Fairness -
A multidimensional approach

Kumulative Arbeit

Inaugural-Dissertation
zur Erlangung der Doktorwürde

der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität
zu Bonn

vorgelegt von

Sabrina Strang

aus Köln

Bonn 2015

Gedruckt mit der Genehmigung der Philosophischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Henning Gibbons

(Vorsitzender)

Prof. Dr. Martin Reuter

(Betreuer und Gutachter)

Prof. Dr. Bernd Weber

(Gutachter)

Prof. Dr. Ulrich Ettinger

(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 07.10.2015

Table of contents

Acknowledgments

I would like to thank some people without whom this dissertation would not have been possible.

I want to thank Martin Reuter for the supervision of my dissertation, his support and guidance was of great value for me. Although not directly involved in any of my dissertation projects, he was always willing to give helpful advice. Further I would like to thank Bernd Weber for offering me the possibility to do functional imaging studies and for his help with analyzing the data and interpreting the results. I am grateful to Armin Falk for his helpful advice and his inexhaustible support. His great enthusiasm for research and his creative and sometimes unconventional way of thinking were impressive and inspiring to me.

A big thank you to my co-authors Armin Falk, Urs Fischbacher, Yang Hu, Arno Riedl, Alexander Sack, Teresa Schuhmann, Bernd Weber, Xenia Grote and especially Verena Utikal, Katarina Kuss and Jörg Groß. You helped me with planning and conducting the experiments, analyzing the data and even more important showed me how much fun research can be.

During the last four years I had the privilege to work with great colleagues: Markus Antony, Michael Böhm, Tilman Derup, Matthias Hampel, Yang Hu, Fabian Kosse, Sebastian Kube, Laura Schinabeck, Peter Trautner, Matthias Wibral, Lijun Yin and especially Holger Gerhardt, Niklas Häusler, Katarina Kuss and Tina Strombach. You were one of the main reasons why I enjoyed my work so much ☺.

I am very thankful for the enormous scientific support that all the above mentioned people provided. However, as a psychologist I know that in addition to this scientific support the emotional support is of extreme importance. Therefore I would like to thank my friends for enduring all my ups and downs, celebrating every success and giving moral uplift when necessary. Without you the last four years would have become difficult. In particular I want to thank Lisa, Miriam, Janina, Valerie, Moritz, Holger and Tina. Furthermore, I want to thank Patrick for his moral support before my defense.

Last but not least I want to thank my parents and my sister for their remarkable support: You never questioned any of my decisions, but assisted me without exception. Thank you for giving me the possibility to realize my dreams.

Summary

The extent to which humans have a preference for fairness is a puzzling phenomenon. Humans share food, goods, power, services and other resources fairly with their family, friends and even unrelated others (Fehr & Rockenbach, 2004; Gintis, 2000; Melis & Semmann, 2010). Further humans have strong preferences for being treated fairly themselves and also for others being treated fairly (Bendor, 1990; Fehr & Fischbacher, 2004). This uniquely human behavior has been object of research in several disciplines. Psychologists, economists, sociologists, neuroscientists, biologists, lawyers as well as anthropologists have tried to understand this astonishing prosocial behavior (Cook & Hardin, 2001; Falk, Fehr, & Fischbacher, 2003; Henrich et al., 2014; Walster, Berscheid, & Walster, 1973; Weiner, Graham, & Reyna, 1997). Fairness can be observed across cultures and is already present in young children (Fehr, Bernhard, & Rockenbach, 2008; Ellickson, 2001; Ostrom 2000) suggesting that fair behavior has evolutionary advantages and a underlying neural basis.

In recent years, there has been substantial progress in understanding neural mechanisms enforcing fair behavior. Psychological as well as economic theories were tested for their neurological plausibility (Dulebohn, Conlon, Sarinopoulos, Davison, & McNamara, 2009; Güroğlu, van den Bos, Rombouts, & Crone, 2010; Hsu, Anen, & Quartz, 2008). For that purpose paradigms from behavioral economics were adapted and tested in the fMRI scanner. Brain areas found to correlate with fair behavior were further tested for their causal involvement by using TMS or tDCS (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Ruff, Ugazio, & Fehr, 2013; A. Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Strang, Gross, et al., 2015; Van't Wout, Kahn, Sanfey, & Aleman, 2005).

In this dissertation I give a systematic overview about psychological and economic theories on fair behavior and their neurological plausibility. I further present four studies on fairness investigated with different neuroscientific methods and using different paradigms.

1 Introduction

Imagine it is your birthday and you are celebrating it with eight friends of yours. One of them prepared your favorite cake and gives it to you as a birthday present. You are very hungry and since you like the cake very much, you would prefer to eat the whole cake on your own. What are you going to do next? Instead of eating the whole cake on your own, you probably share the cake with your friends (Figure 1). There are several options how to divide the cake. Three possible options are the following: 1) You could count the number of people present and divide the cake into eight equally sized pieces (Figure 1A). 2) You could divide the cake into eight pieces but make your piece and the one of the cake-maker larger, since you both deserve more (Figure 1B). 3) Or you divide the cake into two small, two large and four medium sized pieces according to how much everyone needs (Figure 1C). The two small pieces are for your chubby friends, the two large pieces for your skinny friends and the medium sizes for you and the others. Most people would probably consider all options as more or less fair. The question is, why do you share the cake with others while you would prefer to eat it on your own? And what do you consider as a fair split? One answer might be, because as humans we have social norms that tell us what is fair and how to behave in these social situations. And if you do not adhere to these norms, you might find yourself celebrating your next birthday alone.

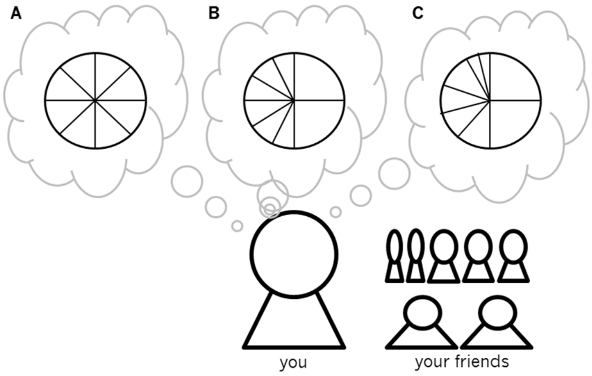

Figure 1. Possible options about how to split a birthday cake. A) Dividing the cake into equally sized pieces according to number of people present; B) dividing the cake into six small and two large pieces; C) dividing the cake into two large pieces, two small pieces and four medium size pieces.

2 Outlook

In this dissertation, I focus on one particular social norm, the norm to be fair. The first part of my dissertation provides an overview of the theoretical background of the topic. First, I will discuss psychological and economical concepts of fairness. Then I will shortly explain current neuroscientific methods that allow investigating brain activity in healthy subjects. I will also give an overview of the brain areas that have been found to be associated with fair behavior. In the second part, I will introduce three studies on fairness. The first one is about the neural correlates of strategic fairness. The others are about restorative fairness from different perspectives, as a transgressor, as a victim and as a third party. In the first study, we used transcranial magnetic stimulation (TMS; Strang, Gross, Schuhmann, Riedl, Weber and Sack, 2015), in the second and third study functional magnetic imaging (fMRI; Strang, Utikal, Fischbacher, Weber, & Falk, 2014 and Hu, Strang and Weber, 2015), and in the last one a purely behavioral paradigm (Strang, Kuss, Grote, Q Park & Weber, in preparation). By investigating fairness from different perspectives and by using a variety of methods, I contribute to obtaining a multidimensional view on fairness and its neural basis.

Part I

1 Social norms

We as humans have a highly complex system of social norm which is uniquely human and essential for the functioning of our society (Ernst Fehr & Rockenbach, 2004; Gintis, 2003; Tomasello & Rakoczy, 2003). Social norms are generally defined as unwritten statements that are based on widely shared beliefs about how individual group members should act in certain situations. They guide our choices in social interactions and thereby facilitate social behavior. They are object of research in various research fields; sociology, anthropology, psychology, law and more recently economics. Although intensively studied, there is no consensus on one common definition between but also within research disciplines investigating them.

Thomas Hobbes was one of the firsts to recognize the need for social norms, or social contract, as he called it. According to Hobbes, people are willing to fight against each other by nature, making life in large groups difficult. But since people dislike this unstable environment, they have a desire for social order. Social norms fulfill this desire, since they allow us to live in large groups of genetically unrelated others by guiding our behavior (Horne, 2001). Although social norms differ between groups, they are omnipresent; every human group or society has a social norm system (Ellickson, 2001). According to Ostrom (2000), humans have a predisposition to learn social norms, and they do this through socialization (Fine, 2001; Gintis, 2003b).

Many of our everyday situations are mixed-motive social dilemmas, like the example in the introduction (Figure 1). These are situations in which immediate self-interests are in conflict with benefits for another individual, the group or society at one's own expense. For sociologists and psychologists these are typical situations in which social norms help to guide behavior. However, according to classical economic theory these situations are no dilemmas; a homo oeconomicus only tries to maximize his own utility and disregards others. Thus he will always choose the selfish option - eating the entire cake. However, fortunately, most people are not purely selfish; they care about the welfare of others. These so called other-regarding preferences are incorporated in more recent economic theories (Falk et al., 2003; Fehr &

Falk, 2002; Fehr & Schmidt, 1999). A variety of research has shown that people often choose the non-selfish option (Camerer & Thaler, 2014; Guth, Schmittberger, & Schwarze, 1982), indicating that people value social norm compliant behavior.

1.1 Maintenance of social norms

People are even willing to punish norm violators. When being victim of a norm violation people invest their own money to punish the transgressor (second party punishment; Egas & Riedl, 2008; Fehr & Gächter, 2002). Interestingly, people are also willing to punish norm violators at their own expense as third parties (Axelrod, 1986; Ellickson, 2001; Fehr & Fischbacher, 2004; Hechter, 1984; Hu, Strang & Weber, 2015). As a third party people only observe a norm violation but are not personally involved in the transgression. In a study by Fehr & Fischbacher (2004b) 50% of the participants punished norm violators as third parties. Thus, although it is costly and they do not have any direct benefits from it, a large fraction of people punishes norm violators, probably in order to enforce norm compliant behavior.

Since social norms are informal, often vaguely defined and hence easy to circumvent, their widespread prevalence is puzzling (Fehr & Fischbacher, 2004; Melis & Semmann, 2010). Indeed, experiments have shown that prosocial behavior declines without credible sanctioning mechanisms (Egas & Riedl, 2008; Gächter, Renner, & Sefton, 2008; Ule, Schram, Riedl, & Cason, 2009). Thus, some people behave norm compliant only due to the expectation that violations will be punished. These people are strategically prosocial. Since several of our social norms are part of our legislative these are enforced by law (Ellickson, 2001), while others are enforced by the social group (Axelrod, 1986; Gächter et al., 2008; Hechter, 1984). Both forms of sanctioning threats help to enforce and thereby maintain social norms in society.

Taken together, previous research has shown that many people behave according to social norms and that they try to enforce norm compliant behavior in others. However, there is considerable variation between people; some do not act as prosocially as others do. Messick and McClintock (1968) recognized these differences and developed the social value orientation measurement to asses those. Participants are asked to make several allocations between themselves and others. Based on this they are categorized as cooperators (maximizing joint payoffs), individualists (maximizing own payoffs), competitors (maximizing the differences between own and other payoffs) and altruists (maximizing other's payoffs). This measurement was shown to correlate with prosocial behavior (Balliet, Parks, & Joireman, 2009; Van Lange, 1999). In line with the concept of other-regarding preferences, the largest fraction of people (ca. 45%; Balliet et al., 2009) is categorized as cooperators, valuing other people's outcomes.

1.2 Deviant social behavior in patients

Some neurological patients with frontal lobe damages have difficulties to behave according to social norms (Damasio, Tranel, & Damasio, 1990; Harlow, 1993; Rudebeck, Bannerman, & Rushworth, 2008). Phineas Gage is probably the most famous of these patients. As for the other patients his brain damage to the frontal lobe resulted in a profound and sudden change of his social behavior (Damasio, Grabowski, Frank, Galaburda, & Damasio, 1994). He had problems to behave according to social norms and was rather impulsive and egoistic. His deficits were constrained to the social domain; he had no learning problems, no deficits in language or perception and a normal IQ (Damasio et al., 1994; Harlow, 1993). Importantly, all patients showed normal social behavior before their injuries. Thus, the data of these patients suggest that there is a specific part in our brain involved in social norm compliant behavior.

2 Fairness

One very prevalent social norm is the norm to cooperate and to do this in a fair way (Cook & Hardin, 2001). As humans we share food, goods, power, services and other resources with our family, friends and unrelated others (Ernst Fehr & Rockenbach, 2004; Gintis, 2000; Melis & Semmann, 2010). Children at the age of 7-8 already prefer to allocate resources in a fair way (equal split; Ernst Fehr, Bernhard, & Rockenbach, 2008), showing that we are able to behave according to fairness norms from an early age on. Further we have strong preferences both for being treated fairly ourselves and also for others being treated fairly (Bendor & Mookherjee, 1990; Fehr & Fischbacher, 2004, Hu et al., 2015). This norm of fair behavior has been the object of research in several disciplines. Psychologists, economists, sociologists, neuroscientists, biologists, lawyers as well as anthropologists have tried to understand this astonishing uniquely human prosocial behavior (Cook & Hardin, 2001; Falk et al., 2003; Gintis, 2000; Henrich et al., 2014; Walster et al., 1973; Weiland et al., 2012; Weiner et al., 1997). However, although studied intensively there is no consensus on one uniform definition or concept of fairness. In the following I will concentrate on economic and psychological concepts of fairness.

2.1 Economic concept of fairness

According to standard economic theory people will not behave fairly since fair behavior usually does not maximize their own utility (Varian, 2010). Thus, to return to the example from the beginning (Figure 1), according to standard economic theory we will eat the whole cake on our own, instead of sharing it with our friends. There are several economic games used to investigate social behavior and to probe standard economic theory. The most popular are the Prisoner's Dilemma, the Dictator Game, the Trust Game and the Ultimatum Game. The major advantage of these games is that they are very simple and mathematically well-specified, in the sense that it is possible to calculate the "optimal" behavior for each player in the game. In this dissertation I will focus on the Ultimatum Game. Güth, Schmittberger and Schwarze (1982) were among the firsts who used the Ultimatum Game to test the predictions of economic standard theory. They used the Ultimatum Game to investigate how people distribute resources.

2.1.1 The Ultimatum Game

The standard Ultimatum Game is a two-person bargaining game. One player, the proposer, receives a certain monetary endowment and is asked to divide it between him/herself and another player, the receiver. The receiver can then decide whether to accept the allocation or to reject it. In case that the receiver accepts the offer from the proposer, both receive the respective amounts of money. However, in case that the receiver rejects the offer, both, proposer and receiver, do not get any money (see Figure 2).

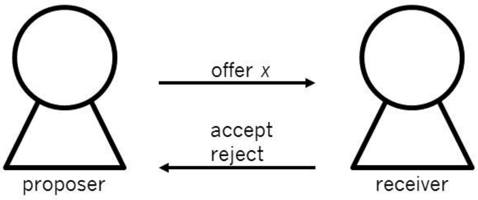

Figure 2. In the Ultimatum Game, the proposer can offer an amount to the receiver and the receiver can either accept or reject the offer. In case the receiver accepts the offer, both players receive their respective payoffs (proposer: endowment – x; receiver: x). In case the receiver rejects the offer, both players do not receive any money.

Standard economic theory predicts that the proposer should offer the smallest possible amount to the receiver. Moreover, since the receiver is supposed to maximize his or her payoffs as well, he or she should accept any amount offered. After all, any amount should be preferred over nothing. In the study by Güth et al. (1982) participants did not behave according to these standard predictions; proposers did offer money to the receivers and receivers rejected unfair offers (here defined as non-equal split). Thus, both players did not exclusively try to maximize their payoffs. Receivers explained their decision in the following way, "If player 1 left a fair amount to me, I will accept. If not and if I do not sacrifice too much, I will punish him by choosing conflict (reject)" (Güth et al., 1982, p. 384). Thus, receivers were willing to punish unfair behavior of proposers. Moreover, proposers in turn seemed to anticipate this, since they stated: "I have to leave at least an amount for player 2 so that he will consider the costs of choosing conflict (rejection) as too high" (Güth et al., 1982, p. 384). The findings of Güth et al. (1982) were replicated in a variety of subsequent studies (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Suleiman, 1996; Van't Wout, Kahn, Sanfey, & Aleman, 2006; Weg & Smith, 1993). Usually, allocations of 40–50% of the proposer's endowment are accepted and allocations below 20% are rejected in about 50% of the cases (Camerer, 2003). Thus, the economic standard hypothesis of purely selfish payoff maximizers can be rejected. When facing a conflict between selfish impulses and fairness norms both players in the Ultimatum Game incorporate fairness norms into their decision (Figure 3).

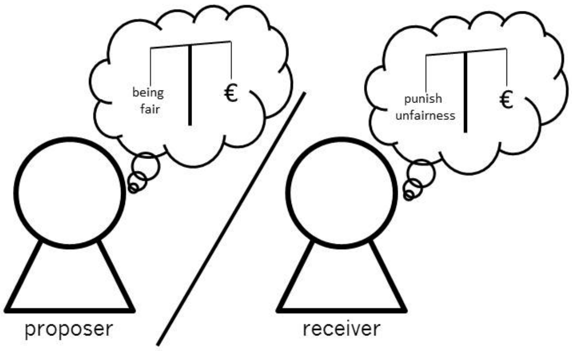

Figure 3. Conflict between the selfish impulse to maximize monetary payoffs and fairness norms in proposer and receiver in the Ultimatum Game.

2.1.2 Fair-share hypothesis and expectation hypothesis of fairness

Subsequent economic models have accounted for people's predisposition for fairness. Two alternative hypothesis to the standard theory have arisen; the fair-share hypothesis (Kahneman, Knetsch, & Thaler, 1986) and the expectation hypothesis (Hoffman, McCabe, Smith, Hoffman, & McCabe, 1996; Weg & Smith, 1993). The fair-share hypothesis predicts that proposers will split their endowment equally between themselves and the receiver (in the introduction example, splitting the cake in equally sized pieces, see Figure 1A). According to this hypothesis people should always make fair offers, even when the receiver does not have the possibility to reject. This hypothesis can be tested by using the Dictator Game. The Dictator Game is, as the Ultimatum Game, a two-person bargaining game. However, in contrast to the Ultimatum Game, the receiver is passive; he or she cannot reject the offer made by the proposer. The fair-share hypothesis predicts that allocations in the Ultimatum Game should not be different from those in the Dictator Game. Hoffman et al. (1996) demonstrated that this prediction is wrong; allocations differ substantially between the two games. Proposers in the Dictator Game transferred less money (but still more than predicted by standard economic theories) to receivers compared to proposers in the Ultimatum Game. These results show that fairness cannot be the only consideration proposers take into account when making their decisions. Hoffman et al. (1996) suggested that it is not a predisposition for fairness but rather strategic reasoning that drives proposer's allocation decision in the Ultimatum Game. People incorporate their expectations about how the receiver will decide into their decisions and act strategically fair. They anticipate that they will be punished for unfair offers and try to avoid that by making fair offers (for more details on strategic fairness see Part II.2).

2.1.3 Inequity aversion

Fehr and Schmidt (1999) proposed an alternative explanation for people's fair behavior in the Ultimatum Game. They suggest that some people are inequity-averse. According to inequity-aversion people's utility decreases when outcomes are unequal and they are therefore willing to make fair offers and to punish inequity, both at their own expenses. Fehr and Schmidt suggest that inequity-averse people dislike all transfers different from 50%. More specifically, they assume that 80% transfers are disliked as much as 20%. In line with the expectation hypothesis they assume that proposers form beliefs about the behavior of the receiver in case they do not know their preferences and behave accordingly. Thus, inequity averse proposers, who believe that the receiver is inequity averse as well, will make fair offers of 50%.

2.1.4 Reciprocal fairness

Another driving factor of fairness seems to be reciprocity (Fehr & Falk, 2002). There are two forms of reciprocity; positive reciprocity and negative reciprocity. The former means that people are willing to sacrifice own resources to be kind to those who were kind to them, while the latter means that people are willing to costly punish those who were unkind. This mechanism is suggested to drive behavior in repeated as well as in one shot situations (Fehr & Falk, 2002). The former situation is straightforward, if someone was kind to you, you will be nice to him or her next time (trial) as well. However, the latter situation is less obvious. According to Fehr and Falk (2002) people anticipate whether the other person will be reciprocal and behave according to this expectation in one-shot situations. Thus people have expectations about other people's reciprocity.

2.1.5 Intentions matter

When judging the fairness of a distribution in addition to the material consequences of an outcome the intention of the proposer matters. Falk et al. (2003) showed that the same allocation from a proposer can be judged as fair and unfair depending on the alternatives he or she had. Participants played four different versions of Mini-Ultimatum Game, meaning that the proposer's choice set was limited to two options. The first option was identical across all versions; 80% for themselves and 20% for the receiver (80/20). The alternative options were: 50% for each (50/50) in condition one, 20% for the proposer and 80% for the receiver (20/80) in condition two, 80% for the proposer and 20% for the receiver (80/20) in condition three and 100% for the proposer and 0% for the receiver (100/0) in condition four. 18% of the receivers rejected the 80/20 allocation when the proposer had actually no choice because the alternative was 80/20 as well. The behavior of these 18% can be explained by inequity aversion. However, when the alternative option is 50/50, even 45% of the receivers reject the 80/20 allocation, significantly more compared to the other conditions. Thus, although objectively identical, receivers rejected offers when the proposer had a fair alternative. The results suggest that intentions of the proposer matter in addition to objective outcomes when evaluating fairness (Falk et al., 2003; Falk, Fehr, & Fischbacher, 2008).

2.1.6 Emotions

Emotions might be another important factor influencing our decision to either accept or reject an unfair offer as a receiver in an Ultimatum Game as well. Several studies using behavioral and psychophysiological measurements have shown that unfair offers elicit negative emotions (Pillutla & Murnighan, 1996; Sanfey et al., 2003; Van't Wout et al., 2006). Straub and Murnighan (1995) suggested that negative emotions lead people to reject offers (for more details see Part II.5). This idea is supported by findings showing that negative emotions are correlated with higher rejection rates (Pillutla & Murnighan, 1996; Sanfey et al., 2003; Van't Wout et al., 2006). Unfair offers induce a conflict between cognitive and emotional processes in the receiver. On the one hand the receiver wants to increase his or her payoffs but on the other hand he or she is upset about the unfair offers and wants to restore fairness but also to restore his or her internal emotional state by punishing the proposer. Since both motives influence each other it is difficult to disentangle the two processes in a behavioral experiment.

In summary, people seem to have a general concern for fairness. However, people are not unconditionally fair, they are strategically fair by incorporating expectations about other's behavior into their decision. Finally, when assessing the fairness of a distribution, it is not only the final outcome that matters, but also the intention of the one who made the distribution.

2.2 Psychological concept of fairness

In the psychological literature 'equitable', 'just' and 'fair' are used interchangeably. Psychological research distinguishes three types of fairness; distributive, procedural and restorative justice (Gilovich, Keltner, & Nisbett, 2006). Distributive justice refers to the fairness of outcomes and procedural justice to the fairness of the process by which outcomes are distributed. Actions meant to restore justice are termed restorative justice. The three types will be explained in more detail in the following.

2.2.1 Distributive justice

There are three principles of distributive justice determining how resources should be allocated (Gilovich et al., 2006). One possibility for a fair distribution is to base the decision on people's contributions; this is called the equity principle. According to this principle outcomes should match inputs. Thus people who contribute more deserve to receive more as well. In the example in the introduction this principle corresponds to option B (Figure 1); the one who made the cake and the one whose birthday it is deserve more than the others. This principle is preferred by people with power and/or wealth, because it makes justification for their resources easier (Cook & Hegtvedt, 1896). Equality, the second principle, in contrast is based on the idea that all people contributing to an outcome should receive the same. Option A from the example in the introduction corresponds to this principle (Figure 1); everyone receives an equally large piece of cake. This principle is very common in team sports; all team members receive the same prize when winning a tournament independent of their individual performance. Finally, resources can be distributed in a way that they match people's needs, allocating most to those with the greatest need. This principle is called principle of need and is most common in families (Gilovich et al., 2006). The principle of need corresponds to option C of the example in the introduction (Figure 1).

2.2.2 Procedural justice

Distributive justice is not the only factor people base their fairness judgments on. In addition to the final outcome people have a concern for the allocation process itself (Brockner & Wiesenfeld, 1996; Folger, 1977; Tyler, 1989). Questions like, 'Who distributes the resource and why?' or 'Who else is involved in the distribution process?' do have an influence on fairness judgments as well. Procedural justice is mostly used in the context of authorities, assuming that some authority decides how to distribute resources (Gilovich et al., 2006). According to Tyler (1989) there are three factors shaping our judgment of procedural justice: neutrality, trust and standing. Standing refers to the status information transmitted by the allocator. Politeness, respect and dignity can for example be communicated in a distribution process.

Trust involves beliefs about the intention and neutrality, the honesty and unbiased view of the allocator. These three factors have an independent impact on judgments about procedural justice (Tyler, 1989).

2.2.3 Restorative justice

As long as people stick to the principles of distributive and procedural justice, a belief in a just world is maintained. A belief in a just world is the conviction that people get what they deserve (Furnham, 2003; Lerner & Miller, 1978). Injustice threatens this belief and motivates restorative actions (Greenberg, 1986; Hafer & Olson, 1993). There are different possibilities to restore justice; the most prevalent are punishment and reconciliation (Gilovich et al., 2006). Punishment of injustice can have two aims, either to revenge or to prevent future injustice (Carlsmith, Darley, & Robinson, 2002). The former is called retributive punishment and the latter utilitarian punishment. Retributive punishment is influenced by emotions, whereas utilitarian punishment is more associated with cognitive processes (Harmon-Jones, Sigelman, Bohlig, & Harmon-Jones, 2003; Lerner, Goldberg, & Tetlock, 1998). Apologies and forgiveness are two components of reconciliation. Apologies promote forgiveness; after an apology people are more likely to forgive the perpetrator of injustice (Abeler, Calaki, Andree, & Basek, 2010; Fischbacher & Utikal, 2010; Strang et al., 2014; for more detail see Part II. 3,4). However, not every transgression is forgiven after an apology, the intention behind the transgression matters. Apologies only promote forgiveness after transgressions committed unintentionally (Fischbacher & Utikal, 2010). Forgiveness is associated with decreased negative emotions, and increased empathy towards the perpetrator and it is an important factor in restoring justice (McCullough, Sandage, & Worthington, 1997; Strang et al., 2014; Walster et al., 1973).

2.2.4 Equity theory

There are different psychological models about why people behave fairly and why they value fair behavior of others (Folger, 1977; Tyler, 1994). For approximately 40 years, the dominant psychological model of fairness has been equity theory (Walster et al., 1973). Equity theory predominantly focuses on distributional justice. Like most economic models equity theory is also based on the assumption that people are selfish and try to maximize their outcomes. However, an additional assumption of equity theory is that groups maximize their collective outcome by allocating outcomes equitably to group members. To make equitable behavior of individuals profitable it will be rewarded, whereas inequitable behavior will be punished by the group. Thus people behave in an equitable way as long as they profit from it. According to equity theory an equitable distribution is one that has identical relative outcomes (Walster et al., 1973). Thus, it is not the total outcome but the ratio of total outcomes and inputs that matters. Inputs are people's contributions to an outcome. This implies that two people who invest differently in an exchange receive different total outcomes, and still this distribution is equitable. Thus, allocating resources according to equity theory follows a similar principle as the equity principle of distributive justice (see section "Distributive justice"). The values attached to inputs and outputs are determined by the social norms of a society, therefore equitability is not a uniform concept but differs between societies.

2.2.5 Relational model of justice

Another psychological model of justice is the relational model (Tyler, 1994). In contrast to equity theory the relational model is focused on procedural justice. The relational model of justice links a concern for social bonds and status with the concern about justice. The main assumption of this model is that people are predisposed to be a member of a social group and that they try to maintain and improve social status within that group. People seek to become and stay a member of a social group because membership provides a source of self-validation (Festinger, 1954). Being accepted by their social group is rewarding whereas being rejected is a form of punishment for most people. Therefore people are constantly concerned about their position in their group. Procedural justice does contain information about this position. As mentioned above the distribution process transfers information about neutrality, trust and standing (Tyler, 1989). All three factors provide group-membership information and are therefore of importance to people of the group.

3 Neuroscience and Fairness

As already mentioned in the Introduction, the first insights about brain areas that are involved in adherence to social norms and fairness in particular came from patients with frontal lobe injuries (Damasio et al., 1990; Harlow, 1993; Rudebeck et al., 2008). Patients with ventromedial prefrontal lobe damage show for example higher rejection rates for unfair offers in the Ultimatum Game, suggesting that they have an altered perception of fairness (Koenigs & Tranel, 2007). Nowadays neuroscience offers a variety of methods to investigate the underlying mechanisms of fairness in healthy participants. Neuroscientific studies on fairness can be grouped into three main categories: genetic studies, hormone studies and neuroimaging/stimulation studies. Since the structure and function of our brain is determined largely by our genes, investigating the relation between genes and fair behavior provides useful insight into the processes underlying fairness (Glahn, Thompson, & Blangero, 2007; Walter, Markett, Montag, & Reuter, 2011). Another determinant of these processes are hormones which have a huge impact on our behavior as well, therefore understanding the impact certain hormones have on fairness improves our understanding of the underlying processes as well (De Dreu, 2012).

However, in this dissertation I will focus on the third category: neuroimaging and stimulation studies. Thanks to new methods in neuroimaging it is nowadays possible to investigate the involvement of a given brain area in fairness in healthy participants. Two frequently used neuroscientific methods are functional magnetic resonance imaging (fMRI) and transcranial magnetic stimulation (TMS). Both methods will be explained in the following.

3.1 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a neuroscientific method which creates images of brain activity. It does so by measuring changes in blood oxygenation that occur in response to neuronal activity; the so called blood oxygenation level dependent signal (BOLD signal).

3.1.1 The BOLD signal

Increased local neuronal activity first leads to an increased oxygen extraction and thereby to an increased relative deoxygenated blood concentration (Huettel, Song, & McCarthy, 2009). This fast response to neuronal activity is called initial dip. After the first short-time deoxygenation, increased local neuronal activity causes a boost in local blood flow and results in an oversupply of oxygenated blood. This response to the increased energy demand has a time-lag of 4-8 seconds and is called blood oxygenation level dependent (BOLD) hemodynamic response (see Figure 4). Oxygenated and deoxygenated hemoglobin have different magnetic properties. While oxygenated blood is diamagnetic, deoxygenated hemoglobin is paramagnetic. Paramagnetic deoxygenated blood creates magnetic field distortions, whereas diamagnetic oxygenated blood leads to a more homogeneous local magnetic field (Huettel et al., 2009). It is this associated change in magnetic field homogeneity that serves as a marker for neural activity.

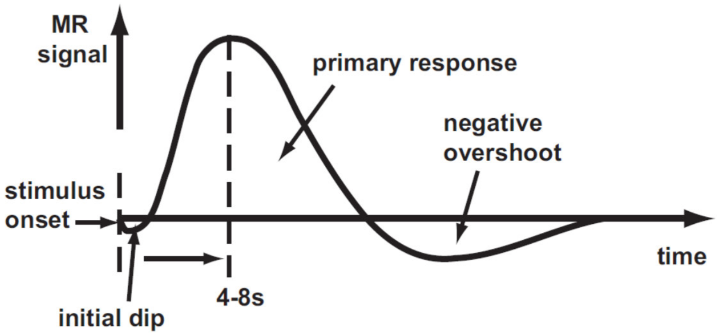

Figure 4. Schematic illustration of the blood oxygen level dependent (BOLD) signal. (Adapted from Kornak, Hall, & Haggard, 2011)

Neuronal activity results in a BOLD signal increase of about 1% to 5% (Huettel et al., 2009), this signal is masked by physical and physiological noise. Henceforth, in order to detect stimulus related effects an appropriate preprocessing and data analysis needs to be performed.

3.1.2 Preprocessing

A series of computational processes in order to remove artifacts is typically performed prior to the statistical analysis. These standard preprocessing steps are: motion correction, slice time correction, normalization, as well as temporal and spatial filtering.

FMRI data quality is strongly influenced by head movements. As a rule of thumb data sets with head movements exceeding 3 mm or 2.5° should be removed. Motion correction upon smaller head movements can substantially improve data quality. For motion correction one functional image is used as a reference to which all the other images are aligned. Head movements can be described by three translation (movements along the x, y and z-axes) and three rotation parameters (rotation around x, y and z-axes). These parameters are estimated by analyzing how an image has to be translated and rotated to better align with the reference image. The process stops when no further improvement can be achieved. The final movement parameters are then applied to the source image to produce a new image replacing the original one. The obtained parameters can also be integrated as an additional regressors in the subsequent data analysis. A perfect motion correction would results in a difference of zero with regard to the voxel by voxel subtraction from reference and source image.

In fMRI data analysis we treat one functional volume as a data set obtained at one certain time point. However, the slices of one functional volume are scanned sequentially and therefore distributed over time. Usually, 30 slices can be acquired within 3s (Huettel et al., 2009). This means, that the last slice is measured 3s later than the first slice. The aim of slice time correction is to preprocess the data in a way that the obtained data can be treated as if all slices were obtained at the same moment in time. For this purpose the time series of individual slices are shifted to match a reference time point. The choice of the reference slice depends on the stimulation protocol, in which slice acquisition can be ascending, descending or interleaved. Since the temporal shift of the slices leads to sampling at time points falling between measurement time points, the new values are estimated by interpolation. After successful slice time correction all slices within one functional volume represent the same time point (Huettel et al., 2009).

After motion correction and slice time correction, brain activity can be localized in time and space within a single participant (first level). The goal of most social neuroscience studies is, however, to analyze data at the group level (second level). Since brains can vary substantially in size and shape, individual data must be normalized in order to compare it across participants (Huettel et al., 2009). This process starts out by first determining the overall size and raw anatomic landmarks of the brain and then continues by stretching, squeezing, and warping the individual images mathematically so that brain shape and size are the same for all participants. This is done by transforming data into a common stereotaxic space, of which the most common ones are the Talairach and MNI space (Huettel et al., 2009).

Finally, the data is temporally and spatially filtered to remove artifacts. The filter choice depends on what kind of artifacts one desires to remove. High-pass temporal filters are commonly used to remove changes of very low frequency that are caused by the scanner. Depending on the paradigm additional filters can be used to remove artifacts due to heart rate, respiration etc. Furthermore the data can be spatially filtered to remove high frequency spatial components. This can be done by convolving the data with a 3D Gaussian kernel. Each voxel is then replaced by a weighted value calculated across neighboring voxels. Shape and width of the kernel determine the weights used to include adjacent voxels.

To sum up, preprocessing can substantially enhance data quality. However, there are no universally accepted criteria for preprocessing. The exact preprocessing steps and parameters rather depend on the paradigm and the subsequent analysis.

3.1.3 First level analysis

The objective of fMRI data analysis is, in its simplest form, to identify brain regions that show differential response to two different conditions (mostly control versus experimental condition). Thus, the null hypothesis (H_0) states that there is no difference in activation between two conditions, whereas the alternative hypothesis (H_1) states that there is a difference. In standard, univariate, fMRI analysis this is tested independently for the time course of each voxel resulting in one statistical value per voxel per participant.

fMRI analysis is usually performed in two steps: first and second level analysis. In the first level analysis differential activations within one participant are computed. The most frequently used first level approach is the general linear model (GLM). The GLM predicts the variation of the observed BOLD time course in terms of a linear combination of several regressors (also called predictors or explanatory variables). All regressors are convolved with the hemodynamic response function (see Figure 4.). β-weights define the contribution of each regressor. It is these β-weights that are estimated in the fMRI analysis. In order to test whether two conditions differ from each other a contrast needs to be calculated. The null hypothesis of this test states that the β values of two regressors do no not differ. Depending on the number of regressors a t or an F test can be used.

3.1.4 Second level analysis

In order to test whether the results are generalizable at population level a random effects analysis is used. The obtained contrasts from the first level analysis per subject are therefore transferred to the second level analysis. In the second level analysis a one sample t-test (in case of one group) is used to determine whether results are significantly different from zero.

One problem of fMRI analysis is the huge number of voxels and the corresponding number of statistical tests. If you had data of only one voxel you could use a conventional significant threshold of 0.05. However, having thousands of statistical tests performed simultaneously, fMRI analysis has a multiple comparison problem resulting in false positives (Type I Error). There are different methods which can be used to correct for the multiple comparison problem, with the simplest one being the Bonferroni correction. This correction adjusts the single-voxel threshold in such a way that an error probability of 0.05 at the global level is retained. Thus, with *N* independent statistical tests, a statistical significance level which is *N* times smaller than usual is used. Since this correction controls the false positives across *all* voxels, it is also called family-wise error correction (FWE-corr.).

3.2 Transcranial Magnetic Stimulation

FMRI is an excellent method for investigating correlations between brain activity and a certain task. However, it does not allow any conclusions about the causal involvement of the brain region found to be activated. Transcranial magnetic stimulation (TMS) in contrast permits establishing causality. TMS stimulates the brain through the skull without causing any pain. It works based on Faraday's law of electromagnetic induction: a change in a magnet field induces an electrical current in a wire located in this magnetic field. A brief electrical current passes through the TMS coil producing a strong magnetic field. Via rapid changes the magnetic field induces, as predicted by Faraday's law, an electrical current. Holding the TMS coil close to the skull the magnetic field passes the skull without causing pain and induces an electrical current in the underlying brain area. The induced electrical current is thought to activate neurons in the cortex (see Figure 5). Using single-pulse TMS, the stimulation induces neuronal activity different from the activity produced naturally. It therefore induces noise to the ongoing neuronal process (Pascual-Leone, Walsh, & Rothwell, 2000; Walsh & Cowey, 2000). With single-pulse TMS brain activation can be disrupted transiently and this makes it possible to assess the causal involvement on a millisecond scale (Hamilton & Pascual-Leone, 1998).

It is important to distinguish between single pulse/event-related and repetitive offline TMS (rTMS) protocols. rTMS induces after-effects. The duration of these after effects depends on the stimulation frequency and intensity. In general frequencies smaller than 1 Hz decrease cortical excitability, whereas frequencies higher than 5 Hz increase it (Sack, 2006).

A rather new stimulation protocol is theta-burst stimulation (TBS); here short bursts of three 50 Hz pulses are applied. With TBS protocols it is possible to create aftereffects of more than 60 min duration (Huang, Edwards, Rounis, Bhatia, & Rothwell, 2005).

In order to investigate cognitive processes TMS can be used as a complementary approach to fMRI. TMS can probe whether task correlated brain activity found in fMRI studies is necessary for successful performance of the task. To this end, individual or group functional coordinates are used as TMS target sites. Using this approach it was shown that regions showing task correlated activity are not necessary for task performance (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Sanfey et al., 2003). TMS localization based on fMRI data outperforms other localization methods (MRI guided TMS and 10-20 EEG position guided TMS) by yielding higher effect sizes (Sack et al., 2009). Furthermore TMS can be combined with fMRI. Using a combination of both methods, the involvement of a given brain region in a network can be investigated (Baumgartner, Knoch, Hotz, Eisenegger, & Fehr, 2011). The methods can be combined online (simultaneously) or offline (subsequently, using the rTMS after-effects). A combination with offline rTMS is easier since no fMRI compatible TMS equipment is needed. The participant is stimulated before the fMRi session and changes due to the TMS after-effects are investigated.

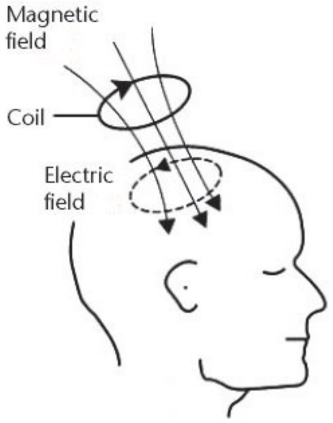

Figure 5. Basic principles of transcranial magnetic stimulation (TMS). The current in the TMS coil generates a changing magnetic field that induces an electric field in the brain (Adapted from Ruohonen & Ilmoniemi, 2002).

3.2.1 Possible confounds

Since cortical excitability varies across participants (Stewart, Walsh, & Rothwell, 2001), TMS stimulation intensity should be adjusted at the beginning of a TMS experiment. The standard is that participant's motor threshold (MT) is determined and used as an index of cortical excitability. In most TMS studies MT is measured as the minimum stimulation intensity needed to produce a muscle movement in the right index finger when stimulating the left motor cortex. In case that the muscles controlling the index finger are relaxed (finger rests on table) during stimulation it is called resting MT, in case that the muscles controlling the index finger are tensed (finger stretched out) it is called active MT. Both MTs are used in cognitive research. Stimulator output during the actual TMS paradigm is then adjusted to the MT. By using multiples of the MTs stimulation intensity can be normalized across subjects. However, there is no broad agreement on what percentage of the MT should be used for TMS stimulation the magnitudes range from 90% to 110% of the MT (Robertson, Théoret, & Pascual-Leone, 2003).

TMS is commonly described as noninvasive and painless. However depending on the target site TMS can cause sensory sensations that might interfere with task performance. Especially at frontal, temporal and occipital regions TMS can induce muscle activation. Furthermore the TMS stimulator produces a loud clicking noise. There are several approaches used to minimize the influence of these confounds to ensure that changes in task performance are specifically due to TMS stimulation.

One approach is to use different control stimulation sites. If the TMS effect is only observed at one specific site, the difference between sites is most likely due to the TMS effect. The bilateral site and the Vertex are often used as control sites. For unilateral processes the contralateral site is a good control site because the sensory effects are very similar. An alternative approach is to use sham TMS. The auditory clicking sound of a sham coil is identical to a normal TMS coil; however, no magnetic field is produced. Using repetitive TMS elegantly removes potential sensory confounds, since participants perform the task in the offline TMS period, thus not during stimulation.

3.3 The neural basis of fairness

In order to investigate the neural basis of fairness and other decision making phenomena behavioral paradigms from economics were combined with neuroscientific techniques. This interdisciplinary research field is called neuroeconomics. The aim of neuroeconomics is to gain further insights into the neural mechanisms underlying decision making in economic and social contexts. Although this field is still quite young the number of publications increased exponentially during the last years (see Figure 6). A deeper understanding of the neural basis of fairness went along with this boost of publications. Researchers have tried to find neural evidence for psychological as well as economic concepts described in the sections "Economic concept of fairness" and "Psychological concept of fairness" by using the methods explained in the section "Functional magnetic resonance imaging" and Transcranial magnetic stimulation" by using the methods described in "Functional Magnetic Imaging" and "Transcranial Magnetic Stimulation". In the following some of these studies will be described and discussed.

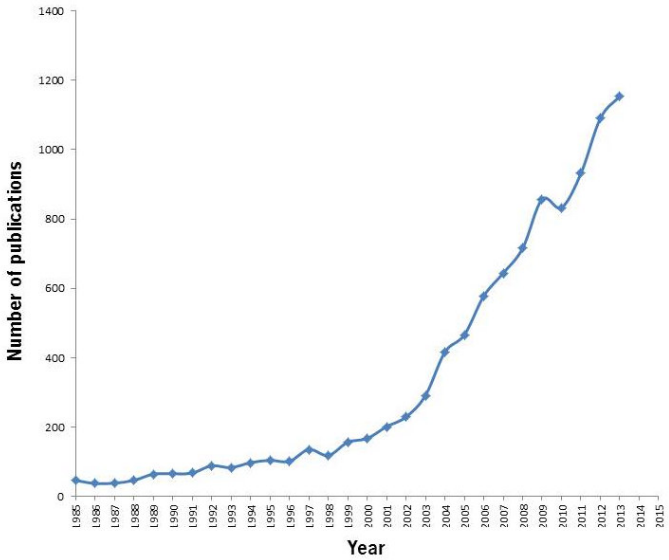

Figure 6. Increase in number of publications in Neuroeconomics using "decision making" and "brain" as search criteria in PubMed (November, 2014).

3.3.1 The neural basis of decision making in the Ultimatum Game

3.3.1.1 Unfairness

In 2003 Sanfey and colleagues were the firsts to investigate the neural mechanisms underlying decision making in the Ultimatum Game (Sanfey et al., 2003). Their aim was to test whether rejections are driven by cognitive or emotional processes and thereby to disentangle a still ongoing debate about the motives to reject unfair offers (see section "Emotions" for further details). Participants were receivers in the Ultimatum Game, they played either with a computer or with a human partner. Half of the offers participants received were fair (even split); the other half was unfair (less than 50%). The anterior insula (AI), dorsolateral prefrontal cortex (DLPFC) and anterior cingulate cortex (ACC) were more activated when participants received unfair compared to fair offers from human partners (Figure 7).

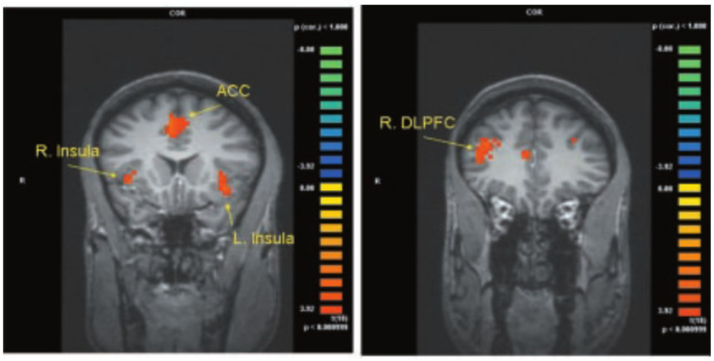

Figure 7. Activity related to unfair versus fair offers in the UG in the human condition (adapted from Sanfey et al., 2003)

These areas were more strongly activated by human than by computer offers, indicating that activation is not only a function of the amount of money. Thus, the social context, the unfair treatment of a *human* partner, evokes activity in AI, DLPFC and ACC. Activity in the AI was further sensitive to unfairness, showing increased activity the more unfair the offers were. It also correlated with rejection rates. Thus, higher activity was related to higher rejection rates. Since the insula is known to be involved in negative emotions (Evans et al., 2002;

Hein, Silani, Preuschoff, Batson, & Singer, 2010) activation in this study might reflect a negative emotional response to unfairness. This interpretation supports the theory that emotional motives drive the decision to reject unfair offers. ACC is known to be involved in conflict monitoring and might encode the conflict between selfish impulses and the emotional impulse to reject. Activity in the DLPFC was not sensitive to the degree of unfairness. Its higher activation for unfair compared to fair offers can have several reasons. The exact involvement of the DLPFC in the Ultimatum Game could not be disentangled in the study by Sanfey et al. (2003).

In order to get a better understanding of the role of the DLPFC in the decision process as a receiver in the Ultimatum Game two research groups conducted similar TMS experiments (Knoch et al., 2006; Van't Wout, Kahn, Sanfey, & Aleman, 2005). Knoch et al. (2006) wanted to test two different hypotheses about the role of the DLPFC. According to them the DLPFC can either be involved in controlling selfish or fairness impulses. The former means that the DLPFC overrides selfish impulses in order to implement fair behavior. Thus, when the DLPFC is disrupted selfish impulses should have a stronger impact on participant’s decisions leading to lower rejection rates. According to the second hypothesis fairness impulses should dominate behavior and rejection rates should increase when TMS is applied over the DLPFC. Both Knoch et al. (2006) and Van’t Wout et al. (2005) found evidence for the first hypothesis; TMS stimulation over the right but not left DLPFC decreased rejection rates. Selfish impulses cannot be controlled anymore and participants attach greater importance on maximizing their payoffs. Participants were further asked to judge the fairness of the offers. Interestingly fairness jugdments did not differ across TMS conditions. Thus, participants knew that the received offers were unfair, but still accepted them in the right DLPFC TMS condition.

These results were replicated with another brain stimulation method: transcranial direct current stimulation (tDCS; Knoch et al., 2008). Cathodal tDCS stimulation increased acceptance rates of unfair offers, while not affecting fairness judgments.

To sum up, ACC, AI and DLPFC show higher activation for unfair compared to fair offers (Sanfey et al., 2003). AI probably encodes the negative emotional response to unfair offers and ACC monitors the conflict between this emotional response and the selfish impulse to take the money. Further Knoch et al. (2006) and Van't Wout et al. (2005) showed that the right DLPFC seems to be involved in overriding selfish impulses.

However, the exact role of this network of regions and the connectivity within this network cannot be investigated with fMRI or TMS alone. Baumgartner et al. (2011) combined the two methods in order to investigate the causal effect of TMS on task related activity in the stimulated brain area (DLPFC) and on the connectivity between DLPFC, AI and ACC. They used the same TMS protocol and paradigm as Knoch et al. (2006) but during the TMS after-effect period participants lay in the scanner while playing the Ultimatum Game. They found that first, the right DLPFC was not activated by unfair offers when it was disrupted by TMS. There was no effect of left DLPFC TMS on activation of the left or right DLPFC. Second, TMS of the right DLPFC but not the left DLPFC decreased activity in the posterior ventromedial prefrontal cortex (pVMPFC). This effect was specific to unfair offers; for fair offers no differential activation in the pVMPFC due to right DLPFC TMS could be observed. Further analysis revealed that the connectivity strength between right DLPFC and pVMPFC correlated with individual rejection rates. Interestingly no differences in activation were observed in AI and ACC.

The results from Baumgartner et al. (2011) suggest that there are no top-down or bottom-up influences between DLPFC, ACC and IA. Although this result may seem surprising at first, the absence of a TMS effect on AI and ACC can explain why the fairness judgments of participants are unaffected when the right DLPFC is disrupted. Thus, while both AI and ACC might be involved in fairness attribution and fairness judgments, they are not causally involved in the behavioral change caused by TMS. DLPFC in contrast and its connectivity to the pVMPFC was shown to be causally involved in behavioral changes. pVMPFC is associated with computing decision values (Smith et al., 2010). In the context of the Ultimatum Game pVPMFC might encode the decision value of rejecting an offer and this computation might be regulated through communication with the right DLPFC in case of an unfair offer.

These three studies demonstrate the advantages of TMS and its combination with fMRI for gaining a better understanding of the role of a certain brain region or a network of brain regions in a given cognitive process. First, the TMS results by Knoch et al. (2006) and Van't Wout et al. (2005) allow for statements about a causal involvement of the right DLPFC in the decision making process and support the hypothesis that it is involved in controlling selfish impulses. Second, they could show that, although both left and right DLPFC activation are correlated with unfair offers, only the right DLPFC is *causally* involved in the decision to accept or reject. Third, by combining TMS and fMRI Baumgartner et al. (2011) demonstrated that the communication between the right DLPFC and pVMPFC is important for making fairness-norm compliant decisions (rejecting unfair offers). And finally, they could demonstrate that although IA, ACC and DLPFC are activated by unfair offers, only the right DLPFC is causally involved in the decision to reject unfair offers and thereby implementing fairness norms.

3.3.1.2 Fairness

A lot of research has been conducted on *un*fairness (unfair offers) in the Ultimatum Game and its neural correlates (Knoch et al., 2006; Sanfey et al., 2003; Van't Wout et al., 2005). Tabibnia and colleagues in contrast investigated the underlying neural process of being treated fairly (fair offers; Tabibnia, Satpute, & Lieberman, 2008). Since fair offers in the Ultimatum Game always go along with higher monetary payoffs for the receiver, it is difficult to disentangle these two factors. Tabibnia et al. (2008) controlled for the potential confound of higher monetary payoffs by varying the proposer's endowment across trials. In this way the same offer can be a fair offer with a small endowment ($5 out of $10) and an unfair offer with a larger endowment ($5 out of 20$). A difference in brain activation between these two trials can thus only be attributed to fairness concerns. Participants had the role of the receiver and lay in the scanner while seeing the offers and deciding to reject or accept. Additionally participants were asked to rate how happy they felt in response to each offer.

Their results indicate that happiness correlates positively with fairness. Moreover they found that fair compared to unfair offers elicited activity in the ventral striatum, the amygdala, the ventromedial prefrontal cortex (VMPFC), the orbitofrontal cortex (OFC) and midbrain regions.

Consistent with Sanfey et al. (2003) they found that the insula was activated during unfair trials that were rejected. The ventral striatum and the VMPFC are known to be involved in reward processing. Therefore the authors suggest that fairness is rewarding. This interpretation is in line with the idea that humans value fairness and that they have a predisposition for it (see the section "Fair-share hypothesis and expectation hypothesis of fairness" for more details).

3.3.2 Procedural and distributive justice – Two different brain processes?

In the psychological literature fairness is subdivided in procedural and distributive justice (for more details see the section "Distributive justice" and "Procedural justice"). However, whether people really distinguish between these two types of fairness could not fully be answered by purely behavioral paradigms. Dulebohn and colleagues conducted an fMRI study to test whether the two types of fairness involve different neural processes (Dulebohn et al., 2009). Their participants lay in the scanner while playing the Ultimatum Game. Before each offer in the Ultimatum Game participants had to solve three math problems. They were told that the participant who solved more problems correctly got the role of the proposer. Procedural justice was manipulated by sometimes violating this rule in different ways and assigning the participant the role of the receiver although he or she should be the dictator. Participants were for example told that they gave a wrong response although the problem was easy and actually correctly solved. Only those trials in which participants were receivers were used for the analysis. Distributive justice was manipulated via the offers participants receive; 40% and 50% of the proposer's endowment was coded as fair and 10% and 20% as unfair.

Their results indicate that distributive injustice recruits different brain areas compared to procedural justice. The dorsolateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC) and anterior insular (AI) are more activated by distributive unfairness, whereas the ventrolateral prefrontal cortex (VLPFC) and the superior temporal sulcus (STS) are more activated by procedural unfairness. The former result is in line with the results of Sanfey et al. (2003) and Tabibnia et al. (2008) who also found the AI, DLPFC and ACC to be activated by unfair offers in the Ultimatum Game. According to the authors their findings show that procedural and distributive unfairness evoke different processes; distributive injustice recruits emotion related brain areas while procedural injustice recruits areas known to be involved in social cognition (Dulebohn et al., 2009).

Thus, the results support the psychological concept of two distinct types of fairness and additionally reveal that the two different types can be linked to different brain processes, namely cognitive and emotional processes respectively.

3.3.3 Neural correlates of Equity and Efficiency

As discussed in the section "Distributive justice" there are several principles according to which resources can be distributed in a fair way. According to the principle of equality, resources should be distributed equally, meaning that everyone gets the same amount independent of their contribution (Gilovich et al., 2006). Economists sometimes use the term equity (although it has a different meaning in psychology; see the section 'Distributive justice') for the psychological principle of equality. They further distinguish equity from another principle, efficiency. Hsu, Anen and Quartz (2008) use the following example to illustrate the difference between the two principles of equity and efficiency: Imagine that there is a truck with fresh food driving to a famine-stricken region. Since it takes a long time to reach everyone, 20% of the food would spoil before everyone receives the same amount of food (equity principle). If the truck driver delivered food only to half of the people, only 5% would spoil (efficiency principle; Hsu et al., 2008). Distributing resources according to the efficiency principle means maximizing the overall good/the sum of individual payoffs, whereas distributing resources according to the equity principle means allocating resources equally even if this might decrease the overall good.

In order to test whether these two principles have different neural correlates a new paradigm was used (Hsu et al., 2008). Participants made decisions about how to allocate meals between three children from northern Uganda. In each trial different children were presented and every child had 24 meals as an endowment. Participants received two options about whom to take away some of the meals. For example they had to decide to either taking away 0 meals from child one, 11 meals from child two and 11 meals from child three or taking away 23 meals from child one, 0 meals from child two and 0 from child three. Someone who prefers equity would probably choose the first option (allocating resources as equal as possible), whereas someone preferring efficiency would choose the second option (maximizing overall good). Hsu et al. (2008) used an inequity aversion model to estimate individual parameters for efficiency and equity. Efficiency was measured by the total number of meals in each option. Equity was measured by the difference between meals.

Their results show that the activity in the putamen correlated with the efficiency parameter, whereas the activity in the insula correlated with the inequity parameter. Furthermore, both parameters correlate with activity in the caudate nucleus. These results suggest that the two different fairness principles, efficiency and equity, are coded in different brain regions, insula and putamen, and that information about both principles is combined in the caudate nucleus. Thus, the psychological and economic division of fairness in different subprinciples could be demonstrated on a neural level as well.

3.3.4 Intentions matter – for our brain as well?

Falk and Fischbacher (2003) suggest that the perception of fairness is influenced by intentionality (see the section “Intentions” for further information). Objectively identical offers are rejected when the intention of the proposer was negative (i.e. when a fair alternative was available, but he chose the unfair one). Güroğlu and colleagues tested whether different brain areas are involved in processing intentionally unfair offers compared to unintentionally unfair offers (Güroğlu et al., 2010). They used a similar task as Falk and Fischbacher (2003): three Mini-Ultimatum Games with two choice options each. One option was always 80% for the proposer and 20% for the receiver and the other one was 50%/50% (fair alternative), 20%/80% (hyperfair alternative) or 80%/20% (no alternative). In both fair alternative and hyperfair alternative conditions the intention of the proposer behind an unfair offer is clear; he or she wants to maximize his or her own payoffs. In the no alternative condition on the contrary intentions are rather ambiguous; the receiver cannot be sure whether the proposer was intentionally or unintentionally fair.

Their results indicated that activity in the insula, anterior cingulate cortex (ACC), and temporo parietal junction (TPJ) were influenced by intentionality. The insula showed higher activity when receivers rejected unfair offers in the no-alternative condition and when they accepted unfair offers in the two fair-alternative and hyperfair-alternative conditions. In the no-alternative condition accepting the unfair offer might be the social norm whereas in the other two fair-alternative and hyperfair-alternative condition rejecting is the social norm. The authors suggest that increased insula activity reflects social norms violations. These results extend the previously described neuroimaging results by Hsu, Anen, & Quartz (2008), Sanfey et al. (2003), Dulebohn et al. (2009) and Tabibnia et al. (2008) who suggested that the insula has a general role in inequity aversion. However, if the insula were involved in inequity aversion, it should be most active when participants reject unfair offers in the fair alternative condition. The results in contrast are rather in line with error signals due to social norm violations (Montague & Lohrenz, 2007) since activity is higher when social norms are violated. Hence, according to these results the insula is rather involved in detecting error signals than inequity.

The TPJ was more activated when unfair offers were rejected in the no-alternative condition compared to the other two conditions. Thus, TPJ activity was specific to the unintentional condition. Since the TPJ is known to be involved in mentalizing, activity in this paradigm probably reflects mentalizing about the intentions of the proposer. In both the fair-alternative and the hyperfair-alternative condition the intentions of the proposer are clear; only the no-alternative condition requires thinking about intentionality and therefore recruits TPJ.

In summary, Güroğlu and colleages could first of all replicate the findings of Falk and Fischbacher (2003) by showing that decision-making on the receiver's side in the Ultimatum Game is modulated by intentionality and second, demonstrate that differential brain activity can be associated with rejecting intentionally and unintentionally (ambigious) unfair offers. They thereby emphasize the importance of intentions in the assessment of fairness in the Ultimatum Game.

Part II

1 Introduction

In the second part of my dissertation I present three studies on fairness using different paradigms and methods. In all three studies fairness is defined according to the psychological principle of equality ('equity' in the economics literature). In the context of the bargaining games used in these experiments, this means an equal split for both players; unequal splits are regarded as unfair. In the first study the causal involvement of the right DPLFC in strategic fairness is investigated using TMS. The second study is an fMRI study exploring the neural basis of forgiving unfair behavior and receiving an apology. The third study investigates the neural correlates of fairness norm maintenance. And finally, in the last study a purely behavioral paradigm is used and the effect of different emotion regulation strategies on negative emotions after being treated unfairly is tested.

2 Study 1: Right DLPFC plays a causal role in strategic fairness

Published in Sabrina Strang, Jörg Gross, Teresa Schuhmann, Arno Riedl, Bernd Weber and Alexander T. Sack (2015). Be nice if you have to – The neurbiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, 10, 790-796.

As described in the section "Maintenance of social norms" the maintenance of social norms often depends on external enforcement. This means that people show increased fair behavior when they are afraid that they will get punished for unfair behavior (Ernst Fehr & Gächter, 2002; Gächter et al., 2008; Ule et al., 2009). These results suggest that humans are able to strategically adapt their behavior, i.e. acting selfishly when sanctioning is not possible and acting fairly when they have to. One prerequisite for fair behavior is the control of selfish impulses. The right DLPFC was shown to be involved in controlling selfish impulses when there is a conflict between selfish impulses and norm compliant fair behavior (Baumgartner et al., 2011; Knoch et al., 2006; Van't Wout et al., 2005; see section "The neural basis of decision making in the Ultimatum Game" for more details). In this study we investigated whether the right DLPFC is involved in the strategic acquisition of this control mechanism. We hypothesized that disruption of the right DLPFC decreases people's ability to act in a strategically fair way.

We used two different types of the Dictator Games in order to measure the TMS effect on strategic fairness, a classical Dictator Game and a Dictator Game with punishment option (similar to the Ultimatum Game). In the latter, receivers had the possibility to punish dictators by subtracting part of the dictator's payoffs. This punishment was costly for the receivers, meaning that they had to invest their own money for punishing the dictator. In the classical Dictator Game, dictators do not face any punishment threat; in the Dictator Game with punishment threat dictators should adapt their behavior by making higher transfers. We used the difference in transfers between the two games as an index for strategic adaptation. Seventeen male participants were invited to three TMS session (left DLPFC TMS, right DLPFC TMS and sham TMS) and played the role of the dictator in the two different Dictator Games. Participants received no direct feedback during the sessions and sessions were separated by at least one week. After the last session participants received the responses of the receivers and both dictator and receiver were paid according to one randomly selected trial per session. In addition to playing the Dictator Games participants judged the fairness of hypothetical offers and were asked about their punishment expectation and own punishment expenses, were they in the role of the recipient. These questions were answered while participants were still within the TMS after effect time window (approximately 7 min TMS after effects).

Our results show that TMS over the right but not left DLPFC increased selfish behavior in the classical Dictator Game. This provides further evidence for a role of the right DLPFC in overriding selfish impulses. Furthermore, the difference in transfers between the classical Dictator Game and the one with the punishment option was smaller when the right DLPFC was disrupted, suggesting that the right DLPFC is involved in a strategic acquisition of controlling selfish impulses. These results cannot be explained by altered fairness perception or punishment expectations. Both fairness perception and punishment expectations did not change across TMS conditions. However, participants indicated that they would use less money in order to punish unfair transfers from others if they would be in the role of the receiver.

To sum up, disruption of the right DLPFC impaired the strategic acquisition of the control of selfish impulses without altering fairness perception and punishment expectations. Participants knew that their transfers in the Dictator Game were unfair and expected to be punished for them but made them anyway. In line with Baumgartner et al. (2011; see section "The neural basis of decision making in the Ultimatum Game") these results provide further evidence for the hypothesis that fairness perceptions and punishment expectations are processed not in the DLPFC but in other brain regions. Moreover, although perceiving small transfers as unfair, when asked to imagine being the receiver participants were less willing to punish unfair behavior of others. This finding is in line with the results of Knoch et al. (2006) and shows that in both roles, as dictators and receivers, participants act more selfishly when the right DLPFC is disrupted. Ultimately, this implies that less unfair behavior would be punished and more selfish behavior would be tolerated, indicating that the right DLPFC is not only crucially involved in the compliance but also in the enforcement of fairness norms.

3 Study 2: The neural correlates of forgiving unfair behavior

Published in Sabrina Strang, Verena Utikal, Urs Fischbacher, Bernd Weber and Armin Falk (2014). Neural correlates of receiving an apology and active forgiveness. *PlosOne,* 5 (12), e14187.

Unfair behavior challenges relationships and people's belief in a just world. One possible process to rebuild relationships is the act of apologizing. People are more willing to forgive unfair behavior after an apology compared to no apology (Fischbacher & Utikal, 2013; McCullough, Fincham, & Tsang, 2003). Therefore, apologizing and forgiving are both important factors of restorative justice (for more details see sections "Restorative justice"). Forgiveness is associated with decreased negative emotions and increased empathy towards the person who behaved unfairly (Fincham, Paleari, & Regalia, 2002; Macaskill, Maltby, & Day, 2002; McCullough et al., 1998). Moreover, forgiveness was found to be associated with several brain regions, amongst others the left ventromedial prefrontal cortex, posterior cingulate cortex and right temporo-parietal junction (Farrow et al., 2001; Hayashi et al., 2010; Young & Saxe, 2009). However, all of these studies used different paradigms to operationalize forgiveness and none of the studies directly investigated brain areas associated with forgiving compared to not forgiving. Additionally, these studies used narrative scenarios, where participants were asked whether they would forgive a fictitious person if they were the victim. Thus, forgiveness was measured in a rather passive way and neither the unfair behavior nor the act of forgiveness had any effect on participants.

In the present experiment we combined a behavioral paradigm and neuroscientific methods to investigate the neural correlates of receiving an apology and of active forgiveness. The game involved two players, player A (transgressor, $N = 38$) and player B (affected person, in the scanner, $N = 32$). We asked players of type A to make decisions which were either fair or unfair. In case of an unfair decision, participants in the role of player A could send a message to participants in the role of player B. Subsequently players B were asked whether they wanted to forgive player A for making an unfair decision. All decision were incentivized, meaning that player A's decision to behave fairly or unfairly had monetary consequences for player B and player B's decision to forgive or not had monetary consequences for player A. In line with the psychological literature (McCullough et al., 1998) we hypothesized that apologies increased activation in empathy-related brain areas. Furthermore, concerning the neural correlates of forgiveness we expected to find an overlap with the results of Young and Saxe (2009) since their paradigm captured our measure of forgiveness most closely.

Our results support both hypotheses. Receiving an apology versus no apology revealed higher activation in the left middle temporal gyrus, left angular gyrus and left inferior frontal gyrus. All three areas have been suggested to be involved in empathy (Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Schulte-Rüther, Markowitsch, Fink, & Piefke, 2007). The results support the theory by McCullough et al. (1998) who suggest that empathy is a mediator between apologies and forgiving.

Forgiving compared to not forgiving increased activation in the right angular gyrus. Exactly the same area was also found to be activated in the study by Young and Saxe (2009). Moreover, this region seems to have a distinct role in social cognition (Carter & Huettel, 2013). Since forgiveness is a highly social process, our results provide additional evidence for an important role of the right angular gyrus in social cognition. We further replicate the finding from Fischbacher and Utikal (2010) that player B was more willing to forgive player A when player A had sent an apology. We showed that apologies likely invoke empathy for the offender and thereby increase the willingness to forgive and that forgiveness recruits a brain area known to be particularly involved in social cognition.

4 Study 3: The neural correlates of fairness norm maintenance

Published in Yang Hu, Sabrina Strang and Bernd Weber (2015). Helping or punishing strangers. Neural correlates of altruistic decisions as third party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, 9 (24), 1-11.

Social norms are a cornerstone of human society. As described in the section "Maintenance of social norms" the maintenance of social norms often depends on external enforcement. When observing a social norm transgression as a third party two enforcement mechanisms are possible: we can either help the victim or punish the violator when social norms are violated (e.g. fairness norm). Punishing the offender is referred to as retributive justice (Hogan and Emler, 1981) and helping the victim is referred to as compensatory justice (Darley and Pittman, 2003). Usually people have to choose whom they want to focus on (i.e., the offender or the victim); they have to decide whether they want the offender to pay for what he or she did, or whether they want to restore the harm done to the victim (Schroeder et al., 2003). Helping a victim as well as punishing a norm violator as a third-party (outside observer) can be regarded as altruistic acts. Both cost people at least time and effort but provide no direct benefits. A recent behavioral study found that third-party help and punishment decisions are modulated by empathy (Leliveld et al., 2012). However, the neural underpinnings of third-party help and punishment and how empathy is involved in these processes still remains unclear. In the present study we used fMRI to address these questions.

Eighty-four participants were in the role of the first and second parties in a dictator game and thirty-six participants were tested as third parties in the scanner. While lying in the scanner participants saw transgression from the first party (i.e. unfair allocations in the Dictator Game) and could decide to either punish the violator (i.e. the first party) or help the victim (i.e. the second party). In a control condition these decisions were made by the computer. Empathy was measured using the Interpersonal Reactivity Index (IRI) scale after scanning.

Both helping the victim and punishing the violator elicited similar activity in reward-related brain regions (i.e. bilateral ventral striatum). The contrast between help and punish yielded no significant activation. Moreover, IRI scores positively correlated with activation in the left lateral prefrontal cortex (LPFC) as well as in the left inferior parietal lobule (IPL) and angular gyrus (AG) for the contrast between help and punishment. The psycho-physiological interaction (PPI) analysis further indicated that the functional connectivity between the right lPFC and the bilateral striatum increased during help decisions, whereas the left lPFC showed enhanced functional connectivity with bilateral striatum during punishment decisions.

These results suggest that the mechanism underlying third-party help and punishment are similar, both are accompanied by activity in reward related areas. Further, it was shown that high empathic people recruit different brain areas compared to low empathic people in order to help or punish. These results provide evidence for understanding the neural basis of social norm enforcement and its between-subject variability.

5 Study 4: Writing regulates negative emotions due to unfair behavior

Will be published as Sabrina Strang, Xenia Grote, Katarina Kuss, Soyoung Q Park and Bernd Weber. Generalized Negative Reciprocity – How to Interrupt the Chain of Unfairness

Being treated unfairly by others elicits negative emotions (Pillutla & Murnighan, 1996; Sanfey et al., 2003; Van't Wout et al., 2006). According to Straub and Murnighan (1995) these negative emotions lead people to make emotion-driven decisions like rejecting unfair offers in the Ultimatum Game (see section "Emotions" for more details). Emotion regulation strategies might help to alter decision making in these situations. Previous research has shown that primarily reappraisal is a strategy successful in regulating negative emotions in social situations (Grecucci, Giorgetta, Van't Wout, Bonini, & Sanfey, 2013). However, most studies on emotion regulation in the Ultimatum Game have two weaknesses. First, rejection rates were used as a measure of emotion regulation success. Since altered rejection rates are only an indirect measure of a change in emotions, this measurement might not reflect the true effect. Second, reappraisal is an antecedent-focused emotion regulation strategy, meaning that the strategy has to be applied before the emotions that need to be controlled are elicited.

In our study we tried to resolve these limitations by, first, using the Dictator Game and a direct emotion measurement and, second, testing a response-focused emotion regulation, namely writing a message. We hypothesized that writing a message decreased negative emotions due to unfair offers in the Dictator Game.

One group of participants (*N* = 24) played the role of the dictator in a Mini-Dictator Game. They had to decide between a fair and an unfair offer. The other group of participants (total *N* = 213, all female) received those offers from the dictators. 80% of the offers were unfair, and only the data of participants receiving one of these unfair offers was analyzed. Subsequently, receivers were asked to write a message to the dictator who made the unfair offer. The Self-Assessment Manikin (SAM, arousal, dominance and pleasure were measured) was used to measure emotions at three time points: at baseline, after the participant received the unfair offer and after emotion regulation.

The following three experimental conditions and a control condition were applied. In one condition the message was forwarded to the dictator, in the second condition participants were asked to write a message to the dictator but it was not forwarded and in the third condition participants were asked to describe an emotionally neutral picture. In the control condition participants were instructed to simply wait for three minutes until the experiment continued.

We found that unfair offers elicited negative emotions. Participants were more aroused and felt less pleasure and dominance after they had received an unfair offer compared to baseline. Writing a message which is forwarded successfully regulated emotion; pleasure ratings were higher after participants had written the messages. Pleasure ratings were not altered by the others conditions compared to baseline. Dominance and arousal ratings did not change in any of the conditions.

The experimental design allowed us to distinguish between several factors that might explain the underlying process. Describing a picture did not have an effect on pleasure; therefore the writing process itself can be ruled out as a crucial factor in the emotion regulation process. Furthermore, a pure time factor cannot explain the effect as well; in the control condition participants waited for 3 minutes but emotion ratings did not change. As hypothesized, writing a message to the one who treated you unfairly is a successful response-focused emotion regulation strategy. Therefore, we conclude that ordering and expression of thoughts while writing a message and forwarding this message are a key factors in emotion regulation strategies.

6 Conclusion

All four studies described in part II of this dissertation investigated fairness, but with different methods and from different perspectives. Together they provide an improved understanding of fair/unfair behavior from a first-party (person who behaves fairly/unfairly) second-party (person who is treated fairly/unfairly) and third-party (person who observes someone else being treated fairly/unfairly) perspective.

Study 1 shows that the right DLPFC is involved in fair behavior in both first and second parties. Previous fMRI as well as TMS studies have investigated responses to fair and selfish behavior (second party) and have shown that the right DLPFC is involved in controlling selfish impulses that lead to the rejection of unfair offers by receivers in the Ultimatum Game (Knoch et al., 2006; Sanfey et al., 2003). We extend these findings and show that even when subjects can actively decide to behave fairly or unfairly as a dictator in the Dictator Game (first party) the right DLPFC controls the selfish impulse to keep as much as possible and thereby enforces fair behavior. We further suggest that this control is used in a strategic way. Conversely, this implies that enhanced right DLPFC activity, thus a better control of selfish impulses, would increase fair behavior. Further research should test whether this hypothesis holds true. For instance, a self-control intervention that fosters cognitive control and increases right DLPFC activation might give rise to more prosocial behavior.

Once you lose your self-control, let your selfish impulse win and are unfair against another person an apology can help to rebuild the relationship to this person. This was shown in study 2. We provided evidence that apologies trigger activity in empathy-related brain areas in the person who was treated unfairly and that apologies increase forgiveness. Moreover, we associated forgiving with activity in the right angular gyrus, an area known to be involved in many social tasks (Carter, Bowling, Reeck, & Huettel, 2012; Carter & Huettel, 2013). It would be interesting to investigate whether differences in trait empathy correlate with forgiveness. If empathy is a mediator between apologies and forgiveness, highly empathic people should tend to forgive more often compared to low empathic people. Furthermore a TMS study could test whether the right angular gyrus is causally involved in the process of forgiveness. If so, disruption of this region by TMS should decrease forgiveness.

In study 3 we show that there are two ways two maintain fairness norms as a third party. People can either punish the norm violator or help the victim, both types of behavior have the same consequence in that they diminish the inequality between violator and victim. We showed that both processes share a common neuronal basis, but that specific networks are additionally involved in the two processes. Individual differences in empathic concern are associated with different networks involved in the two processes. It would be interesting to investigate these individual differences in more detail by testing the two extremes; people very high in empathic concern and people very low in empathic concern. This might provide further insights in individual differences of the neuronal basis of third party help and punishment.

As a second party you have an emotional response to unfairness. In the fourth study we showed that unfairness decreased pleasure and dominance and increased arousal. We could further show that people can regulate their negative emotions by writing a message to the person who has treated them unfairly and that this message does not need to be forwarded in order to be effective. Open questions are, first, whether this emotion regulation has an impact on subsequent decisions concerning the person who was unfair as well as other people and, second, whether this emotion regulation strategy can be used repeatedly. Both should be investigated in further studies.

To sum up, fairness is a highly complex social behavior that requires control of selfish impulses. When disregarded, it should be followed by an apology since it elicits negative emotions in the other person.

7 References

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. *Economics Letters*, *107*(2), 233–235.

Axelrod, R. (1986). An Evolutionary Approach to Norms. *American Political Science Review*, *80*(4), 1095–1111.

Balliet, D., Parks, C., & Joireman, J. (2009). Social Value Orientation and Cooperation in Social Dilemmas: A Meta-Analysis. *Group Processes & Intergroup Relations*, *12*, 533–547.

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

Bendor, J., & Mookherjee, D. (1990). Norms, Third-Party Sanctions, and Cooperation. *Journal of Law, Economics and Organization*, *6*(1), 33–63.

Brockner, J., & Wiesenfeld, B. (1996). An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological Bulletin*, *120*(2), 189–208.

Camerer, C. (2003). Behavioral studies of strategic thinking in games. *Trends in Cognitive Sciences*, *7*(5), 225–231.

Camerer, C., & Thaler, R. (1995). Anomalies Ultimatums, Dictators and Manners. *Journal of Economic Perspectives*, *9*(2), 209–219.

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299.

Carr, L., Iacoboni, M., Dubeau, M., Mazziotta, J., & Lenzi, G. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *PNAS*, *100*(9), 5497–5502.

Carter, R., Bowling, D., Reeck, C., & Huettel, S. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, *337*(6090), 109–111.

Carter, R., & Huettel, S. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, *17*(7), 328–336.

Cook, K., & Hardin, R. (2001). Norms of Cooperativeness and Networks of trust. In M. Hechter & K.-D. Opp (Eds.), *Social Norms* (pp. 327 – 348). New York: Russel Sage Foundation.

Cook, K., & Hegtvedt, K. (1896). Justice and power. In *Justice in social relations* (pp. 19–41). Springer US.

Damasio, A., Tranel, D., & Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioral Brain Research*, *41*, 81–94.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A., & Damasio, A. (1994). The Return of Phineas Gage: Clus About the Brain from the Skull of famous Patient. *Science*, *264*(5162), 1102–1105.

De Dreu, C. (2012). Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda. *Hormones and Behavior*, *61*(3), 419–428.

Dulebohn, J., Conlon, D., Sarinopoulos, I., Davison, R., & McNamara, G. (2009). The biological bases of unfairness: Neuroimaging evidence for the distinctiveness of procedural and distributive justice. *Organizational Behavior and Human Decision Processes*, *110*(2), 140–151.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of Biological Sciences*, *275*(1637), 871–878.

Ellickson, R. (2001). The Evolution of Social Norms: A Perspective from the Legal Academy. In M. Hechter & K.-D. Opp (Eds.), *Social Norms* (pp. 35 – 76). New York: Russel Sage Foundation.

Evans, K., Banzett, R., Adams, L., Mckay, L., Richard, S., Frackowiak, J., & Corfield, D. (2002). BOLD fMRI Identifies Limbic, Paralimbic, and Cerebellar Activation During Air Hunger. *Journal of Neurophysiology*, *88*, 1500–1511.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry*, *41*(1), 20–26.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness - Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303.

Farrow, T., Zheng, Y., Wilkinson, I., Spence, S., Deakin, J., Tarrier, N., … Woodruff, P. (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*, *12*(11), 2433–2438.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*(7208), 1079–1083.

Fehr, E., & Falk, A. (2002). Reciprocal Fairness, Cooperation and Limits to Competition. In E. Fullbrook (Ed.), *Intersubjectivity in Economics: Agents and Structures* (pp. 28– 42). Bury St Edmunds: Tayler & Francis Group.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fehr, E., & Rockenbach, B. (2004). Human altruism: economic, neural, and evolutionary perspectives. *Current Opinion in Neurobiology*, *14*(6), 784–790.

Fehr, E., & Schmidt, K. (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140.

Fincham, F., Paleari, F., & Regalia, C. (2002). Forgiveness in marriage: The role of relationship quality, attributions, and empathy. *Personal Relationships*, *9*(1), 27–37.

Fine, G. A. (2001). Enacting Norms: Mushrooming and the Culture of Expectations and Explanations. In M. Hechter & K.-D. Opp (Eds.), *Social Norms* (pp. 139–165). New York: Russel Sage Foundation.

Fischbacher, U., & Utikal, V. (2013). On the Acceptance of Apologies. *Games and Economic Behavior*, *82*, 592–608.

Folger, R. (1977). Distributive and procedural justice: Combined impact of voice and improvement on experienced inequity. *Journal of Personality and Social Psychology*, *35*(2), 108–119.

Furnham, A. (2003). Belief in a just world: research progress over the past decade. *Personality and Individual Differences*, *34*(5), 795–817.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510.

Gilovich, T., Keltner, D., & Nisbett, R. (2006). *Social Psychology*. (J. Durbi, Ed.) (pp. 554–586). New York: Norton & Company.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, *206*(2), 169–179.

Gintis, H. (2003). The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. *Journal of Theoretical Biology*, *220*(4), 407–418.

Glahn, D., Thompson, P., & Blangero, J. (2007). Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, *28*(6), 488–501.

Grecucci, A., Giorgetta, C., Van't Wout, M., Bonini, N., & Sanfey, A. (2013). Reappraising the ultimatum: an fMRI study of emotion regulation and decision making. *Cerebral Cortex*, *23*(2), 399–410.

Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology*, *71*(2), 340–342.

Güroğlu, B., Van den Bos, W., Rombouts, S., & Crone, E. (2010). Unfair? It depends: neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience*, *5*(4), 414–423.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior & Organization*, *3*, 367–388.

Hafer, C., & Olson, J. (1993). Beliefs in a Just World, Discontent, and Assertive Actions by Working Women. *Personality and Social Psychology Bulletin*, *19*(1), 30–38.

Hamilton, R., & Pascual-Leone, A. (1998). Cortical plasticity associated with Braille learning. *Trends in Cognitive Sciences*, *2*(5), 168–174.

Harlow, J. (1993). Recovery from the passage of an iron bar through the head. *History of Psychiatry*, *4*(14), 274–281.

Harmon-Jones, E., Sigelman, J., Bohlig, A., & Harmon-Jones, C. (2003). Anger, coping, and frontal cortical activity: The effect of coping potential on anger-induced left frontal activity, *17*(1), 1–24.

Hayashi, A., Abe, N., Ueno, A., Shigemune, Y., Mori, E., Tashiro, M., & Fujii, T. (2010). Neural correlates of forgiveness for moral transgressions involving deception. *Brain Research*, *1332*, 90–99.

Hechter, M. (1984). When Actors Comply: Monitoring Costs and the Production of Social Order. *Acta Sociologica*, *27*(3), 161–183.

Hein, G., Silani, G., Preuschoff, K., Batson, C., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron*, *68*(1), 149–160.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Mcelreath, R., … Gintis, H. (2014). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *American Economic Review*, *91*(2), 73–78.

Hoffman, E., McCabe, K., Smith, V., Hoffman, B., & Mccabe, K. (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *American Economic Review*, *86*(3), 653–660.

Hogan, R., & Emler, N. (1981). The Justice Motive in Social Behavior. (M. J. Lerner & S. C. Lerner, Eds.) (pp. 125–143). Boston, MA: Springer US.

Horne, C. (2001). Sociological Perspective on the Emergence of Norms. In M. Hechter & K.-D. Opp (Eds.), *Social Norms* (pp. 3–35). New York: Russel Sage Foundation.

Hsu, M., Anen, C., & Quartz, S. (2008). The right and the good: distributive justice and neural encoding of equity and efficiency. *Science*, *320*(5879), 1092–1095.

Hu, Y., Strang, S. and Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, 9 (24), 1-11.

Huang, Y., Edwards, M., Rounis, E., Bhatia, K., & Rothwell, J. (2005). Theta burst stimulation of the human motor cortex. *Neuron*, *45*(2), 201–206.

Huettel, A., Song, A., & McCarthy, G. (2009). *Functional Magnetic Resonance Imaging, Second Edition*. (Sinauer Associates, Ed.)*book*. Sunderland, Massachusetts.

Darley, J., & Pittman, T. (2003). The Psychology of Compensatory and Retributive Justice. *Personality and Social Psychology Review*, 7(4), 324–336.

Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness and the Assumptions of Economics. *Journal of Business*, *59*(4), 285–300.

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

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

Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *Journal of Neuroscience*, *27*(4), 951–956.

Kornak, J., Hall, D., & Haggard, M. (2011). Spatially extended FMRI signal response to stimulus in non-functionally relevant regions of the human brain: preliminary results. *Open Neuroimaging Journal*, *5*, 24–32.

Leliveld, M., Van Dijk, E., & Van Beest, I. (2012). Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice. *European Journal of Social Psychology*, 140, 135–140.

Lerner, J., Goldberg, J., & Tetlock, P. (1998). Sober Second Thought: The Effect of Accountability, anger, and Authoritariansim on Attributions of Responsibility. *Personality and Scoial Psychology Bulletin*, *24*(6), 563– 574.

Lerner, M., & Miller, D. (1978). Just world research and the attribution process: Looking back and ahead. *Psychological Bulletin*, *85*(5), 1030–1051.

Macaskill, A., Maltby, J., & Day, L. (2002). Forgiveness of self and others and emotional empathy. *Journal of Social Psychology*, *142*(5), 663–665.

McCullough, M., Fincham, F., & Tsang, J. (2003). Forgiveness, forbearance, and time: The temporal unfolding of transgression-related interpersonal motivations. *Journal of Personality and Social Psychology*, *84*(3), 540–557.

McCullough, M., Rachal, K., Sandage, S., Worthington, E., Brown, S., & Hight, T. (1998). Interpersonal forgiving in close relationships: II. Theoretical elaboration and measurement. *Journal of Personality and Social Psychology*, *75*(6), 1586–1603.

McCullough, M., Sandage, S., & Worthington, E. (1997). *To Forgive Is Human: How to Put Your Past in the Past*. Madison: InterVarsityPress.

Melis, A., & Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society of London*, *365*(1553), 2663–2674.

Messick, D., & McClintock, C. (1968). Motivational Bases of Choices in Experimental Games. *Journal of Experimental Social Psychology*, *4*, 1–25.

Montague, P., & Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron*, *56*, 14–18.

Ostrom, E. (2000). Collective Action and the Evolution of Social Norms. *American Economic Association*, *14*(3), 137–158.

Pascual-Leone, A., Walsh, V., & Rothwell, J. (2000). Transcranial magnetic stimulation in cognitive neuroscience-virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology*, *10*(2), 232–237.

Pillutla, M., & Murnighan, J. (1996). Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers. *Organizational Behavior and Human Decision Processes*, *68*(3), 208–224.

Robertson, E., Théoret, H., & Pascual-Leone, A. (2003). Studies in cognition: the problems solved and created by transcranial magnetic stimulation. *Journal of Cognitive Neuroscience*, *15*(7), 948–60.

Rudebeck, P., Bannerman, D., & Rushworth, M. (2008). The contribution of distinct subregions of the ventromedial frontal cortex to emotion, social behavior, and decision making. *Cognitive, Affective & Behavioral Neuroscience*, *8*(4), 485–497.

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

Ruohonen, J., & Ilmoniemi, R. (2002). Physical principles for transcranial magnetic stimulation. In Arnold (Ed.), *Handbook of Transcranial Magnetic Stimulation* (pp. 17–29). New York.

Sack, A. (2006). Transcranial magnetic stimulation, causal structure-function mapping and networks of functional relevance. *Current Opinion in Neurobiology*, *16*(5), 593–599.

Sack, A., Cohen Kadosh, R., Schuhmann, T., Moerel, M., Walsh, V., & Goebel, R. (2009). Optimizing functional accuracy of TMS in cognitive studies: a comparison of methods. *Journal of Cognitive Neuroscience*, *21*(2), 207–221.

Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., & Cohen, J. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, *300*(5626), 1755–1758.

Schroeder, D., Steel, J., Woodell, A., & Bembenek, A. (2003). Justice Within Social Dilemmas. *Personality and Social Psychology Review*, 7(4), 374–387.

Schulte-Rüther, M., Markowitsch, H., Fink, G., & Piefke, M. (2007). Mirror neuron and theory of mind mechanisms involved in face-to-face interactions: a functional magnetic resonance imaging approach to empathy. *Journal of Cognitive Neuroscience*, *19*(8), 1354–1372.

Smith, D., Hayden, B., Truong, T., Song, A., Platt, M., & Huettel, S. (2010). Distinct value signals in anterior and posterior ventromedial prefrontal cortex. *Journal of Neuroscience*, *30*(7), 2490–2495.

Stewart, L., Walsh, V., & Rothwell, J. (2001). Motor and phosphene thresholds: a transcranial magnetic stimulation correlation study. *Neuropsychologia*, *39*(4), 415–419.

Strang, S., Utikal, V., Fischbacher, U., Weber, B., & Falk, A. (2014). Neural correlates of receiving an apology and active forgiveness: an FMRI study. *PloS One*, *9*(2), e87654.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. (2015). Be Nice if You Have to - The Neurobiological Roots of Strategic Fairness. *Social Cognitive and Affective Neuroscience*, 10, 790-796.

Straub, P., & Murnighan, J. (1995). An experimental investigation of ultimatum games: information, fairness, expectations, and lowest acceptable offers. *Journal of Economic Behavior and Organization*, *27*, 345–364.

Suleiman, R. (1996). Expectations and fairness in a modified Ultimatum game. *Journal of Economic Psychology*, *17*(5), 531–554.

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

Tomasello, M., & Rakoczy, H. (2003). What Makes Human Cognition Unique? From Individual to Shared to Collective Intentionality. *Mind and Language*, *18*(2), 121–147.

Tyler, T. (1989). The psychology of procedural justice: A test of the group-value model. *Journal of Personality and Social Psychology*, *57*(5), 830–838.

Tyler, T. (1994). Psychological Models of the Justice Motive: Antecedents of Distributive and Procedural Justice. *Journal of Personality and Social Psychology*, *67*(5), 850–863.

Ule, A., Schram, A., Riedl, A., & Cason, T. (2009). Indirect punishment and generosity toward strangers. *Science*, *326*(5960), 1701–1704.

Van Lange, P. (1999). The Pursuit of Joint Outcomes and Equality in Outcomes: An Integrative Model of Social Value Orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349.

Van't Wout, M., Kahn, R., Sanfey, A., & Aleman, A. (2005). Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport*, *16*(16), 1849–1852.

Van't Wout, M., Kahn, R., Sanfey, A., & Aleman, A. (2006). Affective state and decision-making in the Ultimatum Game. *Experimental Brain Research*, *169*(4), 564–568.

Varian, H. (2010). *Microeconomics - A Modern Approach*. (J. Repcheck, Ed.) (8th ed.). New York: W. W. Norton.

Walsh, V., & Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nature Neuroscience*, *1*(1), 73–79.

Walster, E., Berscheid, E., & Walster, G. (1973). New directions in equity research. *Journal of Personality and Social Psychology*, *25*(2), 151–176.

Walter, N., Markett, S., Montag, C., & Reuter, M. (2011). A genetic contribution to cooperation: dopamine-relevant genes are associated with social facilitation. *Social Neuroscience*, *6*(3), 289–301.

Weg, E., & Smith, V. (1993). On the failure to induce meager offers in ultimatum games. *Journal of Economic Psychology*, *14*, 17–32.

Weiland, S., Hewig, J., Hecht, H., Mussel, P., & Miltner, W. (2012). Neural correlates of fair behavior in interpersonal bargaining. *Social Neuroscience*, *7*(5), 537–551.

Weiner, B., Graham, S., & Reyna, C. (1997). An attributional examination of retributive versus utilitarian philosophies of punishment. *Social Justice Research*, *10*(4), 431–452.

Young, L., & Saxe, R. (2009). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.