DEPARTMENT OF MANAGEMENT

When do people cooperate? The neuroeconomics of prosocial decision making

Carolyn H. Declerck, Christophe Boone & Griet Emonds

UNIVERSITY OF ANTWERP
Faculty of Applied Economics

EARNED EXCELLENCE
AACSB ACCREDITED EST. 1916
THE BEST BUSINESS SCHOOLS IN THE WORLD

Stadscampus
Prinsstraat 13, B.226
BE-2000 Antwerpen
Tel. +32 (0)3 265 40 32
Fax +32 (0)3 265 47 99
http://www.ua.ac.be/tew

FACULTY OF APPLIED ECONOMICS

DEPARTMENT OF MANAGEMENT

When do people cooperate? The neuroeconomics of prosocial decision making

Carolyn H. Declerck, Christophe Boone & Griet Emonds

RESEARCH PAPER 2011-009
JUNE 2011

University of Antwerp, City Campus, Prinsstraat 13, B-2000 Antwerp, Belgium
Research Administration – room B.226
phone: (32) 3 265 40 32
fax: (32) 3 265 47 99
e-mail: joeri.nys@ua.ac.be

The papers can be also found at our website:
www.ua.ac.be/tew (research > working papers) &
www.repec.org/ (Research papers in economics - REPEC)

D/2011/1169/009

When do people cooperate?

The neuroeconomics of prosocial decision making.

Carolyn H. Declerck, Christophe Boone, Griet Emonds

Corresponding author:

Carolyn Declerck
University of Antwerp, Faculty of Applied Economics
Prinsstraat 13, 2000 Antwerp

carolyn.declerck@ua.ac.be

May 31, 2011

When do people cooperate? The neuroeconomics of prosocial decision-making.

Abstract

Understanding the roots of prosocial behavior is an interdisciplinary research endeavor that has generated an abundance of empirical data across many disciplines. This review integrates research findings from different fields into a theoretical framework that can account for when prosocial behavior is likely to occur. Specifically, we propose that the motivation to cooperate is generated by the reward system in the brain (extending from striatum to the ventromedial prefrontal cortex), and that it can be modulated by two neural networks: a cognitive control system (centered on the lateral prefrontal cortex) that processes extrinsic cooperative incentives, and/or a social cognition system (including the superior temporal sulcus, the anterior medial frontal cortex and the amygdala) that processes trust signals. The independent modulatory influence of incentives and trust on the decision to cooperate is substantiated by a growing body of neuroimaging data and reconciles the apparent paradox between economic versus social rationality in the literature, suggesting that we are in fact wired for both. Furthermore, the theoretical framework can account for substantial behavioral heterogeneity in prosocial behavior. Based on the existing data, we further postulate that self-regarding individuals (who are more likely to adopt an economically rational strategy) are more responsive to extrinsic cooperative incentives and therefore rely relatively more on cognitive control to make (un)cooperative decisions, whereas other-regarding individuals (who are more likely to adopt a socially rational strategy) are more sensitive to trust signals to avoid betrayal and recruit relatively more brain activity in the social cognition system.

Key words: neuroeconomics, social dilemmas, social preferences, cooperation, altruistic punishment

When do people cooperate? The neuroeconomics of prosocial decision-making.

Introduction

The high level of voluntary cooperation in social interaction can be considered as one of the unique and defining human features. Teamwork and collective action have undeniably contributed greatly to the success of economy-based societies. But as cooperation often involves a reciprocal exchange of benefits in an interdependent fashion, it also provides an opportunity for exploitation. Some people may be tempted to free-ride on the cooperation of others, and hence profit from cooperative benefits without contributing. Such opportunistic behavior may easily undermine the effectiveness of cooperative action, and poses a challenge to evolutionary and economic explanations of cooperative behavior, especially in exchange situations where one has the choice between a self-interested strategy versus a strategy that benefits the whole group, but at a personal cost. Classic economic theory predicts that a *Homo economicus*, who compares the costs and benefits of different courses of action, will not cooperate given this dilemma. However, if no-one cooperates, everyone is worse off (Dawes & Messick, 2000). Yet an abundance of field and experimental research has revealed that in all cultures people are willing to pay the cost of cooperation, even in anonymous situations where the probability of future repayment is zero (Henrich et al., 2005).

This cooperation dilemma has sparked researchers in many scientific domains to pool their efforts in order to understand why costly prosocial behavior persists despite the high levels of uncertainty intrinsic to many social exchanges. The search for the roots of human cooperation has produced two mostly independent streams of research that have revealed two fundamentally different logics behind prosocial behavior, one claiming that cooperation is economically

rational, the other that it is socially rational. Following *economic rationality*, cooperation is the product of natural selection acting on the individual (Hagen & Hammerstein, 2006). People are naturally motivated to pursue self-interest, but cooperate readily when self-interest coincides with collective interest. Hence this research stresses the importance of extrinsic incentives that align self- and collective interest prompting people to act prosocially to reap personal benefits from cooperative interactions (e.g., Bornstein, 2003; Kollock, 1998).

Adherents of *social rationality* consider cooperation to be compatible with theories of group selection (e.g., Sober & Wilson, 1998; Wilson & Sober 1994). People are intrinsically motivated to cooperate and feel good doing so, because the cooperative default has been selected in evolution due to the benefits that accrue in the group. Showing one's willingness to cooperate is thereby an effective way to strengthen belonging, build social networks, and avoid ostracism (Caporael, Dawes, Orbes, & Vandekragt, 1989). However, a group of cooperating individuals is very vulnerable to invasion by free-riders. Therefore, research on social rationality stresses the importance of trust in social interaction (e.g., Marckoczy, 2004; Haselhuhn & Mellers, 2005; Yamagishi, 1998; Yamagishi & Sato, 1986), and the threat of ostracism which may result from breaches of trust (Williams, 2007).

The purpose of this review is to develop a theoretical framework that integrates these two contrasting views regarding the rationality behind cooperation. Specifically we advance two new and general propositions. First, we propose that economic and social rationality are not contradictory for the brain, but that we are in fact wired for both. Economically and socially rational choices are rooted in different neural networks that operate in concert and independently modulate decision making. Cooperative decisions can be explained as motivated choices that yield either economically valuable or social rewards. However, these choices are contingent on the presence of extrinsic incentives that align self- and collective interest (following economic

rationality), and/or trust signals that minimize the chance of exploitation (consistent with social rationality). Therefore, brain systems that process extrinsic incentives and trust are expected to modulate the willingness to cooperate. The second proposition is that individual differences in self- versus other-regarding preferences coincide with economic and social rationality and influence the relative extent to which these brain networks will be recruited in cooperative decision-making.

To create an overarching model that spans both the external events and the neural networks influencing prosocial behavior (e.g., Beauchamp & Anderson, 2010), we rely on current research in the new and growing field of neuroeconomics, a joined effort of psychologists, neuroscientists, and behavioral economists with the primary purpose of opening the "black box" of human decision making. The neuroeconomic approach combines the rigorous experimental paradigms from game theory with neuroimaging techniques in order to identify the brain regions that are recruited during decision-making (Camerer, Loewenstein, & Prelec, 2005; Camerer, 2008; Kahneman, 2003). Neuroimaging experiments that investigate choice behavior in combination with contextual influences and individual motives are able to reveal the interplay between affective and cognitive processes that influence conscious deliberation. With this approach we hope to fine-tune the "rational choice" behind cooperation and present a framework for solving social dilemmas that incorporates both economically and socially rational motives, resolving the apparent paradox that emerged in the cooperation literature.

This review is organized as follows. We begin by outlining the two contrasting views in the literature – one based on economic, the other on social rationality – that associate cooperative behavior with (respectively) extrinsic incentives or trust. Next, we propose a neural model that incorporates the modulatory role of extrinsic incentives and trust in generating a prosocial decision in the face of uncertainty (e.g., in a social dilemma that offers a cooperative versus a

self- serving choice). We also propose a second model that identifies the brain systems that are recruited when prosocial norms are violated, as this tends to lead to altruistic punishment. The latter is included because of its presumed importance in sustaining long-term cooperation (e.g., Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gachter, 2002), and because it has generated much experimental and neuroeconomic research. We then review the individual differences in cooperative behavior and altruistic punishment. We include studies that have addressed behavior and/or neural correlates of self- versus other-regarding individuals, and investigate if their prosocial choices are associated with different patterns of brain activation. This would indicate that there are different individual drivers for prosocial behavior that correspond to distinct types of rationalities. Finally, we outline avenues for future research.

Economically versus socially rational reasons to cooperate

Extrinsic Incentives

Most of the evidence pointing to the role of extrinsic incentives in cooperation comes from experimental economic games that simulate social dilemmas. Social dilemmas are mixed motive situations in which self-interest and fear of betrayal tend to pull people towards non-cooperation. Extrinsic cooperative incentives are ecologically relevant, context-related incentives that align self-interest with collective interest and thereby remove the temptation to free-ride and motivate people to cooperate. Thus extrinsic incentives objectively transform the pay-off matrix of a social dilemma so that cooperation becomes an economically rational choice yielding tangible rewards. Such economic motives convince even those who are not naturally inclined to cooperate.

Extrinsic incentives come in many different forms. First, realizing the benefits of long-term cooperative relationships is one of the more common reasons that compels someone to pay for an initially costly cooperative act. This is true for the "tit for tat" strategy, where a person in a dyadic interaction starts out cooperating and thereafter reciprocates all responses of the partner. When both partners end up cooperating mutually, profits accrue over time, making "tit for tat" an evolutionary stable strategy and cooperating a rational choice (Axelrod & Hamilton, 1981). Second, when cooperation leads to synergy, as in team tasks where team members possess complementary skills, there is no conflict between self- and collective interest and cooperation becomes beneficial to all. Not surprisingly, in social dilemma games, synergy in pay-offs tends to significantly increase cooperation (Camerer & Fehr, 2006; Boone et al., 2010; Boone, Declerck, & Suetens, 2008).

Third, people cooperate to acquire a reputation for being generous. Conspicuous prosocial behavior may well be a self-presentation strategy to increase one's status in a group and benefit from indirect reciprocity (Nowak & Sigmund, 2005). Corroborating this "competitive altruism" hypothesis, Hardy and Van Vugt (2006) showed that, the more generous experimental participants are in social dilemma games, the more often they are chosen as interaction partners, and the higher their status. In real life situations, generous donations to charity correspond to higher sympathy and trustworthiness scores (Bereczkei, Birkas, & Kerekes, 2007). Not surprisingly, social cues that reduce anonymity and introduce possible audience effects are very effective in increasing prosocial behavior (Kurzban, DeScioli, & O'Brien, 2007; Piazza & Bering, 2008).

Finally, sanctions that substantially increase the cost of non-cooperation make up a powerful class of negative incentives. The threat of verbal criticism, poor reputation, or monetary fines are very effective in encouraging people to abide by prosocial societal norms. Corroborating

this is a large-scale experimental study across 15 different societies indicating that, for each culture, the incidence of punishment in experimental social dilemma games correlates positively with a measure of altruism (Henrich et al., 2006).[i] In experimental public good games (explained in appendix 1) the opportunity to inflict costs on free-riders is essential to maintain reciprocal cooperation (Fehr & Gachter, 2002). When participants of such games are further given the choice between a game without sanctioning opportunity versus one where sanctioning is possible, they prefer the latter (Fehr & Rockenbach, 2004). In real life, it also appears that institutions that make use of punishment enjoy a competitive advantage over institutions that don't (Gurerk, Irlenbusch, & Rockenbach, 2006).

Trust

Cooperation becomes a socially rational choice when it contributes to group success and when people value group belonging. Regardless of whether people are extrinsically motivated to cooperate in the group because they expect explicit rewards from the group, or whether they are intrinsically motivated because it gives them pleasure to see the group succeed, they want to minimize the possibility that others in the group will free-ride given their cooperative decision. Therefore, trust is an essential ingredient of cooperating groups, because it subjectively transforms the fear of betrayal to positive expectations of reciprocity.

The relation between social cues signaling trust and cooperative behavior has been well investigated by social psychologists. To begin with, communication, which removes uncertainty, enhances cooperation in social dilemmas (see the meta-analysis by Balliet, 2010). Subtle cues that reduce social distance increase trust and cooperation (Hoffman, McCabe, & Smith, 1996). This is further revealed by experiments that indicate that one is more cooperative towards people

of flesh and blood than towards computers (Kiesler, Sproull, & Waters, 1996), towards friends than towards strangers (Thompson, Kray, & Lind, 1998; Yamagishi & Sato, 1986), and towards someone named a "partner" compared to an "opponent" (Burnham, McCabe, & Smith, 2000). Cooperation also increases when previously unknown interacting parties merely introduc themselves (Boone et al., 2008), when participants perceive demographic similarity between themselves and their partner (Cole & Teboul, 2004) or when pictures of the interacting person show facial resemblance (DeBruine, 2002).

Section summary

In this section we briefly review evidence from two separate literatures that point to two different logics behind costly cooperative behavior in social dilemmas: an economically-based rationality that is associated with processing extrinsic incentives, and a more socially oriented rationality that promotes group functioning and requires trust. We identified several classes of extrinsic incentives, including accruing benefits over time, synergy, and reputation gains, all of which align self-interest with collective interest. We also describe various ways in which social signals can build trust and reduce the "fear of betrayal." Each in their own way, extrinsic incentives and trust signals make cooperation an attractive and rational choice.

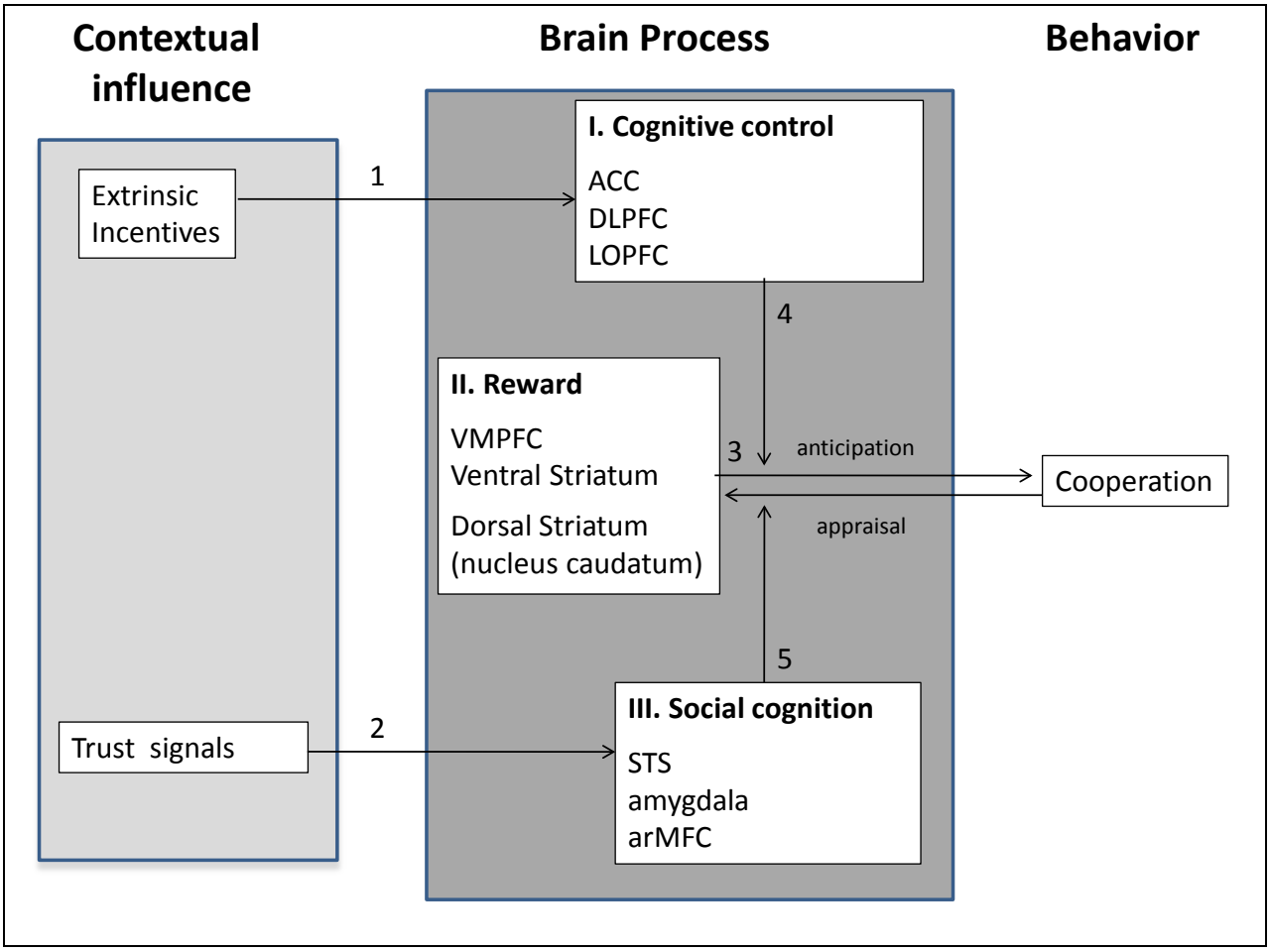
The neural correlates of costly prosocial behavior: Cooperation

Existing reviews in neuroeconomics have revealed that three brain systems are consistently recruited when people face social dilemmas. These are the neural networks dedicated to cognitive control, social cognition, and processing reward (Sanfey, 2007; Fehr & Rockenbach,

2004; Fehr & Camerer, 2007; Tabibnia & Lieberman, 2007; Walter, Abler, Ciaramidaro & Erk, 2005). We propose that these three brain systems are linked together in such a way that a cooperative decision in a social dilemma is the result of the modulatory influences of cognitive control and social cognition on the reward processing system of the brain. We present this modulating mechanism in the model presented in Figure 1.

Figure 1

Theoretical framework identifying the neural networks recruited to solve social dilemmas. Numbered arrow labels are explained in the text.
(ACC = anterior cingulate gyrus; DLPFC = dorsolateral prefrontal cortex; LOPFC = lateral orbitofrontal cortex; VMPFC = ventromedial prefrontal cortex; STS = superior temporal sulcus; arMFC = anterior medial frontal cortex)



From a neuroeconomic point of view, any type of decision-making will activate the mesolimbic reward system (Figure 1, Box II), because components of this network are involved in processing the possible outcomes of the decision, including computing the probability of a rewarding outcome, encoding the saliency of reward, and informing connecting regions to update behavior if rewards are not as expected (O'Doherty, 2004). The orbitofrontal cortex within the VMPFC is specifically involved in forming anticipatory expectations of reward (Knutson et al., 2005), and in appraising whether or not these expectations are met following a decision (Peterson, 2005), (see Figure 1, double arrow 3). But the reward system is impartial as to whether the decision should be (un)cooperative. Our model proposes that the direction of the decision will

depend on activity in the cognitive control system (Figure 1, Box I), which processes prevailing extrinsic incentives (Figure 1, arrow 1), and the social cognition system (Figure 1, Box III), which processes the presence or absence of trust signals and hence minimizes the chance of betrayal (Figure 1, arrow 2). Via these two systems, extrinsic incentives and trust affect the motivation to cooperate generated by the reward system (Figure 1, arrows 4 and 5). When extrinsic incentives predominate, cooperation becomes economically rational. When extrinsic incentives are absent but trust is present, cooperation can still be socially rational.

In this section we review a number of studies which corroborate that indeed these three brain systems are involved in generating prosocial behavior and that they are likely to act together in a modulatory way. A summary of these neuroimaging studies is given in Table 1.

Table 1

Summary of fMRI studies that address cooperative behavior and indicate activity in brain regions associated with processing reward, cognitive control, and social cognition. (VMPFC = ventromedial prefrontal cortex; ACC = anterior cingulated cortex; DLPFC = dorsolateral prefrontal cortex; LOPFC = lateral orbitofrontal cortex; arMFC = anterior medial frontal cortex; STS = superior temporal sulcus).

*Transformed from MNI to Tal coordinates using BrainMap GingerAle 2.0.2 (Eickhoff et al., 2009), and rounded to the nearest whole numeral.

Study	Contrast	Brain region	Side	Talarairch coordinates X, Y, Z		
Studies that show activity in reward system						
Decety et al., 2004	Target pattern building game, cooperation versus competition	VMPFC	L	-12	33	-4*

Rilling et al., 2002	Repeated prisoner's dilemma game with human partner, mutual cooperative outcome versus average of other 3 possible outcomes.	VMPFC	L	-4	46	14*
Rilling et al., 2004a	One-shot prisoner's dilemma game with human partner, reciprocated versus unreciprocated cooperation.	VMPFC	R	8	39	-5
Tabibnia et al., 2008	Ultimatum game, high fairness offers versus low fairness offers	VMPFC	R L L	8 -14 -16	55 29 15	6* -2* -9*
Harbaugh et al., 2007	Regression analysis explaining activations during mandatory transfers as a function of transfer amounts (to subject and charity) in anatomical regions of interest	Ventral striatum	R R L L	6 8 -8 -10	2 8 2 8	8* -0* 8* -1*
Moll et al., 2006	Frequency of charitable (costly) donations correlated with ventral striatum	Ventral striatum	L	-6	11	4
Rilling et al., 2002	Repeated prisoner's dilemma game with human partner, mutual cooperative outcome versus average of other 3 possible outcomes	Ventral striatum	R	3	18	0
Rilling et al., 2004a	One-shot prisoner's dilemma game with human partner, reciprocated versus unreciprocated cooperation	Ventral striatum	R	10	17	-4
Tabibnia et al., 2008	Ultimatum game, high fairness offers versus low fairness offers	Ventral striatum	L	-7	2	4*
Studies that indicate extrinsic incentives influence cognitive control system						
Emonds et al., 2010	Prisoner's dilemma versus coordination game	ACC	R L L	2 -3 -1	18 9 7	44 46 48
Emonds et al., 2010	Prisoner's dilemma versus coordination game	DLPFC	R R L L	44 46 -28 -28	31 21 29 44	27 24 34 5
Spitzer et al., 2007	Dictator game with punishment versus dictator game without punishment	DLPFC	R R L L	34 25 -31 -40	31 20 20 38	33* 28* 31* 25*
Rilling et al.,	Repeated prisoner's dilemma game,	LOPFC	L	-42	27	6

2008	unreciprocated cooperation followed by defection versus unreciprocated cooperation followed by cooperation					
Spitzer et al., 2007	Dictator game with punishment versus dictator game without punishment	LOPFC	R L	40 -38	38 33	3* -6*
Studies that indicate trust signals influence social cognition system						
Fukui et al., 2006	Chicken game with human partner versus computer partner	arMFC	L	-8	43	35
Gallagher et al., 2004	Rock, paper, scissors game, mentalizing versus rule solving	arMFC	R L	8 -10	54 50	12 30
Rilling et al., 2004 b	Human versus computer partners in Ultimatum game and prisoner's dilemma	arMFC	R	3	44	20
Winston et al., 2002	Explicit trustworthiness judgment of face versus judging gender of face	arMFC	L	-27	37	16*
Fuikui et al., 2006	Chicken game with human partner versus computer partner	STS	L	-57	-51	28
Rilling et al. 2004 b	Prisoner's dilemma, picture of human face versus picture of roulette wheel	STS	R	40	-55	32
Singer et al., 2004	Evaluating faces of intentional cooperators versus neutral faces	STS	R L	54 -51	-48 -9	18 -15
Winston et al., 2002	Explicit trustworthiness judgment of face versus judging gender of face	STS	R	51	-43	5*
Studies that indicate modulation of reward system by trust signals						
Baumgartner et al., 2008	Trust game, placebo group versus oxytocin group	Dorsal striatum	R R R L	10 10 7 -9	3 -1 7 6	13* 21* 5* 15*
Delgado et al., 2005	Trust game, positive outcome (gain) versus negative outcome (loss)	Dorsal striatum	R	13	8	1
King-Casas et al., 2005	Benevolent versus malevolent reciprocity	Dorsal striatum	R L	10 -8	20 17	10* 9*

The reward system generates the motivation to cooperate.

The reward system, formed by the mesolimbic dopaminergic pathway running from the ventral striatum to the ventromedial prefrontal cortex (VMPFC) is activated when people receive or expect to receive a reward (Preuschoff, Bossaerts, & Quartz, 2006; Tricomi, Rangel, Camerer, & O'Dohery, 2010). Therefore, we postulate that this brain system will also be activated when people expect a reward following a cooperative decision. This would generate the willingness or the necessary motivation to cooperate. The rewards people expect from cooperation can be lucrative (monetary compensations, prizes, status) or affective, creating hedonic feelings (compliments, satisfaction, inner peace, or a "warm glow"). In a review article, Fehr and Camerer (2007) show that there is extensive overlap between on the one hand the brain circuitry that anticipates or represents extrinsic types of rewards that are purely for one's own, such as anticipated monetary recompenses or other self-serving rewards, and on the other hand the circuitry activated by intrinsic social reward. This suggests that, for the brain, extrinsic lucrative rewards resulting from rationally calculating the costs and benefits of cooperation are equivalent to the intrinsic hedonic rewards stemming from socially motivated cooperative decisions.

To evaluate whether or not these (material or hedonic) expectations are met, dopamine neurons running from the ventral striatum to the VMPFC fire in response to the magnitude of the obtained reward. They code the quality (positive or negative) of the so called dopamine prediction error, i.e. a positive prediction error signal following a better than expected reward, and a negative signal following the omission of an expected reward (Abler et al., 2006; Schultz, 1997). Thus, depending on the value of the dopamine prediction error, cooperative decisions-

making could be sustained or interrupted to be replaced by decisions that may lead to more satisfying outcomes.

Rilling et al. (2002; 2004a) suggested a possible role of the dopamine prediction error in evaluating the outcome of a cooperative decision in an fMRI experiment using the prisoner's dilemma paradigm. The prisoner's dilemma (described in detail in Appendix 1) is a classic non-zero sum game often used to investigate cooperation in evolutionary biology, social psychology, and behavioral economics. In a first experiment, each participant played 10 one-shot games with supposedly 10 different partners and thus with no prospect for future repayments (Rilling et al., 2004a). After each game, the participant received feedback regarding the decision of the other player. Given that human behavior is rarely 100% predictable, each player would have his own estimate of the likelihood that cooperation is reciprocated, which may or may not correspond to the actual feedback. Brain contrasts between reciprocated and unreciprocated cooperation indicated that only reciprocated cooperation activated the ventral striatum and the VMPFC. The authors suggest that this activity reflects a positive prediction error (the discrepancy between reward estimate and actual outcome) from which the player can learn whether to keep on cooperating or switch to a different strategy.

Likewise, in a second experiment where the prisoner's dilemma game was repeatedly played with the same partner (Rilling et al., 2002), mutual cooperation was also associated with increased activity in the VMPFC and the ventral striatum. Most participants reported that they found mutual cooperation the most personally satisfying outcome, despite the fact that the game structure was such that in each round free-riding yielded greater material gains. Interestingly, the activation of the striatum was restricted to mutual cooperation in *social* interactions, and was not observed when participants played the same game with a computer partner. This suggests that

cooperation has rewarding value over and above the material rewards obtained from unilateral defection.

The intrinsically rewarding value of cooperation without remuneration has repeatedly been illustrated with fMRI. Decety et al. (2004) compared the neural correlates of cooperation and competition during a computer game whereby players had to construct a target pattern either alone, with another person (cooperation) or against another person (competition). Compared to competition, cooperation was associated with increased activity in the medial orbitofrontal cortex, anterior frontal cortex, and posterior cingulate cortex (all contained within or touching on the VMPFC). Decety et al. interpret these results to be consistent with prepositions that people seek cooperation because it is socially more rewarding than competing, and that rewarding value of cooperation may stem from the psychological satisfaction of establishing a commonality with a conspecific.

The role of the VMPFC in generating positive affect from social interaction is further corroborated by a study of patients with VMPFC lesions (Krajbich et al., 2009). Compared to normal controls and other brain damaged control patients, the VMPFC patients were less generous and less trustworthy in a battery of economic games, which the authors contributed to their inability to experience social emotions (see also Bechara and Damasio, 2005). Therefore, VMPFC patients may also fail to appreciate the intrinsic rewards that come with acting prosocially.

The hedonic rewards stemming from cooperation may also result from the fact that many people are intrinsically motivated to cooperate out of fairness considerations (Tabibnia & Lieberman, 2007). An fMRI study further illustrates this by examining the neural responses to "fair" and "unfair" monetary offers from a human partner in a dictator game (described in Appendix 1). Compared to "unfair offers" of equal monetary value, "fair offers" led to higher

happiness ratings by the receiver and greater activation of the VMPFC and the ventral striatum (Tabibnia, Satpute, & Lieberman, 2008). A subsequent experiment shows that these same regions are activated when a player under the scanner observes that unfair offers to another person are subsequently reduced by transferring additional money into the account of the underprivileged party (Tricomi et al., 2010).

Finally, pure altruistic acts can be rewarding because they often generate a "warm glow" associated with giving. The neural correlates of this "warm glow of giving" were investigated in an fMRI study of people donating to charity (Harbaugh, Mayr, & Burghart, 2007). Mandatory money transfers (including from the player's own account) into a charity's account corresponded to increased activation in the ventral striatum. Apparently, people like to see a charity receive money. Participants who showed more activation in the ventral striatum were more likely to subsequently contribute voluntarily to the charity, and also reported more satisfaction from giving. In the charitable donations study by Moll et al. (2006), striatal regions were activated both by donating and by receiving monetary rewards, once again corroborating that obtaining economic and social rewards share anatomical brain systems, and that giving is rewarding in and of itself. In addition, decisions to donate activated the subgenual area, a brain region connected to the mesolimbic dopamine reward system and implicated in social attachment formation. Thus, when cooperation is not remunerative (and therefore not economically rational), its rewarding value is strengthened by personal feelings of social affiliation.

So far the role of the *VMPFC* and *ventral striatum* of the reward system have been discussed in shaping and appraising expectations of reward, which holds regardless of whether the rewards are material (economic) or hedonic (often social) in nature. We proposed earlier that contextual information regarding the likelihood of attaining these rewards (extrinsic incentives and trust) should be able to alter (or modulate) decision-making, a process which is dependent on

activity in the *dorsal striatum* (Figure 1, box II). Especially the caudate nucleus, located in the dorsal striatum and previously linked to learning and memory, is an important component in the reinforcement of actions potentially leading to reward (Tricomi, Delgado, & Fiez, 2004). It is considered part of the reward system because it is involved in updating information to either modify future behavior, or to keep the status quo as long as expectations are met (Delgado, 2007; O'Doherty, 2004).

How does the reward system register outside information? Functional connectivity analyses has confirmed that the VMPFC and the ventral and dorsal striatum integrate inputs from other networks, including those of cognitive control (DLPFC and LOPFC; see Staudinger, Erk, Abler, & Walter, 2009), and social cognition (the STS, see Hare, Camerer, Knoepfle, & Rangel, 2010). Interestingly, activity in the caudate nucleus can be altered by information regarding the trustworthiness of others (Baumgartner et al., 2008; Delgado, Frank, & Phelps, 2005; King-Casas et al., 2005), a finding we return to in detail later. Thus, at the neural level it is possible that the motivation to cooperate (or not) driven by the reward system can be altered by activity in brain regions that process outside information, such as extrinsic incentives and/or social signals indicating trust (see arrows 4 and 5 in Figure 1). We turn to this modulation next.

The cognitive control system processes extrinsic incentives and affects cooperative motivation.

Appreciating that cooperation is economically rational and that it may have the potential for compensation when there is also a strong temptation to defect, requires cognitive resources. Some authors (e.g. Stevens & Hauser, 2004) believe that reciprocal cooperation is rare among animals because the cognitive capacity to weigh the benefits of cooperative versus non-

cooperative decisions emerged only late in evolution with the great anatomical expansion of the dorsolateral prefrontal cortex in hominid primates (Previc, 1999). As an important component of working memory and executive functions (Miller & Cohen, 2001), the dorsolateral prefrontal cortex, or DLPFC (Figure 1, box I), may provide the cognitive capacities needed to resolve the conflict generated by the mixed motives of a social dilemma (Carter & van Veen, 2007). Therefore, the DLPFC may be especially important in calculating the benefits (or lack there-off) of a decision in the presence or absence of cooperative incentives. An additional function of the DLPFC is to provide the impulse control to resist immediate selfish urges in order to realize greater cooperative benefits at a later time (McClure, Laibson, Loewenstein, & Cohen, 2004).

In a long-term relationship or an endlessly played prisoner's dilemma, the iterated benefits generate incentives that turn the temptation to free ride into a game of eliciting reciprocity. Presumably people understand clearly that a long-term, mutually cooperative relationship is economically more beneficial (and therefore rational) compared to unilateral defection. Even when uncertainty is high, it may still be worth to pay the initial cost of cooperation rather than putting future reciprocity at stake.[ii] To do so, however, requires overcoming a putative bias that humans and other animals have, named time discounting: people tend to weigh the attractiveness of a reward in inverse proportion to its delay. McClure et al. (2004) hypothesized that the discrepancy between short-term and long-term choices in human time discounting reflects the differential activation of distinguishable neural systems. They measured brain activity of participants as they made a series of inter-temporal choices between early and later monetary rewards. The early rewards always had a lower (undiscounted) value than the later ones. Their fMRI results indicate that long-term patience is mediated by the lateral prefrontal cortex including the right DLPFC, the right ventrolateral prefrontal cortex and right

lateral orbitofrontal cortex.[iii] The degree of engagement of these regions predicts the ability to defer gratification.

Working in concert with the DLPFC, the anterior cingulate cortex (ACC, Figure 1, box 1) is involved in conflict monitoring whenever there are competing motives (Carter, Botvinick, & Cohen, 1999; Rilling et al., 2002). When cooperation leads to synergy and all cooperating parties benefit equally, there is no temptation to free-ride, and hence no conflict between self- and collective interest. Thus, the presence of synergetic incentives should relax the need for cognitive control. This was tested in an fMRI experiment by Emonds et al., (2011a) that contrasted brain activity elicited in a one-shot prisoner's dilemma game and a coordination game (see Appendix 1). The main difference between these two games is that the coordination game offers synergetic incentives to cooperate while the prisoner's dilemma generates conflicting motives. Because the experiment offered no chance to develop a long-term partnership, the best response in this one-shot PD was to defect. As expected, playing a prisoner's dilemma game generated more defect decisions and activated the DLPFC and ACC significantly more than playing a coordination game. Cognitive control registers the absence of synergetic incentives and modulates decision towards the economically rational best response, which, in the case of a one-shot prisoner's dilemma game, is not cooperating.

The involvement of cognitive control when cooperation leads to reputation formation was investigated using repetitive transcranial magnetic stimulation (rTMS). This method allowed researchers to block out the lateral PFC in participants playing a repeated trust game with different partners (Knoch et al., 2009). The details of the trust game are explained in Appendix 1. A trustee in this game could show gratitude towards an investor by back-transferring a portion of the money allotted to him. When information regarding the behavior of the trustee was available to future investors in the game (the reputation condition), back-transfers from the trustee to the

investor were double of what they were in anonymous transactions, at least in the no-rTMS condition. When, however, the right DLPFC is disrupted with rTMS, the trustee lacks the necessary control to overcome the temptation to defect. The back-transfers in the reputation condition were therefore significantly lower under rTMS, even when trustees knew that defecting was irrational because it would put their reputation at stake.

How sanctioning affects the neural mechanism of prosocial behavior was illustrated in an fMRI study by Spitzer et al. (2007). Participants played two versions of a dictator game. In the "no punishment" version, player 1 (inside the scanner) received a sum of money and was asked to split it in any proportion between himself and another player (outside the scanner). In the "punishment" condition, player 2 could impose a monetary sanction if the allotment was perceived as unfair. On average, the presence of the punishment threat doubled the level of generosity. Contrasting brain activity between the two conditions revealed that the threat of punishment was associated with increased activity bilaterally in the DLPFC, the LOPFC, and the ventrolateral cortex. The role of the LOPFC in evaluating punishing stimuli (as opposed to rewarding stimuli) has been demonstrated before (Kringelbach & Rolls, 2004).

The sensitivity of the LOPFC to punishment and its importance in solving social dilemmas was further shown by Rilling et al. (2008). In a repeated prisoner's dilemma game, they found that activity in the LOPFC following a defect decision of the partner could predict a subsequent defect decision. A cooperative person who experiences that cooperation is not reciprocated needs to switch strategies. We postulate that this is only possible if the LOPFC informs the caudate nucleus in the reward system (dorsal striatum) that there is no longer an incentive to cooperate and that the current behavior should be updated (O'Doherty, 2004).

The social cognition system processes trust signals and affects cooperative motivation.

Social cognition refers to understanding other people, and includes mental processes such as social perception, mind reading, and face recognition. Social cognition became essential in the evolutionary history of humans when they adopted group living and hence needed a way of bookkeeping the (un)honorable activities of fellow group members (Dunbar, 1998). A neural basis for the social cognition system was first proposed by Brothers (1990). Based on animal lesion studies, single cell recordings and neurological studies, she described social cognition as a function of three regions, collectively called the "social brain:" the amygdala, the medial frontal cortex (MFC), and the superior temporal sulcus (STS) and gyrus. Numerous later studies have confirmed that these regions are involved in forming expectations of others during social interactions (for the amygdala, see Baron-Cohen et al., 2000; Stone et al., 2003, for the anterior part of the MFC, or arMFC,[iv] see Amodio & Frith, 2006; Frith & Frith, 2003; for the STS see Gallagher & Frith, 2003; Frith & Frith, 2006). Furthermore, fMRI studies of social dilemmas in particular have revealed increased arMFC activity when players were asked to detect the intentional stance in the children's game "rock-paper-scissors" (Gallagher, Jack, Roepstorff, & Frith, 2002), and STS activity in a game of "chicken" (see Appendix 1) that requires mindreading (Fukui et al., 2006).

Regardless of whether people are motivated to cooperate for economically rational reasons (when there are explicit incentives indicating material gains) or whether they are intrinsically motivated to do so for social, affective reasons, confronting betrayal remains an issue in a social dilemma. In Figure 1 we suggest that the "social cognition system" (Box III) processes trust signals which helps to temper the fear of betrayal so that the goal of attaining a rewarding outcome can be pursued safely (Acevedo & Krueger, 2005; Boone et al., 2008; Simpson, 2004).

The neural correlates of trustworthiness judgments were first investigated in an fMRI study by Winston, Strange, O'Doherty, and Dolan (2002) and proved to be very similar as those processing intentions or other mindreading tasks. Participants viewed pictures of faces while under the scanner. When asked to explicitly judge trustworthiness, the right STS showed increased activation. In addition, untrustworthy faces (more than trustworthy faces) elicited right amygdala activation, regardless of whether the judgment was made explicitly, or implicitly while performing another task. Automatic right amygdala response during implicit evaluation of untrustworthy faces is also reported in the study by Engell, Haxby, and Todorov (2007). These findings are consistent with reports that the amygdala may have a role in enhancing perceptual processing of threat stimuli (Anderson & Phelps, 2001), alerting a decision-makers that they may be in a precarious situation.

Implicit evaluation of the trustworthiness of partners in a (one-shot) prisoner's dilemma game was investigated in an fMRI study by Rilling et al. (2004b). They compared brain activity during the decision making phase in two conditions: a social condition where the pictures of the faces of human partners were shown to participants under the scanner, and an asocial condition where the partner was depicted with pictures of roulette wheels. The human faces elicited more activity in the STS and in the paracingulate cortex (within the arMFC) compared to the roulette wheels.

Singer et al. (2004) used fMRI to investigate implicit social judgments of people who differ in acquired moral status. To manipulate the moral status, the initial phase of the experiment consisted of letting participants play a series of sequentially played prisoner's dilemma games (see Appendix 1) with several identifiable partners who would either intentionally or unintentionally reciprocate or betray a cooperative participant. In the second phase of the experiment, the participant was shown pictures of the faces of these partners while in the scanner.

The authors predicted that seeing the face of a partner who had acted intentionally would elicit automatic emotional responses reflecting trustworthiness. Indeed, seeing faces of intentional cooperators (as opposed to non-intentional cooperators) was associated with increased activity in the (bilateral) posterior STS, the lateral orbitofrontal cortex, the left amygdala and the ventral striatum.

A number of experiments have shown that trust signals may specifically affect cooperative decision making by modulating activity in the caudate nucleus within the dorsal striatum of the reward system. Because of its role in contingency learning, the caudate nucleus is well suited to process (affective) feedback information related to the likelihood of others' continued cooperative behavior (Tricomi, Delgado, & Fiez, 2004). King-Casas et al. (2005) performed a hyperscan fMRI study whereby they concurrently scanned two participants playing a repeated trust game (see Appendix 1). As player 1 (the investor) increases his investments, player 2 (the trustee) increases repayments. Increases in investments were associated with greater activation of the caudate nucleus in the trustee's brain. Additionally, an interesting pattern in the timing of caudate nucleus activation emerged. At first, the increase in caudate nucleus activation in the trustee's brain occurred after the investment of the investor was revealed to the trustee (reactive activity). As the experiment progressed, the caudate nucleus activation shifted to a time prior to knowing the investor's investment (anticipatory activity). Based on these results, the authors suggest that the caudate nucleus receives or computes information about both the fairness of a social partner's decision and the intention to repay this decision with trust. Thus, through reinforcement learning in the caudate nucleus, the reputation of the partner is developed over time and serves as input for future decision making.

Delgado et al. (2005) further investigated the effect of contextual social information on the neural correlates of decision making during a repeated trust game. The player in the scanner

received an amount of money which could be "kept" on any given trial, or "shared" with one of several partners outside the scanner. Prior to the scan session, the players were first given vivid descriptions of each partner's character. These could be neutral, praiseworthy, or morally suspect. "Good partners" always elicited more "sharing" from players than either neutral or bad partners, even after frequent violations of the expectations that were experimentally created in the game. At the neural level, activity in the caudate nucleus showed normal differential responses between positive and negative feedback from a "neutral" partner. However, when the partner was initially perceived as "good" or "bad," the differential neural response to feedback in the caudate nucleus was not as robust or completely nonexistent. The "good" ("bad") moral profile thus not only created positive (negative) expectations of reciprocity, it led players to discount feedback information and it prevented them from adapting their false beliefs and their behavioral choices according to past experiences. Thus, experimentally induced trustworthiness inhibits the caudate nucleus to respond adequately to feedback information and may lead to economically unwarranted, but socially motivated cooperative decisions.

The hypothesis that the caudate nucleus is involved in guiding and adjusting cooperative decision making on the basis of trust is further confirmed by a study investigating the neural effects of the neuropeptide oxytocin on this brain structure. Oxytocin is a nanopeptide hormone that typically facilitates reproduction in almost all mammalian taxa, and has been reported to increase trust among humans (Kosfeld, et al., 2005). It is naturally released in the brain in response to trust signals (Zak, Kurzban, & Matzner, 2004) or human touch in combination with trust (Morhenn, Park, Piper, & Zak, 2008). Baumgartner et al. (2008) used fMRI to show how this neuropeptide affects adapting to feedback information. Participants received either nasal oxytocin or a placebo and were scanned as they decided to invest or not in a number of trust games (Appendix 1). Of those who received oxytocin, none changed their willingness to invest

after being informed that their initial investments were repaid only 50 % of the time. In contrast, those who received a placebo significantly decreased their trusting behavior in subsequent games after receiving the same information. Consistent with these behavioral results, the oxytocin group showed reduced activation in the caudate nucleus compared to the placebo group. This indicates that subjects who receive oxytocin act as if they implicitly 'know' that they can trust their partners, and that they do not have to rely on feedback information. Thus, without the input of the caudate nucleus, no contingency between decision and feedback is perceived, and decision making will no longer be updated.

Section summary

The neuroeconomic studies we have reviewed so far lead us to infer that the reward system is central to prosocial decision-making under uncertainty. Through reward anticipation, the neural network running from the striatum to the VMPFC generates both the motivation to cooperate and a means to adapt decision-making to feedback or changing conditions. This neural network is impartial as to whether the rewards are lucrative (economically rational), or affective (socially rational). The motivation to cooperate is, however, contingent upon activity in other brain regions that are functionally connected to the reward system. Consistent with economic rationality, cooperative decisions occur when there are extrinsic incentives (processed by the brain's cognitive control system) making cooperation economically desirable. When cooperation is intrinsically motivated because of its social or hedonic value, the reward system still needs to respond to trust signals (processed by the brain's social cognition system) indicating that cooperation is a "safe" decision with minimal chance of betrayal. Several studies confirm that

contextual information regarding the trustworthiness of the partner alters activity in the caudate nucleus which in turn affects cooperative decision making.

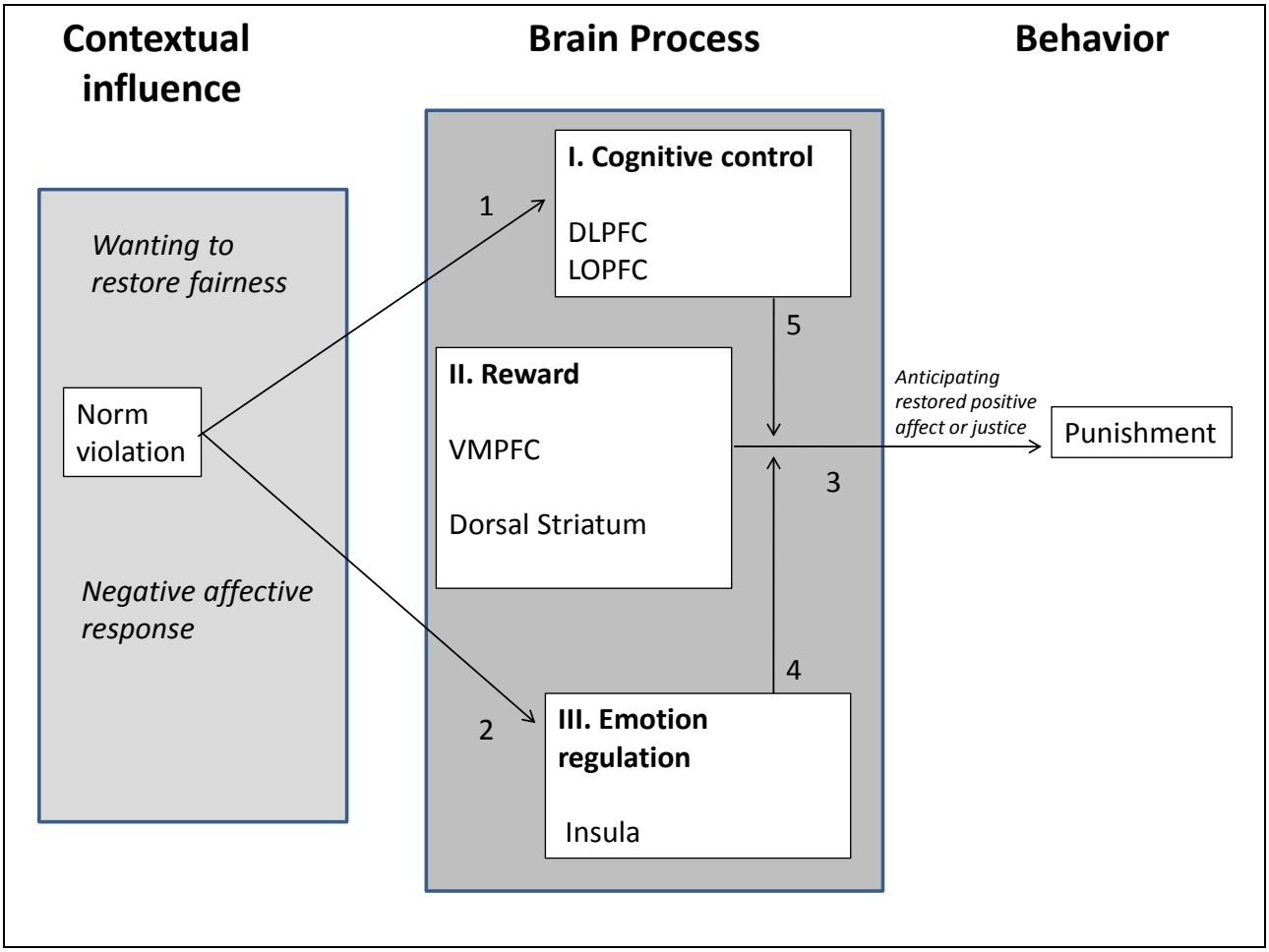
The neural correlates of costly prosocial behavior: Altruistic punishment

Because cooperation is such a vital part of society, sanctioning institutions are often established to discourage free-riding and to punish norm violation. However, even without institutions, people readily punish a breach of trust, even when it is costly and anonymous (without possible reputation gains), and hence unlikely to appear economically rational (Gintis et al., 2003). Because punishing norm violators is often considered a stabilizing act that sustains long-term cooperation (Boyd et al., 2003), it is considered a prosocial decision consistent with social rationality.

In Figure 2 we indicate which brain regions are involved when norm violation leads to punishment. Three regions are consistently associated with altruistic punishment: the lateral prefrontal cortex (DLPFC and LOPFC implementing cognitive control, see Figure 2, box I), the insula associated with an emotional response (Figure 2, box III), and again, the reward system to generate the "willingness to punish" (Figure 2, Box II, arrow 3). We propose that the decision to punish (arrow 3) can be the result of emotion regulation to undo negative emotions (arrows 2 and 4), or the result of exerting cognitive control in order to restore fairness (arrows 1 and 5).

Figure 2

Theoretical framework identifying the neural networks recruited in altruistic punishment. Numbered arrow labels are explained in the text.
(DLPFC = dorsolateral prefrontal cortex; LOPFC = lateral orbitofrontal cortex; VMPFC = ventromedial prefrontal cortex)



The reward system motivates costly punishment

Within the reward system, the VMPFC (Figure 2, box II) appears to be crucially involved in the decision to punish partners who break the fairness norm. Patients with VMPFC lesions tend to *not punish*, and accept abnormally low unfair offers in the ultimatum game (see Appendix 1) compared to healthy and brain-damaged comparison groups (Koenigs & Tranel, 2007). In addition, the dorsal striatum (particularly the caudate nucleus, see Figure 2, box II), is activated when punishing occurs with the anticipation of restoring the status quo, and re-establishing positive affect.

Evidence for the "sweet taste of revenge," insinuating that justified punishing feels rewarding, comes from a positron emission tomography (PET) study by De Quervain et al. (2004). The authors tested the hypothesis that people derive satisfaction when they punish someone who has intentionally abused their trust. Their experiment involved a trust game (see Appendix 1) where player 1 (in the scanner) has the opportunity to punish player 2 (an anonymous partner outside the scanner) if he finds that his trust has been betrayed. There were three punishment options: free (it did not cost anything to player 1 to impose a monetary fine on player 2), costly (player 1 had to pay to punish, but the fine imposed on player 2 was larger than the cost for player 1), and symbolic (neither player lost anything). Symbolic punishment was not effective and was therefore unlikely to feel satisfactory. Free and costly punishment on the other hand were both effective. As expected, contrasting brain activity of punishing decisions made in both the free and the costly condition with the decisions made in the symbolic condition showed increased activation in the caudate nucleus within the dorsal striatum. Questionnaire data confirmed that players who felt betrayed indeed experienced a strong desire to punish. In the free punishment condition, all participants punished maximally. In the costly punishment condition, players who showed more dorsal striatum activity punished more. Interestingly, players who showed more dorsal striatum activity in the free condition were also more likely to punish in the costly condition. Because all participants punished maximally in the free condition, the difference in striatum activation among the participants cannot be due to different levels of punishment, but rather reflects the expected satisfaction that will be derived from punishing.[v]

The proposition that it is the anticipated satisfaction of punishing that activates the dorsal striatum is compatible with social psychologists' view of emotion regulation (Tice, Baumeister, & Zang, 2004). Negative emotions impair self-regulation through a priority shift: when people are sufficiently distressed, they give priority to regulating affect. Pleasures and satisfactions are

the key to feeling good again. Therefore, distressed individuals punish (a behavior they normally want to avoid) because they anticipate it will make them feel better.

Anticipating positive affect is by itself not a sufficient reason to punish, but is contingent on an individual's emotional response to betrayal, and/or the need to restore equality. Therefore, as we proposed for cooperative decision-making, activity in the reward system leading to a potential punishing decision will be modulated by other brain regions involved in emotion regulation (the insula, see Figure 2, arrow 4) or cognitive control (specifically the DLPFC, see Figure 2, arrow 5). From functional connectivity analyses we know that the anterior insula is connected to the mesolimbic reward system (Preuschoff, Quartz, & Bossaerts, 2008), and that it can modulate the value signal in the VMPFC (Hare, Camerer, Knoepfle, & Rangel, 2010). Similarly, the DLPFC is highly connected to the dorsal striatum via the nigrostriatal dopamine pathway (Grahn, Parkinson, & Rangel, 2008).

Negative affect triggers emotion regulation and leads to punishment

A breach of trust or a violation of the cooperative norm by the partner is likely to cause negative emotions which may leave the betrayed person with the urge to punish in order to restore positive emotionality (Seip, van Dijk, & Rotteveel, 2009). Xiao and Houser (2005) showed that punishment in an ultimatum game (see Appendix 1) specifically serves to express negative emotions. Participants who were given the opportunity to write nasty messages to a partner that treated them unfairly were more willing to accept the unfair offers compared to participants who did not write notes. Hence the written expression relieved the participant from negative emotions and replaced the urge to punish with an economically rational decision-making scheme. When no such alternative emotion-regulation outlet existed, negative emotions drove the

decision to punish. This decision may be costly, but serves a socio-emotional function by re-establishing positive affect after betrayal.

Van't Wout and colleagues substantiated the behavioral findings that emotions can drive costly punishment with physiological data (Van't Wout, Kahn, Sanfey, & Aleman, 2006). These authors measured skin conductance of players interacting in an ultimatum game, and they found that it was significantly increased in response to unfair offers. More importantly, the increase was associated with greater punishment rates. Because skin conductance is often used as a proxy for affective arousal (Bechara & Damasio, 2005), this study convincingly indicates that negative emotions associated with breaking the fairness norm can in fact drive costly punishment.

Neurologically, the brain region that appears to be crucially involved in emotion-driven punishment is the insula. The insula is often activated when experiencing disgust, as when exposed to an obnoxious smell (e.g. Phillips et al., 1997). Being a key structure in the awareness of visceral, autonomic feedback stimuli, it could thereby serve as a "homeostatic alert" with enhanced activity indicating that some physical or emotional state is out of balance. The insula has previously been associated with aversive social stimuli and negative social interactions (e.g. Eisenberger, Lieberman, & Williams, 2003). Not surprisingly, an fMRI study of the ultimatum game (Sanfey et al., 2003) showed increased bilateral anterior insula activation in response to unfair offers. Furthermore, regions of the insula appeared to be sensitive to the actual degree of unfairness of the offers, so that insula activation was positively correlated with rejection rates. The positive relationship between activation of the anterior insula and rejection of unfair offers in an ultimatum game was later replicated by Tabibnia et al. (2008). Rilling et al. (2008) also specifically reported anterior insula activation in response to unreciprocated cooperation in a repeated prisoner's dilemma game.

If punishment via insula activation is an emotional response to restore socio-emotional functions, we can wonder if there are brain structures that are acting on the basis of economic rationality that suppress these emotions and resist punishment in order to avoid the costs associated with it. In their fMRI study with the ultimatum game, Tabibnia et al. (2008) reported that rejection was inversely correlated with right ventrolateral PFC (VLPFC) activation. This correlation was more pronounced for participants who accepted a higher proportion of offers they perceived as unfair. The VLPFC has previously been associated with down-regulation of activity in regions supporting negative affect (e.g. Eisenberger et al., 2003; Hariri et al., 2003) and has also been linked to rational decision-making in economic games where people are asked to maximize their expected outcomes (De Martino, Kumaran, Seymour, & Dolan, 2006).

Wanting to restore fairness triggers cognitive control and leads to punishment

Costly punishment can also be a deliberate choice meant to restore equality, in which case it is not an emotional response but an ethical decision surrounding the fairness norm. Restoring equality of pay-offs has been shown to be a valid motive for punishment (Price, Cosmides, & Tooby, 2002). Restoring the fairness norm, however, requires conscious suppression of the self-interest motive which dictates not paying for the cost to punish. The DLPFC provides the necessary self-control to suppress this selfish impulse and to punish despite its cost (Miller & Cohen, 2001). It is strongly connected to the dorsal striatum via the nigrostriatal dopamine pathway (Grahn et al., 2008), which suggests that deliberate punishment may go hand in hand with information update generated by the reward system.

A study by Knoch et al. (2006) investigated the hypothesis that the DLPFC controls the fundamental self-interest motive in favor of implementing culturally adaptive normative or moral

behavior. Their hypothesis was partially based on clinical reports of patients with DLPFC lesions showing aberrant normative behavior (Damasio, 2005) and experiments on impulse control (McClure et al., 2004) that illustrates the DLPFC's role in controlling selfish urges. Using transcranial magnetic stimulation Knoch et al. (2006) knocked out the DLPFC of participants prior to playing an ultimatum game. Compared to an untreated control group, the (right) DLPFC-deficient participants accepted more unfair offers, suggesting that the DLPFC drives rejection when there is a conflict between self-interest and fairness. Because the DLPFC is a driver of conscious deliberation, this assumes that these participants in fact valued fairness and wanted to restore it out of moral or ethical concern. When the DLPFC is knocked out, this deliberation can no longer occur. As a result, individuals may give in to the selfish impulse and accept any positive offer.

Contrasting the findings of Knoch et al. (2006), Sanfey et al. (2003) have previously reported that the intensity of (bilateral) DLPFC activation of the receivers in an ultimatum game was unrelated to their subsequent decision. In this study both the insula and DLPFC were activated following unfair offers, but the DLPFC activation was not related to the type of response (reject or accept) to an unfair offer. Only the heightened insula activation was associated with rejection, pointing to the role of emotions and downplaying the importance of the DLPFC with respect to punishment. We will return to this intriguing paradoxical role of the DLPFC (driving rejection in one study and being unrelated to rejection behavior in another study) when discussing avenues for future research.

The role of the DLPFC in altruistic punishment is further investigated by modifying the intentionality of the norm violator. Intentionality does matter in the ultimatum game, as responders tolerate inequality even less and reject more when proposers intentionally ignore an even-split alternative (Falk, Fehr, & Fischbacher, 2008). Consistent with this, Knoch et al. (2006)

reported that disruption of the DLPFC by transcranial magnetic stimulation had no behavioral effect when responders were told that offers had been computer-generated (and therefore were could not have been intentional). Blameworthiness of the transgressor also appeared to influence DLPFC activity during punishment in the study by Buckholtz et al. (2008). Participants who acted as "third party" observers of crimes were asked to determine appropriate punishments. The crimes varied in severity as well as assumed responsibility. As expected, the DLPFC was more activated for scenarios describing the perpetrators to be responsible for their crimes compared to scenarios were the perpetrator's responsibility was diminished.

Section summary

Similar to cooperative decision making, altruistic punishment in response to breaches of trust is the result of modulatory influences on the brain's reward system. A punishing decision occurs when the cognitive control system is evaluating the ethical concerns regarding reinstating justice, or when the brain's emotion regulation system becomes involved in restoring positive affect following betrayal.

The involvement of the cognitive control system in punishment suggests that people are able to consciously deliberate when to act prosocially, and that they understand when and how fairness can be restored through punishment. Given that punishment does not lead to immediate material benefits, it has stronger ties with social than with economic rationality. When it comes to punishment, both cognitive control and emotion regulation drive socially rational decisions that benefit emotional and social functioning in the group. They induce people who are naturally inclined to cooperate to also punish those that violate their trust or prosocial norms in general.

Individual differences

We have claimed so far that two different rationalities to cooperate – one economical and one social - do not have to be paradoxical if the brain is wired for both. However, these two rationalities do not have to be expressed equally in all individuals. In this section we propose that they correspond to individual differences in social preferences and that individuals differ in their patterns of neural activation. We suggest that self-regarding individuals who favor economically rational decisions will respond primarily to extrinsic incentives and recruit brain systems associated with cognitive control. Consequently, other-regarding individuals will favor socially rational decisions and will rely more heavily on the brain's social cognition system to process trust signals.

There is considerable behavioral heterogeneity in the extent to which people exhibit prosocial tendencies. People differ in how much they value cooperativeness, and this influences the strategies they will adopt in social dilemmas (e.g., Bogaert, Boone, & Declerck, 2008; Camerer & Fehr, 2006; Kurzban & Houser, 2005; Van Lange, 1999; 2000). Both agent-based simulations and laboratory experiments have shown that multiple strategies can coexist in a population at equilibrium. In a study by Kurzban and Houser (2005) 63 % of the players were classified as conditional cooperators, 13 % as altruists, and 20 % as free-riders. Similar proportions were reported by Fischbacher, Gachter, and Fehr (2001), namely 50 % conditional cooperators and one third of free-riders. Such individual variation in cooperative strategies appears to be a consistent individual trait, (Kurzban & Houser, 2001; 2005) with measurable heritability (Cesarini et al., 2009).

The willingness to cooperate in one-shot social dilemmas has further been related to several stable personality traits. For example, agreeable and extraverted people tend to cooperate

more (Kurzban & Houser, 2001; Lu & Argyle, 1991), while people scoring high on Machiavellianism tend to cooperate less (Wilson, Near, & Miller, 1998). Similarly, social psychologists have long paid attention to individual differences in so-called social value orientation (SVO), a stable trait that reflects how people evaluate outcomes for self and others (Messick & McClintock, 1968; Van Lange, 2000). People with a prosocial value orientation value fairness and tend to maximize joint outcomes. They are by default cooperative. People with a proself inclination are either self-maximizing or competitive and will only cooperate when it is in their self-interest to do so (Van Lange, 1999; 2000). Their cooperative behavior is purely strategic.

Considerable heterogeneity in punishing behavior has been reported as well (Fehr & Fischbacher, 2004), and this too appears to represent a stable disposition, as a substantial portion of the variance (> 40 %) has been attributed to additive genetic effects based on a twin study (Wallace, Cesarini, Lichtenstein, & Johannesson, 2007). It is not unlikely that punishing behavior should predominate among individuals with other-regarding preferences (Gintis et al., 2003). In line with this, cooperators in a social dilemma were more likely to punish intentional defectors compared to non-cooperators (Kiyonari, Declerck, Boone, & Pollet, 2011). Not surprisingly, the term "strong reciprocator," commonly used by economists, refers to those individuals that are inclined to both cooperate in anonymous, one-shot interactions, and punish those who do not. Strong reciprocators are even willing to pay the cost to punish a defector when they themselves were merely third party observers of an unfair interaction (e.g. Fehr & Fischbacher, 2004).

The findings that some individuals punish readily echoes the work of Frank (1988), who describes the anger following betrayal as a "commitment" device that makes the threat of punishment real. Anticipating that an angered person is "committed" to some type of revenge, a potential defector may think twice before defecting. If emotions are "commitment devices"

supporting prosocial behavior in social interactions (Frank, 1988), the committed cooperators (strong reciprocators) should use their emotions to communicate their intentions. That is, by involuntary expressing emotions (by blushing, perspiring, and facial expressivity) cooperators reveal their honest motivational intentions (to cooperate and to punish) which serve to attract potential interaction partners and deter defection. Because emotions are difficult to fake (e.g. the Duchenne smile), emotional expressivity could be a true indicator of trustworthiness (Boone and Buck, 2003). Although data is sparse, a recent experiment seems to confirm this. After being treated fairly and unfairly in an economic game, cooperators express emotions, regardless of valence, more frequently than do non-cooperators (Schug et al., 2010). Thus emotional expressivity and commitment (shown by punishing behavior) may be traits that are principally associated with other-regarding preferences and social rationality.

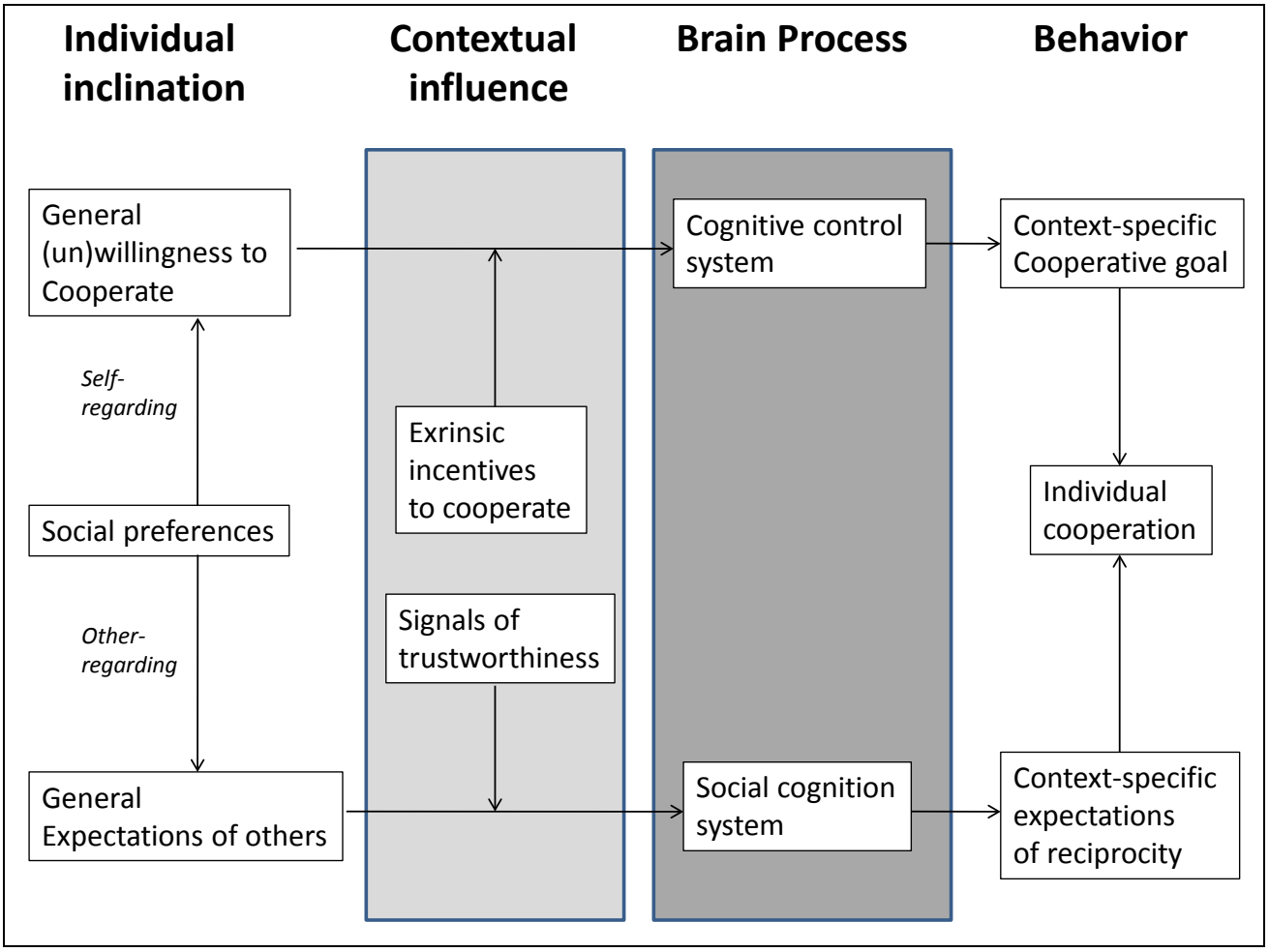
A generalized model for individual differences in prosocial behavior

Based on a comprehensive literature review Bogaert et al. (2008) proposed that cooperative decisions can be attained by two different psychological routes that correspond to individual differences in preferences: individuals with self-regarding preferences will be induced to cooperate when there are incentives that change their general unwillingness to cooperate to a context-specific cooperative goal. Extrinsic incentives that align self-interest with the larger collective transform their greed into cooperative motivation. On the other hand, individuals with other-regarding preferences will cooperate if there is sufficient trust to translate their generalized expectations of others to context-specific expectations of reciprocity. Based on the current review, we hereby add that brain networks involved in cognitive control are necessary to compute the context-specific benefits of cooperation and to realize goal alignment of individuals with self-

regarding preferences, and that social cognition is necessary to realize the context-specific expectations of reciprocity for individuals with other-regarding preferences. This framework for individual differences in cooperative behavior (adapted from Bogaert et al., 2008) is shown in Figure 3.

Figure 3

Theoretical framework accounting for individual differences in cooperative behavior (Adapted from Bogaert et al., 2008).



A recent large scale experiment (Boone, et al., 2010) corroborated the behavioral aspects of the model presented in Figure 3. The experiment compared cooperative decisions of prosocials versus proselfs and manipulated the presence of incentives and trust. Proselfs cooperated significantly more when cooperative incentives accentuated the win-win nature of the interaction while these same incentives had less impact on prosocials. Conversely, generalized trust as well as induced trust signals were found to facilitate cooperative behavior in prosocials, but they had no impact on proselfs.

There are many other examples showing that generalized or dispositional trust correlates with individual differences in cooperative behavior. Many personality types such as wariness of reciprocation, social anxiety, and borderline personality disorders, fail to exhibit normal trust levels and consistently cooperate less in social dilemmas (Cotterell, Eisenberger, & Speicher, 1992; Seres, Unoka, & Keri, 2009; Shore, Bommer, Rao, & Seo, 2009; Sripada et al., 2009). As a corollary in a field study, prosocially oriented employees perform better when they perceive a trustworthy manager (Grant & Summanth, 2009).

Experimental evidence further suggests that extrinsic incentives do not influence everybody alike. Stable individual differences in the behavioral tendency to strategically increase cooperative decisions when interactions become repeated, and decrease them in the last round, have been reported several times (Boone et al., 1999; Reuben & Suetens, 2009; Selten & Stoecker, 1986). Reputation gains as well appear to be an incentive to cooperate, primarily for self-regarding individuals. Simpson and Willer (2008) have shown that, when playing a prisoner's dilemma, individuals classified a priori as egoists respond strategically to reputation incentives, whereas those classified a priori as altruists are less affected. Furthermore, the egoists are aware of their strategic behavior, as they tend to discount other's prosocial behavior when it is

made public and hence lending itself to strategic self enhancement. Finally, the threat of punishment (a negative incentive) also appears to have a greater influence on the cooperative behavior of self-regarding individuals. Spitzer et al. (2007) report that especially individuals with a Machiavellian personality become more cooperative when there is a punishment threat.

The neural correlates of individual differences in prosocial behavior

If extrinsic incentives are particularly important to achieve cooperation of self-regarding individuals, and trust is more important to other-regarding individuals, we suggest that this should be noted by relative differences in the activation of cognitive control and social cognition when making cooperative decisions. As suggested by Figure 1, processing incentives recruits the cognitive capacities of the DLPFC, ACC, and LOPFC, while processing trust signals recruits the arMFC, STS, and amygdala. We next review the studies that have addressed individual differences in the neural correlates of social preferences to see if they reveal this hypothesized pattern in neural activation.

Direct evidence

A first and direct method to investigate the moderating influence of individual differences through fMRI is to divide the participants into a "cooperating" versus "non-cooperating" group based on the decisions they make during the experiment, and then compare the brain activation patterns between these groups. McCabe et al. (2001) compared participants that consistently cooperated with participants that also defected in a repeated trust game. The game was played against a human and a computer partner. For the consistent cooperators (whom we consider to have other-regarding preferences), the human partner elicited significantly more activation of the

arMFC. This difference was not found for the non-cooperating individuals. Apparently, social cognition matters less to those who do not value cooperativeness.

Similarly, Krueger et al. (2007) investigated individual differences in the neural correlates of trust. They compared brain activity of people using different strategies in a repeated trust game (Appendix 1). Defectors were those players who were not trustworthy and assumed a self-interested partner just as they were themselves, while the trustworthy players assumed their partner was also trustworthy and positively reciprocated trust. The latter group showed more activation in the paracingulate cortex (arMFC) and the septal area (an area that is associated with encoding good will which is needed to maintain a trusting relationship (Krueger et al., 2007)).

Just as trust is essential for other-regarding individuals, we propose that strategic thinking is typical for the self-regarding individuals. Bhatt and Camerer (2005) compared the brain patterns of participants who varied in so-called "strategic IQ." They derived this strategic aptitude measure from the higher-order beliefs people form during a repeated coordination game. Money could be earned in this game by matching one's choices to the choices of an anonymous partner who plays the game concurrently. Strategic IQ was computed by standardizing participants' expectations of their earnings based on their actual choices and the accuracy of their beliefs regarding the choices of their partners. Inferring the choices of others in order to compute one's own pay-off is compatible with a self-regarding inclination. Although other-regarding individuals also infer other's beliefs before making decisions in an interdependent game, they use this information in order to avoid betrayal. To them, the actual pay-off is less relevant, as long as it is fair (see also Stouten De Cremer, & Van Dijk, 2005). Hence we believe that the "strategic IQ" measure in Bhatt and Camerer's experiment indicates a self-regarding preference. The authors predicted that high strategic IQ would correlate with increased activity in brain regions involved in theory of mind (especially arMFC and posterior cingulated cortex), reflecting a

careful consideration of what other people do. However, consistent with our preposition that self-regarding individuals rely less on social cognition, no such link was found. Instead, regressing brain activation on strategic IQ revealed a positive correlation between strategic IQ and the caudate nucleus. This indicates that highly strategic people reinforce their own decisions when they can derive the strategies of others with a high degree of certainty and thus also obtain a greater certainty regarding the expected rewards. In addition, there was a negative correlation between strategic IQ and activation of the insula. This is compatible with the idea that insula activation associated with emotional responses is less prominent for self-regarding individuals.

A final indication that strategically processing incentives recruits brain regions involved in cognitive control comes from the study of Emonds et al. (2011a). In a blocked design fMRI experiment, they compared decision-making in a prisoner's dilemma and a coordination game (see Appendix 1). Only the latter contains extrinsic cooperative incentives. Then they contrasted brain activity of individuals who adjusted their behavior strategically between games (those who cooperated in the coordination game and defected in the prisoner's dilemma) with those who did not. The strategic players showed more activation in the DLPFC and the ACC, corroborating that strategic thinking requires cognitive control to adapt behavior according to incentives.

Indirect evidence

An alternative method to assess the modulating role of individual differences on cooperative behavior through fMRI is to compare brain activity of people who vary in stable personality traits that have been determined through questionnaires independent of their behavior during the experiment.

One example of a stable personality trait that is particularly relevant to social interaction is Machiavellianism. People that score high on Machiavellianism display a combination of

selfishness and opportunism (e.g., Wilson, Near, & Miller, 1998). Spitzer et al. (2007) report that people with high Machiavelli scores tend to adjust their behavior depending on the presence or absence of sanctions. In a regular dictator game (Appendix 1), they kept more money for themselves, but they became more generous when the game changed slightly to include a punishment threat. At the neural level, Machiavelli scores correlated positively with brain activations in two regions when the games with and without punishment were contrasted. First, higher activation in the LOPFC suggests that the adjustment behavior associated with Machiavellianism is driven by the evaluation of punishment threats (Kringelbach & Rolls, 2004). Second, correlations with insula activation indicate that the punishment threat leads to enhanced representation of emotional states for people with high Machiavelli scores. Hence punishment by other-regarding individuals (strong reciprocators) may truly serve as an emotional deterrent to self-regarding individuals.

Another personality trait for which the neural correlates of cooperative behavior have been investigated is Psychopathy. Psychopaths can neither be classified as self-regarding or other-regarding, because they are neither committed, nor strategic when it comes to cooperation. Their behavior tends to be erratic and often driven by vengeance rather than (economic or social) rationality. Consequently, they also show erratic patterns of brain activation when confronted with social dilemmas. Rilling et al. (2007) scanned normal participants with fMRI while they played a repeated prisoner's dilemma game with human confederates. Participants scoring high on Psychopathy (measured by self-report questionnaires) defected more in the prisoner's dilemma and were less likely to build a long term cooperative relationship with one partner. Therefore they also experienced more game outcomes in which their cooperative move was not reciprocated. Interestingly, in response to such trials, high-psychopaths showed less amygdala activation than low-psychopaths, which suggests that the former show weaker aversive

conditioning to unreciprocated cooperation. High-psychopaths further showed weaker activation of the VMPFC when cooperating. Hence they may lack the feelings of reward associated with cooperation that are typical of low-psychopaths. Finally, high-psychopaths showed a decrease in DLPFC activation, and in the rostral anterior cingulate cortex (arMFC). Thus, as Rilling et al. (2007) suggest, they may possibly experience less conflict when deciding to defect, and consequently need less cognitive control to act on this decision. In addition, they do not take other people into account, reducing the need for social cognition.

Borderline Personality Disorder has been studied under fMRI as well. These patients are typically self-centered and have little concern for others, and they lack the insula response of healthy matched controls when the reciprocation norm has been broken in a repeated trust game (King-Casas et al., 2008). Their insensitivity of the insula suggests that these borderline patients do not perceive low offers to be a violation of a social norm. Hence they fail to differentiate between generous and selfish partners and update expectations accordingly. Because of their typical negative expectations of social partners and their inability to coax generous behavior in general, cooperation in dyads containing a borderline patient usually breaks down quickly. They express significantly lower levels of self-reported trust relative to healthy controls (King-Casas et al., 2008).

Self- versus other regarding preferences are perhaps best captured by the personality trait Social Value Orientation, which distinguishes between prosocials and proselfs (Bogaert, et al., 2008; Van Lange et al., 2000). A comparison of brain activity of prosocials and proselfs while they solved social dilemmas under the fMRI scanner revealed that they indeed rely on different brain regions (Emonds et al., 2011b). Proselfs showed more activation in the DLPFC, precuneus, and the posterior STS, suggesting that their social strategies are more calculative and adapted to the situation at hand. Prosocials, on the other hand, showed more activation in the LOPFC and

anterior STS. In addition to its role in evaluating sanctions (as we described earlier), the LOPFC also reflects generalized normative behavior (perhaps as a result of wanting to avoid social disapproval). Because the experiment of Emonds et al. did not involve sanctions, we interpret this result to indicate that prosocials are more norm compliant. This same interpretation holds for the anterior STS, which in previous research has been described as "involved in previously resolved routine moral judgment that requires more semantically based representational knowledge." (Borg et al., 2006, p. 811).

The neural correlates of inequity aversion of prosocials versus proselfs was investigated by Haruno and Frith (2009). Correlating the absolute value of a reward difference between self and others with neural activity indicates that prosocials, but not proselfs, show increased activity in the dorsal amygdala when the absolute value of the reward difference is large. These authors suggest that the intuitive aversion for an inequitable division of resources of prosocials is characterized by neurologically grounded automatic and emotional processing.

Finally, the level of altruism also appears to show individual differences at the neural level. In an fMRI experiment described earlier, Moll et al. (2006) correlated the brain activity of costly charity donations with participants' self-reported engagement in real-life voluntary activity. Costly donations were associated with increased activation in several regions of the prefrontal cortex associated with reward, social cognition, and cognitive control. It appears that these regions are even more activated in people who volunteer much in real life. Altruistic decisions, however, can be made by both self- and other-regarding individuals, but the underlying motivation will depend on the presence of reputation concerns (Bereczkei, et al. 2007), hence we cannot clearly distinguish between the independent roles of cognitive control and social cognition in this case.

Section Summary

People vary in social preferences and hence in the extent to which they adhere to economically versus socially rational motives to cooperate. Such individual differences do moderate the neural correlates of cooperative decision making, often by enhancing activity in brain regions associated with processing social signals or extrinsic incentives. Because humans also differ in their degree to which their decisions in dilemma situations are influenced by emotion versus cognition (e.g. Gray, 2004; Hamann & Canli, 2004), it is perhaps not surprising that activity in regions as the insula, amygdala and DLPFC appeared to be easily affected by individual differences. At least the literature reviewed here suggests that self-regarding preferences are more associated with cognitive processing relying on increased activity in the ACC and DLPFC (Emonds et al., 2010a and 2010b), LOPFC (Spitzer et al., 2007), and less activity in the insula (Bhatt & Camerer, 2005; King-Casas et al., 2008). Other-regarding preferences appear to be more associated with automatic, emotional processing of the amygdala (Haruno and Frith, 2009), and social cognition recruiting the arMFC (Krueger et al., 2007; McCabe et al., 2001) and the anterior STS (Emonds et al., 2010b).

Concluding remarks and directions for future research

Prosociality includes a wide array of behavior, including mutual cooperation, pure altruism, and the costly act of punishing norm violators. Neurologically, these behaviors are all motivated by neural networks dedicated to reward, indicating that prosocial acts (such as cooperating in a social dilemma) are carried out because they were desired and feel good. However, the underlying reasons for the pleasant feelings associated with cooperative

behavior may differ. First, cooperation may be valued because of accruing benefits, making it economically rational. This route to cooperation is made possible through brain regions in the lateral frontal cortex that generate cognitive control and process the presence or absence of extrinsic cooperative incentives. Second, consistent with proponents of social rationality, cooperation can also occur when people expect to experience reward through a "warm glow of giving." Such intrinsically motivated cooperation yields collective benefits from which all group members may eventually benefit, but it can only be sustained when it exists in concert with a mechanism to detect and deter free-riding. Hence socially rational cooperation is facilitated by a neural network dedicated to social cognition that processes trust signals.

The combined influence of (independently operating) cognitive control and social cognition on motivated decision-making in the brain indicates that economic and social rationality are not contradictory but in fact exist next to each other. Such a double information-processing system may be reminiscent of Plato's metaphor describing the human mind as a charioteer driving a chariot pulled by two horses, a white horse of reason and a black horse of passion. The onset of neuroeconomic research has replaced the horses in the metaphor by an elephant, referring to the brain's reward system that steers motivated decisions to obtain immediate gratification, and a smart pony, referring to the advisory role of the cognitive control system (Camerer et al., 2005). This review suggests yet another revision of the metaphor: The elephant is in fact blind and without insight or foresight. It is guided along by two pony's with long-term vision: one is economically rational and calculates benefits based on extrinsic incentives (cognitive control), the other one is socially rational and assesses the likelihood of reciprocity versus betrayal by interpreting contextual trust signals. By computing iterated benefits of mutual cooperative acts (relying on the DLPFC, LOPFC, and ACC), by interpreting other people's intentions (relying on the amygdala, STS, and arMFC), and by

reinforcement learning based on outcomes of past interactions (evaluated by the caudatum in the reward system), this integrated neurological network underlying cooperation explains how human collaborations could have extended in evolution beyond simple exchanges and grow into fruitful long-term relationships that function on mutual trust.

An additional postulate that emerges from this review is that individual differences in cooperative motivation correspond to the relative extent to which brain systems related to cognitive control versus social cognition are recruited during decision-making. For self-regarding individuals, cooperation is rational when it is economically beneficial and immune to free-riders. If they calculate that cooperation will (safely) yield them money or status, or if they anticipate that non-cooperation will be sanctioned, they will act prosocially. Especially the brain's cognitive control system will be activated in this case. In contrast, other-regarding individuals who have internalized the cooperative norm need to rely more heavily on the brain's social cognition system in order to avoid betrayal by self-regarding others. Cooperation is rational for an other-regarding individual when the collective good is attained and free-riding is successfully avoided. When trust is abused, they are more likely than self-regarding individuals to pay the cost to punish the wrong-doer. They are the committed, strong reciprocators who are, on the one hand, emotionally outraged at non-cooperation by others and motivated to restore equality, and, on the other hand, they are hedonically pleased by mutual cooperation.

However, we stress that these two drivers of cooperation (one based on economic and the other on social rationality) are not mutually exclusive, and that individual differences only correspond to the degree to which brain regions involved in cognitive control and social cognition are activated. We do not imply that different types respond in an all-or-nothing fashion to incentives or trust. Other-regarding types may also adapt their choices to extrinsic incentives, even if that is not their primary motive for cooperation (Boone et al., 2010). And, while we

suggested that self-regarding types may be less inclined to rely on contextual social signals in their decision-making strategy (see also Boone et al., 2010), they may still appreciate trustworthiness and value mutual cooperation for its own sake (e.g., Rilling et al., 2010). The great extent to which people enjoy hedonic rewards from cooperation, and the general finding that rewards that are the result of human cooperation appear more valuable than mere material compensations of the same value (Elliott et al., 2006), substantiate that selfishness is not likely the prevalent norm in our species. This conclusion was also reached in the study by Tricomi et al. (2010) who investigated the neural correlates of inequality-averse social preferences. When overpaid individuals were scanned with fMRI, they showed increased activity in the reward system of the brain (VMPFC and ventral striatum) when they observed that underpaid partners receive a supplemental monetary transfer to reduce the inequality. This neural pattern remained consistent across individuals, and even for those who stated they would have preferred to have received the additional monetary transfer themselves. Thus, even some of the self-regarding types cannot deny that their nervous system is wired to enjoy fairness and, along with it, the social benefits of selfless acts.

A major contribution of this review is that the theoretical framework, summarized in Figures 1-3, opens up many avenues for future research. First and foremost, the neural mechanisms by which contextual information regarding incentives and trust modulate the reward system should be studied in more depth to further substantiate that the economic versus social drivers of cooperation are the result of different, independently operating neural networks. In this respect the role of neurotransmitters deserves more attention. Oxytocin (OT), for example, is released upon receiving extraneous signals of trust (Zak et al., 2004). We have already described how OT attenuates activity in the caudate nucleus, prompting people to take on a trusting attitude despite existing feedback that might indicate otherwise (Baumgartner, et al., 2008). OT may not

only enhance cooperative behavior by boosting trust, but also by making cooperation more appealing. There are already several reasons to believe that OT may facilitate mutual cooperation by enhancing the reward value of social encounters. In children diagnosed with autism spectrum, plasma OT administration tends to improve social information processing and increase the number of self-initiated social interactions (Hollander et al., 2007). After inhalation of OT, these children also respond in a more socially appropriate way to cooperative initiatives, are better able to differentiate between "good" and "bad" partners, and increasingly maintain eye contact (Andari et al., 2010). OT is highly connected to limbic areas, particularly the amygdala, fusiform gyrus (processing facial stimuli), and the ventral striatum (Petrovic, Kalisch, Singer, & Dolan, 2008). Based on animal studies, it appears that OT stimulates dopamine release in the nucleus accumbens shell and thereby modulates activity in the mesolimbic reward system (Depue & Morrone-Strupinsky, 2005). Therefore, some authors suggest that OT can positively enhance social approach by linking it to the capacity to experience reward from social interaction (Campbell, 2008; Guastella, Mitchell, & Mathews, 2008; Hammock & Young, 2006; Insel, 2003). Future research should aim to validate the modulatory influence of limbic oxytocin release in humans on activity in the reward system. By increasing the rewarding value of social interaction and at the same time decreasing social anxiety, oxytocin is bound to be a key neural building block of socially rational cooperation.

Serotonin is another neurotransmitter for which a clear link with cooperation has been established. In an experiment where participants were deprived of L-tryptophan (a dietary substance that is necessary for serotonin synthesis) cooperative decisions in a repeated prisoner's dilemma game were reduced, even following a mutually cooperative outcome (Wood et al., 2006). This suggests that tryptophan depletion interferes with reward reinforcement and is consistent with serotonin's known modulatory role in reward processing, including the circuitry

encompassing the orbitofrontal cortex and ventral striatum (Robbins & Everitt, 1996; Sasaki-Adams & Kelley, 2001). Other data have suggested that L-tryptophan depletion (interrupting serotonin synthesis) impairs the capacity to delay gratification (Denk et al., 2005). This implies that serotonin supports economic rationality by helping people to control the impulse to prefer short-term rewards over the long-term benefits from mutual cooperation. To complement this preliminary evidence, systematically studying the effects of different neurotransmitters on cooperation while manipulating contextual factors could be useful in unraveling the multiple ways by which reciprocally beneficial relationships become established and maintained. We hope the models in this review serve as a template for such endeavors.

A second research direction that is specified by the model presented in Figure 3 is to further distinguish the neurological differences between self- and other-regarding preferences. Therefore, future fMRI or imaging studies should increasingly take into account individual differences when studying the neural correlates of prosocial behavior. This approach might help to resolve inconsistencies in the literature. An eye-catching example relates to the role of the DLPFC in punishing behavior reported by both Sanfey et al. (2003) and Knoch et al. (2006). Their findings indicate, on the one hand, that DLPFC activity is not related to rejection when a person (under the scanner) is confronted with an unfair offer in an ultimatum game (Sanfey et al., 2003), and, on the other hand, that it is essential in order to reject an unfair offer (Knoch et al., 2006). These conflicting results can be reconciliated by invoking individual differences in self- versus other-regarding preferences. Rejecting an unfair offer is the prosocial response in the ultimating game, while accepting is considered selfish. Unfair offers are more difficult to respond to compared to fair offers, and may therefore place higher cognitive demands on everyone (Sanfey et al., 2003). However, rationality for the self-regarding types requires the function of the DLPFC to suppress anger and accept an unfair offer, thereby fulfilling their selfish needs. In

contrast, to be rational and internally consistent, other-regarding individuals may need the same inhibitory control function of the DLPFC to suppress the selfishness urge in order to reject an unfair offer and implement fairness. This implies or suggests that the findings of Sanfey et al. (2003) and Knoch et al. (2006) may be sample-specific. Future research could unravel if DLPFC activation during rejection depends on the personality and/or social preferences of the selected participants.

A more general approach to the neuroeconomic study of individual differences is to investigate how they arise in ontogeny and evolution. On a developmental level, intrinsically motivated (committed) cooperation may only become possible when people have had sufficient positive experiences with prosocial acts, either directly, through social interaction, or vicariously, through social learning. The growing body of knowledge reviewed in this paper indicates that the subjective experience of reward signals in the striatum and VMPFC can be modulated by a variety of factors, ranging from economic benefits to status and even fairness considerations. How exactly these factors interact with pre-existing tendencies (experience-based or genotype-driven) to diminish or enhance individual cooperative tendencies is a promising research domain. Already a specific polymorphism in the oxytocin receptor gene (rs53576) appears to correlate significantly with prosocial temperament, empathy, and social cognition (Rodrigues et al., 2009; Park et al., 2010, Tost et al., 2010). Individuals who are homozygous for one of the variants of the rs53576 polymorphism are reported to be at risk for social dysfunctions and showed decreased amygdala activation during an emotion recognition task under fMRI (Tost et al., 2010). The genes that gave rise to neural mechanisms by which people acquire the ability to learn the rewarding value of cooperation and thereby become sensitized towards cooperative behavior have likely been selected over evolutionary time, and may lie at the root of much of the social rationality we have discussed. Substantiating this are the recent findings from developmental

psychology that even infants have representations of prosociality (Carey, 2009). Understanding how these innate representations are combined with the emergent ability to mind read (or lack there-of), with the control functions of the DLPFC, and with life-long experiences may yield insights into how adult differences in social preferences arise.

Finally, future research should also pay attention to the limits and potentials of neuroimaging studies. Since its introduction in the early 1990's, fMRI has generated an enormous amount of imaging data, but according to some authors (see Miller, 2008), it may now be experiencing growing pains. The most common problem with interpreting fMRI data seems to be reverse inferencing, or assuming a very specific mental state from activation of a particular brain region. The misconception here is that there is not necessarily a one-on-one relation between brain and behavior. We believe this review has further put the one-on-one relationship between brain and behavior into question, as the current literature shows that cooperative motivation may be modulated by different brain systems that are differentially sensitive to context-specific incentives or trust. As a result, activation of specific brain regions might be related to either cooperative or selfish behavior, depending on the context or the person's preferences. When different motives for the same behavior co-exist, neural reactivity can still be recorded even in the absence of clear behavioral differences. For example, while selfishness and fear of betrayal may both account for non-cooperation in a social dilemma, they are likely associated with very distinct brain regions.

The absence of a one-to one relation between brain and behavior makes fMRI not very suitable for exploratory research. Instead it is better suited as a tool to test hypotheses when there are converging pieces of evidence and different techniques available (Miller, 2008). Claiming that a mental state follows from activation of a particular brain region hinges on previous research that has shown a link between this region of interest and other experimental conditions

that have elicited the same mental state. Having established a clear association between brain and behavior, other techniques can be used in follow up studies. Diffusion tensor imaging (DTI) can further determine whether areas that fire together are physically wired together to establish anatomical connectivity. Transcranial magnetic stimulation, a technique which temporarily blocks a brain function, can assess causal links between brain activity and subsequent behavior.

Given these potentials of the neuro-economic research tools, some scholars have begun to look for practical applications of neuroeconomics and attempt to reveal truth-telling and the potential for free-riding at an individual level (Krajbich, Camerer, Ledyard, and Rangel, 2009). An individual's pattern of brain activation is used to predict the underlying motivation and the subsequent likelihood of certain behaviors. These authors believe that combining such results with carefully designed decision-making schemes in real life could help create better collaborative institutions and a more optimal distribution of public goods. Although such practical paradigms are still in a very introductory phase and prone to being a two-edged sword, researchers as well as policy-makers should pay attention to the progress made in such endeavors in order to fully realize the positive potentials of neuroeconomics and promptly thwart erroneous conclusions or damaging consequences.

Acknowledgments

This work was supported by grants NOI 1044 and ID-BOF 1931 from the University of Antwerp.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., & Spitzer, M. (2006). Prediction error as a linear function of reward probability is coded in human nucleus accumbens. *Neuroimage, 31,* 790-795. DOI: 10.1016/j.neuroimage.2006.01.001.

Acevedo, M., and J. I. Krueger. (2005). Evidential reasoning in the Prisoner's Dilemma. *American Journal of Psychology 118,* 431-457.

Anderson, A. K., & Phelps, E. A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature, 411,* 305-309. DOI:10.1038/35077083

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

Andari E., Duhamel, J.R., Zalla, T., Herbrecht, E., Leboyer, M., & Sirigu, A. (2010). Promoting social behavior with oxytocin in high functioning autism spectrum disorders. *Proceedings of the National Academy of Sciences of the U.S.A., 107,* 4389-4394. DOI: 10. 1073/pnas.0910249107. DOI: 10.1073/pnas.0910249107

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science, 211,* 1390-1396.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution, 54,* 39-57. DOI: 10.1177/0022002709352443

Baron-Cohen, S., Ring, H. A., Bullmore, E. T., Wheelwright, S., Ashwin, C., & Williams, S. C. R. (2000). The amygdala theory of autism. *Neuroscience and Biobehavioral Reviews, 24,* 355-364.

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

Beauchamp, M. H., & Anderson, V. (2010). SOCIAL: An integrative framework for the development of social skills. Psychological Bulleting, 136, 39-64. DOI: 10.1037/a0017768

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52,* 336-372. DOI: 10.1016/j.geb.2004.06.010

Bereczkei, T., Birkas, B., and Kerekes, Z. (2007). Public charity offer as a proximate factor of evolved reputation-building strategy: an experimental analysis of a real-life situation. *Evolution and Human Behavior,28,* 277-284. DOI: 10.1016/j.evolhumbehav.2007.04.002

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10,* 122-142.

Bhatt, M., & Camerer, C. F. (2005). Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior, 52,* 424-459. DOI: 10.1016/j.geb.2005.03.007

Bogaert, S., Boone, C., & Declerck, C. H. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology, 47,* 453-480. DOI: 10.1348/014466607X244970

Boone , R. T., & Buck, R. (2003). Emotional expressivity and trustworthiness: the role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior,* 27, 163-182. DOI: 10.1023/A:1025341931128

Boone C., De Brabander B., Van Witteloostuijn A. (1999). Locus of control and strategic behaviour in a prisoner's dilemma game. *Personality and Individual Differences, 20,* 695-706. DOI: 10.1016/S0191-8869(98)00269-4

Boone, C., Declerck, C. H., & Kiyonari, T. (2010). Inducing cooperative behavior among proselfs versus prosocials: the moderating role of incentives and trust. *Journal of Conflict Resolution,* In press. DOI: 10.1177/0022002710372329

Boone, C., Declerck, C. H., & Suetens, S. (2008). Subtle social cues, explicit incentives and cooperation in social dilemmas. E*volution and Human Behavior, 29,* 179-188. DOI: 10.1016/j.evolhumbehav.2007.12.005.

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience, 18,* 803-817. DOI: 10.1162/jocn.2006.18.5.803

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review, 7,* 129-145. DOI: 10.1207/S15327957PSPR0702_129-145

Boyd, R., Gintis, H., Bowles, S. & Richerson, P. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Science of the USA, 100,* 3531-3535. DOI: 10.1073/pnas.0630443100

Brothers, L. (1990). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neurocience,1,* 27-51.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron, 60,* 930-940. DOI: 10.1016/j.neuron.2008.10.016.

Burnham, T., McCabe, K., & Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization, 43,* 57-73.

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

Camerer, C. F. (2008). Neuroeconomics: Opening the gray box. *Neuron, 60,* 416-419. DOI: 10.1016/j.neuron.2008.10.027

Camerer, C. F., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science, 311,* 47-52. DOI: 10.1126/science.1110600

Camerer, C. F., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 43,* 9-64.

Campbell, A. (2008). Attachment, aggression and affiliation: The role of oxytocin in female social behavior. *Biological Psychology, 7*7, 1-10. DOI: 10.1016/j.biopsycho.2007.09.001

Carey, S. (2009). On the Origin of Concepts. Oxford Series in Cognitive Development. New York: Oxford University Press.

Caporael, L. R., Dawes, R. M., Orbell, J. M., & Vandekragt, A. J. C. (1989). Selfishness examined - cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences, 12,* 683-699. DOI: 10.1017/S0140525X00025292

Carter, C. S., Botvinick, M. M., & Cohen, J. D. (1999). The contribution of the anterior cingulate cortex to executive processes in cognition. *Reviews in the Neurosciences, 10,* 49-57.

Carter, C. S., & van Veen, V. (2007). Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive Affective & Behavioral Neuroscience, 7,* 367-379. DOI: 10.3758/CABN.7.4.367

Cesarini, D., Dawes, C.T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Experimental Game Theory and Behavior Genetics. Values, Empathy, and Fairness across Social Barriers. *Annals of the New York Academy of Science, 1167,* 66-75. DOI: 10.1111/j.1749-6632.2009.04505.x

Cole, T., & Teboul, J. C. B. (2004). Non-zero-sum collaboration, reciprocity, and the preference for similarity: Developing an adaptive model of close relational functioning. P*ersonal Relationships, 11,* 135-160. DOI: 10.1111/j.1475-6811.2004.00075.x

Cotterell, N., Eisenberger, R. & Speicher, H. (1992). Inhibiting effects of reciprocation on interpersonal relationships. *Journal of Personality and Social Psychology*, 62, 658-668.

Damasio, A. R. (2005). Descartes' Error (10th Anniversary ed.). New York, New York: Penguin Books.

Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology, 25*, 111-116.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science, 313*, 684-687. DOI: 10.1126/science.1128356

de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schelthammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254-1258. DOI: 10.1126/science.1100735

DeBruine, L. M. (2002). Facial resemblance enhances trust. Proceedings of the Royal Society of London Series B-Biological *Sciences, 269*, 1307-1312. DOI: 10.1098/rspb.2002.2034

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & 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. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences, 1104*, 70-88. DOI: 10.1196/annals.1390.002

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

Denk, F., Walton, M. E., Jennings, K. A., Sharp, T., Rushworth, M. F. S., & Bannerman, D. M. (2005). Differential involvement of serotonin and dopamine systems in cost-benefit decisions about delay or effort. *Psychopharmacology, 179*, 587-596. DOI: **1**0.1007/s00213-004-2059-4

Depue, R. A., & Morrone-Strupinsky, J. V. (2005). A neurobehavioral model of affiliative bonding: Implications for conceptualizing a human trait of affiliation. *Behavioral and Brain Sciences, 28*, 313-395. DOI: 10.1017/S0140525X05000063

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6, 178-190. DOI: 10.1002/(SICI)1520-6505(1998)

Eickhoff SB, Laird AR, Grefkes C, Wang LE, Zilles K, Fox PT (2009) Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping 30*, 2907-2926. DOI: 10.1002/hbm.20718

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*, 290-292. DOI: 10.1126/science.1089134

Elliott, R., Voellm, B., Drury, A., McKie, S., Richardson, P., & Deakin, J. F. W. (2006). Co-operation with another player in a financially rewarded guessing game activates regions implicated in theory of mind. *Social Neuroscience, 1*, 385-395. DOI:

10.1080/17470910601041358

Emonds, G., Declerck, C.H., Boone, C., Vandervliet, E., & Parizel, P. (2011a). The cognitive demands on cooperation in social dilemmas. An fMRI study. Manuscript submitted.

Emonds, G., Declerck, C.H., Boone, C., Vandervliet, E., & Parizel, P. (2011b). Comparing the neural basis of strategic decision-making in people with different social preferences, a fMRI study. *Journal of Neuroscience, Psychology, and Economics, 4*(1), 11-24. DOI: 10.1037/a0020151

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19,* 1508-1519.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness - intentions matter. *Games and Economic Behavior, 62,* 287-303. DOI: 10.1016/j.geb.2007.06.001

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences, 11,* 419-427. DOI: 10.1016/j.tics.2007.09.002

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25,* 63-87. DOI:10.1016/S1090-5138(04)00005-4

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature, 415,* 137-140. DOI: 10.1038/nature03257

Fehr, E., & Rockenbach, B. (2004). Human altruism: Economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology, 14,* 784-790. DOI: 10.1016/j.conb.2004.10.007

Fischbacher, U., Gachter, S., & Fehr, E. (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters, 71,* 397-404. DOI: 10.1016/S0165-1765(01)00394-9

Frank, R. H. (1988). Passions within reason: The strategic role of the emotions. New York: Norton.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research,* 1079, 36-46. DOI: 10.1016/j.brainres.2005.12.126

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B-Biological Sciences, 358,* 459-473. DOI: 10.1098/rstb.2002.1218

Fukui, H., Murai, T. Shinozaki, J., Aso, T., Fukuyama, H., Hayashi, T., Hanakawa, T. (2006). The neural basis of social tactics: An fMRI study. *NeuroImage, 32,* 913-920. DOI: 10.1016/j.neuroimage.2006.03.039

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7,* 77-83. DOI: 10.1016/S1364-6613(02)00025-6

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

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior, 24,* 153-172. DOI: 10.1016/S1090-5138(02)00157-5

Grahn, J. A., Parkinson, J. A., & Rangel, A. (2008). The cognitive functions of the caudate nucleus. *Progress in Neurobiology, 86,* 141–155. DOI:10.1016/j.pneurobio.2008.09.004

Grant, A. M., & Sumanth, J. J. (2009). Mission Possible? The performance of prosocially motivated employees depend on manager trustworthiness. *Journal of Applied Psychology, 94,* 927-944. DOI: 10.1037/a0014391

Gray, J. R. (2004). Integration of emotion and cognitive control. *Current Directions in Psychological Scienc*e,13, 46-48. DOI: 10.1111/j.0963-7214.2004.00272.x

Guastella, A. J., Mitchell, P. B., & Mathews, F. (2008). Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry, 64,* 256-258. DOI: 10.1016/j.biopsych.2008.02.008

Gurerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science, 312,* 108-111. DOI: 10.1126/science.1123633

Hagen, E. H., & Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretation os experimental games. Theoretical Population Biology, 69, 339-348. DOI: 10.1016/j.tpb.2005.09.005

Hamann, S., & Canli, T. (2004). Individual differences in emotion processing. *Current Opinion in Neurobiology, 14,* 233-238. DOI: 10.1016/j.conb.2004.03.010

Hammock, E. A. D., & Young, L. J. (2006). Oxytocin, vasopressin and pair bonding: Implications for autism. *Philosophical Transactions of the Royal Society B-Biological Sciences, 361,* 2187-2198. DOI: 10.1098/rstb.2006.1939

Harbaugh, W. T., Mayr, U., & 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

Hardy, C. L. & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32,* 1402-1413. DOI: 10.1177/0146167206291006

Hariri, A. R., Mattay, V. S., Tessitore, A., Fera, F., & Weinberger, D. R. (2003). Neocortical modulation of the amygdala response to fearful stimuli. *Biological Psychiatry, 53,* 494-501. DOI: 10.1016/S0002-3223(03)01786-9

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience, 30,* 583-590. DOI: 10.1523/jneurosci.4089-09.2010

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

Haselhuhn, M. P., & Mellers, B. A. (2005). Emotions and cooperation in economic games. *Cognitive Brain Research, 23,* 24-33. DOI: 10.1016/j.cogbrainres.2005.01.005

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78,* 81-91. DOI: 10.1017/S0140525X09991440

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28,* 795-855. DOI: 10.1017/S0140525X05000142

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science, 312,* 1767-1770. DOI: 10.1126/science.1127333

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science, 319,* 1362-1367. DOI: 1°.1126/science. 1153808

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review, 86,* 653-660.

Hollander, E., Bartz, J., Chaplin, W., Phillips, A., Sumner, J., Soorya, L., et al. (2007). Oxytocin increases retention of social cognition in autism. *Biological Psychiatry, 61,* 498-503. DOI: 10.1016/j.bipsych.2006.05.030

Insel, T. R. (2003). Is social attachment an addictive disorder? *Physiology & Behavior, 79,* 351-357. DOI: 10.1016/S0031-9384(03)00148-3

Kahneman, D. (2003) Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93,* 1449-1475. DOI: 10.1257/000282803322655392

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology, 70,* 47-65.

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

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

Kiyonari, T., Declerck, C.H., Boone, C., & Pollet, T (2011) Costly punishment and group cooperation: the role of behavioral heterogeneity and apparent free-riding intentions. Manuscript submitted.

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & 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., & Fehr, E. (2009). Disrupting the prefrontal cortex diminishes the human ability to build a good reputation. *Proceedings of the National Academy of Sciences, 106,* 20895-20899. DOI: 10.1073/pnas.0911619106

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

Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *The Journal of Neuroscience, 27,* 951-956. DOI: 10.1523/jneurosci.4606-06.2007

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology, 24,* 183-214. DOI: 10.1146/annurev.soc.24.1.183

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & 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., Camerer, C. F. (2009). Economic Games quantify diminished sense of guilt in patients with damage to the prefrontal cortex. *Journal of Neuroscience, 29,* 2188-2192. DOI: 10.1523/JNEUROSCI.5086-08.2009

Krajbich, I., Camerer, C. F., Ledyard, J., & Rangel, A. (2009). Using neural measures of economic values to solve the public goods free-rider problem. *Science, 326,* 596-599. DOI: 10.1126:science.1177302

Kringelbach, M. L., & Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology, 72,* 341-372. DOI: 10.1016/j.pneurobio.2004.03.006

Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., et al. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 20084-20089. DOI: 10.1073/pnas.0710103104

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior, 28,* 75-84. DOI: 10.1016/j.evolhumbehav.2006.06.001

Kurzban, R., & Houser, D. (2001). Individual differences in cooperation in a circular public goods game. *European Journal of Personality, 15,* S37-S52. DOI: 10.1002/per.420

Kurzban, R. & Houser, D. (2005). Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 1803-1807. DOI: 10.1073/pnas.0408759102

Lu, L., & Argyle, M. (1991). Happiness and cooperation. *Personality and Individual Differences, 12,* 1019-1030. DOI: 10.1016/0191-8869(91)90032-7

Markoczy, L. (2004). Multiple motives behind single acts of cooperation. *International Journal of Human Resource Management, 15,* 1018-1039. DOI: 10.1080/09585190410001677296

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences of the United States of America, 98,* 11832-11835. DOI: 10.1073/pnas.211415698

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306,* 503-507. DOI: 10.1126/science.1100907

Messick, D. M., & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology, 4,* 1-25. DOI: 10.1016/0022-1031(68)90046-2

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

Miller, G. (2008). Growing pains for fMRI. *Science,* 320, 1412-1414. DOI: 10.1126/science.320.5882.1412

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souzat, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 15623-15628. DOI: 10.1073/pnas.0604475103

Morhenn, V. B., Park, J. W., Piper, E., & Zak, P. J. (2008). Monetary sacrifice among strangers is mediated by endogenous oxytocin release after physical contact. *Evolution and Human Behavior, 29,* 375-383. DOI: 10.1016/j.evolhumbehav.2008.04.004

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437,* 1291-1298. DOI: 10.1038/nature04131

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14,* 769-776. DOI: 10.1016/j.conb.2004.10.016

Park, J., Willmott, M., Vetuz, G., Troye, C., Kirley, A., Hawi, Z., Brookes, K. J., Gill, M., Kent, L. (2010). Evidence that genetic variation in the oxytocin receptor (OXTR) gene influences social cognition in ADHD. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 34, 697-702. DOI: 10.1016/j.pnpbp.2010.03.029

Peterson, R. L. (2005). The neuroscience of investing: fMRI of the reward system. *Brain Research Bulletin, 67,* 391-397. DOI: 10.1016/j.brainresbull.2005.06.015

Petrovic, P., Kalisch, R., Singer, T., & Dolan, R. J. (2008). Oxytocin attenuates affective evaluations of conditioned faces and amygdala activity. *Journal of Neuroscience, 28,* 6607-6615. DOI: 10.1523/JNEUROSCI.4572-07.2008

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature, 389,* 495-498. DOI: 10.1038/39051

Piazza, J., and Bering, J. M. (2008). Concerns about reputation via gossip promote generous allocations in an economic game. *Evolution and Human Behavior, 29,* 172-178. DOI: 10.1016/j.evolhumbehav.2007.12.002

Preuschoff, P. K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron, 51,* 381-390. DOI: 10.1016/j.neuron.2006.06.024

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience, 28,* 2745-2752. DOI: 10.1523/jneurosci.4286-07.2008

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior, 23,* 203-231. DOI: 10.1016/S1090-5138(01)00093-9

Previc, F. H. (1999). Dopamine and the origins of human intelligence. *Brain and Cognition, 41,* 299-350. DOI: 10.1006/brcg.1999.1129

Reuben, E., & Suetens, S. (2009). Revisiting Strategic versus non-strategic cooperation. CentER working paper, No. 2009-22, Tilburg University.

Rilling, J. K., Glenn, A. L., Jairam, M. R., Pagnoni, G., Goldsmith, D. R., Elfenbein, H. A., et al. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. *Biological Psychiatry, 61*, 1260-1271. DOI: 10.1016/j.biopsych.2006.07.021

Rilling, J. K., Goldsmith, D. R., Glenn, A., Jairam, M. R., Elfenbein, H. A., Dagenais, J. E., Murdock, C. D., & Pagnoni, G. (2008). The neural correlates of affective response to unreciprocated cooperation. *Neuropsychologia,* 46, 1256-1266. DOI: 10.1016/j.neuropsyhologia.2007.11.033

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & 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., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004a). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *Neuroreport, 15,* 2539-2543. DOI: 10.1097/00001756-200411150-00022

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

Robbins, T. W., & Everitt, B. J. (1996). Neurobehavioural mechanisms of reward and motivation. *Current Opinion in Neurobiology, 6*, 228-236. DOI: 10.1016/S0959-4388(96)80077-8

Rodrigues, S. M., Saslow, L. R., Garcia, N., John, O., & Keltner, D. (2009). Oxytocin receptor genetic variation relates to empathy and stress reactivity in humans. *Proceedings of the National Academy of Sciences,* 106, 21437-21441. DOI: 10.1073/pnas.0909579106

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., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*, 1755-1758. DOI: 10.1126/science.1082976

Sasaki-Adams, D. M., & Kelley, A. E. (2001). Serotonin-dopamine interactions in the control of conditioned reinforcement and motor behavior. *Neuropsychopharmacology, 25*, 440-452. DOI: 10.1038/S0893-133X(01)00240-8

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior, 31*, 87-94. DOI: 10.1016/j.evolhumbehav.2009.09.06

Schultz, W. (1997). Dopamine neurons and their role in reward mechanisms. *Current Opinion in Neurobiology, 7,* 191-197. DOI: 10.1016/S0959-4388(97)80007-4

Seip, E. C., van Dijk, W.W., & Rotteveel, M. (2009). On hotheads and dirty harries. The primacy of anger in altruistic punishment. Values, Empathy and Fairness Across Social Barriers. *Annals of the New York Academy of Sciences, 11670,* 190-196. DOI: 10.1111/j. 1249-6632.2009.04503

Selten, R., & Stoecker, R. (1986). End behavior in sequence of finite prisoner's dilemma supergames. *Journal of Economic Behavior & Organisation 3,* 47-70. DOI: 10.1016/0167-2681(86)90021-1

Seres, I., Unoka, Z., & Keri, S. (2009). The broken trust and cooperation in borderline personality disorder. *Neuroreport, 20,* 388. 392. DOI 10.1097/WNR.0B013e328324eb4d

Shore, L. M., Bommer, W. H., Rao, A. N., & Seo, J. (2009). Social and economic exchange in the employee-organization relationship: the moderating role of reciprocation wariness. *Journal of Managerial Psychology, 24*, 701-721. DOI: 10.1108/02683940910996752

Simpson, B. (2004) Social values, subjective transformations, and cooperation in social dilemmas. *Social Psychology Quarterly 67*, 385-395.

Simpson, B., & Willer, R. (2008). Altruism and indirect reciprocity: The interaction of person and situation in prosocial behavior. *Social Psychology Quarterly, 71*, 37-52. DOI: 10.1177/019027250807100106

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41,* 653-662. DOI: 10.1016/S0896-6273(04)00014-5

Singer, T., Seymour, B., O'Doherty, J., Stephan, K. E., Dolan, R. J., & Froth, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature, 439,* 466-469. DOI: 10.1038/nature04271

Sober, E., & Wilson, D. S. (1998). Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press.

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

Sripada, C. S., Angstadt, M., Banks, S., Nathan, P. J., Liberzon, I., Phan, K. L. (2009). Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. *Neuroreport, 20,* 984-989. DOI: 10.1097/WNR;0b013e32832d0a67

Staudinger, M. R., Erk, S., Abler, B., & Walter, H. (2009). Cognitive reappraisal modulates expected value and prediction error encoding in the ventral striatum. *Neuroimage, 45,* 713-721. DOI: 10.1016/j.neuroimage.2009.04.095

Stevens, J. R., & Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences, 8,* 60-65. DOI: 10.1016/j.tics.2003.12.003

Stone, V. E., Baron-Cohen, S., Calder, A. J., Keane, J., & Young, A. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia, 41,* 209-220. DOI: 10.1016/S0028-3932(02)00151-3

Stouten, J., De Cremer, D., & Van Dijk, E. (2005). All is well that ends well, at least for proselfs: Emotional reactions to equality violation as a function of social value orientation. *European Journal of Social Psychology, 35*(6), 767-783. DOI: 10.1002/ejsp.276

Tabibnia, G., & Lieberman, M. D. (2007). Fairness and cooperation are rewarding - evidence from social cognitive neuroscience. *Annals of the New York Academy of Sciences, 1118,* 90-101. DOI: 10.1196/annals.1412.001

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness - preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science, 19,* 339-347. DOI: 10.1111/j.1467-9280.2008.02091.x

Thaler, R. H. (1988). Anomalies - the Ultimatum Game. Journal of Economic Perspectives, 2(4), 195-206.

Thompson, L., Kray, L. J., & Lind, E. A. (1998). Cohesion and respect: An examination of group decision making in social and escalation dilemmas. *Journal of Experimental Social Psychology, 34,* 289-311. DOI: 10.1006/jesp.1998.1351

Tice, D., Baumeister, R. F., & Zang, L. (2004). The role of emotion in self regulation: Differing role of positive and negative emotions. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tost, H., Kolachana, B., Hakimi, S., Lemaitre, H., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., Myer-Lindenberg, A. (2010). A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function. *Proceedings of the National Academy of Sciences,*107, 13936-13941. DOI: 10.1073/pnas.1003296107

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron, 41,* 281-292. DOI:10.1016/S0896-6273(03)00848-1

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.

Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research, 169,* 564-568. DOI: 10.1007/s00221-006-0346-5

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77,* 337-349. DOI: 10.1037/0022-3514.77.2.337

Van Lange , P. A. M. (2000). Beyond self-interest: A set of propositions relevant to interpersonal orientations. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (Vol. 11, pp. 297-331). New York: Wiley. DOI: 10.1080/14792772043000068

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences of the United States of America,* 104, 15631-15634. DOI: 10.1073/pnas.076642104

Walter, H., Abler, B., Ciaramidaro, A., & Erk, S. (2005). Motivating forces of human actions - neuroimaging reward and social interaction. *Brain Research Bulletin, 67*, 368-381. DOI: 10.1016/j.brainresbull.2005.06.016

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology, 58*, 425 - 452. DOI: 10.1146/annurev.psych.58.110405.085641

Wilson, D. S., Near, D. C., & Miller, R. R. (1998). Individual differences in Machiavellianism as a mix of cooperative and exploitative strategies. *Evolution and Human Behavior, 19*, 203-212. DOI: 10.1016/S1090-5138(98)00011-7

Wilson, D. S., & Sober, E. (1994). Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences, 17,* 585-608. DOI: 10.1017/S0140525X00036104

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience, 5*, 277-283. DOI: 10.1038/nn816

Wood, R. M., Rilling, J. K., Sanfey, A. G., Bhagwagar, Z., & Rogers, R. D. (2006). Effects of tryptophan depletion on the performance of an iterated Prisoner's Dilemma game in healthy adults. *Neuropsychopharmacology, 31*, 1075-1084. DOI: 10.1038/sj.npp.1300932

Xiao, E., & Houser, D. (2005). Emotion, expression in human punishment behavior. *Proceedings of the National Academy of Sciences, 102,* 7389-7401. DOI: 10.1073/pnas.0502399102

Yamagishi, T. (1998). The structure of trust. An evolutionary game of mind and society. Tokyo: Tokyo University Press.

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity: Ingroup favoritism and ingroup boasting. *Advances in Group Processes, 16*, 161-197.

Yamagishi, T., & Sato, K. (1986). Motivational bases of the public goods problem. *Journal of Personality and Social Psychology, 50*, 67-73. DOI: 10.1037/0022-3514.50.1.67

Zak, P. J., Kurzban, R., & Matzner, W. T. (2004). The neurobiology of trust. *Annual Reviews of the New York Academy of Sciences, 1032,* 224-227. DOI: 10.1196/annals.1314.025

Appendix 1.

Description of economic games.

Prisoner's dilemma

In the classic prisoner's dilemma, two players have to independently choose to cooperate (C) or defect (D) and will be awarded a sum of money in function of the choices they both make. The players are usually given a pay-off matrix from which they can compute their possible gains. There are four possible outcomes: both players cooperate (CC), player 1 cooperates and player 2 defects (CD), player 1 defects and player 2 cooperates (DC), or both players defect (DD). The pay-offs for the outcomes are such that DC > CC > DD > CD. This particular pay-off structure generates mixed motives: a greedy person may be tempted to defect because DC > CC. On the other hand, fear of betrayal may also lead to defection because DD > CD. Therefore, when this game is played only once, defect is the dominant strategy. Irrespective of the strategy of the other person, a rational decision-maker can always increase his or her pay-off by choosing to defect. The DD outcome is therefore the Nash equilibrium.

In a sequentially played prisoner's dilemma, the decision of the first player is revealed to the second player, removing the risk of betrayal for the second player.

Coordination game

The coordination game, also called assurance game or Stag Hunt, is a dilemma game in which the temptation to free-ride has been removed. This is accomplished by lowering the pay-off for unilateral defection (DC) to the same pay-off as for mutual cooperation (CC). There are

two Nash equilibria: mutual cooperation is the pay-off dominant equilibrium, whereas mutual defect the risk-averse equilibrium. Although the game is very similar to the prisoner's dilemma, its dual Nash equilibria generate incentives to cooperate. Cooperative behavior is an economically rational choice if one expects the other party to cooperate.

Chicken game

The chicken game is another variant of the dyadic exchange game whereby both parties have to independently choose between a competitive (D) and a submissive (C or "chicken") option. The game is fatal if both people compete. It is also called an anti-coordination game, because the best strategy is to choose the opposite strategy of the partner. There are again two Nash equilibria (CD and DC) and players are best off to alternate between the two.

Dictator game

In a dictator game, a participant receives an amount of money and is then given the option to offer any part of it to an anonymous other. The partner has no power and has to accept any offer that is given to him/her.

Ultimatum Game

In this variant of the dictator game, a first player (proposer) receives a specific monetary endowment which can be shared in any proportion with the second player (the receiver). The second player has the choice to either accept the offer, in which case the split is implemented, or reject, in which case neither player receives anything. The proposer has the economic incentive to give the smallest possible amount of money and keep the rest. But knowing that a receiver who

values fairness might reject such offer, she is compelled to give more. A rich amount of empirical data collected on all continents indicates that proposers are likely to give away 50 % of their endowment, that offers of 20 % or less are rejected more than half the time, and that rejection rates increase as offers become smaller (Camerer, 2003). Because a receiver should value any positive amount more than no money at all, rejection involves a personal cost. Rejection is therefore considered an act of punishment, as it deprives the unfair proposer of benefits. A common explanation is that the responders' rejection expresses their taste for "fairness" and that they rather forgo some money than be treated unfairly (Thaler, 1988).

Trust Game

In this game (first described by Berg, Dickhaut & McCabe, 1995), player 1 (the investor) receives a small amount of money of which any portion can be "invested" with player 2 (the trustee). The transferred money is tripled. Player 2 is then given the choice to keep the tripled endowment entirely, or to give part of it back to player 1. Hence the decision of player 1 to invest the money relies heavily on the expectation that player 2 is trustworthy.

Public Goods Game

In a Public Good Game, participants typically receive an initial endowment and are asked to contribute a fraction of it to a common pool (the public good). The contributions are then multiplied by a factor, and redistributed equally among the participants. The rational choice is to contribute nothing because people who contribute zero will still receive a portion of the profits thanks to the contributions of the other participant. But, if all participants contribute zero, all will be worse off.

Endnotes

[i] Recent research shows that punishment only enhances cooperation in societies that have strong social norms. In societies that have a high degree of "anti-social punishment" (punishing cooperators due to vengeance), punishment may have no beneficial effect on cooperation (Herrmann, Thoni, & Gachter, 2008).

[ii] This logic is captured by error management theory, see Haselton and Buss (2000); Yamagishi, Jin, and Kiyonari (1999).

[iii] Short-term impulsiveness is driven primarily by the limbic system (including the ventral striatum and left posterior hippocampus), the prMFC (posterior cingulate cortex), and the medial orbitofrontal cortex. These regions respond preferentially to the immediate expectation of reward and are less sensitive to the value of future rewards.

[iv] In this review we refer specifically to the role of the arMFC (rather than the more encompassing MFC) in social cognition based on the meta-analysis of Amodio & Frith (2006). These authors delineate the arMFC as the region in between the orbitofrontal MFC (the most inferior part) and the posterior MFC. The arMFC includes the paracingulate cortex and is functionally the most diverse of the three regions. Its task-related neural activity appears to be the most associated with mentalizing tasks involving self-knowledge and person perception.

[v] There may be gender differences related to experiencing pleasure from punishment. Singer et al. (2006) investigated the empathic neural response of men and women under fMRI. Participants

were first engaged in an economic game with a confederate who either played fairly or unfairly. Afterwards, their brain activity was measured while they were observing the confederates receiving pain. The normal empathic neural response for fair players was significantly reduced when observing the unfair players receiving pain, and this was especially true for men. In addition, only in men was there a noted increase in activity in the reward system of the brain (ventral striatum and nucleus accumbens), which correlated with an expressed desire for revenge. Future research will need to unravel whether men's greater pleasure derived from observing punishment is a general phenomenon, or whether it is more specifically related to the nature of the punishment (i.e., inflicting physical pain) used in this particular study.