The Neural Basis of and a Common Neural Circuitry in Different Types of Pro-social Behavior

Jun Luo*

Neuro & Behavior EconLab, School of Economics, Center for Economic Behavior and Decision-Making, Zhejiang University of Finance & Economics, Hangzhou, China

Pro-social behaviors are voluntary behaviors that benefit other people or society as a whole, such as charitable donations, cooperation, trust, altruistic punishment, and fairness. These behaviors have been widely described through non self-interest decision-making in behavioral experimental studies and are thought to be increased by social preference motives. Importantly, recent studies using a combination of neuroimaging and brain stimulation, designed to reveal the neural mechanisms of pro-social behaviors, have found that a wide range of brain areas, specifically the prefrontal cortex, anterior insula, anterior cingulate cortex, and amygdala, are correlated or causally related with pro-social behaviors. In this review, we summarize the research on the neural basis of various kinds of pro-social behaviors and describe a common shared neural circuitry of these pro-social behaviors. We introduce several general ways in which experimental economics and neuroscience can be combined to develop important contributions to understanding social decision-making and pro-social behaviors. Future research should attempt to explore the neural circuitry between the frontal lobes and deeper brain areas.

Keywords: pro-social behaviors, neural basis, neural circuitry, functional magnetic resonance imaging, transcranial direct current stimulation

INTRODUCTION

Humans are the most successful species at restraining their self-interest motives, even in interactions with unfamiliar strangers, through the development and enforcement of social norms (Fehr and Fischbacher, 2003; Boyd and Richerson, 2005). This human behavioral feature is thought to be a social adaptation that underlies our evolutionary success (Hrdy, 2009; de Waal, 2010). Pro-social behaviors, in particular, play a crucial role in social life across many cultures (Henrich et al., 2001). They represent a broad category of acts that are defined by significant regions of society as generally beneficial to other people or one's group (Penner et al., 2005). Pro-social behavior involves trade-offs between our own well-being and the well-being of others, including a donation to charity, reciprocal exchange, interpersonal trust, mutual cooperation, costly punishment of norm violations. Pro-social behaviors that are exhibited in game tasks have been found and replicated under controlled environments in many behavioral experiments; players like to share wealth with strangers (Eckel and Grossman, 1996; Fehr and Fischbacher, 2003), punish defectors at a cost (Fehr and Gächter, 2002; Fehr and Fischbacher, 2003; Dawes et al., 2007), invest money in a stranger (Berg et al., 1995; Kosfeld et al., 2005) and reject unfair divisions of a sum of money (Güth et al., 1982; Camerer, 2003).

In this paper, we review studies on the neural activity of going against pure self-interest behaviors. This evidence is based on neuroimaging and brain stimulation approaches that provide a micro-foundation of pro-social behaviors with regard to the underlying neural networks. These studies that involve social preferences are based on neuroscientific methods that include the neural networks and motivational forces involved in charitable donations, rejections to unfair divisions, punishments for non-cooperation behavior at a cost, or decisions to trust in an investment game. The combination of economic game models with modern neuroscientific methods enables researchers to investigate the neural mechanisms of pro-social behaviors and to advance theoretical models of how we make decisions in a social context.

There has been a gradual appearance of studies that reveal the mechanisms of action of social preferences on the brain's reward system, the role that affective factors play in economic decisions, and the neural model of the capacity to infer an actor's mental state during a strategic game. According to these neuroscientific findings, we thus propose an integrated model for a common shared neural circuitry for various kinds of pro-social behaviors, involving the theory of mind network, the reward system, emotion-related brain regions and prefrontal cortical areas. Indeed, this review would be a fruitful starting point for future studies on a model of the neural circuitry involved in pro-social behaviors, by describing the relationship between the behavioral patterns of social preferences and the empirically verified parameters of the brain model. This will bring about an improved model of social decisions and a better understanding of the nature of pro-social behaviors.

EMPATHY/CHARITABLE GIVING

We can empathize with others, that is, understand and share their emotions, feelings, motivations without any exogenous emotional stimulation. This crucial phenomenon of human social interactions occurs in various situations. Prior work from cognitive and behavioral psychology reveals the complex emotion process of empathy, including cognitive appraisal, cognitive perspective taking, and affect sharing (Decety and Jackson, 2004; Lamm et al., 2007; Olsson and Ochsner, 2008; Hein and Singer, 2008).

In accordance with these studies, advances in neuroscience enable us to gain new insights into the neural basis of empathy (Eisenberg, 2000; Hoffman, 2001; Preston and de Waal, 2002; de Vignemont and Singer, 2006; Batson, 2009). First, neuroscientific experiments about empathy indicate that the same neural circuits underlying both affective and cognitive processes are activated when we have a feeling and when others have this feeling. Preston and de Waal (2002) proposed a neuroscientific model of empathy, which specifically states that attended perception to another person in an emotional state automatically activates the participant's representation of that state and that activation of these representations are associated with autonomic and somatic responses.

Moreover, imaging studies have also investigated the brain activity of empathic responses in the field of touch, smell, and pain. Wicker et al. (2003) have performed a functional magnetic resonance imaging (fMRI) study that reveal the same brain regions are activated when observing a facial expression of disgust and when inhaling disgusting odorants. Keysers et al. (2004) have found that there are similar neural mechanisms involved when participants are touched and when they observe someone else being touched by objects. Another study has assessed the brain activity associated with empathy for pain (Singer et al., 2004, 2006). They indicated that activity in the anterior cingulate cortex (ACC) and anterior insula (AI) was observed when participants either felt pain or observed pain in someone else. These brain areas compose the affective pain circuits that represent our responses to pain and our understanding of how others feel pain. Further studies have investigated the temporal dynamics of the neural mechanisms underlying empathy for pain using event-related brain potentials (ERPs). These results showed that the early and late responses to empathy are separately adjusted by the situational reality of the stimuli, and these results support the hypothesis that empathy for pain consists of early emotional sharing and later cognition evaluations (Fan and Han, 2008).

In addition, responses in DMPFC regions while mentalizing with others who have similar and dissimilar thoughts and beliefs have also been shown to predict empathy (Zaki et al., 2009; Majdandžić et al., 2016). Neuroimaging studies have also examined the relations between activation in specific brain areas related to social preferences and self-reported empathy and willingness to help (Tankersley et al., 2007; Mathur et al., 2010; Powers et al., 2015), and found the correlations between the reflexive engagement of neural mechanisms of mentalizing and altruistic behaviors for monetary allocation and time spent helping others (Waytz et al., 2012). In fact, additional research has also demonstrated that the brain activation of brain areas involved in empathy predicts pro-social behaviors toward social exclusion (Masten et al., 2011) and that such activation occurs when participants make decisions to donate money to their family members (Telzer et al., 2011); thus, the neural basis of empathy during in tasks involving charitable donations has received much attention.

A prior attempt on the neural basis of giving showed that the mesolimbic reward system, including ventral tegmental (VTA) and striatal areas were both engaged by receiving money and by anonymous donations to charitable organizations, suggesting that giving has its own reward (Moll et al., 2006). Further study has clarified that there are different neural mechanisms for purely altruistic and warm-glow motives for charitable giving (Harbaugh et al., 2007). To test these two motives, researchers have assessed fMRI while participants played a dictator game in which participants were required to make decisions about whether to give money to a charitable organization. All the participants were randomly assigned to mandatory and voluntary conditions. In the mandatory condition, participants observed money being transferred tax-like to a charitable organization. In the voluntary condition, subjects could make transfers voluntarily to the charity. Similar neural substrates linked to reward processing were elicited while participants received

money themselves, when they performed free transfers, and when they observed the charity receiving money. However, this neural activation was higher when charitable giving was voluntary rather than mandatory.

In another study, the motivational mechanisms of charitable giving were identified by multivariate decoding techniques (Tusche et al., 2016). Neural responses in the AI predicted affective empathy for beneficiaries, while temporoparietal junction (TPJ) activity was associated with the degree of cognitive perspective taking, suggesting that these distinct paths of social cognition and psychological mechanisms differentially lead to intraindividual and interindividual heterogeneities in charitable giving. Indeed, there was specific neural evidence of a correlation between individual differences in helpful decisions and the neural activation of AI, ACC, and TPJ (Greening et al., 2014), and neural mechanisms of individual differences in empathy and pro-social behaviors were further revealed by reinforcement learning theory (Lockwood et al., 2016). However, how affective empathy is linked to pro-social behaviors in charitable giving and the neural circuitry underlying empathy in terms of multi-faceted cognitive and emotional process remain poorly understand. Thus, one possible direction is to integrate various constructs of the neural mechanisms of empathy and provide connections between the neural responses to empathy and charitable giving in future studies.

FAIRNESS/INEQUITY AVERSION

People tend to helped those who helped them, and to hurt those who hurt them. Consequences that represent such preferences are called fairness equilibria (Rabin, 1993). This fairness effect has also been recognized in formal theory models of reciprocal fairness (Rabin, 1993) and inequity aversion (Fehr and Schmidt, 1999), both of which assume that there is a trade-off between fairness and individual benefits. To examine decisions about fairness, an ultimatum game (UG) has been proposed (Güth et al., 1982) involving strategic interaction behaviors. As the hypothesis of self-interest motivation, the responder in the UG should accept any non-zero offer from the other party. The proposer can expect this self-interest response, and then will give a smallest non-zero offer to responder.

However, a number of studies have found that offers are commonly around 50% of the sum amount no matter of the total monetary, and lower than 20% of the total offers have more than 50% probability of being rejected (Güth et al., 1982; Roth et al., 1991; Bolton and Zwick, 1995; Henrich et al., 2001). Strong evidence indicates that many subjects reject low offers from proposers in the UG (Henrich et al., 2001; Camerer, 2003). It is thus clear that the actual decisions in the game do not agree with the behaviors of the model predicted to be driven by self-interest motivation, and neuroscience research has begun to provide evidence for the mechanism underlying these decisions in an UG.

An fMRI study first investigated the neural basis of response decisions in an UG (Sanfey et al., 2003). They found that unfair proposals elicited neural activity in brain regions involved in both the processing of cognition (DLPFC) and emotion (bilateral AI); these areas showed greater activation with an unfair offer that was subsequently rejected, whereas a greater response was seen in the DLPFC when an unfair offer was accepted. Further, there was significantly stronger activity in the AI when a participant received an unfair offer from another human compared to the same offer from a computer partner. Finally, the unfair offer was also related to heightened activity in ACC, and may imply the conflict between cognitive and emotion process in the response decision-making for the unfair offer of UG. Thus, receiving unfair offers in an UG was weakly associated with increased activity in these brain areas (see Gabay et al., 2014; Feng et al., 2015 for meta-analyses). Indeed, activation of the AI region involved in emotional arousal and measured as an autonomic index of affective status, indicated that skin conductance responses were stronger for unfair offers and related to the rejection rate of unfair offers in an UG (van't Wout et al., 2006).

Compared to unfair offers in an UG, fair offers led to greater activation in the VMPFC region. Importantly, the choice to reject unfair transfers is associated with improved activity in the AI region (Tabibnia et al., 2008). The key role of the ventromedial prefrontal cortex (VMPFC) in response decisions involving fairness preferences of the UG is also supported by neural evidence (Koenigs and Tranel, 2007) that patients with brain injuries in the VMPFC reject unfair offers in the UG more frequently than healthy participants, implying that the cost of declining non-zero offers is of less concern in the response decisions of the UG when the VMPFC is damaged. An ERP study (Boksem and De Cremer, 2010) showed that medial frontal negativity amplitude was greater for unfair offers than fair offers. Moreover, this effect was shown to be the greatest for responders with high fairness concerns.

To distinguish the functions of different brain areas in response decision-making in an UG, Knoch et al. (2006) used repetitive transcranial magnetic stimulation (rTMS) to inhibit the activation of the right DLPFC (rDLPFC) when responders in an UG faced unfair offers and observed a reduction in responders' willingness to reject unfair offers from proposers, which suggests that participants are more unable to resist the temptation to accept unfair offers from partners. However, participants did not change their judgment for such offers to be unfair after receiving rTMS, which reveals that the rDLPFC is crucial in implementing fairness-related decisions. In terms of transfer decisions from a proposer, another rTMS study indicated that reducing the activity of the right lateral PFC (rLPFC) led to a significant decrease in transfers in the UG, but neither the expected rejection from responders nor the fairness judgments were changed by rTMS (Strang et al., 2014). To modulate the neural excitability (activate or reduce) of specific regions, Ruff et al. (2013) employed transcranial direct current stimulation (tDCS) to demonstrate whether fairness-related decisions in the UG rely causally on neural activation of the rLPFC region. This study revealed that anodal tDCS in rLPFC caused transfers improvement significantly while cathodal tDCS to the rLPFC decreased transfers in the UG compared to sham stimulation. Together, these results provide strong

causal evidence for the rLPFC in the implementation of fairness preference.

A pervasive notion in social science is that people have a preference to reduce inequality gaps in wealth distribution (Fehr and Schmidt, 1999). Studies have thus used inequality aversion to represent a fairness motive. To explore the tendency for inequity aversion in distributive decisions, participants performed in a distribution task (similar to UG) while scanning fMRI (Hsu et al., 2008). The experimental results suggested that the putamen encodes efficiency, whereas the insula represents inequity, and the caudate/septal subgenual area responds to a trade-off in efficiency and inequity. Strikingly, the choice about inequitable allocation was related to greater insula activity.

Neural evidence for preference of inequality aversion in distributive decision was also revealed by Tricomi et al. (2010). They have employed fMRI to demonstrate the existence of inequality aversion preferences in the brain. Inequality was created in experiments by recruiting pairs of participants and giving one of them an endowment. The participant who received the endowment showed greater neural reward activation while providing transfers to "other" rather than "self," whereas the participants who did not receive endowment showed a significantly greater activation in reward areas while providing transfers to "self" rather than "other." These results suggest that people are rewarded for reductions in the wealth gap, and the neural mechanisms of reward are strongly related to both advantageous and disadvantageous inequality. Civai et al. (2012) were more concerned with the differential roles of the AI and MPFC in equality versus self-interest in distributive decisions, especially for disadvantageous unequal offers and consequent rejections. The researchers found that the AI region was active during unequal offers, whereas the activity of the MPFC was negatively associated with rejection decisions. When inequity and efficiency were in conflict, participants showed greater activity in a simplified prefrontal network, including the rDLPFC, VMPFC, and the connectivity between them, according to fMRI signals (Baumgartner et al., 2011). Individual differences in inequity aversion were predicted by the blood-oxygen-level dependent (BOLD) signals of the amygdala (AMYG) during a resource-sharing task involving inequitable distributions to one's self and others (Haruno and Frith, 2010).

Taken together, social interactions with inequitable outcomes are linked to neural systems, including the AI, AMYG and prefrontal cortex, that are associated with affective and emotional signaling that alter distribution decisions by modulating fairness perceptions. In addition, inequity may induce a punishment action; thus, the neural networks implicated in inequity aversion could lead to the decision to punish at a cost to the punisher. On the other hand, the preference for inequity aversion may reflect how the neural processes that conform to inequity detection are influenced and further processed through emotional circuitry. Converging evidence indeed suggests that decision tasks related to inequity normally activate brain regions involved in affective processing. Further studies are needed to determine how these signals transform the decision to punish.

COSTLY PUNISHMENT

Across cultures, human always engage in individual costs in readiness to punish violators (Henrich et al., 2001; Fehr and Fischbacher, 2004; Bernhard et al., 2006), who propose an unfair offer during monetary allocation or take a self-interest strategy during a social exchange (Fehr and Gächter, 2002; Egas and Riedl, 2008). Why would humans punish defectors of universally maintained rules while diminishing their personal benefits? The view from evolutionary economics (Boyd et al., 2003; Bowles, 2009) indicates that human behavior in costly punishment has profound evolutionary foundation, and promoting pro-social behaviors, such as reciprocity and cooperation (Nakamaru and Iwasa, 2006; Rand et al., 2010; Rand and Nowak, 2013; Peysakhovich et al., 2014). These suggest that sanction at the cost of personal gain evolved as a spontaneous mechanism rather than as an intended or deliberate pattern; people thus feel satisfaction when punishing norm defectors. It is obvious that costly punishment brings a huge array of discusses about its behavioral mechanism, and start to focus on neural basis of costly punishment in recent years to further explain why we have willingness to costly punish.

A Neuroimaging research (de Quervain et al., 2004) first provided essential insight into the neural networks that shape such costly punishment actions. They designed a context of economic exchange in which investors transferred endowments to agents, but agents did not send back money to investors. This action of non-reciprocity was observed by a third party. Subjects could choose to punish these violators, and symbolic and effective punishments were available. Symbolic punishments did not influence the material benefits of the violator, while effective punishments did decrease the violator's payoff. They used positron emission tomography (PET) to scan the third party's brains while they confronted with the defection and determined the sanction. The neuroimaging results suggested that punishing defections effectively instead of symbolically activated the dorsal striatum (DS) region, which plays an important role in the processing of reward. Furthermore, subjects with higher activity in the DS were ready to pay more costs to punish. These findings proved the hypothesis that humans may achieve satisfaction from the action of punishing violators, even when this punishment causes a monetary loss to themselves.

To further assess this satisfaction through punishing defectors, another neuroimaging study using fMRI scanned the brain reward regions of participants during two-person economic game involving costly punishments (Strobel et al., 2011). They found that, indeed, brain reward areas such as the nucleus accumbens (NAC) and DLPFC were activated by the action of punishment. In addition, this activation was similarly affected by genetic variation of dopamine turnover during both first player and third party punishments. Overall, these results suggest that the interactive network of cognition, affect and motivation form the driving force in costly punishments.

A recent study has also investigated the brain mediation mechanisms during costly punishments based on the BOLD responses in related brain areas (White et al., 2014). Subjects showed greater modulation of BOLD signals based on the

level of costly punishment in regions of the reward neural network, for example, the AI cortex and caudate, whereas subjects showed negative modulation of BOLD signals as a level of costly punishment within posterior cingulate cortex (PCC) and VMPFC regions. Converging evidence seems to indicate a transform via the reward circuitry in mediating costly punishment. In addition, the neurobiological determinants have been found an influence in decisions of punishing costly (Crockett et al., 2013). Manipulating the serotonin system of participants during economic exchange game alters the possibility of punishment through modulating the activity of striatum, indicating that serotonin may create the sensitivity threshold for punishment processing.

Some brain stimulation studies provide a causal evidence of prefrontal cortex regions on decisions of costly punishment through changing the activity of prefrontal cortex (van't Wout et al., 2005; Knoch et al., 2006, 2007). Subjects have a lower propensity to punish unfair behavior at a personal cost when rLPFC activity is restrained compared with the sham condition (Knoch et al., 2007). Based on this result, it can be expected that distinctions in the brain functions of the prefrontal cortex could illustrate individual variations in the willingness to punish, that is, the higher the individual baseline level of rLPFC activity, the greater the punishment behavior performed by the individual. To demonstrate whether individual differences in the activity levels of the rLPFC region predict participants' willingness to provide costly punishments to other people, a neuroscience study measured participants' resting-state electroencephalography (EEG) activity (Knoch et al., 2010) before they executed punishments for unfair proposals. A positive relationship was found between resting alpha activity in the rLPFC and the likelihood of a costly punishment. It is well known that the bilateral LPFC was associated with implementing of self-control and cognition processes (Miller and Cohen, 2001; Knoch et al., 2006; Cohen and Lieberman, 2010).

Another brain stimulation study on sanctions (Buckholtz et al., 2015) combined rTMS with fMRI to verify the explicit role of the DLPFC in pro-social behaviors induced by blame and punishment. The participants reduced punishments for violation activities when their brain activity in the DLPFC was inhibited by rTMS, but these participants' blameworthiness ratings were not influenced. The researchers also used fMRI to observe punishment-selective DLPFC region recruitment. These results indicated that these two aspects of decisions are neurobiologically dissociable and confirm a selective causal effect of the DLPFC on punishment behavior. Thus, brain stimulation to related brain regions has a significant effect on norm compliance induced by social punishment threats, whereas stimulation left beliefs of what the norm regulated and subject expectations about social sanctions unaffected.

However, perhaps it is still unclear what the fundamental driving force for neural responses in decisions for costly punishments is. Du and Chang (2015) concluded that three main cognitive and affective functions occur in costly punishment contexts that might have a crucial effect on activating neural regions, such as cost-benefit calculations, inequity aversions and social reference frames. The previous studies show that these three cognitive and affective functions have different neural

circuitries underlying the complicated decision process of costly punishments. Furthermore, these neural mechanisms, involving distinct cognitive and affective processes, are likely to interact with one another during the decision to punish at a cost, and such interactions may lead to individual deliberations on the execution of this decision. Therefore, how to differentiate the neural circuitries of these cognitive and affective functions during decision-making in costly punishment is a key issue that needs to be solved.

COOPERATION

Humans often cooperate with each other in society, even with irrelevant strangers and people they will never meet again. This behavioral feature in human is considered as a social adaptation, implying human success in evolutionary progress (Hrdy, 2009; de Waal, 2010). Some behavioral studies focused on what motivation promoted evolution of human cooperation, such as costly punishment, altruistic rewarding and strong reciprocity (Fehr and Gächter, 2002; Boyd et al., 2003; Bowles and Gintis, 2004). Cooperation behavior has also been illustrated extensively in the economic exchange game, for example the prisoner's dilemma game (PDG) (Sally, 1995). In the standard PDG, two players' payoffs depend on interaction of their decisions. The player can get the most payoff if she or he choose to defect and the partner choose to cooperate, while the least to the player happens if she or he choose to cooperate and the partner choose to defect. In addition, mutual cooperation takes a modest amount to each player, whereas mutual defection leads to a lesser payoff to the two players.

Neuroscience methods combined with the paradigms of game theory have examined the neural basis of cooperative behaviors. In two neuroimaging studies (Rilling et al., 2002, 2004), it was revealed that the ventral stratum was activated when playing in mutual cooperation with a partner in a game, as compared to playing with a computer partner. Rilling et al. (2002) first employed fMRI to scan subjects when they played with a paired partner in a repeated PDG to explore the neural substrates of cooperative behavior. They found that the activity of related brain regions, such as NAC, rostral ACC, orbitofrontal cortex (OFC) and the caudate nucleus, involved in reward processing were associated with cooperative behavior, and a crucial role of the striatum in mutual cooperative behavior was demonstrated. Participants' mutual cooperative behavior leads to a higher BOLD signal in the related neural network during a PDG but results in a lower BOLD signal in the same regions if the partner defects. In subsequent research (Rilling et al., 2004), the reward neural network was also activated during cooperation in a sequential PDG, and subjects showed higher anterior paracingulate cortex and posterior STS activity when playing with person rather than with a computer. Cooperation following the defection of a partner would be characterized as an action against one's anticipation of the reciprocity norm and, thus, increase activation of the left AMYG and the bilateral AI (Rilling et al., 2008).

In another experimental paradigm, pairs of subjects were required to perform the same estimation task and received

a monetary reward for right answer (Fliessbach et al., 2007). A higher activation of the ventral striatum was linked with the amount of reward earned by the subject, while a lower activation of the same area was linked with the amount of reward paid to the partner. That is, when people are assessed and rewarded by an identical standard, the ventral striatum activity is more closely related to personal relative earnings than payments to the partner. This finding indicates the likelihood that the striatal involvement in rewarding processes seems to vary depend on whether a social exchange was considered to be competition or cooperation. Similarly, when participants were asked to play with a partner competitively or cooperatively during a board game, differential brain regions were activated in two distinct patterns of interaction. The results showed that cooperation caused higher activity in the medial orbitofrontal cortex (MOFC) and anterior frontal cortex (AFC) compared to competition (Decety et al., 2004; Babiloni et al., 2007). However, whether and how these cortical regions are linked to the striatal activity involved in cooperation are currently unknown.

To further elucidate the neural mechanisms of cooperation, King-Casas et al. (2008) recruited subjects suffering from borderline personality disorder (BPD) to play an iterative social interaction game with healthy subjects. Healthy subjects exhibited a linear correlation between AI activity and both the amount of monetary payoff received from the partner and the magnitude of money sent back to their partner. In contrast, subjects with BDP only showed a relationship between AI activity and the amount of money repaid to the partner, not the amount of money received from their partner. These results are evidence that individuals with BPD show impaired AI activity that leads to an inhibition of their ability to benefit from mutual cooperation. Thus, the insula and the ventral striatum track the social interaction decision of the partner of whether to reciprocate cooperation, representing an encoding of the reward processes for the satisfaction gained through mutual cooperation (Sanfey, 2007). In addition, a computational model of social value was provided to predict individual cooperative behavior, which indicated that people receive a signal of social value reward for mutual cooperation (Fareri et al., 2015). This signal of social value was strongly associated with greater activation of the ventral striatum and MPFC, which suggests that this signal predicts cooperative behavior in an iterative social exchange game.

In summary, the implications and motivations behind pro-social behaviors in economic games have been widely discussed, and scans of related brain areas when people play social interaction economic games with a partner could reveal individual differences in cooperative behavior. However, why are people willing to cooperate with the other people in a game? What are motives driving this behavior for all humans? Whereas some economic games have been used mainly to investigate cooperative behavioral consistency (Yamagishi et al., 2013; Peysakhovich et al., 2014), manipulating different economic games with the same subjects could also enable researchers to isolate within-subject motives in order to more accurately examine the nature of cooperative decisions (Brañas-Garza et al., 2014). Studies that have used this methodology have indicated that cooperative behavior is always multi-determined and can be assigned to completely different motives.

Prior work using imaging tools such as fMRI have allowed the identification of the neural networks involved in cooperative behavior. Nonetheless, these tools can provide only limited support to this ambitious purpose as they lack temporal resolution. In addition, they do not permit an on-line, real-life social exchange environment. However, social interaction is an essential part of cooperative behavior. Therefore, how our brains specifically exploit social cues and contexts when considering whether to cooperate remains unclear (Jahng et al., 2017). To account for the complexity of this event, the hyperscanning approach supports a high temporal resolution that allows the capture of simultaneous recordings of brain activity as a possible research direction.

TRUST/TRUSTWORTHINESS

It is well known that trust penetrates into many aspects of our life, including working relations, friendships, and family relations. Interpersonal trust is also a core element for deeply understanding economic among people and the loss of trust between exchange partners seriously hinders market exchange. Thus, there are many reasons for researchers to concern the decision of trust. To investigate the decision of trust in economic interaction, Berg et al. (1995) firstly constructed a trust game, in which a player (the investor) has to decide how many amounts of endowment to invest with the other player (the trustee), and then the trustee can choose whether to give back and how much money return to the investor. As the model hypothesis of rational and self-interested people, the trustee will never return money to the investor. The investor can expect this rational decision from the trustee, and should never invest any amounts of money with the trustee.

Despite the predictions of game theory, in fact, most of the investors are still quite willing to transfer considerable amounts of money to a partner, and the trustees often repay some amount of money to the investor. Extensive studies have also discussed the potential factors that induce both trusting and trustworthiness behaviors among people that are not consistent with the hypothesis of the Homo economicus (Cook and Cooper, 2003; Bohnet and Zeckhauser, 2004; Cox, 2004; Ashraf et al., 2006; Schechter, 2007). In laboratory experiments, the subjects robustly showed behavior of trust although with completely strangers, or even when reputation is absent (McCabe and Smith, 2000; King-Casas et al., 2005).

Based on the results of behavioral studies, neuroscientists have attempted to provide the neural basis of trusting behavior. Krueger et al. (2007) employed hyper fMRI to scan pairs of subjects while they were playing against each other in a trust game. According to the within-brain and between brains analyses, several lines of evidence from functional brain activity indicated that the differential activation of related neural systems involves two trust strategies. First, the paracingulate cortex region is linked with the building of a relation of trust by inferring the partner's intentions to predict the subsequent decision. Second,

the more recently evolved brain areas could be distinctly involved in interactions with more primitive neural systems developing conditional and unconditional trust relations. Conditional trust decisions significantly activated the VTA region, associated with the estimation of expected rewards, while unconditional trust decisions activated the septal region, associated with social interaction behaviors.

Interestingly, the evidence from neuroendocrinology shows that humans can secrete two hormones in opposite ways that are linked to establish a subtle balance in adjustable trust behaviors. A hormone that promotes social trust is oxytocin (OT). There is evidence that the brain differentiates between interpersonal trust and risk-seeking, derived from a study where the synthetic neuropeptide OT was injected intranasally to subjects while playing a trust game (Kosfeld et al., 2005). The hypothesis was that the betrayal aversion reducing effect of OT might result in decreased activity in the AMYG, suggesting that OT reduces AMYG activity. AMYG function has been demonstrated to be involved in evaluating the trustworthiness of faces (Winston et al., 2002; Adolphs et al., 2005) and ambiguous incidents (Hsu et al., 2005), which both have relevance with decisions in a trust game. It should be noted that these effects of OT might not extend to all people, because a neuropathological study found that OT can inversely inhibit trust behavior in individuals with BPD (Bartz et al., 2010; Bos et al., 2010). These results demonstrate the necessity of taking into account personal heterogeneity while reporting the effects of hormones on individual behaviors (Bartz et al., 2011).

Similarly, in a testosterone administration and placebo-controlled experiment, Bos et al. (2012) used fMRI to provide insights into the neural mechanisms involved in the effect of testosterone on trusting behavior. They found that testosterone improved social vigilance to untrustworthy faces by affecting neuropeptide systems in the central AMYG region, enhancing the communication between the AMYG and brainstem areas. However, testosterone can also change the functional connectivity between the OFC and AMG while judging unfamiliar faces, which then induces an improvement in social vigilance by decreasing top–down control over the AMYG. Although speculative, a neurobiological interpretation based on these results is that testosterone leads to the continuous reduction, in an uncertain social interaction, of the connectivity between the OFC and AMG via a prefrontal-dopaminergic mechanism, which results in more vigilant AMYG responses to signals of untrustworthiness.

Other studies have specifically examined the neural basis of trustworthiness of trustees, and the mechanisms involved in decision factors such as risks, benefits, and reputation have been examined (Baumgartner et al., 2009; Knoch et al., 2009; van den Bos et al., 2009; Aimone et al., 2014). Specifically, researchers have paid attention to the correlation between altruism and trustworthiness in neuroscience studies. A clinical lesion example indicated that patients with injuries to the VMPFC region offer less in a dictator game and show less trustworthy behaviors in a trust game, suggesting that the VMPFC plays an indispensable role in both altruism and trustworthiness decisions (Krajbich et al., 2009; Moretto et al., 2013). There was evidence that the

trustee playing the game showed activation in the VMPFC, the posterior cingulated cortex (PCC), the lateral OFC, and the right AMYG, and that the VMPFC response was linked with altruistic behavior (Li et al., 2009). Previous studies in neural cognition have also illustrated that the VMPFC is crucial for evaluating social information and that impairments in the VMPFC caused serious disruptions of emotion and resulted in impaired to decision making, behavior regulation and planning (Damasio, 1994; Anderson et al., 2006).

In addition to the VMPFC region, the roles that other brain areas play when realizing a partner's trustworthiness have also been tested in other studies. In one study, the caudate nucleus activity predicted whether the trustee showed trustworthiness for a partner in a trust game (King-Casas et al., 2005). In a second study, researchers used the same paradigm of a trust game to investigate specializations of the cingulate cortex in encoding trustworthy decisions in a social domain (Tomlin et al., 2006). A further study arranged investors in a trust game to sequentially face three trustees and provided profiles that made them seem morally positive, neutral or negative in order to instill a prior belief about trustworthiness (Delgado et al., 2005). The researchers found that the caudate nucleus activity in the investors was involved in the decision of whether the trustees were weakened when the investors depended on the trustees' information about moral character. Irrespective of the exact neural mechanisms, fMRI results indeed indicate that the AMYG is also associated with facial detections of trustworthiness (Winston et al., 2002; Engell et al., 2007; Todorov et al., 2008a,b; Said et al., 2009), and patients with damage to the bilateral AMYG show more facial evaluations of trustworthiness compared to healthy participants (Adolphs et al., 1998). The ACC and insula are also responsible to the processing of interpersonal trust and social threat (Rushworth et al., 2007).

Prior neuroimaging studies on trustworthiness have led to a well-founded discussion about the correlation between trustworthiness and empathy (Adolphs, 2002; Winston et al., 2002; Engell et al., 2007; Said et al., 2009). In particular, two meta-analyses studies have summarized the differential neural circuits linked to the process of identifying trustworthy and untrustworthy faces (Bzdok et al., 2011; Mende-Siedlecki et al., 2013). Notably, faces considered to be trustworthy primarily involve activity in reward-related brain areas, while faces considered to be untrustworthy primarily engage activation of the ventral AMYG, which quickly responds to a potential threat. A recent ERP study showed that faces perceived as trustworthy are implicitly evaluated even during an unrelated task, as when subjects must memorize characteristics, as seen by the modulation of neural activity linked with visual working memory that processes faces (Meconi et al., 2014).

However, these neuroimaging studies have failed to provide a direct causal effect between the activity in related brain areas and behavioral decisions. In contrast, recent tDCS studies, by affecting brain activity non-invasively, have established causal links between brain activity and trust or trustworthiness decisions. Colzato et al. (2015) used tDCS over the VMPFC region while participants played a trust game and did not found a correlation between VMPFC activity and trust behavior. Other

tDCS studies showed that the modulation of activity in several brain areas, such as the OFC, DLPFC and VMPFC may change the subjects' trustworthy behavior (Nihonsugi et al., 2015; Wang et al., 2016; Zheng et al., 2016).

On the whole, these results provide insight into understanding how related brain areas work together when subjects exhibit reciprocal trusting by showing how these neural substrates are distinctly derived from reciprocated trust, betrayal aversion, risk preferences and perspective-taking motives. However, it is still not clear which mechanisms connect these neural substrates that underlie the different motivations in trust behavior, or which neural structures factors determine individual levels of trust behavior. Additionally, identifying the effect of social factors, such as social status, social information and peer influence, on trust behavior based on neural results is necessary in further studies.

In summary, we describe previous experimental studies that used neuroscience methods to explore the role of related brain areas in various pro-social behaviors (see **Table 1**).

NEURAL CIRCUITRIES OF PRO-SOCIAL BEHAVIORS

Based on previous neuroscience studies of different types of pro-social behaviors, it can be seen that there are some similar properties in the neural basis of these behaviors, which all activate related brain areas, including the theory of mind network, reward system, and prefrontal cortex. This implies there may be specific connections between the functions of these areas that lead to people often making decisions according to their other-regarding preferences. Our knowledge of this connection can help us better understand the neural mechanisms of pro-social behaviors. Thus, it is important to find the shared, common neural substrates that link these different types of pro-social behaviors. Here, an integrated model is proposed that depicts the neural circuits by which decision making is firstly primed in the theory of mind network while receiving input from other's information, then the social cognition signal activates reward system, and this action then activates the brain areas associated with emotion to reinforce the reward experience, and, finally, the decision is reflected upon

TABLE 1 | Summary of the study for the neural basis of pro-social behaviors.

Study	Technology	Experimental Task	Pro-social behaviors	Brain areas	Experimental design	Sample size
Rilling et al., 2002	fMRI	Prisoner's Dilemma Game	Cooperation	Ventral striatum	Between (human versus computer) and within (iterated game) subjects.	36
Sanfey et al., 2003	fMRI	Ultimatum game	Fairness	AI and DLPFC	30 rounds in all, 10 playing the game with a human, 10 with a computer, and a further 10 control rounds.	19
de Quervain et al., 2004	PET	Third-party punishment game	Costly punishment	Dorsal striatum	Participants experienced four different conditions.	14
Rilling et al., 2004	fMRI	UG and PDG	Fairness	aPCC and posterior STS.	Between (human versus computer) and within (iterated game) subjects.	19
Decety et al., 2004	fMRI	Computer game	Cooperation	OFC and MPFC	Between (alone, cooperation, or against) subjects.	12
Moll et al., 2006	fMRI	Charitable donation	Altruistic behavior	VTA and STR	Different payoff types were designed: (*i*) pure monetary reward, (*ii*) non-costly donation, and (*iii*) costly donation.	19
Knoch et al., 2006	rTMS	Ultimatum game	Fairness	DLPFC	Applied rTMS to the right or to the left DLPFC and a control group.	52
Knoch et al., 2010	EEG	Ultimatum game	Costly punishment	rPFC	Responder played with 12 different proposers.	20
Spitzer et al., 2007	fMRI	Dictator game with the sanction threat	Costly punishment	OFC and DLPFC	Control and punishment conditions	45
Krueger et al., 2007	hyperfMRI	Trust game	Trust	pACC and VTA	Sequential decisions for monetary payoffs (low, medium, or high).	44
Emonds et al., 2011	fMRI	PDG and coordination game	Reciprocity	DLPFC and STS	Between (proself or prosocial) and within (two games) subjects.	28
Bos et al., 2012	fMRI	Rate facial pictures	Trustworthiness	OFC and AMYG	A randomized, counterbalanced, placebo-controlled, testosterone administration paradigm.	16
Ruff et al., 2013	tDCS	Ultimatum game	Fairness	rLPFC	Randomly assigned to one of three groups: anodal, sham, or cathodal.	64
Aimone et al., 2014	fMRI	Trust game	Betrayal aversion	AI	Both within- and between-subject	30
Strang et al., 2014	TMS	Dictator game	Strategic fairness	DLPFC	Randomly assigned to one of three groups: anodal, sham, or cathodal.	17
Zheng et al., 2016	tDCS	Trust game	Trustworthiness	VMPFC	Randomly assigned to one of three groups: anodal, sham, or cathodal.	60

in the prefrontal cortex to execute a pro-social behavior (see **Figure 1**).

Pro-social behaviors surely requires theory of mind during cognition signal input, as it is the neural network underlying our ability to attribute other's mental states, providing us with a prediction of their intentions and actions. Theory of mind is likely to provide our other-regarding, which allows the ability to share others' thoughts and feelings and, therefore, motivates pro-social behaviors (de Waal, 2008). A number of studies in cognitive neuroscience have discovered the neural network for considering another person's thoughts, comprising the precuneus, bilateral TPJ and right superior temporal sulcus (RSTS) (Rilling et al., 2004; Saxe, 2006; Young et al., 2007; van Overwalle and Baetens, 2009; Ye et al., 2015). In particular, the TPJ shows increased activity when participants read about a person's beliefs in non-moral (Saxe and Powell, 2006) and moral (Young et al., 2007) contexts. Accordingly, it has also been suggested that the ACC might play an important role in representing the mental states of others (Gallagher et al., 2002; Gallagher and Frith, 2003; Rilling et al., 2004). This brain area is involved not only when mentalizing about the thoughts, intentions or beliefs of others but also when people are attending to their own states. Frith and Frith (2003) suggest that this area subserves the formation of decoupled representations of beliefs about the world. In addition, it has been shown that activity in this area is strongly associated with the level of an individual's pro-social behavior (Tankersley et al., 2007). These results address long-standing discussions about the sources of social decision making by indicating that pro-social behaviors might first derive from the theory of mind network, with its the proclivity toward social-cognitive thoughts of other's mental states.

Pro-social behaviors are always accompanied by the activation of the reward system, and these behaviors were marked and intensified after other-regarding thoughts of mental states. The reward system is a neural network responsible for incentive salience (i.e., craving, motivation, or desiring for a reward), related learning (mainly positive reinforcement for actions), and the activation of emotions, especially ones that involve pleasure as a central constituent (e.g., happiness, joy, and euphoria). Importantly, correlations between different kinds of pro-social behaviors and the reward system have been shown in many previous neuroscience studies. Moll et al. (2006) found that the reward system, such as VTA and striatum areas were both activated when participants give donations to the charity. A significant activation in the reward system in response to inequity aversion has also been found (Tricomi et al., 2010). Neuroimaging results suggest that punishment defection in an economic game activates the participants' brain reward regions, such as the NACs and thalamus (de Quervain et al., 2004; Strobel et al., 2011). It was also demonstrated that the activities of related brain regions, such as NACs and the striatum area involved in reward processing were strongly correlated with cooperation behavior (Rilling et al., 2002, 2004; Sanfey, 2007). In addition, trust behaviors are primarily linked to reward-related brain areas when participants identify trustworthy faces (Meconi et al., 2014). These findings provide strong evidence of the rewarding process of different types of pro-social behaviors, which in return

demonstrates the crucial role the reward system plays in the shared neural substrates of pro-social behaviors.

Pro-social behaviors inevitably activate emotion-related regions in the brain when these decisions are rewarded, and these reward experiences must be stored in memory. The emotion-related system is a group of neural structures that are primarily involved in many of our feelings and motivations, including fear and disgust. The AI is thought to process convergent information to create a relevant context for the emotions involved in a sensory experience. Certain structures in this system are involved in memory processing as well. The AMYG is responsible for determining where memories are stored in the brain and which the memories are stored. It is believed that this determination is based on how great an emotional response the action invokes. The hippocampus sends out memories to related brain regions for long-term storage. In sum, this system supports various functions, such as interpreting emotional signals, regulating hormones, storing memories and processing motivation. Therefore, it is easy to infer that these neural structures are related with pro-social behaviors. A recent study has also found that the BOLD signals from the AI predicted affective empathy and helpful decisions (Greening et al., 2014; Tusche et al., 2016), and the BOLD signals from the AMYG region can be used to assess individual differences in fairness (Haruno and Frith, 2010). In fact, neural responses in the AI region associated with emotional arousal through the measurement of an autonomic index in affective status indicated that there was a strong correlation between AI activity and inequity aversion (van't Wout et al., 2006; Civai et al., 2012), and healthy participants exhibited a strong relationship between trusting behaviors and activity in this neural system including AI, AMYG, and PCC regions (Adolphs et al., 2005; Krueger et al., 2007; King-Casas et al., 2008; Watanabe et al., 2014). Bos et al. (2012) explored the neural mechanisms regarding the cause effect of testosterone on trusting behaviors via neuropeptide systems in the central AMYG region. In addition, it was shown that participants showed a significant modulation of neural responses within the PCC as a function of costly punishments and trusting behaviors (Li et al., 2009; White et al., 2014). Rilling et al. (2008) has also found that cooperative behaviors increase AMYG and AI activity, and subjects showed a higher anterior paracingulate cortex activity when playing cooperation strategies (Rilling et al., 2004). In conclusion, there is enough evidence to indicate that these neural structures, triggered by the need for arousing emotions and long-term memory play an important role in the production and reinforcement of pro-social behaviors.

Pro-social behavior truly needs to be controlled and planned as a whole while balancing various motivations, and the prefrontal cortex region is considered to be the control center of pro-social behavior and is the key part of the common shared neural substrates of different types of pro-social behaviors. The prefrontal cortex region is known to be associated with planning and modulating pro-social behaviors (Yang and Raine, 2009). In terms of psychology, the most typical functions carried out through the prefrontal cortex are executive functions (Shimamura, 2000). Executive functions involve the abilities to make decisions among different conflicting considerations
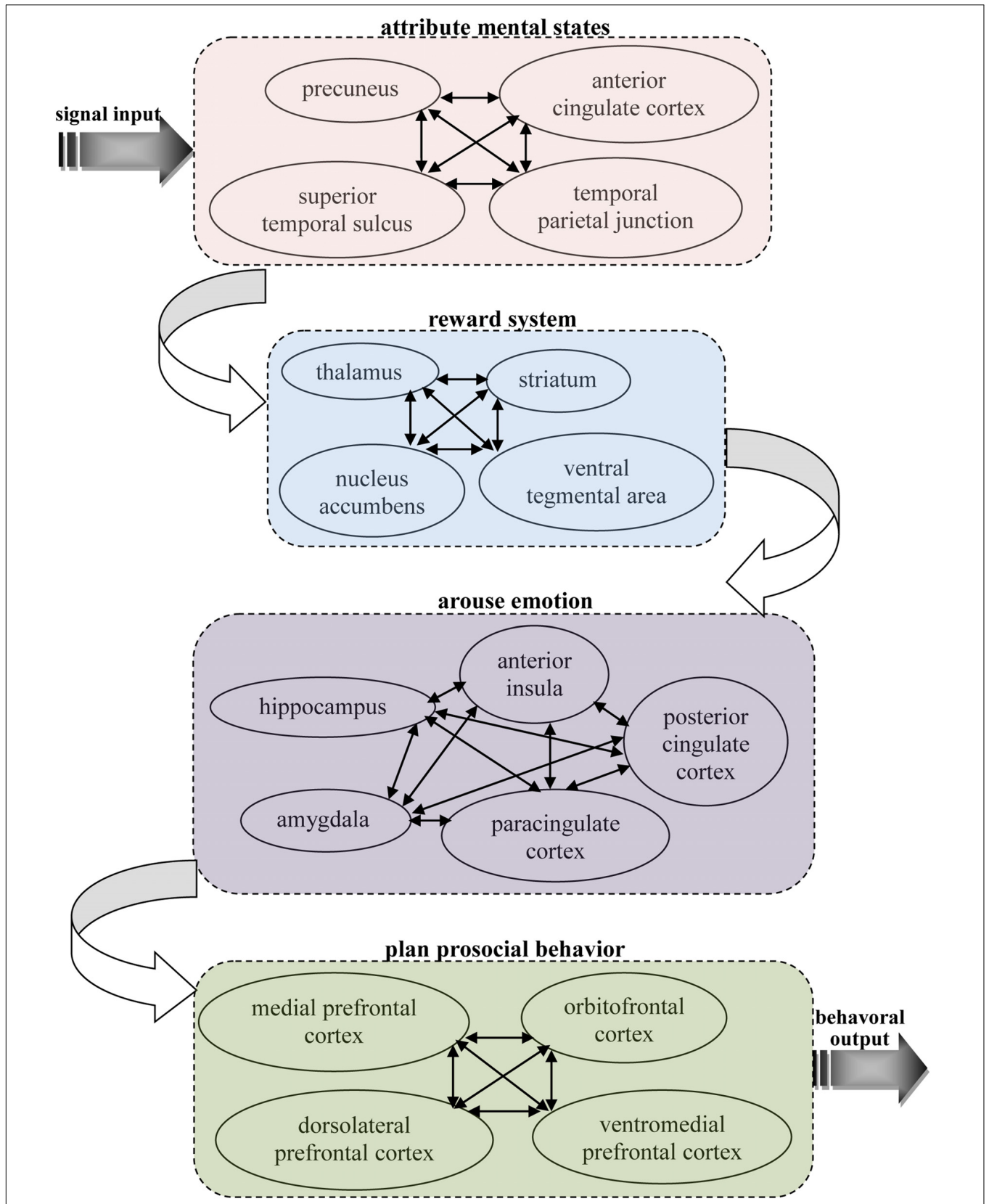
FIGURE 1 | A possible common neural circuitry associated with different types of pro-social behavior in the human brain.

(Goldberg, 2002); determine good or bad, self-interest or other-regarding, and same or different; create expectations according to events; create predictions of consequences and future outcomes of current actions; and enable social "control" (the capability to inhibit desires that, if not inhibited, could cause socially unacceptable consequences). The effect of this neural network on carrying out pro-social behaviors has been demonstrated in many neuroscience studies. For example, several studies have shown that individual differences in signals from the MPFC are correlated with empathy and altruistic behaviors (Zaki et al., 2009; Wagner et al., 2011; Powers et al., 2015). The indispensable roles of the VMPFC in the decision making involved in fairness, altruism and trustworthiness are also supported by previous neural evidence (Koenigs and Tranel, 2007; Krajbich et al., 2009; Moretto et al., 2013). A fMRI study (Spitzer et al., 2007) investigated the relationship between costly punishments and activity in the lateral OFC and rDLPFC and the causal effect of activity in the LPFC region on decisions of costly punishment by altering the activation of this region (van't Wout et al., 2005; Knoch et al., 2006, 2007). Interestingly, other brain stimulation studies have indicated that reducing the activation of the rLPFC leads to a significant change in fairness-related behaviors, but neither expected punishment from others nor fairness norms were altered (Ruff et al., 2013; Strang et al., 2014). These findings reveal that activity in the prefrontal cortex is a crucial biological prerequisite for an important aspect of evolutionary and social human behavior. These findings suggest that the prefrontal cortex is responsible for behavioral control although it is dissociated from the neural structures that enable people to anticipate social norms and attribute other's mental states. The structural connectivity and situation-dependent functions of prefrontal cortex (Duncan, 2010) make it possible to integrate and coordinate activation of the neural networks related to pro-social behavior during action control.

DISCUSSION

This review introduces several widely used methods that combined the game theory of economics with neuroscience technologies to develop new insights into understanding pro-social behaviors. These findings contribute important progress for measuring the neural mechanisms involved in pro-social behaviors and yield the guarantee of identifying and accurately characterizing both the mechanisms and the factors that influence interactions and engagements in pro-social behavior.

The economic game approach has several some advantages over typical paradigms of decision-making, not only for use within real, sequential, social interactions that make it possible to study complex exchange contexts such as fairness, trust, cooperation, and norm compliance. On the other hand, neuroscience methods of assessing pro-social behavior have obtained several prominent achievements in examining how the utility of parameters of behavior are represented in neural systems (Knutson et al., 2005; Sugrue et al., 2005; Padoa-Schioppa and Assad, 2006). Thus, neuroscience studies of pro-social behaviors could explore the neural circuitries of parameters that game

models both expect (such as strategic payoffs) and do not expect (such as social preference). In addition, behavioral and neural data created through this method can confirm significance in offer special restraints, based on neural networks, for any theory that attempts to build precise models of pro-social behavior.

However, there are important challenges to address for any novel approach. These challenges are involved in disciplines that manipulate diverse analysis perspectives and have distinct theoretical assumptions. In particular, there are crucial differences in study methodologies, such as the use of deception, which is strictly prohibited in economics but widely used in neuroscience and psychology. Furthermore, it is important to be prudent when understanding brain activity measured by neuroimaging methods. For instance, the correlation of a brain area with either reward encoding or emotion processing in prior studies does not necessarily mean that this brain region's activity during at social interaction game can automatically be considered, respectively, as involved in reward or punishments (Sanfey, 2007). Therefore, one should be cautious in this field and support these conclusions by collecting evidence from other methodologies or, at a minimum, illustrating that behavioral results are consistent with the identified neural activities, for example, high levels of activation of the rewarding neural network being associated with an individual's preference for social decisions (de Quervain et al., 2004).

Although neuroimaging data cannot provide causal inferences, it is likely close to causality by predicting decision making during a treatment based on brain region activity during another treatment. For example, individual heterogeneity in the activity of the caudate nucleus when punishment is free predicts the level of willingness to punish when the punishing is costly (de Quervain et al., 2004). Similarly, individual differences in the activation of striatal regions when donations are mandatory are associated with participants' willingness to give money when this is a voluntary behavior (Harbaugh et al., 2007). These results further provide evidence of the rewarding process of pro-social behavior, which in return supports the hypothesis of shared neural circuitries of social reward and other primary and secondary reward (Montague and Berns, 2002).

Future studies in neuroscience should exploit the wide range of available tools, by using multiple measurement methods simultaneously (such as fMRI and rTMS, tDCS, or hormone measurement) along with the valuable behavioral parameters and theoretical predictions from complex game models. It is well known that non-invasive brain stimulation can establish causal correlations between related brain activities and individual behaviors by altering these neural processes and subsequent individual behaviorally expressed preferences. These methods have been used to not only support a biological basis for a mathematical characterizations in pro-social behaviors that are based on neural systems but also provide predictions of brain activity in social interactions and economic exchanges and how these behaviors can be transformed when manipulated by rTMS, tDCS, and other tools.

Nevertheless, brain stimulation technology is not the only approach to establish a causal inference between identified neural circuits and individuals' pro-social behaviors. Several

pharmacological studies show great potential in this domain. For example, testosterone can increase the fairness transfers of a proposer (Eisenegger et al., 2010) and the probability that a responder rejects unfair offers (Burnham, 2007) in an UG; the neurohormone oxytocin improves behaviors of trust but not trustworthiness (Kosfeld et al., 2005); the depletion of the neurotransmitter serotonin increases the number of rejection decisions for unfair offers in an UG (Crockett et al., 2008; Crockett, 2009); and benzodiazepine can decrease the number of decisions to reject (Gospic et al., 2011). These studies of pharmacological interventions combined with fMRI allow observations of how neural systems causally influence behavioral changes (Baumgartner et al., 2008; Gospic et al., 2011).

In addition, several neuroscience studies have began to consider how the neural systems involved in pro-social behaviors are influenced by various factors. One factor is "social image," that is, how does knowing that other people are watching you influence your decisions and brain activity? This topic has been a concern of economists (Andreoni and Bernheim, 2009) and is important because social image could be changed by different organizational norms in information and institutions. An fMRI study indicated that the bilateral striatum was more highly activated when participants' charitable donations were observed than in the control treatment (Izuma et al., 2008), which supports the hypothesis that a social image rooted in charitable giving is rewarding. Consistent with a wide range of inequity aversion, one study was concerned with whether an awareness of high-status people suffering a failure would create a positive reward. Activity in the ventral striatum was found in response to these hypothetical contexts, and the BOLD signal predicted self-rated decisions (Takahashi et al., 2009). Emotions can also have an impact on pro-social behaviors. A neuroimaging study exploring this topic was based on real crime cases with "mitigating circumstances" (Yamada et al., 2012) and found that the activity in the insula, an identified neural correlate of empathy, was related to the level of sentence reduction.

However, our current understanding of neural mechanisms of pro-social behavior is still limited. This understanding will be improved if we obtain additional interpretations of the genetic and neurophysiological mechanisms of information processing in neural reward systems. Previous studies have revealed that social reward generally activate the ventral or DS, and there is a substantial overlap between the activity in these regions and the activity observed in studies about anticipated monetary reward or reinforcement learning (Fehr and Camerer, 2007; Fehr, 2009). This overlap is in accordance with the hypothesis that social preferences are similar to preferences for physical reward in terms of brain activity, which supports the theory about which decisions reflect a tradeoff between one's own benefits and the benefits of others.

More importantly, our brain has to compare social welfare and individual benefits and solve a conflict between these aspects when we exhibit pro-social behaviors. Previous studies have demonstrated that the prefrontal cortical regions that evolved lately (in evolutionary perspective) play crucial roles in this process of conflict resolution. For example, the VMPFC region is more activated when subjects can punish defectors at a personal

cost than when punishment is free (de Quervain et al., 2004). Both the VMPFC and dorsal ACC regions show high activation levels when participants give charitable donations involving a cost (Moll et al., 2006). The ACC is known that have an important effect on conflict monitoring (Botvinick et al., 2001), so the activation of this area aligns with the presence of a conflict between pro-social motivations and self-interest incentives. In addition, the value of the response of the VMPFC is influenced by other responses of the posterior superior temporal cortex (PSTC) that have been shown to be crucial in overwhelming egocentricity prejudice, suggesting that the activity in both the VMPFC and the PSTC are important constituents of the neural network of pro-social behaviors.

In addition, the crucial role of DLFPC region in the processing of pro-social behavior has also been demonstrated (Sanfey et al., 2003). This study investigated the neural networks involved in the response decisions of an UG in which a rejection of unfair transfers indicates a balance between a self-interest motive and a fairness motive. Indeed, a function of the DLPFC may be enabling an individual to make choices for their long-term benefit for a good reputation in social interactions rather than the short-term benefit of the individual (van den Bos et al., 2009). The effect of the DLPFC on overwhelming short-term self-interest has also been investigated, and results show that the behavior of compliance with norms under the threat of punishment is positively related with the level of activity in the DLPFC (Spitzer et al., 2007).

More importantly, neuroscientists have indicated that we cannot consider brain regions as separate mini-brains but rather as widely interconnected regions. Notably, the frontal lobes are more linked to other brain regions than any other parts of the brain (Goldberg, 2002). In addition, the frontal lobes perform the most complex and developed functions of all parts of the brain, namely, the executive functions. This region is involved in complex, purposeful, and intentional decision making. However, previous studies also show that other brain areas, especially activation of brain–stem structures and the prefrontal cortex, are commonly associated with pro-social behaviors, such as the AI and the rostral ACC, which are activated when subjects empathize with other people experiencing pain (Singer et al., 2004); the ventral striatum, whose activity is strongly correlated with the amount of money given to charity (Harbaugh et al., 2007); the amygdale, whose BOLD signals can predict individual differences in aversion to inequity (Haruno and Frith, 2010); and both the DLPFC and caudate nucleus, whose activities are high when individuals receive unfair transfers from partners (Harbaugh et al., 2007). We also propose a common shared neural network of pro-social behaviors involving the theory of mind network, emotion-related regions, the reward system and the prefrontal cortex. In this neural circuitry, the control function of the prefrontal cortex plays a key role in coordinating human rationales, emotion, perception, cognition, motivation and reinforcement learning.

Among all species, humans are unique in terms of the ways in which they govern social life by executing pro-social behaviors. In addition to having the longer period during which brain development has been shaped by living environment,

human beings change their environment, which shapes brains to an unprecedented extent among other species (Wexler, 2006). We thus suggest that the evolutionary consequence of promoting pro-social behaviors is the development of the prefrontal cortex, which plays a crucial role in the neural circuitry of social preferences. The prefrontal cortex connected with other brain regions when executing social cognition functions, but the region-to-region interaction mechanisms between the prefrontal cortex and deeper brain areas that represent pro-social behaviors are still not clear.

Finally, the overall target of this effort is the identification of a complete and general model of the neural network of decisions involved in pro-social preferences. There have been previous attempts from both "cognitive neuroscience in social decisions" and "neuroeconomics" to interpret the social brain and the related moral emotions (Adolphs, 2001, 2003; Greene et al., 2001; McCabe et al., 2001; Moll et al., 2002; Rilling et al., 2002; Sanfey et al., 2003; Singer et al., 2004). In addition, a paradigm with high sensitivity may indicate how affective and cognitive responses in the brain diverge or converge throughout decision making. Future studies combining fMRI with another method with a higher temporal resolution, such as EEG or functional near-infrared spectroscopy (fNIRS), may describe novel data on the region-to-region interactions between neural activities associated with cost-benefit calculations, social preferences, and the processing of information across self and others. Interestingly,

finding a non-human primate model for pro-social behaviors can complement studies in humans by demonstrating specific neuronal circuitries in the core neural processes using single-unit recordings and pharmacological interventions in particular populations of neurons. Another important aspect of concern in the neuroscientific study of pro-social behavior is the social context in which pro-social behavior occurs. Understanding how the social context transforms the neural processes involved in cost-benefit calculations, social preferences, and the process of information between self and others will lead to better interpretations of the complex events behind human pro-social behaviors.

AUTHOR CONTRIBUTIONS

JL drew the table, wrote and revised the manuscript, and finally approved the version to be published.

FUNDING

REFERENCES

Adolphs, R. (2001). The neurobiology of social cognition. *Curr. Opin. Neurobiol* 11, 231–239. doi: 10.1016/S0959-4388(00)00202-6

Adolphs, R. (2002). Trust in the brain. *Nat. Neurosci.* 5, 192–193. doi: 10.1038/nn0302-192

Adolphs, R. (2003). Cognitive neuroscience of human social behaviour. *Nat. Rev. Neurosci.* 4, 165–178. doi: 10.1038/nrn1056

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., and Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature* 433, 68–72. doi: 10.1038/nature03086

Adolphs, R., Tranel, D., and Damasio, A. R. (1998). The human amygdala in social judgment. *Nature* 393, 470–474. doi: 10.1038/30982

Aimone, J. A., Houser, D., and Weber, B. (2014). Neural signatures of betrayal aversion: an fMRI study of trust. *Proc. R. Soc. B.* 281:20132127. doi: 10.1098/rspb.2013.2127

Anderson, S. W., Barrash, J., Bechara, A., and Tranel, D. (2006). Impairments of emotion and real-world complex behavior following childhood-or adult-onset damage to ventromedial prefrontal cortex. *J. Int. Neuropsych. Soc.* 12, 224–235. doi: 10.1017/S1355617706060346

Andreoni, J., and Bernheim, B. D. (2009). Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. *Econometrica* 77, 1607–1636. doi: 10.3982/ECTA7384

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Exp. Econ.* 9, 193–208. doi: 10.1007/s10683-006-9122-4

Babiloni, F., Astolfi, L., Cincotti, F., Mattia, D., Tocci, A., Tarantino, A., et al. (2007). "Engineering in medicine and biology society, EMBS 2007," *Proceedings of the 29th Annual International Conference of the IEEE*, Paris, 4953–4956.

Bartz, J., Simeon, D., Hamilton, H., Kim, S., Crystal, S., Braun, A., et al. (2010). Oxytocin can hinder trust and cooperation in borderline personality disorder. *Soc. Cogn. Affect. Neur.* 6, 556–563. doi: 10.1093/scan/nsq085

Bartz, J. A., Zaki, J., Bolger, N., and Ochsner, K. N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends. Cogn. Sci.* 15, 301–309. doi: 10.1016/j.tics.2011.05.002

Batson, C. D. (2009). "These things called empathy: eight related but distinct phenomena," in *The Social Neuroscience of Empathy*, ed. J. Decety, and W. Ickes (Cambridge, MA: MIT Press).

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

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

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

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Game. Econ. Behav.* 10, 122–142. doi: 10.1006/game.1995.1027

Bernhard, H., Fischbacher, U., and Fehr, E. (2006). Parochial altruism in humans. *Nature* 442, 912–915. doi: 10.1038/nature04981

Bohnet, I., and Zeckhauser, R. (2004). Trust, risk and betrayal. *J. Econ. Behav. Organ.* 55, 467–484. doi: 10.1016/j.jebo.2003.11.004

Boksem, M. A., and De Cremer, D. (2010). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. *Soc. Neurosci.* 5, 118–128. doi: 10.1080/17470910903202666

Bolton, G. E., and Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Game. Econ. Behav.* 10, 95–121. doi: 10.1006/game.1995.1026

Bos, P. A., Hermans, E. J., Ramsey, N. F., and Van Honk, J. (2012). The neural mechanisms by which testosterone acts on interpersonal trust. *Neuroimage.* 61, 730–737. doi: 10.1016/j.neuroimage.2012.04.002

Bos, P. A., Terburg, D., and Van Honk, J. (2010). Testosterone decreases trust in socially naive humans. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9991–9995. doi: 10.1073/pnas.0911700107

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.* 108, 624–652. doi: 10.1037/0033-295X.108.3.624

Bowles, S. (2009). *Microeconomics: Behavior, Institutions, and Evolution*. Princeton, NJ: Princeton University Press.

Bowles, S., and Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* 65, 17–28. doi: 10.1016/j.tpb.2003.07.001

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. (2003). The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3531–3535. doi: 10.1073/pnas.0630443100

Boyd, R., and Richerson, P. J. (2005). *The Origin and Evolution of Cultures.* Oxford: Oxford University Press.

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., and Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Sci. Rep.* 4:6025. doi: 10.1038/srep06025

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., et al. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron* 87, 1369–1380. doi: 10.1016/j.neuron.2015.08.023

Burnham, T. C. (2007). High-testosterone men reject low ultimatum game offers. *Proc. R. Soc. B.* 274, 2327–2330. doi: 10.1098/rspb.2007.0546

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., et al. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain. Struct. Funct.* 215, 209–223. doi: 10.1007/s00429-010-0287-4

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends Cogn. Sci.* 7, 225–231. doi: 10.1016/S1364-6613(03)00094-9

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

Cohen, J. R., and Lieberman, M. D. (2010). "The common neural basis of exerting self-control in multiple domains," in *Self Control in Society, Mind, and Brain*, Vol. 9, eds R. Hassin, K. Ochsner, and Y. Trope, 141–162.

Colzato, L. S., Sellaro, R., Van den Wildenberg, W. P. M., and Hommel, B. (2015). tDCS of medial prefrontal cortex does not enhance interpersonal trust. *J. Psychophysiol.* 29, 131–134. doi: 10.1027/0269-8803/a000144

Cook, K. S., and Cooper, R. M. (2003). "Experimental studies of cooperation, trust, and social exchange," in *Trust and Reciprocity: Interdisciplinary Lessons for Experimental Research* (New York, NY: Russell Sage), 209–244.

Cox, J. C. (2004). How to identify trust and reciprocity. *Game. Econ. Behav.* 46, 260–281. doi: 10.1016/S0899-8256(03)00119-2

Crockett, M. J. (2009). The neurochemistry of fairness. *Ann. Ny. Acad. Sci.* 1167, 76–86. doi: 10.1111/j.1749-6632.2009.04506.x

Crockett, M. J., Apergis-Schoute, A. M., Herrmann, B., Lieberman, M. D., Müller, U., Robbins, T. W., et al. (2013). Serotonin modulates striatal responses to fairness and retaliation in humans. *J. Neurosci.* 33, 3505–3513. doi: 10.1523/JNEUROSCI.2761-12.2013

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

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain.* New York, NY: Quill.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., and Smirnov, O. (2007). Egalitarian motives in humans. *Nature* 446, 794–796. doi: 10.1038/nature05651

de Quervain, D. J., Fischbacher, U., Treyer, V., and Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science* 305:1254. doi: 10.1126/science.1100735

de Vignemont, F., and Singer, T. (2006). The empathic brain: how, when and why? *Trends. Cogn. Sci.* 10, 435–441.

de Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annu. Rev. Psychol.* 59, 279–300. doi: 10.1146/annurev.psych.59.103006.093625

de Waal, F. B. (2010). *The Age of Empathy: Nature's Lessons for a Kinder Society.* New York, NY: Broadway Books.

Decety, J., and Jackson, P. L. (2004). The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* 3, 71–100. doi: 10.1177/1534582304267187

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., and Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *Neuroimage* 23, 744–751. doi: 10.1016/j.neuroimage.2004.05.025

Delgado, M. R., Frank, R. H., and Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nat. Neurosci.* 8, 1611–1618. doi: 10.1038/nn1575

Du, E., and Chang, S. W. (2015). Neural components of altruistic punishment. *Front. Neurosci.* 9:26. doi: 10.3389/fnins.2015.00026

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends Cogn. Sci.* 14, 172–179. doi: 10.1016/j.tics.2010.01.004

Eckel, C. C., and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Game. Econ. Behav.* 16, 181–191. doi: 10.1006/game.1996.0081

Egas, M., and Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. B.* 275, 871–878. doi: 10.1098/rspb.2007.1558

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annu. Rev. Psychol.* 51, 665–697. doi: 10.1146/annurev.psych.51.1.665

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

Emonds, G., Declerck, C. H., Boone, C., Vandervliet, E. J., and Parizel, P. M. (2011). Comparing the neural basis of decision making in social dilemmas of people with different social value orientations, a fMRI study. *J. Neurosci. Psychol. Econ.* 4, 11–24. doi: 10.1037/a0020151

Engell, A. D., Haxby, J. V., and Todorov, A. (2007). Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J. Cognitive. Neurosci.* 19, 1508–1519. doi: 10.1162/jocn.2007.19.9.1508

Fan, Y., and Han, S. (2008). Temporal dynamic of neural mechanisms involved in empathy for pain: an event-related brain potential study. *Neuropsychologia* 46, 160–173. doi: 10.1016/j.neuropsychologia.2007.07.023

Fareri, D. S., Chang, L. J., and Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *J. Neurosci.* 35, 8170–8180. doi: 10.1523/JNEUROSCI.4775-14.2015

Fehr, E. (2009). On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7, 235–266. doi: 10.1162/JEEA.2009.7.2-3.235

Fehr, E., and Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends. Cogn. Sci.* 11, 419–427. doi: 10.1016/j.tics.2007.09.002

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. *Nature* 425, 785–791. doi: 10.1038/nature02043

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

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868. doi: 10.1098/rspb.2015.0392

Feng, C., Luo, Y. J., and Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis. *Hum. Brain. Mapp.* 36, 591–602. doi: 10.1002/hbm.22649

Fliessbach, K., Weber, B., Trautner, P., Dohmen, T., Sunde, U., Elger, C. E., et al. (2007). Social comparison affects reward-related brain activity in the human ventral striatum. *Science* 318, 1305–1308. doi: 10.1126/science.1145876

Frith, U., and Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philos. T. R. Soc. B.* 358, 459–473. doi: 10.1098/rstb.2002.1218

Gabay, A. S., Radua, J., Kempton, M. J., and Mehta, M. A. (2014). The Ultimatum Game and the brain: a meta-analysis of neuroimaging studies. *Neurosc. Biobehav. R.* 47, 549–558. doi: 10.1016/j.neubiorev.2014.10.014

Gallagher, H. L., and Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends. Cogn. Sci.* 7, 77–83. doi: 10.1016/S1364-6613(02)00025-6

Gallagher, H. L., Jack, A. I., Roepstorff, A., and Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage.* 16, 814–821. doi: 10.1006/nimg.2002.1117

Goldberg, E. (2002). *The Executive Brain: Frontal Lobes and the Civilized Mind.* New York, NY: Oxford University Press.

Gospic, K., Mohlin, E., Fransson, P., Petrovic, P., Johannesson, M., and Ingvar, M. (2011). Limbic justice—amygdala involvement in immediate rejection in the ultimatum game. *PLoS Biol.* 9:e1001054. doi: 10.1371/journal.pbio.1001054

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Greening, S., Norton, L., Virani, K., Ty, A., Mitchell, D., and Finger, E. (2014). Individual differences in the anterior insula are associated with the likelihood of financially helping versus harming others. *Cogn. Affect. Behav. Neurosci.* 14, 266–277. doi: 10.3758/s13415-013-0213-3

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.* 3, 367–388. doi: 10.1016/0167-2681(82)90011-7

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

Haruno, M., and Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nat. Neurosci.* 13, 160–161. doi: 10.1038/nn.2468

Hein, G., and Singer, T. (2008). I feel how you feel but not always: the empathic brain and its modulation. *Curr. Opin. Neurobiol.* 18, 153–158. doi: 10.1016/j.conb.2008.07.012

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91, 73–78. doi: 10.1257/aer.91.2.73

Hoffman, M. L. (2001). *Empathy and Moral Development: Implications for Caring and Justice.* Cambridge, MA: Cambridge University Press.

Hrdy, S. B. (2009). *The Woman That Never Evolved: With a New Preface and Bibliographical Updates.* Cambridge, MA: Harvard University Press.

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

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science* 310, 1680–1683. doi: 10.1126/science.1115327

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

Jahng, J., Kralik, J. D., Hwang, D. U., and Jeong, J. (2017). Neural dynamics of two players when using nonverbal cues to gauge intentions to cooperate during the Prisoner's Dilemma Game. *Neuroimage.* 157, 263–274. doi: 10.1016/j.neuroimage.2017.06.024

Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., and Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron* 42, 335–346. doi: 10.1016/S0896-6273(04)00156-4

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., and Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science* 321, 806–810. doi: 10.1126/science.1156902

King-Casas, B., Tomlin, D., Anen, C., Camerer, C., Quartz, S. R., and Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science* 308, 78–83. doi: 10.1126/science.1108062

Knoch, D., Gianotti, L. R., Baumgartner, T., and Fehr, E. (2010). A neural marker of costly punishment behavior. *Psychol. Sci.* 21, 337–342. doi: 10.1177/0956797609360750

Knoch, D., Nitsche, M. A., Fischbacher, U., Eisenegger, C., Pascual-Leone, A., and Fehr, E. (2007). Studying the neurobiology of social interaction with transcranial direct current stimulation—the example of punishing unfairness. *Cereb. Cortex.* 18, 1987–1990. doi: 10.1093/cercor/bhm237

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

Knoch, D., Schneider, F., Schunk, D., Hohmann, M., and Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20895–20899. doi: 10.1073/pnas.0911619106

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., and Glover, G. (2005). Distributed neural representation of expected value. *J. Neurosci.* 25, 4806–4812. doi: 10.1523/JNEUROSCI.0642-05.2005

Koenigs, M., and Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *J. Neurosci.* 27, 951–956. doi: 10.1523/JNEUROSCI.4606-06.2007

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. *Nature* 435, 673–676. doi: 10.1038/nature03701

Krajbich, I., Adolphs, R., Tranel, D., Denburg, N. L., and Camerer, C. F. (2009). Economic games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *J. Neurosci.* 29, 2188–2192. doi: 10.1523/JNEUROSCI.5086-08.2009

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., et al. (2007). Neural correlates of trust. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20084–20089. doi: 10.1073/pnas.0710103104

Lamm, C., Batson, C. D., and Decety, J. (2007). The neural substrate of human empathy: effects of perspective-taking and cognitive appraisal. *J. Cognitive. Neurosci.* 19, 42–58. doi: 10.1162/jocn.2007.19.1.42

Li, J., Xiao, E., Houser, D., and Montague, P. R. (2009). Neural responses to sanction threats in two-party economic exchange. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16835–16840. doi: 10.1073/pnas.0908855106

Lockwood, P. L., Apps, M. A., Valton, V., Viding, E., and Roiser, J. P. (2016). Neurocomputational mechanisms of prosocial learning and links to empathy. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9763–9768. doi: 10.1073/pnas.1603198113

Majdandžić, J., Amashaufer, S., Hummer, A., Windischberger, C., and Lamm, C. (2016). The selfless mind: How prefrontal involvement in mentalizing with similar and dissimilar others shapes empathy and prosocial behavior. *Cognition* 157, 24–38. doi: 10.1016/j.cognition.2016.08.003

Masten, C. L., Morelli, S. A., and Eisenberger, N. I. (2011). An fMRI investigation of empathy for 'social pain' and subsequent prosocial behavior. *Neuroimage* 55, 381–388. doi: 10.1016/j.neuroimage.2010.11.060

Mathur, V. A., Harada, T., Lipke, T., and Chiao, J. Y. (2010). Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage* 51, 1468–1475. doi: 10.1016/j.neuroimage.2010.03.025

McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U.S.A.* 98, 11832–11835. doi: 10.1073/pnas.211415698

McCabe, K. A., and Smith, V. L. (2000). A comparison of naive and sophisticated subject behavior with game theoretic predictions. *Proc. Natl. Acad. Sci. U.S.A.* 97, 3777–3781. doi: 10.1073/pnas.97.7.3777

Meconi, F., Luria, R., and Sessa, P. (2014). Individual differences in anxiety predict neural measures of visual working memory for untrustworthy faces. *Soc. Cogn. Affect. Neur.* 9, 1872–1879. doi: 10.1093/scan/nst189

Mende-Siedlecki, P., Verosky, S. C., Turk-Browne, N. B., and Todorov, A. (2013). Robust selectivity for faces in the human amygdala in the absence of expressions. *J. Cognitive. Neurosci.* 25, 2086–2106. doi: 10.1162/jocn_a_00469

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., et al. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *J. Neurosci.* 22, 2730–2736. doi: 10.1523/JNEUROSCI.22-07-02730.2002

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., and Grafman, J. (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 15623–15628. doi: 10.1073/pnas.0604475103

Montague, P. R., and Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron* 36, 265–284. doi: 10.1016/S0896-6273(02)00974-1

Moretto, G., Sellitto, M., and di Pellegrino, G. (2013). Investment and repayment in a trust game after ventromedial prefrontal damage. *Front. Hum. Neurosci.* 7:593. doi: 10.3389/fnhum.2013.00593

Nakamaru, M., and Iwasa, Y. (2006). The coevolution of altruism and punishment: role of the selfish punisher. *J. Theor. Biol.* 240, 475–488. doi: 10.1016/j.jtbi.2005.10.011

Nihonsugi, T., Ihara, A., and Haruno, M. (2015). Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *J. Neurosci.* 35, 3412–3419. doi: 10.1523/JNEUROSCI.3885-14.2015

Olsson, A., and Ochsner, K. N. (2008). The role of social cognition in emotion. *Trends Cogn. Sci.* 12, 65–71. doi: 10.1016/j.tics.2007.11.010

Padoa-Schioppa, C., and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226. doi: 10.1038/nature04676

Penner, L. A., Dovidio, J. F., Piliavin, J. A., and Schroeder, D. A. (2005). Prosocial behavior: multilevel perspectives. *Annu. Rev. Psychol.* 56, 365–392. doi: 10.1146/annurev.psych.56.091103.070141

Peysakhovich, A., Nowak, M. A., and Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nat. Commun.* 5:4939. doi: 10.1038/ncomms5939

Powers, K. E., Chavez, R. S., and Heatherton, T. F. (2015). Individual differences in response of dorsomedial prefrontal cortex predict daily social behavior. *Soc. Cogn. Affect. Neur.* 11, 121–126. doi: 10.1093/scan/nsv096

Preston, S. D., and de Waal, F. B. (2002). Empathy: its ultimate and proximate bases. *Behav. Brain. Sci.* 25, 1–20.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *Am. Econ. Rev.* 83, 1281–1302.

Rand, D. G., Armao, J. J. I. V., Nakamaru, M., and Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *J. Theor. Biol.* 265, 624–632. doi: 10.1016/j.jtbi.2010.06.010

Rand, D. G., and Nowak, M. A. (2013). Human cooperation. *Trends. Cogn. Sci.* 17, 413–425. doi: 10.1016/j.tics.2013.06.003

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., and Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron* 35, 395–405. doi: 10.1016/S0896-6273(02)00755-9

Rilling, J. K., King-Casas, B., and Sanfey, A. G. (2008). The neurobiology of social decision-making. *Curr. Opin. Neurobiol.* 18, 159–165. doi: 10.1016/j.conb.2008.06.003

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., and Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* 22, 1694–1703. doi: 10.1016/j.neuroimage.2004.04.015

Roth, A., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *Am. Econ. Rev.* 5, 1068–1095.

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

Rushworth, M. F., Behrens, T. E., Rudebeck, P. H., and Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trend. Cogn. Sci.* 11, 168–176. doi: 10.1016/j.tics.2007.01.004

Said, C. P., Baron, S. G., and Todorov, A. (2009). Nonlinear amygdala response to face trustworthiness: contributions of high and low spatial frequency information. *J. Cognitive. Neurosci.* 21, 519–528. doi: 10.1162/jocn.2009.21041

Sally, D. (1995). Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Ration. Soc.* 7, 58–92. doi: 10.1177/1043463195007001004

Sanfey, A. G. (2007). Social decision-making: insights from game theory and neuroscience. *Science* 318, 598–602. doi: 10.1126/science.1142996

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

Saxe, R. (2006). Why and how to study theory of mind with fMRI. *Brain. Res.* 1079, 57–65. doi: 10.1016/j.brainres.2006.01.001

Saxe, R., and Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychol. Sci.* 17, 692–699. doi: 10.1111/j.1467-9280.2006.01768.x

Schechter, L. (2007). Traditional trust measurement and the risk confound: an experiment in rural Paraguay. *J. Econ. Behav. Organ.* 62, 272–292. doi: 10.1016/j.jebo.2005.03.006

Shimamura, A. P. (2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology* 28, 207–218.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science* 303, 1157–1162. doi: 10.1126/science.1093535

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature* 439, 466–469. doi: 10.1038/nature04271

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., and Fehr, E. (2007). The neural signature of social norm compliance. *Neuron* 56, 185–196. doi: 10.1016/j.neuron.2007.09.011

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., and Sack, A. T. (2014). Be nice if you have to—the neurobiological roots of strategic fairness. *Soc. Cogn. Affect. Neur.* 10, 790–796. doi: 10.1093/scan/nsu114

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., et al. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *Neuroimage* 54, 671–680. doi: 10.1016/j.neuroimage.2010.07.051

Sugrue, L. P., Corrado, G. S., and Newsome, W. T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. *Nat. Rev. Neurosci.* 6, 363–375. doi: 10.1038/nrn1666

Tabibnia, G., Satpute, A. B., and 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, 339–347. doi: 10.1111/j.1467-9280.2008.02091.x

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

Tankersley, D., Stowe, C. J., and Huettel, S. A. (2007). Altruism is associated with an increased neural response to agency. *Nat. Neurosci.* 10, 150–151. doi: 10.1038/nn1833

Telzer, E. H., Masten, C. L., Berkman, E. T., Lieberman, M. D., and Fuligni, A. J. (2011). Neural regions associated with self control and mentalizing are recruited during prosocial behaviors towards the family. *Neuroimage* 58, 242–249. doi: 10.1016/j.neuroimage.2011.06.013

Todorov, A., Baron, S. G., and Oosterhof, N. N. (2008a). Evaluating face trustworthiness: a model based approach. *Soc. Cogn. Affect. Neur.* 3, 119–127. doi: 10.1093/scan/nsn009

Todorov, A., Said, C. P., Engell, A. D., and Oosterhof, N. N. (2008b). Understanding evaluation of faces on social dimensions. *Trend. Cogn. Sci.* 12, 455–460. doi: 10.1016/j.tics.2008.10.001

Tomlin, D., Kayali, M. A., King-Casas, B., Anen, C., Camerer, C. F., Quartz, S. R., et al. (2006). Agent-specific responses in the cingulate cortex during economic exchanges. *Science* 312, 1047–1050. doi: 10.1126/science.1125596

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

Tusche, A., Böckler, A., Kanske, P., Trautwein, F. M., and Singer, T. (2016). Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. *J. Neur.* 36, 4719–4732. doi: 10.1523/JNEUROSCI.3392-15.2016

van den Bos, W., van Dijk, E., Westenberg, M., Rombouts, S. A., and Crone, E. A. (2009). What motivates repayment? Neural correlates of reciprocity in the Trust Game. *Soc. Cogn. Affect. Neur.* 4, 294–304. doi: 10.1093/scan/nsp009

van Overwalle, F., and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48, 564–584. doi: 10.1016/j.neuroimage.2009.06.009

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

van't Wout, M., Kahn, R. S., Sanfey, A. G., and Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Exp. Brain. Res.* 169, 564–568. doi: 10.1007/s00221-006-0346-5

Wagner, D. D., Kelley, W. M., and Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cereb. Cortex.* 21, 2788–2796. doi: 10.1093/cercor/bhr074

Wang, G., Li, J., Yin, X., Li, S., and Wei, M. (2016). Modulating activity in the orbitofrontal cortex changes trustees' cooperation: a transcranial direct current stimulation study. *Behav. Brain. Res.* 303, 71–75. doi: 10.1016/j.bbr.2016.01.047

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, 3990–3995. doi: 10.1073/pnas.1318570111

Waytz, A., Zaki, J., and Mitchell, J. P. (2012). Response of dorsomedial prefrontal cortex predicts altruistic behavior. *J. Neurosci.* 32, 7646–7650. doi: 10.1523/JNEUROSCI.6193-11.2012

Wexler, B. E. (2006). *Brain and Culture: Neurobiology, Ideology, and Social Change.* Cambridge, MA: MIT Press.

White, S. F., Brislin, S. J., Sinclair, S., and Blair, J. R. (2014). Punishing unfairness: rewarding or the organization of a reactively aggressive response? *Hum. Brain. Mapp.* 35, 2137–2147. doi: 10.1002/hbm.22316

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., and Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron* 40, 655–664. doi: 10.1016/S0896-6273(03)00679-2

Winston, J. S., Strange, B. A., O'Doherty, J., and Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat. Neurosci.* 5, 277–283. doi: 10.1038/nn816

Yamada, M., Camerer, C. F., Fujie, S., Kato, M., Matsuda, T., Takano, H., et al. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nat. Commun.* 3:759. doi: 10.1038/ncomms1757

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., et al. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organ. Behav. Hum. Decis. Process.* 120, 260–271. doi: 10.1016/j.obhdp.2012.06.002

Yang, Y., and Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiat. Res-Neuroim.* 174, 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Ye, H., Chen, S., Huang, D., Zheng, H., Jia, Y., and Luo, J. (2015). Modulation of neural activity in the temporoparietal junction with transcranial direct current stimulation changes the role of beliefs in moral judgment. *Front. Hum. Neurosci.* 9:659. doi: 10.3389/fnhum.2015.00659

Young, L., Cushman, F., Hauser, M., and Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8235–8240. doi: 10.1073/pnas.0701408104

Zaki, J., Weber, J., Bolger, N., and Ochsner, K. (2009). The neural bases of empathic accuracy. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11382–11387. doi: 10.1073/pnas.0902666106

Zheng, H., Huang, D., Chen, S., Wang, S., Guo, W., Luo, J., et al. (2016). Modulating the activity of ventromedial prefrontal cortex by anodal tDCS enhances the trustee's repayment through altruism. *Front. Psychol.* 7:1437. doi: 10.3389/fpsyg.2016.01437