 2009. Neural responses to sanction threats in two-party economic exchange. Proc. Natl. Acad. Sci. U. S. A. 106 (39), 16835–16840.

Liao, S.M., 2016. Morality and neuroscience: past and future. In: Liao, S.M. (Ed.), Moral Brains: The Neuroscience of Morality. Oxford University Press, New York, pp. 1–42.

Lisofsky, N., Kazzer, P., Heekeren, H.R., Prehn, K., 2014. Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. Neuropsychologia 61, 113–122.

Lockwood, P.L., Apps, M.A.J., Chang, S.W.C., 2020. Is there a 'social' brain? Implementations and algorithms. Trends Cogn. Sci. 24 (10), 802–813.

Lockwood, P.L., Apps, M.A., Valton, V., Viding, E., Roiser, J.P., 2016. Neurocomputational mechanisms of prosocial learning and links to empathy. Proc. Natl. Acad. Sci. U. S. A. 113 (35), 9763–9768.

Loewenstein, G.F., Thompson, L., Bazerman, M.H., 1989. Social utility and decision making in interpersonal contexts. J. Pers. Soc. Psychol. 57 (3), 426.

Ma, Q., Hu, Y., Jiang, S., Meng, L., 2015. The undermining effect of facial attractiveness on brain responses to fairness in the Ultimatum Game: an ERP study. Front. Neurosci. 9, 77.

Malle, B.F., Guglielmo, S., Monroe, A.E., 2014. A theory of blame. Psychol. Inq. 25 (2), 147–186.

Maréchal, M.A., Cohn, A., Ugazio, G., Ruff, C.C., 2017. Increasing honesty in humans with noninvasive brain stimulation. Proc. Natl. Acad. Sci. U. S. A. 114 (17), 4360–4364.

Masserman, J.H., Wechkin, S., Terris, W., 1964. "Altruistic" behavior in rhesus monkeys. Am. J. Psychiatr. 121 (6), 584–585.

McAuliffe, K., Blake, P.R., Steinbeis, N., Warneken, F., 2017. The developmental foundations of human fairness. Nat. Human Behav. 1 (2), 0042.

McCullough, M.E., Kilpatrick, S.D., Emmons, R.A., Larson, D.B., 2001. Is gratitude a moral affect? Psychol. Bull. 127 (2), 249.

Moll, J., De Oliveira-Souza, R., Zahn, R., 2008. The neural basis of moral cognition: sentiments, concepts, and values. Ann. N. Y. Acad. Sci. 1124 (1), 161–180.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., Grafman, J., 2006. Human fronto–mesolimbic networks guide decisions about charitable donation. Proc. Natl. Acad. Sci. U. S. A. 103 (42), 15623–15628.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., Grafman, J., 2005. The neural basis of human moral cognition. Nat. Rev. Neurosci. 6 (10), 799.

Moretto, G., Sellitto, M., di Pellegrino, G., 2013. Investment and repayment in a trust game after ventromedial prefrontal damage. Front. Hum. Neurosci. 7, 593.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C., Fehr, E., 2012. Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. Neuron 75 (1), 73–79.

Mothes, H., Enge, S., Strobel, A., 2016. The interplay between feedback-related negativity and individual differences in altruistic punishment: an EEG study. Cognit. Affect Behav. Neurosci. 16 (2), 276–288.

Namkung, H., Kim, S.H., Sawa, A., 2017. The insula: an underestimated brain area in clinical neuroscience, psychiatry, and neurology. Trends Neurosci. 40 (4), 200–207.

Nihonsugi, T., Ihara, A., Haruno, M., 2015. Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. J. Neurosci. 35 (8), 3412–3419.

Nowak, M.A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437 (7063), 1291–1298. https://doi.org/10.1038/nature04131.

O'Doherty, J.P., Cockburn, J., Pauli, W.M., 2017. Learning, reward, and decision making. Annu. Rev. Psychol. 68, 73–100.

Oberliessen, L., Kalenscher, T., 2019. Social and non-social mechanisms of inequity aversion in non-human animals. Front. Behav. Neurosci. 13, 133.

Obeso, I., Moisa, M., Ruff, C.C., Dreher, J.-C., 2018. A causal role for right temporo-parietal junction in signaling moral conflict. Elife 7, e40671.

Padoa-Schioppa, C., 2011. Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. 34, 333–359.

Patil, I., Zanon, M., Novembre, G., Zangrando, N., Chittaro, L., Silani, G., 2018. Neuroanatomical basis of concern-based altruism in virtual environment. Neuropsychologia 116, 34–43.

Perach-Barzilay, N., Tauber, A., Klein, E., Chistyakov, A., Ne'eman, R., Shamay-Tsoory, S., 2013. Asymmetry in the dorsolateral prefrontal cortex and aggressive behavior: a continuous theta-burst magnetic stimulation study. Soc. Neurosci. 8 (2), 178–188.

Platt, M.L., Plassmann, H., 2014. Multistage valuation signals and common neural currencies. Neuroeconomics 237–258.

Pornpattananangkul, N., Zhen, S., Yu, R., 2018. Common and distinct neural correlates of self-serving and prosocial dishonesty. Hum. Brain Mapp. 39 (7), 3086–3103.

Qu, C., Météreau, E., Butera, L., Villeval, M.C., Dreher, J.-C., 2019. Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. PLoS Biol. 17 (6), e3000283.

Qu, C., Wang, Y., Huang, Y., 2013. Social exclusion modulates fairness consideration in the ultimatum game: an ERP study. Front. Hum. Neurosci. 7, 505.

Quaresima, V., Ferrari, M., 2019. Functional near-infrared spectroscopy (fNIRS) for assessing cerebral cortex function during human behavior in natural/social situations: a concise review. Organ. Res. Methods 22 (1), 46–68.

Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 1281–1302.

Rachman, S., 2010. Betrayal: a psychological analysis. Behav. Res. Ther. 48 (4), 304–311.

Rand, D.G., Greene, J.D., Nowak, M.A., 2012. Spontaneous giving and calculated greed. Nature 489 (7416), 427–430.

Rangel, A., Camerer, C., Montague, P.R., 2008. A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. 9 (7), 545.

Rangel, A., Hare, T., 2010. Neural computations associated with goal-directed choice. Curr. Opin. Neurobiol. 20 (2), 262–270.

Ruff, C.C., Fehr, E., 2014. The neurobiology of rewards and values in social decision making. Nat. Rev. Neurosci. 15 (8), 549–562.

Ruff, C.C., Ugazio, G., Fehr, E., 2013. Changing social norm compliance with noninvasive brain stimulation. Science 342 (6157), 482–484.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., Cohen, J.D., 2003. The neural basis of economic decision-making in the ultimatum game. Science 300 (5626), 1755–1758.

Schaafsma, S.M., Pfaff, D.W., Spunt, R.P., Adolphs, R., 2014. Deconstructing and reconstructing theory of mind. Trends Cognit. Sci. 19 (2), 65–72.

Schein, C., Gray, K., 2018. The theory of dyadic morality: reinventing moral judgment by redefining harm. Pers. Soc. Psychol. Rev. 22 (1), 32–70.

Schultz, Wolfram, 2015. Neuronal reward and decision signals: from theories to data. Physiol. Rev. 95 (3), 853.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34.

Shalvi, S., Gino, F., Barkan, R., Ayal, S., 2015. Self-serving justifications: doing wrong and feeling moral. Curr. Dir. Psychol. Sci. 24 (2), 125–130.

Siegel, J.Z., Mathys, C., Rutledge, R.B., Crockett, M.J., 2018. Beliefs about bad people are volatile. Nat. Human Behav. 2 (10), 750.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., Fehr, E., 2007. The neural signature of social norm compliance. Neuron 56 (1), 185–196.

Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C.K., Sanfey, A.G., 2018. Neurobiological mechanisms of responding to injustice. J. Neurosci. 38 (12), 2944–2954.

Stanley, D.A., 2016. Getting to know you: general and specific neural computations for learning about people. Soc. Cognit. Affect Neurosci. 11 (4), 525–536.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., Sack, A., 2014. Be nice if you have to-The neurobiological roots of strategic fairness. Soc. Cognit. Affect Neurosci. nsu114.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., Kirsch, P., 2011. Beyond revenge: neural and genetic bases of altruistic punishment. Neuroimage 54 (1), 671–680.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I.I., Tobler, P.N., Kalenscher, T., 2015. Social discounting involves modulation of neural value signals by temporoparietal junction. Proc. Natl. Acad. Sci. U. S. A. 112 (5), 1619–1624.

Sul, S., Güroğlu, B., Crone, E.A., Chang, L.J., 2017. Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. Sci. Rep. 7.

Sul, S., Tobler, P.N., Hein, G., Leiberg, S., Jung, D., Fehr, E., Kim, H., 2015. Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. Proc. Natl. Acad. Sci. U. S. A. 112 (25), 7851–7856.

Sun, L., Tan, P., Cheng, Y., Chen, J., Qu, C., 2015. The effect of altruistic tendency on fairness in third-party punishment. Front. Psychol. 6, 820.

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction, second ed. MIT Press.

Tabibnia, G., Satpute, A.B., Lieberman, M.D., 2008. The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychol. Sci. 19 (4), 339–347.

Taylor, S.P., 1967. Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. J. Pers. 35 (2), 297–310.

Tricomi, E., Rangel, A., Camerer, C.F., O'Doherty, J.P., 2010. Neural evidence for inequality-averse social preferences. Nature 463 (7284), 1089–1091.

Turiel, E., 1998. The development of morality. In: Eisenberg, N. (Ed.), Handbook of Child Psychology, vol. 3. University of California, Berkeley, CA, pp. 789–857.

Turiel, E., Killen, M., Helwig, C.C., 1987. Morality: its structure, functions, and vagaries. In: The Emergence of Morality in Young Children, pp. 155–243.

Tusche, A., Böckler, A., Kanske, P., Trautwein, F.-M., Singer, T., 2016. Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. J. Neurosci. 36 (17), 4719–4732.

van Baar, J.M., Chang, L.J., Sanfey, A.G., 2019. The computational and neural substrates of moral strategies in social decision-making. Nat. Commun. 10 (1), 1483.

van Den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S.A., Crone, E.A., 2009. What motivates repayment? Neural correlates of reciprocity in the Trust Game. Soc. Cognit. Affect Neurosci. 4 (3), 294–304.

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S.A., Crone, E.A., 2011. Changing brains, changing perspectives: the neurocognitive development of reciprocity. Psychol. Sci. 22 (1), 60–70.

Volz, K.G., Vogeley, K., Tittgemeyer, M., von Cramon, D.Y., Sutter, M., 2015. The neural basis of deception in strategic interactions. Front. Behav. Neurosci. 9, 27.

Watanabe, T., Takezawa, M., Nakawake, Y., Kunimatsu, A., Yamasue, H., Nakamura, M., et al., 2014. Two distinct neural mechanisms underlying indirect reciprocity. Proc. Natl. Acad. Sci. U. S. A. 111 (11), 3990–3995.

Wojciszke, B., Parzuchowski, M., Bocian, K., 2015. Moral judgments and impressions. Curr. Opin. Psychol. 6, 50–54.

Woo, C.-W., Koban, L., Kross, E., Lindquist, M.A., Banich, M.T., Ruzic, L., Andrews-Hanna, J.R., Wager, T.D., 2014. Separate neural representations for physical pain and social rejection. Nat. Communi. 5, 5380.

Wrangham, R.W., 2018. Two types of aggression in human evolution. Proc. Natl. Acad. Sci. U. S. A. 115 (2), 245–253.

Wu, Y., Leliveld, M.C., Zhou, X., 2011. Social distance modulates recipient's fairness consideration in the dictator game: an ERP study. Biol. Psychol. 88 (2–3), 253–262.

Wu, Y., Yu, H., Shen, B., Yu, R., Zhou, Z., Zhang, G., et al., 2014. Neural basis of increased costly norm enforcement under adversity. Soc. Cognit. Affect Neurosci. nst187.

Wu, Y., Zang, Y., Yuan, B., Tian, X., 2015. Neural correlates of decision making after unfair treatment. Front. Hum. Neurosci. 9.

Wu, Y., Zhou, Y., van Dijk, E., Leliveld, M.C., Zhou, X., 2011. Social comparison affects brain responses to fairness in asset division: an ERP study with the ultimatum game. Front. Hum. Neurosci. 5, 131.

Xiang, T., Lohrenz, T., Montague, P.R., 2013. Computational substrates of norms and their violations during social exchange. J. Neurosci. 33 (3), 1099–1108.

Yin, L., Hu, Y., Dynowski, D., Li, J., Weber, B., 2017. The good lies: altruistic goals modulate processing of deception in the anterior insula. Hum. Brain Mapp. 38 (7), 3675–3690. https://doi.org/10.1002/hbm.23623.

Yin, L., Reuter, M., Weber, B., 2016. Let the man choose what to do: neural correlates of spontaneous lying and truth-telling. Brain Cognit. 102, 13–25.

Yin, L., Weber, B., 2018. I lie, why don't you: neural mechanisms of individual differences in self-serving lying. Hum. Brain Mapp. 40 (4), 1–13. https://doi.org/10.1002/hbm.24432.

Yin, Y., Yu, H., Su, Z., Zhang, Y., Zhou, X., 2017. Lateral prefrontal/orbitofrontal cortex has different roles in norm compliance in gain and loss domains: a transcranial direct current stimulation study. Eur. J. Neurosci. 46 (5), 2088–2095.

Yu, H., Cai, Q., Shen, B., Gao, X., Zhou, X., 2017. Neural substrates and social consequences of interpersonal gratitude: intention matters. Emotion 17 (4), 589.

Yu, H., Gao, X., Zhou, Y., Zhou, X., 2018. Decomposing gratitude: representation and integration of cognitive antecedents of gratitude in the brain. J. Neurosci. 38 (21), 4886–4898.

Yu, H., Li, J., Zhou, X., 2015. Neural substrates of intention–consequence integration and its impact on reactive punishment in interpersonal transgression. J. Neurosci. 35 (12), 4917–4925.

Yu, H., Siegel, J.Z., Crockett, M.J., 2019. Modeling morality in 3-D: decision-making, judgment, and inference. Topics Cogn. Sci. 11 (2), 409–432.

Yu, R., Calder, A.J., Mobbs, D., 2014. Overlapping and distinct representations of advantageous and disadvantageous inequality. Hum. Brain Mapp. 35 (7), 3290–3301.

Yu, R., Hu, P., Zhang, P., 2015. Social distance and anonymity modulate fairness consideration: an ERP study. Sci. Rep. 5, 13452.

Zhang, Y., Yu, H., Yin, Y., Zhou, X., 2016. Intention modulates the effect of punishment threat in norm enforcement via the lateral orbitofrontal cortex. J. Neurosci. 36 (35), 9217–9226.

Zheng, L., Guo, X., Zhu, L., Li, J., Chen, L., Dienes, Z., 2014. Whether others were treated equally affects neural responses to unfairness in the Ultimatum Game. Soc. Cognit. Affect Neurosci. 10 (3), 461–466.

Zhou, X., Wu, Y., 2011. Sharing losses and sharing gains: increased demand for fairness under adversity. J. Exp. Soc. Psychol. 47, 582–588.

Zhu, L., Jenkins, A.C., Set, E., Scabini, D., Knight, R.T., Chiu, P.H., et al., 2014. Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. Nat. Neurosci. 17 (10), 1319–1321.