# Neural and Cognitive Basis of Third-Party Altruistic Decision-Making and Its Modulators

Inaugural-Dissertation
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von

**Yang Hu, M.Sc.**
aus
Xi'an, China

Bonn 2017

Gedruckt mit der Genehmigung der Philosophischen Fakultät
der Rheinischen Friedrich Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:
Prof. Dr. Martin Reuter
(Vorsitzender)
Prof. Dr. Ulrich Ettinger
(Betreuer und Gutachter)
Prof. Dr. Bernd Weber
(Gutachter)
Prof. Dr. Dr. René Hurlemann
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 03.03.2017

# Acknowledgments

This dissertation can never be completed without the great help of the following people. First and foremost, I would like to thank both of my supervisors, namely Prof. Dr. Ulrich Ettinger and Prof. Dr. Bernd Weber, for their support and encouragement during the past three years of my PhD study along with their efforts in reviewing this dissertation. Second, I would like to thank all of my collaborators, including Dr. Ben Becker, Bastian David, Dr. Susann Fiedler, Dr. Holger Gerhardt, Prof. Dr. René Hurlemann, Prof. Dr. Frank Krüger, Dr. Dirk Scheele, Dr. Sabrina Strang, and Georg Voos, for their important contributions to studies involved in this dissertation. Third, I would like to thank other colleagues in Center for Economics and Neuroscience (CENs) as well as Life and Brain Center, especially Markus Antony, Marcel Bartling, Laura Enax, Xenia Grote, Laura Schinabeck, Dr. Peter Trautner, Dr. Matthias Wibral, Thorben Woelk, and Lijun Yin, for their constructive feedbacks as well as assistance in either designing or conducting the experiments. Fourth, I would like to thank Hanna Braun, Bastian David, Alexander N. Häusler, Mike Irvine, and Christina Walz for their kind help in proof reading and German translation. Fifth, I would also like to thank Dr. Holger Gerhardt for his assistance in the format. Last but never the least, I would like to show my sincere gratitude to the China Scholarship Council (CSC) for providing me with the financial support during my PhD study here in Bonn.

The way to achieve the PhD degree in Germany is full of difficulty especially for a foreign student, but I am very fortunate to be accompanied with my families, relatives and friends who always support me during the past three years. In particular, I would like to give the special thanks to my parents and Ziyan Luan, who always care my concerns the most and share with the happiness of my success as well as the pain of my failure.

# Contents

# Abstract

Human beings live in a world full of social connections. Favoring by the evolution, humans could survive the challenges of nature by not only maximizing their own interests (i.e., selfish motives) but also by considering the well fare of others even at a cost to their own resources (i.e., altruistic motives). Beyond the kindness between relatives and direct reciprocity between friends, humans, as third-party bystanders, will sometimes engage in a costly situation where social norms are violated, to achieve justice via either punishing the unknown offender or compensating the anonymous victim, even when such a violation does not directly affect their own interests and the costs incurred by them will not be paid back. Why do unaffected third parties intervene at a personal cost and what might be the underlying neural as well as cognitive mechanism? What factors might influence their decisions in such situations? To address these questions, the present dissertation used four studies by adopting a modified third-party economic paradigm to capture the third-party altruistic behaviors (i.e., third-party help and punishment) in response to an unfair situation, with the help of the technique of functional magnetic resonance imaging (fMRI; Studies 1-3) and eye-tracking (Study 4). By mainly investigating neural correlates during altruistic decision-making of third parties, Study 1 showed that signals in the bilateral striatum (esp. the ventral part) were stronger when third-party deciders chose to either help the victim or punish the selfish offender. Further analyses revealed an association between either choice of altruistic behavior, or its neural activation, and the empathic concern level, a personality trait closely related with altruism (esp. helping behavior). Studies 2-4 further tested whether, and how, other factors modulate third-party decision-making and the underlying neural or cognitive processes. In particular, Studies 2A and 2B focused on oxytocin, a so-called "pro-social" hormone, and tested whether its effect on other altruistic behaviors extended to the third-party context. As revealed by Study 2A, and replicated by Study 2B, we observed that intranasal oxytocin affects neither type of third-party altruistic decisions; rather, it modulated neural processing, especially via enhancing activity in the temporo-parietal junction (TPJ), a region shown to support mentalizing ability, during the perception of helping decision made by a computer (Study 2A). Study 3 manipulated the attention focus on different aspects of the norm violation (i.e., asking participants to consider either the unfairness of the offender or the feelings of the victim), and showed not only an effect on third-party altruistic choice behavior, but also confirmed the role of TPJ and control-related regions in such modulation. Replicating the effects of empathic concern (Study 1) and attention focus on choice behavior (Study 3), Study 4 provided the first empirical evidence that eye-movement pattern during third-party altruistic decision-making could also be biased by both factors and their interaction, shedding light on the cognitive mechanism underlying attention and information searching. Limitations of the studies and future research directions were also discussed.

# Zusammenfassung

Menschen leben in einer Welt voller sozialer Beziehungen. Im Rahmen evolutionärer Anpassungen haben Menschen gelernt nicht nur ihre eigenen Interessen zu maximieren (d.h. selbstsüchtige Motive zu verfolgen), sondern auch das Wohl anderer, selbst auf Kosten ihrer eigenen Ressourcen, zu berücksichtigen. Über das kooperative Verhalten zwischen Verwandten und die direkte Reziprozität zwischen Freunden hinaus, involvieren sich unbeteiligte Beobachter manchmal auch in Situationen, in denen soziale Normen verletzt werden. Um Gerechtigkeit zu erreichen bzw. wiederherzustellen, bestrafen sie als unbeteiligte Dritte die Täter oder unterstützen die Opfer, auch wenn sich der Verstoß der sozialen Normen nicht unmittelbar auf ihre eigenen Interessen auswirkt und die Kosten, die dadurch entstehen, nicht zurückgezahlt werden. Warum greifen unbeteiligte Dritte unter Inkaufnahme persönlicher Kosten in solche Situationen ein und was sind die zugrundeliegenden neuronalen und kognitiven Mechanismen? Welche Faktoren könnten die Entscheidungen in solchen Situationen beeinflussen? Um diese Fragen zu beantworten, wurden im Rahmen der vorliegenden Dissertation vier Studien durchgeführt, die auf einem modifizierten ökonomischen „third-party“ Paradigma basieren, um das „altruistische“ Verhalten von Dritten (d.h. Hilfe und Bestrafung von Dritten) als Reaktion auf eine ungerechte Situation mittels der funktionellen Magnetresonanztomographie (fMRI; Studien 1-3) und Eye-Tracking (Studie 4) zu erfassen. Studie 1, in der hauptsächlich neuronale Korrelate während der „altruistischen“ Entscheidungsfindung von unbeteiligten Beobachtern untersucht wurden, zeigte, dass Signale im bilateralen Striatum (insbesondere im ventralen Teil) stärker waren, wenn die unbeteiligten Beobachter sich entweder dazu entschieden dem Opfer zu helfen oder den egoistischen Täter zu bestrafen. Weitere Analysen zeigten eine Assoziation zwischen der Wahl des „altruistischen“ Verhaltens oder ihrer neuronalen Aktivierung und dem Ausmaß empathischen Empfindens, einem Persönlichkeitsmerkmal, das eng mit Altruismus zusammenhängt (insbesondere helfendem Verhalten). In den Studien 2-4 wurde weiterhin geprüft, ob und wie andere Faktoren den Zusammenhang zwischen den Entscheidungen von unbeteiligten Dritten und den zugrundeliegenden neuronalen oder kognitiven Prozessen modulieren. Insbesondere konzentrierten sich die Studien 2A und 2B auf Oxytocin, ein so genanntes „prosoziales“ Hormon, und prüften, ob die Wirkung, die es auf andere altruistische Verhaltensweisen hat, auch für den „third-party“ Kontext gilt. In Studie 2A konnte gezeigt und in Studie 2B repliziert werden, dass intranasales Oxytocin keine Art der Entscheidungen von unbeteiligten Beobachtern beeinflusst; stattdessen modulierte es die neuronale Verarbeitung, insbesondere durch verstärkte Aktivität im tempoparietalen Übergang (TPJ), einer Region, die die Mentalisierungsfähigkeit unterstützt, während der Wahrnehmung der Entscheidungshilfe durch einen Computer (Studie 2A). In Studie 3 wurde der Aufmerksamkeitsfokus auf verschiedene Aspekte der Normverletzung gelenkt (d.h. die Teilnehmer sollten entweder die Ungerechtigkeit des Täters oder die Gefühle des Opfers berücksichtigen). Dabei konnte nicht nur eine Wirkung auf das altruistische Entscheidungsverhalten von Dritten gezeigt, sondern auch die Rolle des TPJ und anderen Regionen, die mit Kontrollmechanismen in Verbindung gebracht werden, in einer solchen Modulation bestätigt werden. Studie 4 replizierte nicht nur den Effekt des Ausmaßes empathischen Befindens (Studie 1) und des Aufmerksamkeitsfokusses auf das Entscheidungsverhalten (Studie 3), sondern lieferte auch erste empirische Evidenz dafür, dass das Augenbewegungsmuster bei der altruistischen Entscheidungsfindung unbeteiligter Dritter von beiden Faktoren sowie deren Interaktionen beeinflusst werden kann. Diese Erkenntnis gibt Aufschluss über den kognitiven

Mechanismus, der Aufmerksamkeit und Informationssuche zugrunde liegt. Einschränkungen der vorliegenden Studien sowie zukünftige Forschungsrichtungen werden diskutiert.

*“Justice will not be served until those who are unaffected are as outraged as those who are.”*

*— Benjamin Franklin*

*“Let no one ever come to you without leaving better and happier.”*

*— Mother Teresa*

# 1 Introduction

Imagine a situation from everyday life: one day, you walked in a quiet forest and no others were around. Suddenly, you heard a sound nearby and then saw that a man robbing a girl's wallet. The man pushed the girl down and was about to run away; neither of them was acquainted with you before. At this moment, what would you do? If you are selfish and cold-blooded, you could always witness such a situation and step away from it, since it had nothing to do with you. However, you could also engage in this situation, even though such an intervention might cost your energy, time, and money, and even run the risk of getting hurt. Given limitations on ability and resources, usually you could only choose from one of two altruistic actions, namely to stop and fight the robber, or to take care of the girl. To leave (observe) or to engage in such a situation leads a moral dilemma. More interestingly, to punish or to help, were you to choose to engage, represents another conundrum regarding which altruistic action to take.

A couple of interesting research questions stem from the above example. For example, why do some bystanders choose to help, while others prefer to mete out punishment in response to the same situation? Under what conditions will third-party deciders change their choice preference? Within the fields of social psychology and behavioral economics, there are already numerous researches purposing to answer the above questions. However, at the moment, we still have limited knowledge about the neural and cognitive mechanisms that drive a third party to intervene norm violation at the cost to themselves and how such underlying processes, together with corresponding behaviors, may be modulated by other factors. Answering these questions constitutes the main goal of the studies included in the present dissertation.

Before taking a further look at novel studies and findings, it is always best to introduce the existing research (on third-party altruistic behaviors in the present context; i.e., help and punishment), simply because this helps us understand the research topic better. In the following section, I will start by talking about the key concepts closely related with third-party altruistic behaviors, to give the potential readers a clear, overall outline of the origin and development of the research topic. Then I will provide an overview that focuses on previous literature on this topic, so that the reader is cognizant of what has (not) already been done in this field. After that, I will introduce the motivations behind each study to conclude this section.

## 1.1 Relevant Concepts

### 1.1.1 Pro-Social Behavior and Altruism

The concepts of pro-social behavior and altruism always appear together in textbooks and research literature on social psychology, evolutionary psychology, and behavioral economics. Although they are quite similar, there has been long-standing debate, concerning the definition of these two concepts, between researchers from different fields with disparate perspectives. Therefore, it is very important to list various definitions and try to clarify the similarities and differences between them.

Generally speaking, prosocial behavior refers to a wide range of acts that are intended to benefit other people (one or more) besides oneself; usually, prosocial behavior includes the following: such as comforting, helping, and sharing, as well as more complex behaviors such as cooperating (Batson & Powell, 2003). Similarly, Penner and colleagues (2005) added another point, which is that prosocial behaviors "are defined by some significant segment of society and/or one's social group" (Penner, Dovidio, Piliavin, & Schroeder, 2005). More specifically, they decomposed prosocial behaviors into three levels, based on the scope of the research: 1) the micro level, concerning the neural and evolutionary origins of prosocial tendencies and the etiology of individual differences in these tendencies, 2) the meso level, concerning the context-specific behaviors of helper-recipient dyads (esp. helping), and 3) the macro level, concerning the actions that occur within large groups or organizations (e.g., cooperation).

Further controversy arises from the way in which people define the concept of altruism. From the perspective of behavior, altruism is usually defined as behaviors that are costly to the actor and beneficial (esp. bringing economic benefits) to the recipient (Fehr & Fischbacher, 2003; Kurzban, Burton-Chellew, & West, 2015). Both prosocial behavior and altruism mention benefiting others' (the recipient/s) welfare; however, altruism highlights the cost to the self (the actor), which leads to the view that altruism is a special type of prosocial behavior. However, there is a trend whereby recent literature mixes these two concepts together, for example, by also addressing the cost when defining prosocial behaviors (Gęsiarz & Crockett, 2015).

Disagreeing with the behavior-based definition, Batson and colleagues argued that altruism should be viewed as a motivational concept, i.e., the motivation to increase others' welfare instead of one's own welfare, in contrast to egoism (Batson, 2014; Batson & Powell, 2003). From this perspective, the concepts of altruism and prosocial behavior have different dimensions and are thus independent of each other, so that altruism (altruistic motivation) does not necessarily pro-

duce prosocial behavior, which is also not necessarily triggered by altruism (altruistic motivation).

Given that the current dissertation does not aim to address the divergence among definitions, either of prosocial behavior or altruism, or within the concept of altruism, we instead rely on a more concise (and also more popular) concept of altruism defined from the behavioral perspective in all of studies included within this dissertation.

### 1.1.2 From Kin-Based Altruism to Direct Reciprocity

Why altruism exists in human society remains a big and enduring mystery to science. In the past few decades, evolutionary biologists, anthropologists and psychologists have tried their best to find a plausible evolutionary explanation for the psychological mechanism that is designed to benefit others. By and large, these evolutionary explanations cover the following two facets of altruism (Gintis, Bowles, Boyd, & Fehr, 2003; Kurzban, et al., 2015). The first focus is on explaining why human beings, similar to many other species, desire to aid relatives (e.g., parenting behavior): namely, the role of kinship in human altruism. Based on the gene-centric view of evolution, Hamilton (1964) proposed the idea that kin-based altruism will be favored by selection if the product of the genetic relatedness between the actor and the recipient, and the fitness benefit to the recipient, is larger than the fitness cost to the actor. In other words, by delivering benefits to others who carry the same genes (i.e., relatives), genes can cause copies of themselves to increase in subsequent generations (Hamilton, 1964). However, given the fact that human parents take care of their children is so obvious and axiomatic (Cosmides & Tooby, 1994), psychologists and behavioral economists do not focus much on kin-based human altruism (Fehr & Fischbacher, 2003; Fehr & Rockenbach, 2004; Kurzban, et al., 2015).

A more intriguing and challenging question is why people would also desire to benefit non-genetically related others at cost to themselves, which is very rare in the animal kingdom (Hauser, Chen, Chen, & Chuang, 2003; Seyfarth & Cheney, 2012). For example, it is quite common in modern human society for people to establish long-term non-kin-based friendships (Hruschka, 2010) and even two strangers prefer to cooperate with, instead of defect to, each other in a repeated social context (Andreoni & Miller, 1993). The most famous theory explaining the above phenomena is the Theory of Reciprocity (Trivers, 1971). In particular, people help, or cooperate with, others at initial cost, but such altruistic behavior is still favored as the actor can benefit more through a mutual, sequential exchange of aid in the long term. The crucial point is that the exchange of altruistic acts occurs repeatedly between the same two persons, which explains the meaning of

"direct". In term of this theory, the ultimate goal of reciprocal altruism can be regarded as an instrumental means of achieving self-benefit, namely egoism.

### 1.1.3 Third-Party Reciprocity: A Type of Indirect Reciprocity

Although the Theory of Reciprocity is very powerful, it cannot fully cover and explain the more complex forms of altruism that exist exclusively in human society. For instance, a third-party observer will expend effort to reward the person (i.e., actor) who once kindly gave a seat to another person (i.e., recipient) or chase and fight with a thief (i.e., actor) who once stole money from the other person (i.e., recipient). In the above cases, neither the actor nor the recipient is known to the third party and the behavior of the actor does not directly affect the interests of the third party. Moreover, the three persons in this context are supposed to interact at most only once with each other. These critical features characterize indirect reciprocal altruism or indirect reciprocity (Nowak & Sigmund, 2005).

Formally, indirect reciprocity includes the following two types: 1) pay-it-forward (or generalized) reciprocity (also called upstream reciprocity): here, the agent first receives a benefit from one anonymous person, and then continues to benefit the other stranger. Such reciprocity is based on a recent positive experience, but is hard to understand from an evolutionary perspective (Boyd & Richerson, 1989; Pfeiffer, Rutte, Killingback, Taborsky, & Bonhoeffer, 2005), although it is often observed in the experiments (Gray, Ward, & Norton, 2014; Strang, Grote, Kuss, Park, & Weber, 2016); and 2) third-party reciprocity (also called downstream reciprocity, and exemplified above): here, the agent (i.e., third-party observer) first observes the actions of an actor towards a recipient, and then helps/rewards (if the actor performs a good action) or punishes[1] (if the actor performs a bad action) the actor. In other words, "whereby my actions toward you also depend on your behavior toward others" (Rand & Nowak, 2013). Such reciprocity is based on reputation and is more stable in evolutionarily terms (Nowak & Sigmund, 1998).

[1] It is still debatable at whether to include third-party punishment in the concept of indirect (downstream) reciprocity. In terms of the underlying motive, third-party punishment is also usually regarded as an important form of strong (negative) reciprocity (Fehr, Fischbacher, & Gächter, 2002; Gintis, 2000), which shares the key features of indirect (downstream) reciprocity (i.e., it is costly and brings no benefit, either immediately or in the future, for the actor ) but is not limited to a three-person context (e.g., a two-person sequential dilemma context).

### 1.1.4 Third-Party Altruistic Decision-Making

#### 1.1.4.1 Social Norm Violation and Third-Party Punishment

The concept of norms is one of the most important terms in the field of sociology. Despite there being various definitions, norms are widely defined as statements loaded with enforcement mechanisms that are used to regulate behaviors (Horne, 2001). More specifically, social norms refer to standards of behavior based on broadly accepted beliefs about how individuals within a group (i.e., from a family to society overall) should behave in a certain situation (Fehr & Fischbacher, 2004a). Social norms play a crucial role in constructing the basis of human society and facilitating the evolution of human altruism (e.g., enhancing interpersonal cooperation).

As implied by their definition, social norms are protected and enforced by certain mechanisms, such that social norms persist rather than decay. One of the most important enforcement mechanisms is punishment (or sanctions) imposed on behaviors that violate the social norms[2] (Bendor & Mookherjee, 1990; Fehr & Fischbacher, 2004a). The individual who punishes can be the "second party", whose (economic) welfare is directly influenced by the norm violation. The most widely used example is the ultimatum game (Güth, Schmittberger, & Schwarze, 1982). Two roles are involved in this game, namely those of a proposer and a recipient. A proposer is endowed with a sum of money (i.e., 10 €) and proposes a distribution offer to an anonymous recipient (i.e., a selfish offer: 9/1; or a fair offer: 5/5; the previous number refers to the payoff of the proposer and the latter refers to that of the recipient), who can either accept or reject the proposal. Importantly, both the proposer and the recipient receive nothing once the recipient rejects the offer. Surprisingly at first glance, the recipient always rejects offers with a share percentage lower than 25%, whereas the proposer often proposes a quasi-equal split (e.g., with the share percentage on average around 30-40%) to make sure that the offer will be accepted (Camerer & Thaler, 1995; Fehr & Schmidt, 1999). As fairness is one of the most important social norms, rejection by the recipient due to violation of the fairness norm can be regarded as an altruistic punishment, which may then cause the proposer to be more likely to abide by the fairness norm in the future.

However, a rather limited number of social norms can be enforced merely by the second-party punishment, given that the consequence of one's own (second-party) punishment of the norm violator is relatively weak in most cases. Let us reconsider the example of the thief we mentioned in the previous section. Assuming this time that there is a group of thieves instead of only one, the female victim

[2] For more details on theories and researches on social norm violation, or social injustice, please see the textbooks by (Hechter & Opp, 2001; Sabbagh & Schmitt, 2016).

who was robbed will spend a lot of energy in chasing and fighting with the thieves; however, she finally not only fails to get her belongings or property back, but is also assaulted by the assailants. As a consequence, the social norm is not enforced at all and faces potential breakdown in the future. Therefore, we need another type of enforcement mechanism, namely third-party punishment (or sanctions). This refers to costly punishment of the social norm violator meted out by the unaffected third-party observer, which could be characterized as a specific case of third-party (indirect) reciprocity. Apparently, third-party punishment can greatly increase the scope of social norms, in fact representing the essence of the norm (Fehr & Fischbacher, 2004b). More importantly, third-party punishment also has the advantage over second-party punishment that it is a necessary condition to keep maintaining a cooperative state, from an evolutionary angle (Bendor & Swistak, 2001). Last but not least, third-party punishment is only widely observed in human society; it never happens in other species, even chimpanzees, one of the closest living relatives of humans (Riedl, Jensen, Call, & Tomasello, 2012).

#### 1.1.4.2 Beyond Punishment: Third-Party Helping (Compensation)

In the context of social norm violation, the norm violator, despite being more salient, is never the only target person of the third-party observer. Rather, it is also important to lend a helping hand to the victim (e.g., spend time comforting them, or help them to call the police in the example mentioned above). This altruistic behavior not only occurs in our everyday life, but also is existent in the field of law. In particular, there usually are two ways to achieve justice against people's wrong doing (Darley & Pittman, 2003). Besides retributive justice in which addresses the punishment of offenders (Tyler & Boeckmann, 1997), restorative justice focuses more on how to aid the victim while also taking the community and offender into consideration (Bazemore, 1998).

To sum up, both punishing the perpetrator and helping (compensating) the victim, via the unaffected third-party observer, whose decisions will only bring a cost, and no benefit, to him- or herself, are regarded as altruistic responses to norm violation, which operationalizes the concept of third-party altruistic decision-making (see Figure 1 for summary and illustration).

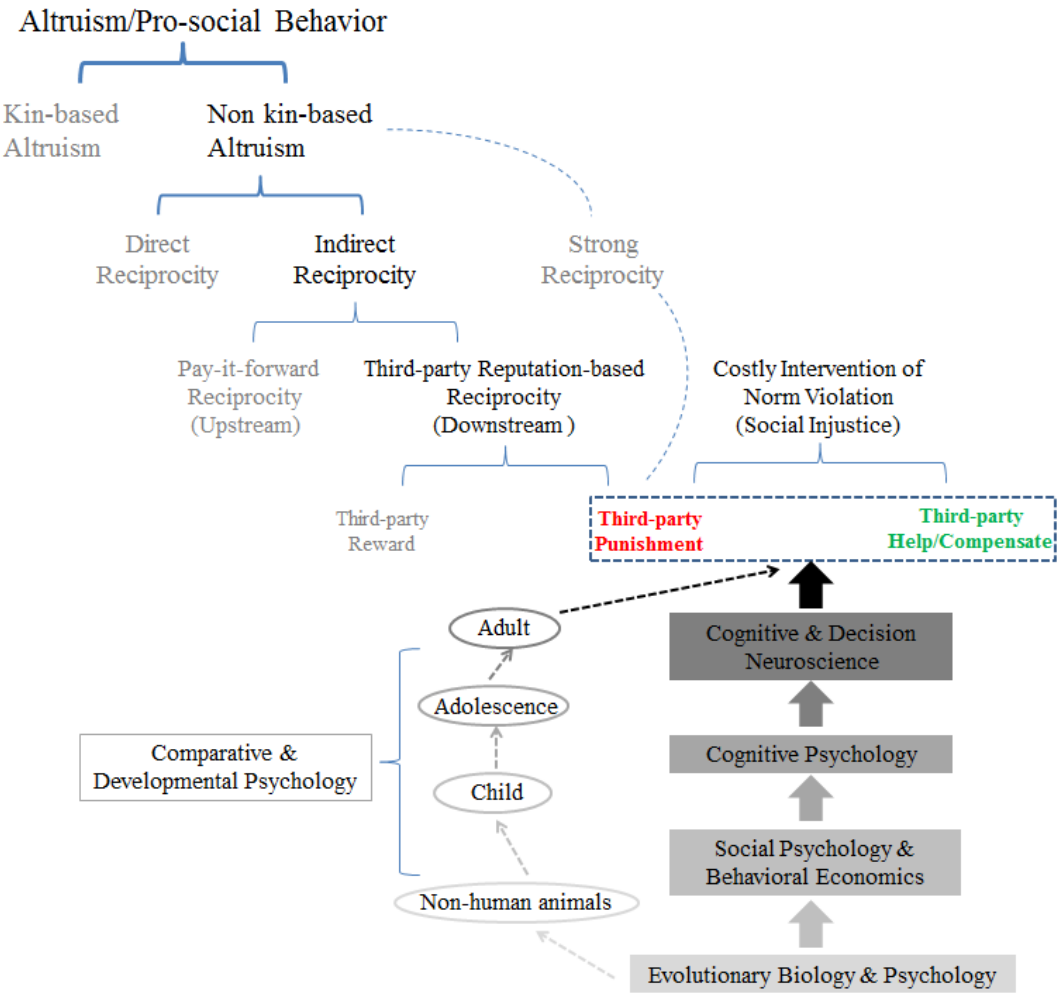


Figure 1. Key concepts relevant to and the inter-disciplinary feature of third-party help and punishment behavior.

## 1.2 Literature Review of Studies on Third-Party Altruistic Decision-making

In general, third-party punishment has been more studied and some empirical evidence has accumulated (esp. behavioral studies). In comparison, studies that take third-party help (or compensation) into consideration are rare. In my personal view, two factors might give rise to such an imbalance of research on these two types of altruistic decision: 1) As already mentioned, third-party punishment is common regarded as key to enforce and maintain a social norm (Bendor & Mookherjee, 1990). Due to its mysterious evolution and huge potential for explaining practical issues in the real life (e.g., protests, or military interventions to

keep the peace in another country), third-party punishment can always attract investigators from various fields, such as theoretical biology, psychology, and economics; and 2) Researches on helping or compensating behavior might not need a clearly defined perpetrator. For instance, there are only two roles, proposer and recipient, involved in dictator game, perhaps the most famous economic paradigm investigating giving behavior (Camerer, 2003; Forsythe, Horowitz, Savin, & Sefton, 1994; Kahneman, Knetsch, & Thaler, 1986). In social psychology studies, the situation is usually described in a way that focuses only on the emergent need of the hypothetical character (Coke, Batson, & McDavis, 1978; Toi & Batson, 1982).

In order to show a clear trajectory of previous research in third-party altruistic decision-making, the rest of this section is organized as follows: first, I will introduce the behavioral studies on third-party punishment, and then those on third-party help; then, I will introduce relevant studies that used human neuroscience techniques (esp. fMRI studies).

## 1.2.1 Behavioral Evidence[3]

### 1.2.1.1 Third-Party Punishment

#### *1.2.1.1.1 The original research*

Compared with a large amount of literature on direct reciprocity (esp. second-party punishment), researches on third-party punishment did not attract sufficient attention of the academic field (Bendor & Mookherjee, 1990; Turillo, Folger, Lavelle, Umphress, & Gee, 2002) until a crucial paper (Fehr & Fischbacher, 2004b) that systematically addressed third-party punishment, via experimental economic paradigms, was published at the beginning of this century.

Fehr and Fischbacher (2004) investigated third-party punishment in the context of two forms of norm violation. In the first study, they created the fairness norm violation via the dictator game. Participants were invited to the study and part of them was randomly assigned to either the role of Player A or Player B, in order to play the dictator game in the context of a money splitting task. Specifically, Player A was endowed with 100 monetary units (MU; 1 MU ≈ CHF 0.3) and could decide to transfer one of the following amounts from their own endowment to the anonymous Player B: 0, 10, 20, 30, 40, or 50 MU. Player B had no money at first, and could only accept an offer from the Player A that matched with him/her (i.e., Player B). The remaining part of participants was labeled as Player C, namely the third-party. Endowed with 50 MU, Player C was presented with the

[3] Here I mainly focused on behavioral studies conducted in the lab. For applied studies under an organizational setting, please see (Skarlicki, O'Reilly, & Kulik, 2015) for a comprehensive review.

choice made by an anonymous Player A, and then freely decided how much they would like to deduct from Player A's payoff with their own endowment (where the minimum amount equals 0 and the maximum amount equals 50, which leads to a possible loses of money for Player A). Importantly, the strategy method was implemented so that Player C had to respond with the amount he/she would like to use according to each possible choice by Player A (i.e., investigators would elicit six responses from each Player C). Moreover, the cost ratio for third-party punishment in this case was set to 1:3; i.e., Player C could use 1 MU from his/her endowment to deduct 3 MU from Player A's final payoff. To rule out the potential confounding effect of demand characteristics, Player C's behavior was framed as a deduction instead of a sanction or punishment. Besides, both Player A and Player B were informed of the third-party context and Player B was also asked to estimate how much Player C would punish Player A, given each possible choice made by Player A (although they cannot influence any other player's payoff). Contrary to the selfish hypothesis, which assumes that third parties would not care about another's payoff and instead always maximize their own payoff, approximately 60% of third parties deducted at least 1 MU from their own endowment to punish the selfish Player A, given their unfair choice. They also observed that the amount Player C transferred, to deduct from Player A's payoff, increased linearly with the level of inequality between the payoff of Player A and that of Player B. Intriguingly, Player B not only expected Player C to costly punish the unfair Player A, but even indicated a higher amount that they hoped Player C could use to punish Player A than the actual amount transferred by Player C, especially in extremely unfair cases (i.e., the payoff of Player A was at least twice as much as that of Player B).

Fehr and Fischbacher (2004) also tested third-party punishment in the context of violation of the cooperation norm, with the prison-dilemma paradigm. Similarly, participants were randomly assigned to the role of either Player A or Player B. Both players were endowed with 10 MU and they had the chance to interact with each other which could affect both of their payoffs. In particular, if both players cooperated, namely transferring their money to the other, their payoff would be tripled by the experimenter (i.e., the final payoff for both would be 30 MU). However, if one of them cooperated and one defected (i.e., retaining his/her original 10 MU), the cooperative player would have nothing left whereas the traitor could ultimately earn 40 MU (i.e., 30 MU tripled from the 10 MU transferred from the other, plus original 10 MU endowment). The last possible situation was that both sides chose to defect, which did not affect their payoff at all (i.e., remained on 10 MU). Third parties, again labeled as Player C, observed the interaction above and were endowed with 40 MU. Player C could use up to 20 MU to subtract the payoff from Player A or B, which was known to all Player A and Player B before-

hand. Consistent with the first study, nearly half of the Player C (45.8%) chose to punish the defector if his/her partner cooperated, which also led to the most severe punishment (≈ 3.4 MU). Last but not least, they also showed that the punishment behavior in both contexts could be predicted by negative emotions, which hinted at an underlying basis rooted in affect. Taken together, these findings supported the notion of indirect reciprocity, namely that people robustly engage in costly altruistic behavior, even if their payoffs are not directly affected by the norm violation.

*1.2.1.1.2 Follow-up studies: factors modulating third-party punishment*

*1.2.1.1.2.1 Emotion*

Investigators further looked at the factors that can modulate third-party punishment decisions. Enlightened by the findings on the relationship between emotion and third-party punishment reported by Fehr and colleagues (2004), Nelissen et al. (2009) extended the previous study and systematically evaluated how moral emotions, especially anger and guilt, can influence third-party punishment toward unfairness (Nelissen & Zeelenberg, 2009). In a similar context of inequality to that induced by the dictator game, third parties participants, facing the only unfair situation (i.e., 80/20 split of money) in the game, were randomly assigned to one of the following conditions: an unfair decision made intentionally, or not, by the proposer (i.e., randomly determined by a computer or the proposer); or the third party's decision being joint (i.e., two other participants were also assigned the role of the third party) or not (i.e., only the third-party participant decided to punish, or not). The first treatment manipulated the intention variable, with the aim of eliciting the variance in anger; the second treatment manipulated the responsibility variable, with the aim of eliciting the variance in guilt. Basically, the study showed that third parties punished significantly more when only one third party made the decision, and there was a trend toward more punishment when the unfair decision was made by the proposer. In the second study, a noise-manipulation was adopted (Van Lange, Ouwerkerk, & Tazelaar, 2002) that independently inhibited anger and guilt, instead of eliciting each emotion (i.e., the noise here means the random choice by the computer). In detail, the manipulation of a positive noise changed the original highly unequal offer to a less unequal offer (i.e., from 80/20 to 80/52; in MU), with the aim of inhibiting only guilt; whereas a negative noise referred to increasing the unequal offer (i.e., from 50/50 to 50/18; in MU), with the aim of inhibiting only anger. Consistent with their prediction, the third party punished less in both conditions, compared with the control treatment, which supported the contribution of both anger and guilt in driving third-party punishment. By focusing only on the emotion of anger, a recent study adopting a similar design to that of Fehr & Fischbacher (2004) showed that angry third parties (i.e., with anger

elicited via writing a past event that made them furious) punished more for the selfish dictator only when the emotion of anger was sustained (i.e., waiting for 3 min) instead of being distracted (i.e., playing a computer game for 3 min) before the third-party punishment task (Gummerum, Van Dillen, Van Dijk, & López-Pérez, 2016).

Inconsistent with the above study, Pedersen *et al* (2013) argued that besides anger and guilt, there is another important moral emotion, envy, which also plays a key role in predicting third-party punishment decisions (Pedersen, Kurzban, & McCullough, 2013). Due to the methodological limitations of the standard paradigm of third-party punishment, they modified the design in the following two ways: 1) no strategy method was implemented due to its impact on the affective system during decision-making, instead, third-party participants just needed to respond once to the terms of proposer's actual choice; 2) participants could be either the second-party receiver or third-party witness (randomly determined), which unfixed the pre-determined role and reduced errors in the punishment measures. Surprisingly, the third-party witness did not show the expected punishment, nor did they feel more anger (with the envy score controlled) towards the selfish (vs. fair) proposer. However, they were more envious (with the anger score controlled) of the selfish (vs. fair) proposer due to the disadvantageous payoff. Pedersen and colleagues (2013) argued that such an emotional difference was responsible for the fact that the third-party witness punished less severely and more rarely; i.e., they were more envious, but less angry, about the unfairness, which might cause them to be unwilling to punish the dictator with their own endowment.

However, the role of envy in driving third-party punishment was questioned in a recent paper (Jordan, McAuliffe, & Rand, 2014). In each of the two studies, participants, as a third-party, were asked to report how angry and envious they felt, and also how angry and envious they expected the recipient to feel. By using linear regression analyses on punishment behavior (i.e., the amount of MU third parties transferred), the study only found a significant effect of the third-party's anger, and not envy (or any vicarious affective feeling), in positively predicting their punishment behavior. Given the above findings, it seems that the third-party's own anger could consistently drive third-party punishment.

*1.2.1.1.2.2 Strategy method and endowment size*

Careful readers might note two features of the standard third-party punishment paradigm. The first feature is the use of strategy method (Mitzkewitz & Nagel, 1993; Selten, 1965), whereby participants, prior to knowing the real choice, need to respond in terms of each possible choice of the proposer. Despite its popularity implemented in studies of third-party punishment (Fehr & Fischbacher, 2004b)

and other behavioral economics studies (Falk, Fehr, & Fischbacher, 2005; Fischbacher, Gächter, & Quercia, 2012), a recent meta-analysis showed that strategy method might somehow reduce the punishment behavior, especially direct punishment meted out by the second party (Brandts & Charness, 2011). Another feature is that the initial endowment of the third party (i.e., 50 MU) is lower than that of the first party (i.e., 100 MU), thereby leading to the alternative explanation for the third-party punishment as being driven by the self-focused envy elicited by the disadvantageous inequality aversion, instead of other-regarding indirect reciprocity.

To further investigate the effect of the above two factors on third-party punishment, Jordan *et al* (2014) adopted a 2 × 2 design to systematically manipulate the decision-making type (i.e., the so-called "cold" strategy method, or the so-called "hot" specific response method) and endowment size of the third party (i.e., equal to the proposer, namely 50 MU, or less than the proposer, namely 25 MU). Despite the endowment size affecting the envy felt by third parties, it did not alter their punishment behavior contingent either on the strategy method or the endowment size. To further check the robustness of the non-significant effect of endowment size, Jordan and colleagues (2014) ran a follow-up study that extended the endowment size condition (i.e., endowment of first/third party: 100/100, high/high; 50/50, low/low; 100/50, high/low) and varied the proposer's behavior (i.e., from a binary fair/unfair response to a continuous spectrum, namely 100/0, 90/10, 80/20, 70/30, 60/40, 50/50). In line with the first study, the results still showed that third-party punishment was independent of initial endowment. All in all, these findings provide strong support for the assumption that third-party punishment served as indirect reciprocity rather than being a byproduct of self-focused envy in the face of inequality.

*1.2.1.1.2.3 Group*

In the real world, third parties are often not objective in their responses to norm violation. Rather, they may respond differently to norm violation committed by offenders from different social or racial groups. For instance, participants usually judge a crime scenario more harshly if it is violated by an outgroup versus in-group perpetrator (Sommers & Ellsworth, 2000). Is such group bias also existent in third-party punishment? If so, what mechanism drives such group bias? To address the above questions, Schiller *et al* (2014) tested how third parties behaved in a social context where the cooperation norm was violated, when the perpetrator was either an in-group member, outgroup member, or unaffiliated person (Schiller, Baumgartner, & Knoch, 2014). To increase salience of the group factor, participants were asked to report their interest in soccer (Hein, Silani, Preuschoff, Batson, & Singer, 2010) or politics (Koopmans & Rebers, 2009) so that the sup-

porter, as well as the corresponding rival, could be defined as in- or outgroup members with respect to the third party. They found that third parties punished most severely when the perpetrator was from a different group, whereas they were more lenient to in-group offenders (both compared with the unaffiliated violator) for trials in which the perpetrator defected while the victim cooperated. Furthermore, they also found that either outgroup discrimination (i.e., the difference in punishment severity meted out an outgroup perpetrator versus unaffiliated offender) or the in-group favoritism (i.e., the difference in punishment severity meted out an unaffiliated offender versus ingroup perpetrator) was positively correlated with the corresponding difference in retribution motive. This result suggests that negative affect toward offenders could explain both outgroup discrimination and in-group favoritism, which could drive the group bias in third-party punishment.

More recently, another study (Yudkin, Rothmund, Twardawski, Thalla, & Van Bavel, 2016) tested the cognitive mechanism underlying the in-group bias in third-party punishment (i.e., punish in-group offenders less severely than the outgroup offenders) from the aspect of dual-process theory (Schneider & Shiffrin, 1977; Smith & DeCoster, 2000). Yudkin and colleagues (2016) first showed that third parties responding more quickly showed more in-group bias than those responding more slowly. In follow-up studies, they directly manipulated the cognitive load and found that the punishment meted out by third parties operating under higher cognitive load (i.e., remembering a letter string) was more biased by the group membership. These findings further demonstrate that in-group bias in third-party punishment is reflexive rather than reflective.

##### *1.2.1.1.2.4 Beyond students samples: evidence from other strata of human societies and compassionate mediators*

As might be noticed, all evidence of third-party punishment has relied on student samples, which are not representative of all people. Is costly third-party punishment also seen in other strata of human society? A striking anthropological study tried to address this by applying the third-party punishment paradigm to 1,762 adult participants sampled from among 15 different populations located in five different continents (Henrich et al., 2006). These societies varied broadly in natural environment (e.g., from urban to tropical forest), economic base (e.g., from wage work to horticulture) and residence type (e.g., from sedentary to nomadic), providing a basis for a high degree of generalizability. It was found that third parties in all societies punished less as the offers increased to 50%, despite with huge inter-group variance. Moreover, the mean maximum acceptable offer in the third-party punishment game was positively correlated the mean offer provided in the dictator game across populations. These results suggest that such norm-enhancing unfair-sensitive costly behavior is widely existent in human society, which is con-

sistent with the gene-cultural co-evolution of human altruism (Boyd, Gintis, Bowles, & Richerson, 2003; Boyd & Richerson, 2002).

McCall and colleagues (2014) applied the third-party paradigm, and other altruistic-relevant paradigms, to long-term mediation practitioners with several years' worth of training in compassion or altruism (McCall, Steinbeis, Ricard, & Singer, 2014). Compared with the control group, long-term mediators, despite not reducing the degree of punishment on average, meted out less punishment with decreasing inequality between the payoff of the proposer and the recipient. Consistently, they felt much less angry about unfair offers, especially with the increasing level of inequality. These findings indicate that social preferences are not fixed and can be changed through experience (training) as well as learning.

*1.2.1.1.2.5 Age and species: a developmental and evolutionary perspective*

We know from the above evidence that third-party punishment is widely observed in human adults. A natural question then arises: how does third-party punishment develop within human beings? Moreover, does it originate from other species?

Given the fact that children at age 5-6 years pay a cost to prevent themselves from being disadvantaged relative to their peers (Blake & McAuliffe, 2011), McAuliffe *et al* (2015) investigated at which age (i.e., 5 or 6 years) children would also punish the unfair proposer and prevent another peer from being unfairly treated at the cost of their own resources. Due to their being in the primary stage of cognitive ability and to their having scant experience with money, the paradigm adopted to study children is different from the standard third-party punishment paradigm. Particularly, children as third parties were made to believe in a fake scenario whereby one peer divided six Skittles (candy) between him-/herself and another peer either in a fair (i.e., proposer/recipient: 3/3) or selfish (i.e., proposer/recipient: 6/0) way on the previous day, as described on a card. Third parties were also informed that their decisions in the current game could affect the final payoffs (i.e., the number of Skittles) and were instructed on how to respond (i.e., by pulling the handle in either the green direction to accept, or the red direction to reject, the Skittle allocation). The key manipulation in this study was whether third parties costly reject (i.e., punish) or not. In particular, if they were assigned in the cost condition, they had to pay one Skittle from their own endowment (i.e., 25 Skittles for the entire game), if and only if they chose to reject. In the free condition, however, they did not have to pay for either decision. The results showed that, although children in both age groups were more likely to punish the proposer in the free condition, only children of 6 years old were also more likely to punish the unfair proposer in both the cost and free conditions. To rule out the possibility that children punished merely because of inequality aversion, they also invited the 6-year old group to participate in a follow-up study that was exactly the same ex-

cept that all of the selfish offers were replaced with the generous offer (i.e., proposer/recipient: 0/6). Although third parties were still more likely to reject in the free condition and to reject the unequal (but generous) offers, the regression analyses of the pooled dataset (i.e., with both experiments, including selfish and generous trials) showed a strong interaction between distribution (i.e., fair or unfair) and inequity (i.e., selfish or generous). Post-hoc analyses further revealed that third parties punished more for selfish (vs. generous) offers, compared with fair offers. The above evidence, in sum, showed that the costly third-party punishment in humans emerges as early as 6 years of age.

Using a similar paradigm, Jordan and colleagues (2015) further investigated at which age (i.e., 6 or 8 years) children showed in-group bias in the context of costly third-party punishment (Jordan, McAuliffe, & Warneken, 2014), as seen in adult samples (Schiller, et al., 2014). Unlike the previous study, they adopted a minimal group paradigm, which is a weak-in-effect but cleaner method commonly used in the field of social psychology (Tajfel, Billig, Bundy, & Flament, 1971), to randomly categorize third parties into “blue” or “yellow” team. In the later decision task, third parties were presented the four combinations based on the group membership of the first peer (i.e., proposer: in-/out-group) and second peer (i.e., recipient: in-/out-group). Replicating the results whereby third parties punished the selfish proposer in both age groups at cost to themselves, they further showed that 6-year old third parties were not only more likely to punish the outgroup proposer, but also more likely to punish when the in-group recipient was treated poorly. The 8 year-old third parties only showed bias in punishing based on the group membership of the proposer, rather than that of the recipient. This interesting interaction suggested that the group bias in third-party punishment, despite emerging at an early stage, might be reduced with development.

By recruiting 8-, 12-, and 15-year-old group as well as an adult group (mean age = 22 yrs), a recent study (Gummerum & Chu, 2014) looked in more details at the following two questions: 1) whether the intention (and also outcome) can influence third-party punishment and if so 2) when this influence emerges. Similar to the standard economic paradigm, third parties always saw a pair of possible choices that could be made by the proposer: one was always 8/2, the alternative was either 5/5, 2/8, 8/2, or 10/0. A strategy method was adopted so that each third party needed to respond twice to each of the four possible pairs. Focusing on the default option (i.e., 8/2), the results showed that only adults punished less frequently and with less points when they were presented with a worse alternative (i.e., 10/0) versus better alternative (e.g., 5/5, 2/8). Although adolescent groups (i.e., the 12- and 15-year-olds) showed a similar response in the second-party punishment game to that of the adult group, they failed to consider intention in the third-party punishment game. The 8-year-old group showed fairness-sensitive

punishment based only on outcome. These findings provide further insight into the origin of the cognitive mechanism underlying third-party punishment.

Concerning the second question, a recent study investigated whether chimpanzees, one of humans' closest relatives, could also show third-party punishment behavior (Riedl, et al., 2012). A norm violation case was created, whereby an offender could steal the food of a victim via pulling the food tray away once the victim had caused the food to drop on a tray. Having witnessed such a scenario, the third-party chimpanzee could decide whether to "punish" the offender by collapsing the trapdoor (within two minutes) to prevent the thief from obtaining the food, which would nevertheless not benefit the third party. Although the chimpanzee punished the thief (vs. other control conditions) more when it was the direct victim, third-party chimpanzees did not punish often when another victim was stolen from, even when it was genetically related to the third party. In sum, these results indicate the unique feature of third-party punishment in human beings versus any other species in the animal kingdom.

#### 1.2.1.2 Third-Party Help (Compensation)

As mentioned above, punishment is not the only altruistic behavior associated with norm violation by third parties. Rather, it is also possible for them to help (compensate) the victim. One of the early papers focusing on third-party help was that by Leliveld *et al* (2012), which also tested the role of empathic concern (see later section for an introduction to this concept) in third-party altruistic decisions (Leliveld, Dijk, & Beest, 2012). In the first study, third parties were presented with a series of (un)fair choices (i.e., payoff between the proposer and the recipient: 100/0, 90/10, 80/20, 70/30, 60/40, 50/50, in MU) made by an anonymous proposer in a hypothetical dictator game (i.e., deception is used in this study). With the strategy method, third parties were asked how many MU they would like to transfer from their initial endowment to compensate the victim (i.e., 50 MU; cost rate = 1:3). Instead of finding a main effect of the offer on transfer amount for compensation, they detected a significant interaction between offer and individual empathic concern level, measured by the empathic concern subscale of Ithe nterpersonal Reactivity Index (IRI; (Davis, 1983)). In particular, the more unequal the offer was, the stronger the positive relationship between empathic concern and compensation amount. To further investigate whether empathic concern can modulate a third-party's choice preference, they ran a follow-up study (see Figure 2 for the design illustration) in which all third parties were only presented with one unfair situation (i.e., 80/20). Importantly, they were provided the help and punishment (together with the keep) choice at the same time, so that they could voluntarily choose among of the options. If they chose either one of the two altruistic choices, they were then asked to indicate the exact amount. Intriguingly, it was

found that participants with different empathic concern level displayed different choice preference. Specifically, more empathic persons were more likely to compensate the victim, whereas less empathic persons were more likely to punish the selfish proposer. Taken together, the above evidence suggests that help, as well as punishment, is also a common and reasonable choice for third parties when facing a norm violation and empathic concern can bias the choice preference.

In a similar, but more complex, study (Chavez & Bicchieri, 2013), participants as third parties were randomly assigned to two conditions. In one condition, they could either add or deduct the payoff of the proposer or the recipient respectively (i.e., the all-adjustment condition); in the other condition, they were only allowed to deduct the payoff of the proposer or the recipient (i.e., the deduct-only condition). Although third parties punished the selfish proposer at cost to themselves only if they could punish, they preferred to spend their own money to compensate the unfairly treated recipient, consistent with a previous study that used a similar paradigm (Lotz, Okimoto, Schlösser, & Fetchenhauer, 2011). These findings again demonstrate that helping is always the most common, or even the favorite, choice for third parties dealing with a norm violation (see general discussion section for a possible explanation for this phenomenon).

Similar to third-party punishment, emotion (esp. anger) is also a crucial factor that affects third-party helping behavior. A recent study tested the causal relationship between anger and third-party compensation in either an attentive or distracted condition: angry third parties (vs. those with neural emotion) gave much less to a victim treated unfairly when their anger was sustained rather than when they were distracted. Moreover, the study further distinguished other-focused anger (i.e., recall a past event where a victim was harmed so that they felt angry towards a norm transgressor) from the self-focused anger (i.e., recall a past event where they felt angry because they were badly treated) and showed that in the attentive condition, third parties with other-focused (vs. self-focused) anger compensated the victim to a large extent. These findings clarify the differential role of distinct forms of anger in third-party helping behavior (Gummerum, et al., 2016).

In order to ascertain the developmental changes in third-party altruistic decisions, a recent study (Will, Crone, van den Bos, & Güroğlu, 2013) used the modified third-party help/punishment paradigm in different groups of adolescents, including 9- (i.e., pre-adolescence), 11-, 14-, 16-, as well as 22-year-olds (i.e., young adults). Instead of being presented with the fairness norm violation, third parties observed a situation of social exclusion, which was regarded as an example of norm violation and peer victimization salient to adolescents (Blakemore & Mills, 2014). In detail, participants themselves played a ball-tossing game (Williams, Cheung, & Choi, 2000) with two other anonymous partner (i.e., includers) who passed the ball to each of the other two with equal frequency in the

first stage. Next, they observed another ball-tossing game in which three novel partners were involved. Critically, two of them (i.e., excluder) intentionally excluded the other partner (i.e., victim), who only received the ball once at the beginning but never again until the end of the game. After that, participants had the chance to influence the payoff of each of the five partners they interacted with (i.e., recipients: two includers, two excluders, one victim). Each time participants and the target other were endowed with 10 MU. Participants could choose one from among seven options (i.e., payoff of the self/other: 7/19, 8/16, 9/13, 10/10, 9/7, 8/4, 7/1, in MU) and the cost rate was set to 1:3. The results revealed a strong interaction between age group and recipient in terms of the MU that participants spent. Specifically, 9-year-old children showed a stronger preference for compensating the victim compared with the other recipients, but they did not transfer different amounts to compensate between the includers and the excluders. Participants of 14-year-olds compensated the excluders less well versus either the includers or the victim. Only the elder groups of third parties showed different compensatory behavior to different recipients: i.e., giving the most to the victim, followed by the includers, with the least given to the excluders. Despite not providing less information on punishment (i.e., all participants seem to choose to compensate the recipient by increasing their payoff), this study provided the first evidence on how development affects both third-party helping and punishment behaviors.

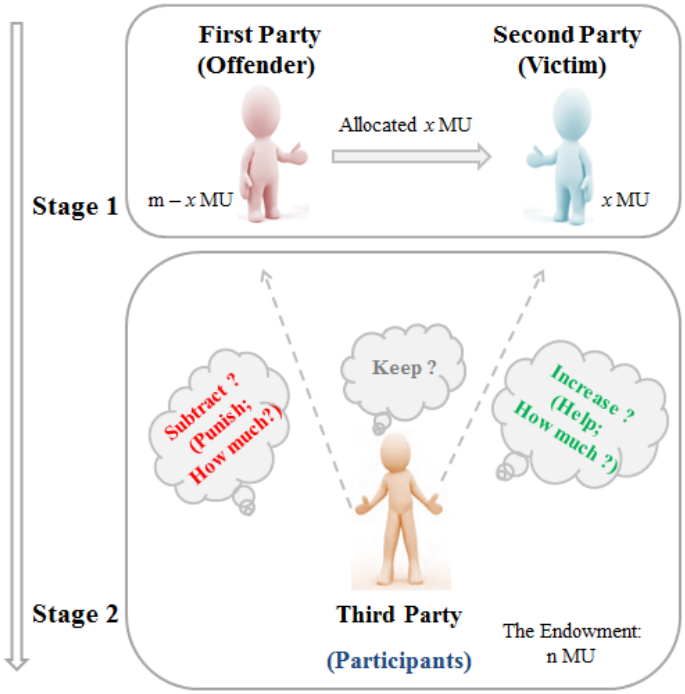


Figure 2. Illustration of the third-party economic paradigm. In Stage 1, several pairs of the first (i.e., offender) and second party (i.e., victim) were invited (either online or to the behavioral lab) and played a dictator game, namely the first party could voluntarily split a certain amount of money (i.e., x MU) from his/her endowment (i.e., m MU) to the second party. Usually x took less than half of m, causing the inequality (unfair) situation. In Stage 2, participants, as the third party, were endowed with a certain amount of money (i.e., n MU) and presented with the unequal split. They could freely decide to either punish the first party (i.e., subtract money from him/her) or help/compensate the second party (i.e., add money to him/her) and then indicate the exact amount, with the cost of their own endowment. Besides they could also choose to keep the endowment (i.e., not costly intervene). For third-party punishment game, the only difference is that participants are not allowed to help/compensate the second party. Abbreviations: MU = monetary unit.

### 1.2.2 Human Neuroscience Evidence

For cognitive neuroscientists (esp. those who are interested in topics centering on economic and social decision-making), it is far from sufficient to only acquire behavioral evidence of third-party altruistic decision-making. Their ultimate research goal is to uncover the neural mechanisms underlying such behaviors. With the increasing popularity of applying human neuroscience techniques to cognitive tasks, there are several such studies focusing on third-party altruism (esp. punishment), which extend our understanding of its underlying neural basis. In order to increase the understanding of potential readers outside of the field of cognitive

neuroscience, I will give a brief overview of the methods commonly adopted by human neuroscience studies before I introduce the neural evidence on third-party altruistic decision-making. Given that the majority of such studies included in the current thesis are only specifically relevant to the technique of functional magnetic resonance imaging (fMRI), I will focus on fMRI in the following overview.

#### 1.2.2.1 A Brief Overview of Techniques in Human Neuroscience Researches

In general, human neuroscience techniques (see Figure 3A for comparisons among different techniques) can be categorized into two major types: measurement and manipulation techniques (Ruff & Huettel, 2013). Measurement techniques refer to those that measure direct or indirect information transmission by neurons. In particular, this includes neurophysiological techniques (i.e., invasive single-unit recording and electrocorticography (ECoG), usually applied to patients with neurological or psychiatric disorders; non-invasive electroencephalography (EEG), magnetoencephalography (MEG), usually applied to healthy participants) and metabolic neuroimaging techniques (i.e., invasive positron emission tomography (PET) and non-invasive functional magnetic resonance imaging (fMRI)). By and large, neurophysiological techniques are much better at providing temporal resolution (i.e., capturing neural signal changes in the unit of milliseconds), and are therefore widely used in studies focusing on the time course of neural activity changes during perceptual or cognitive tasks. In contrast, neuroimaging methods are known for their high spatial resolution (i.e., the neural signal change can be differentiated in the order of millimeters), which can then help to demystify the link between brain regions and specific cognitive functions (Poldrack & Farah, 2015).

Undoubtedly, the fMRI technique (see Figure 3B), among all the aforementioned measurement methods, has been the most widely used in the field of cognitive neuroscience (Bandettini, 2012; Poldrack & Farah, 2015) and especially in neuroeconomics (Camerer, Loewenstein, & Prelec, 2005; Fehr & Camerer, 2007; Glimcher & Fehr, 2013; Konovalov & Krajbich, 2016; Loewenstein, Rick, & Cohen, 2008) since it first appeared in researches on human brain function nearly 25 years ago (Bandettini, Wong, Hinks, Tikofsky, & Hyde, 1992; Kwong et al., 1992; Ogawa et al., 1992). The physics, as well as the biophysics principles, behind MRI are quite complex (Huettel, Song, & McCarthy, 2004), and goes far beyond the scope of the dissertation. One point to highlight is that standard MRI cannot provide any information for understanding brain function, although it can markedly improve the visualization of anatomical structures in any part of our body (e.g., brain, heart, spine), which greatly benefits clinical diagnoses. Functional MRI actually measures changes in microvasculature oxygenation, namely the Blood Oxygenation Level Dependent (BOLD) contrasts (Ogawa, Lee, Kay, &

Tank, 1990), which are devised according to the interrelationship between neuronal activity, oxygen and glucose consumption, as well as the MR signal. BOLD-fMRI thus laid a solid foundation for the majority of later fMRI cognitive neuroscience studies.

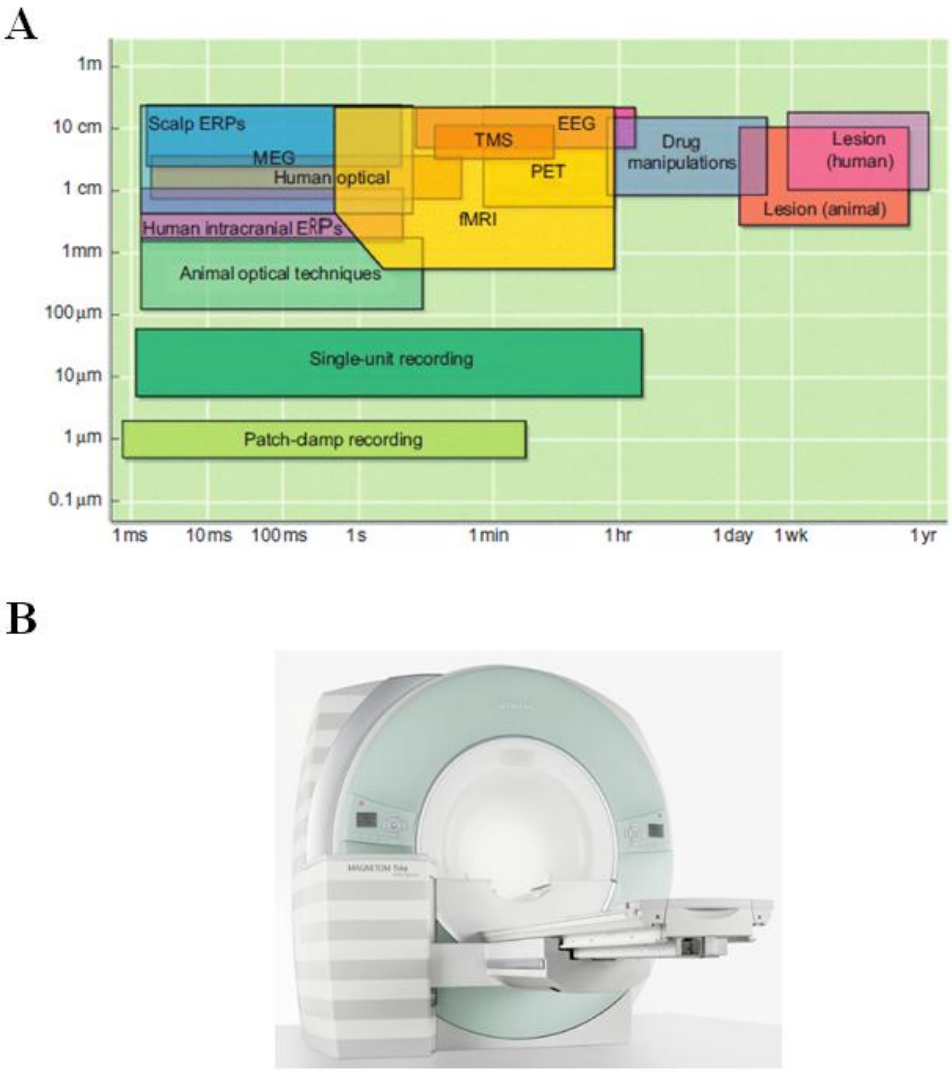


Figure 3. (A) Temporal and spatial features of different neuroscience techniques. The horizontal axis represents the temporal resolution; the vertical axis represents the spatial resolution. Abbreviations: EEG = electroencephalography, ERP = event-related potential, fMRI = functional magnetic resonance imaging, MEG = magnetoencephalography, PET = positron emission tomography, TMS = transcranial magnetic stimulation. This figure is obtained from Glimcher and Fehr (2014) with small adaptations. (B) Illustration of the Siemens Trio 3T scanner. Figure source: https://www.healthcare.siemens.ch/magnetic-resonance-imaging/for-installed-base-business-only-do-not-publish/magnetom-trio-tim.

Compared with knowledge on the physics and biophysics principles of fMRI, it is more important for cognitive neuroscientists to know how to apply this technique to a cognitive task appropriately. I will briefly summarize the key procedures (or points) in detail as follows:

1. Conducting an fMRI experiment: To ensure the fMRI study runs smoothly, it is always necessary to perform several preparatory steps, listed as follows, before running the fMRI experiment. First and foremost, it is important to make sure that the fMRI research proposal has been approved by the local ethics committee. Second, it is crucial to confirm whether the participants you recruit are fit for the MRI environment. Unlike behavioral tests, participants in fMRI studies make their response to the task while lying in a scanner. Given the powerful magnetic field in the scanner (e.g., usually 3 Tesla; the earth's magnetic field is around $5 \times 10^{-5}$ Tesla), it produces a strong gravity which can cause harm to participants with metal implants and permanent pacemakers. Thus, participants will usually be asked to fill out a safety-check questionnaire to rule out any potential harm from participating in the fMRI study. Third, researchers should take care regarding the signal synchrony between the task program and the MRI scanner; otherwise the measured BOLD signal may not reflect the neural activity changes during the cognitive stage of interest. Fourth, it is always recommended to control the length of the paradigm, for example by making it last less than 40 minutes, which can protect the participants from fatigue and distraction. Last but by no means the least, it is an issue of substantial importance to provide the warning button to participants, and to stop the scanning as soon as it is pressed, at any time during the experiment (e.g., due to claustrophobia, uncomfortable feelings and so on).

2. fMRI data analyses (see Figure 4): Generally speaking, the fMRI data analyses adopted in the studies included in the current dissertation consist of the following three major steps (Poldrack, Mumford, & Nichols, 2011). The first major step is preprocessing. In detail, the raw data (i.e., EPI images) usually need to be corrected in the time domain (i.e., slice timing) and the space domain (i.e., head motion correction). After that, the data should be normalized to the standard coordinate space (e.g., Montreal Neurological Institute space, MNI) and spatially, as well temporally (i.e., high-pass filter), smoothed. The second major step is fixed-effect analysis at the individual level via the general linear model (GLM; (Karl Friston et al., 1994)). After this step, we can obtain the parameter estimates for each regressor (i.e., onset time of each condition of interest and other nuisance effects, such as head motion parameters) for each voxel (i.e., the minimum spatial unit in fMRI studies), which then allows us to build contrast images between different conditions. The third step is random-

effect analysis at the group level. Several different statistical models can be adopted given a specific goal, ranging from t-tests to multi-factor ANOVA. After this step, we can obtain the neural correlates of a specific cognitive process with other relevant processes being controlled for. With the rapid development of statistical methods, recent fMRI studies do not limit analyses to the GLM, but rather extend to complex analyses including psycho-physiological interaction (PPI; (K Friston et al., 1997)), dynamic causal modeling (DCM; (Karl Friston, Harrison, & Penny, 2003)), representational similarity analysis (RSA; (Kriegeskorte, Mur, & Bandettini, 2008)), multi-voxel pattern analysis (MVPA; (Norman, Polyn, Detre, & Haxby, 2006)) and so on.

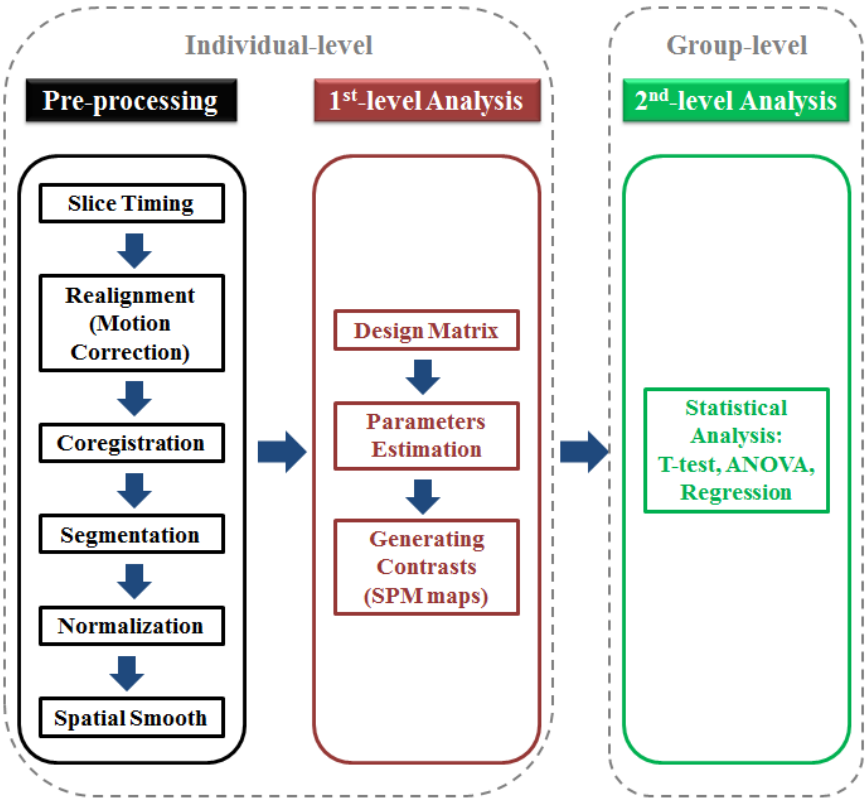


Figure 4. Pipeline for analyzing the fMRI data in a traditional way. Abbreviations: SPM = statistical parametric mapping, ANOVA = analysis of variance.

The main disadvantage for all measurement techniques is that they can only provide correlational results. Thus it is always necessary to be cautious when draw conclusions from these studies (esp. those with GLM analyses), otherwise it is very easy to fall into the reverse inference trap (Poldrack, 2006), namely to infer

the cognitive function based on the neural correlates (e.g., "The participant feels fear because his/her amygdala is activated "). Despite the development of several new statistical methods (e.g., MVPA) to try to make inferences more causal, the direct causal evidence is still not produced, although this could be addressed by manipulation techniques, namely the transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS). TMS can influence the activity of neurons at a specific part of the brain via electromagnetic induction (Hallett, 2000). tDCS affect neuronal firing via a weak but constant electrical current between two electrodes attached to the scalp (Nitsche & Paulus, 2000). For more details on these two techniques, please refer to the corresponding citations.

#### 1.2.2.2 Third-Party Punishment

##### *1.2.2.2.1 The original research*

This first fMRI study on third-party punishment did not surprisingly rely on the standard economic paradigm, and instead was conducted from the perspectives of law, justice and legal decision-making (Buckholtz et al., 2008). While in the scanner, participants were presented with a series of scenarios involving the actions of a protagonist, and were then asked to indicate how much punishment (i.e., the penalty deserved) the protagonist should receive according to a 10-point Likert scale (i.e., 0 = "no punishment", 9 = "extreme punishment"). To identify the neural processes relevant to responsibility and consequence, the investigators categorized the scenarios into three groups, namely those in which the protagonist committed a crime with full responsibility (i.e., Responsibility), those in which the protagonist committed a crime with less responsibility (i.e., Diminished-Responsibility), and those in which the protagonist did not commit a crime (i.e., No Crime). The subjective rating showed a strong modulatory effect of the scenario, with the highest degree of punishment being meted out in those in which the protagonist was completely responsible for the crime. At the neural level, the right dorso-lateral prefrontal cortex (DLPFC), a region also crucial for modulating second-party punishment (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Ruff, Ugazio, & Fehr, 2013; Strang et al., 2014), showed higher activity in response to the Responsibility (vs. Diminished-Responsibility) scenarios, whereas the bilateral temporo-parietal junction (TPJ) displayed higher activity in response to the Diminished-Responsibility (vs. Responsibility) scenarios. Furthermore, the neural activity in the right amygdala was positively associated with punishment in the Responsibility condition, and the relationship remained after controlling for the influence of the Diminished-Responsibility condition. Overall, these findings provided the first neural evidence on third-party punishment and suggested a common neural basis underlying both second- and third-party punishment.

#### *1.2.2.2.2 Follow-up studies: factors modulating third-party punishment*

##### *1.2.2.2.2.1 Evidence from fMRI studies*

With the standard paradigm in which the fairness norm is violated, Strobel *et al* (2011) conducted a more complicated study in order to further investigate 1) the difference between second- and third-party punishment at the neural level and 2) whether the cost rate (i.e., cost 2 MU to reduce 1 MU, weak punishment; cost 1MU to reduce 4 MU, strong punishment) and 3) relevant genes (i.e., COMT genotype, Met/Met, Val/Met, Val/Val) can influence punishment behavior and its neural correlates (Strobel et al., 2011). They showed that third-party punishment was even stronger in an unfair case than second-party punishment, despite there being no difference in other cases. With the region of interest (ROI) approach, they found that the anterior and posterior part of cingulate cortex together with the nucleus accumbens (NAcc), displayed lower activities in third-party (vs. second-party) context. More interestingly, the left DLPFC showed higher activity only during punishment (vs. no punishment) for third parties. Last but not least, they showed that genotype also modulated the third-party punishment-relevant activity in affective regions, including the cingulate cortex, insula and NAcc.

Baumgartner and colleagues (2012) investigated the impact of group membership on neural correlates of third-party punishment (Baumgartner, Götte, Gügler, & Fehr, 2012). To induce in-group bias, participants were randomly assigned to real social groups in the Swiss Army and trained exclusively with their group members for four weeks. On the scanning day, participants, as third parties, were presented with the results of a sequential prison dilemma game (see previous section for details) between two anonymous players (i.e., first and second parties) with only the group membership shown. Then, they indicated the amount by which the first party should be punished. Behaviorally, third parties punished the outgroup perpetrator, who defected especially when the partner cooperated (DC), more harshly than the in-group one. At a neural level, the right orbital frontal cortex (OFC), lateral prefrontal cortex (LPFC) and caudate displayed higher activities for the out-group (vs. in-group) perpetrator during the case of DC. Interestingly, the functional connectivity between the right OFC and LPFC positively correlated with the amount of punishment meted out to the outgroup perpetrator in the case of DC. However, the theory-of-mind network (Schaafsma, Pfaff, Spunt, & Adolphs, 2014; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014), including the dorsal medial prefrontal cortex (DMPFC) and bilateral TPJ, showed higher activation for the in-group (vs. out-group) perpetrator. The functional connectivity between the DMPFC and left TPJ was negatively correlated with the amount of punishment meted out to the in-group perpetrator in the case of DC. These find-

ings suggest differential neural mechanisms interact in the parochial third-party punishment decisions.

Belief in free will (BFW) is also thought to affect judgments in the criminal law. A recent study investigated how BFW influences third-party punishment of the offender in a hypothetical criminal context (Krueger, Hoffman, Walter, & Grafman, 2014). Participants were divided into two groups based on the score of a validated psychological test that measures BFW and scientific determinism (Paulhus & Carey, 2011), namely "Libertarians", who have higher BFW and believe people are morally responsible for their wrongdoings, and "Determinists", who have lower BFW and believe that the material antecedents, instead of the self, should be responsible for an action. Participants were asked to read about the criminal scenarios that, consisted of either high- or low-affective offenses, and to rate how much punishment the protagonist deserved (on a scale ranging from 0 to 100) while in a scanner. Libertarians only punished more than determinists in scenarios with low-affective contents. Consistently, the right TPJ also showed higher activity only for low-affective criminals, in libertarians vs. determinists, during punishment decisions. These findings indicate that the modulatory effect of BFW on third-party punishment was highly context-sensitive.

Another factor that might influence third-party punishment is diffusion of responsibility, which is a hot topic among social psychologists (Latané & Nida, 1981). A recent study addressed this question, in which third parties were either told to decide alone (i.e., Alone condition) or decide simultaneously with four other putative partners (i.e., Group condition), in an unfair situation (Feng et al., 2016). As predicted, third parties felt less responsibility for reducing the selfish proposer's payoff and punished less in the Group condition. Neuroimaging results showed that signals in the bilateral anterior insula (AI) were higher during deciding alone (vs. Group condition) only for the unfair case, which also positively correlated with the difference in punishment amount between the two conditions (i.e., Alone vs. Group condition). However, the medial parts, including dorsal and ventral medial prefrontal cortex (i.e., DLPFC and VMPFC respectively) as well as the precuneus, showed the opposite activation and correlation pattern. Moreover, effective connectivity analyses via Granger causality mapping (Deshpande, LaConte, James, Peltier, & Hu, 2009) showed that the left AI and DMPFC acted as a driver for the other regions mentioned above in the Alone and Group conditions respectively.

To further dissociate the effects of intention and harm (consequence) on third-party punishment at both the behavioral and neural level, a recent fMRI study adopted the criminal-justice judgement paradigm used in Buckholtz *et al* (2008) but with a refined design (Ginther et al., 2016). In particular, the whole decision-making procedure was divided into four cognitive stages (i.e., reading about the

violation event, judging the harm and intention separately with the order counter balanced, and deciding on the punishment), separated by another irrelevant task (i.e., simple mathematical calculation) to intercept the mutual influences between the stages. Behavioral analyses showed that the interaction between intention and harm significantly modulated the punishment intensity. At the neural level, they successfully matched different brain to the corresponding cognitive stages; i.e., the mentalizing network (e.g., TPJ, DMPFC) encoded the judgement of intention, the affective regions (e.g., insula) encoded the judgment of harm, and the information was integrated in the medial prefrontal and posterior cingulate cortices, together with amygdala, which finally informed the right DLPFC to initiate punishment. By adopting advanced connectivity analysis from the perspective of a brain network (e.g., multivariate Granger causality analysis), another recent study, which used a similar paradigm, showed that the DMPFC worked as a hub not only by sending information to the TPJ and VMPFC, but also by connecting with the DLPFC in correlation with the punishment degree (Bellucci et al., 2016). These findings support the previous studies and characterize the possible neural network underlying third-party punishment (Krueger & Hoffman, 2016).

*1.2.2.2.2.2 Evidence from TMS studies*

To our knowledge, two studies have used TMS technique to provide causal evidence for third-party punishment. By adopting the standard paradigm, Brüne *et al* (2012) found that only third parties with repetitive TMS inhibition in the right DLPFC, rather than the left DLPFC or a sham condition, significantly increased the punishment amount toward the unfair proposer (esp. the 8:2 case, with the first and second number being the payoffs for the proposer and recipient) (Brüne et al., 2012).

In a more recent study, Buckholtz and colleagues (2015) also applied repetitive TMS to inhibit the bilateral DLPFC (i.e., active condition, together with a sham condition as the control) and then presented participants with vignettes about criminal scenarios that varied according to harm (e.g., from property theft to murder) and culpability (i.e., responsibility or diminished responsibility of the protagonist). Participants were asked to indicate either how much punishment the offender deserved (i.e., punishment) or how morally responsible the offender was for his actions (i.e., blameworthy) on 10-point Likert scales. Although participants did not show differences in blameworthiness judgments between the active and sham conditions, they differed in terms of punishment: the punishment degree was significantly reduced after the DLPFC was inhibited. The mediation analyses further revealed that inhibition of DLPFC via repetitive TMS influenced the integration of harm and culpability judgments, which then led to altered punishment behavior. An additional fMRI task also showed higher activation in the right DLPFC

during punishment (vs. blameworthiness) assessments, which again corroborated the key role of the DLPFC in driving third-party punishment during judicial decision-making.

*1.2.2.2.2.3 Evidence from patient studies*

Studies based on patients with brain lesions or neurological disorders can always inform and supplement studies of healthy population (i.e., provide stronger evidence for the necessary condition at the neural level initiating a certain behavior); the field of third-party (punishment) decision-making is no exception. In the hypothetical legal justice judgment paradigm (Buckholtz, et al., 2008), a recent study revealed that patients with penetrating traumatic brain injury (pTBI), with specific lesions in regions including the MPFC and DLPFC, punished less than the normal controls (Glass, Moody, Grafman, & Krueger, 2015). Another study first investigated third-party punishment, as well as moral judgment towards norm violations, in a sample of patients with multiple sclerosis (MS) (Patil, Young, Sinay, & Gleichgerrcht, 2016), a demyelinating disease involving deformation of anatomical structures (i.e., inflammation or degeneration in brain or spinal cord) which is associated with several neuropsychological impairments in both non-social (Feinstein, Magalhaes, Richard, Audet, & Moore, 2014; Rocca et al., 2015) and social domains (Charvet, Cleary, Vazquez, Belman, & Krupp, 2014). The results showed that MS patients made harsher punishment as well as judgment than normal controls across different types of violations. Taken together, these findings additionally extend our knowledge of third-party punishment to different clinical populations and contribute to our understanding of the underlying neural mechanism.

## 1.3 Current Studies

Based on the above literature review, we know that third-party altruistic behavior (esp. punishment) has been one of the central topics in behavioral economics, social psychology and decision neuroscience for more than a decade. However, as far as we know, no study has simultaneously considered the neural correlates of both third-party punishment and helping behavior within the same paradigm. This knowledge gap provides the basic motivation for the studies included in the current dissertation. In the rest of this section, I will clarify the motivation behind, and goals of each of the studies in more detail.

### 1.3.1 Motivations and Goals

#### 1.3.1.1 Study 1

The first and foremost goal of Study 1 is quite straightforward; namely, to investigate the common and differential processing during third-party help and punishment at the neural level. Since no previous human neuroscience studies investigated the helping behavior with a third-party paradigm, we need to find clues from studies using paradigms involving direct helping behavior, which has been found to be closely associated with positive emotional experiences (Aknin et al., 2013; Dunn, Aknin, & Norton, 2008, 2014). For example, Dunn and colleagues (2008) found that people's happiness can only be predicted by the money they spend on others (e.g., buying gifts for others or donating to a charity) rather than themselves, after controlling for income. To further test the causal relationship between prosocial spending and happiness, they designed an experiment in which participants were asked to either spending their endowment (i.e., $5 or $10) on themselves or on someone else (or charity). Again, participants felt happier after spending money on others. Consistent with behavioral findings, neuroimaging studies also show that helping others is associated with reward-relevant brain areas, especially the ventral striatum (Haber & Knutson, 2009). In particular, the more participants helped other out-group members (i.e., those in support of the opposing soccer team as the participants) by taking half of the other's pain on oneself, the higher neural activity in the striatum displayed during observation of other out-group members in high pain (Hein, et al., 2010). Striatal activation was also observed in another study, in which participants invested their own endowment in a charity improving the everyday quality of life of African orphans (Genevsky, Västfjäll, Slovic, & Knutson, 2013). Interestingly, the same region is also involved in altruistic punishment (De Quervain et al., 2004; Strobel, et al., 2011). In an earlier PET study, increased activity in the striatum was observed when participants costly punished an anonymous partner who defected in the trust game (De Quervain, et al., 2004). Moreover, Strobel and colleagues (2011) directly compare the neural correlates of second- and third-party punishment in an unfair situation, finding that participants displayed an enhanced response in the striatum during punishment (vs. non-punishment) in both tasks, despite the striatal-relevant reward effect being stronger for the second-party punishment. Based on these findings, it seems that both helping behavior and punishment in the third-party context might elicit positive emotions, which connects with activation in the striatum.

Moreover, we also would like to know why different third parties sometimes show different altruistic choices in the same context. According to a behavioral study by Leliveld *et al* (2012), third-party deciders diverged in their choice prefer-

ences according to empathic concern; i.e., more empathic people preferred to compensate the victim, whereas those with lower empathic concern were in favor of punishing the offender. To be specific, empathic concern is a personality trait, which is defined as other-oriented concern for those who suffer or are in need (Coke, et al., 1978; Toi & Batson, 1982), considered as a reliable precursor for altruistic behavior, especially helping behavior (Batson & Powell, 2003; De Waal, 2008; Eisenberg & Miller, 1987). Usually, empathic concern is measured by the empathic concern subscale of the Interpersonal Reactivity Index (IRI; (Davis, 1983)), as a stable trait. Thus, in Study 1, we will focus on the potential moderating effect of empathic concern on the third party's choice, as well as its neural correlates, in a context whereby helping behavior and punishment are both altruistic options.

#### 1.3.1.2 Studies 2A and 2B

In Studies 2A and 2B, we would like to answer an important and interesting question, whether third-party altruistic decision-making can be influenced by a spray of oxytocin (OXT), a hormone famous for its prosocial effect (Striepens, Kendrick, Maier, & Hurlemann, 2011). Generated in the hypothalamus, the peptide OXT has long been known for its effect in lactation and production in females (Carter, 2014; Gimpl & Fahrenholz, 2001). However, OXT has becomes a central topic in social and affective neuroscience due to a study that revealed its role in enhancing human altruism in a social context (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). In that study, male participants were asked to play a trust game, in the role of "investor", with an anonymous partner as a "trustee". The task for the "investor" was to decide how much he would like to bequeath from his own endowment to the "trustee" (i.e., this amount were tripled by the experimenter), who could transfer the money back to benefit both of them. Surprisingly, participants that received an intranasal spray of OXT significantly increased their investment to the unknown "trustee" compared with the placebo (PLC) condition (but see (Nave, Camerer, & McCullough, 2015) for a different finding). To rule out the alternative explanation that OXT just influences the risk aversion in general, they also ran a control experiment in which participants made similar decisions except that the final payback was randomly determined by the computer, which failed to show the same results. Inspired by this study, a series of researches investigated whether intranasal OXT could influence other aspects of human altruism, including empathy (Hurlemann et al., 2010), generosity (Zak, Stanton, & Ahmadi, 2007) and cooperation (De Dreu et al., 2010; Rilling et al., 2012) in different paradigms. However, no study, to our knowledge, has examined

whether and how OXT affects third-party altruistic decisions in healthy males, which serves as the motivation for Studies 2A and 2B[4].

Moreover, Study 2A had another crucial goal of investigating how OXT exerts an impact on neural processing during altruistic decision-making in the third-party context. Previous neuroimaging studies have already shown that reward-relevant brain regions, mainly the ventral striatum and nucleus accumbens (NAcc), might be involved during costly help (Genevsky, et al., 2013) and punishment decisions (De Quervain, et al., 2004) made in social contexts. Apart from that, another crucial cognitive prerequisite for making third-party decisions is theory-of-mind (ToM) or mentalizing, defined as the ability to understand others' specific (affective) states, beliefs, and intentions (Baron-Cohen, Leslie, & Frith, 1985; De Waal, 2008), which is strongly connected with regions including the temporo-parietal junction (TPJ) and medial prefrontal cortex (MPFC) in terms of multiple fMRI evidences (Frith & Frith, 2006; Schaafsma, et al., 2014; Schurz, et al., 2014). Importantly, either the reward neural circuitry or mentalizing ability could be modulated by OXT. For instance, a recent fMRI study detected stronger neural responses in regions like the striatum and NAcc when healthy males viewed their female partners' faces with OXT treatment compared with a PLC condition (Scheele et al., 2013). Concerning mentalizing ability, one behavioral study documented that male participants' performance in a ToM task was enhanced by intranasal OXT (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007), which also extends to other associated domains such as empathy (Hurlemann, et al., 2010).

Furthermore, a recent study found that OXT selectively enhanced perception of harm to the victim, but not the perceived deservedness of the offender punishment, in a hypothetical criminal judgment task (Krueger et al., 2013). This indicated that OXT might not only affect the decision-making process per se, but also the perceptions accompanying the process. This finding inspired our additional research question, namely regarding how OXT affects perceptions during third-party decisions at the neural level. To disentangle perceptions from the decision-making process, Study 2A included another condition, in which participants were asked to only observe either the offender being punished or the victim being helped by the computer (this condition was the control condition in Study 1; see Study 1 for details). To be consistent, we will focus on the same regions of interest (e.g., NAcc and TPJ) for both research questions.

[4] Study 2A serves as a discovery study; Study 2B can be regarded as a behavioral replication study with a similar, but slightly different, paradigm and design (see corresponding empirical chapter for details).

#### 1.3.1.3 Study 3

As mentioned above, unselfish third parties can intervene in norm violations via either punishing the offender or compensating the victim. Those different altruistic behaviors correspond with two basic types of justice goal, namely retributive and restorative justice (Darley & Pittman, 2003; Gromet & Darley, 2009). In brief, the former goal highlights only punishment, whereas the latter one takes the victim into account. One previous study found that participants, as third-party judges in a simulated context, were less likely to select the way merely addressing punishing the offender but instead preferred the sanction also considering the restorative justice (e.g., helping the victim) after they were asked to think about how the victim was affected by the offender in a given criminal situation (Gromet & Darley, 2009). This suggests a potential cognitive basis underlying the altruistic decision-making process in the third-party context. In particular, two types of other-regarding attention focus (i.e., one concerning the offender's behavior and the other concerning the victim's feelings) highlighting different types of justice goal (i.e., retributive and restorative justice) compete with each other, with the prevailing goal driving the subsequent altruistic decision (i.e., punishment or help). If there is an external cue that highlights one of the foci, to makes it more salient, this will help that specific concern to outweigh the other one and thereby shift the decision in a direction consistent with that concern.

Several recent findings strengthen the validity of the proposed coupling between external attention focus and choice behavior across different tasks. For instance, participants improved their dietary choices (i.e., choosing more healthy food items) when they focused on the healthiness rather than tastiness of the food items (Hare, Malmaud, & Rangel, 2011). Similarly in a social decision task, participants, in the role of recipients, were more likely to reject (accept) an unfair offer from an anonymous proposer while considering the fairness (their own interests) in the ultimatum game (Makwana, Polania, & Hare, 2014). In another study, which used the paradigm of the dictator game, participants, in the role of proposers, were more generous to unknown recipients when considering what was the right thing to do, or their partner's feelings compared with the baseline condition (Hutcherson & Rangel, 2014).

Hence we designed Study 3 to further test the hypothesis of there being a causal relationship between exogenous attention focus and changes in altruistic behaviors in the third-party context. Two key characteristics distinguish this study from a previous study by Gromet & Darley (2009). First and foremost, it adopts an incentivized context with a modified behavioral economic paradigm. Similar to Study 1, participants as third-party deciders are presented with the offer made by the anonymous proposer to the recipient, and they need to decide whether to en-

gage (i.e., punish the offender or help the victim) at self-cost and, if so, how much they will pay. In addition, we experimentally manipulate the other-regarding focus by instructing participants to either consider the offender's behavior, the victim's feelings, or to decide naturally when making their decisions. All decisions made by participants are costly for themselves, and they are also told to believe the consequences of their decisions for the other people involved (i.e., the offender and the victim). Unlike the non-costly choice and hypothetical context used in Gromet & Darley (2009), decisions in the current study have higher ecological validity and can thus reflect real life situations pertaining to morality (FeldmanHall et al., 2012). Moreover, we also employ fMRI to record neural signals during the decision period in the different focus conditions, which could provide insights for understanding the neural mechanisms underlying the attention-induced decision changes of third parties.

Given that the decision-making process (esp. during other-regarding focus conditions) is highly likely to recruit mentalization, we again focus on the TPJ (Schaafsma, et al., 2014; Schurz, et al., 2014). In addition, participants might feel more cognitive conflict when making specific altruistic choice (e.g., help) under certain focus conditions (e.g., focusing on the unfairness of the offender) versus the baseline condition, since the justice goal hinted at by the choice (e.g., help choice hints at restorative justice goal that takes the victim into account more) goes against the goal indicated by the focus (e.g., focusing on unfairness highlights the retributive justice goal that takes the offender into account more). Thus regions related with cognitive control (e.g., the anterior cingulate and inferior frontal cortex) are also assumed to play a role.

#### 1.3.1.4 Study 4

As we mentioned earlier (see literature overview and motivation for Study 1), the altruistic choice preference of third-party deciders could be influenced by individual difference in empathic concern level. Based on this finding, it is natural to ask a follow-up question: what is the underlying cognitive basis driving such an effect?

Traditional behavioral studies cannot answer this question, since they do not capture subtle changes during cognitive or decision-making processes in the temporal domain. However, the eye-tracking technique can offset such a methodological disadvantage and provide a refined measure of eye-movement to describe the general information-searching depth or distribution of attention towards a specific-piece of information (Orquin & Loose, 2013). For example, fixation number is usually adopted to measure the general depth of information searching during decision-making, and is regarded as a better index of decision time (Fiedler & Glöckner, 2012; Fiedler, Glöckner, Nicklisch, & Dickert, 2013; Glöckner &

Herbold, 2011), Another important and widely used index of gaze behaviors is the fixation proportion, which is a reliable measure of how attention is distributed over different pieces of information during cognitive processing (Fiedler & Glöckner, 2015; Orquin & Loose, 2013). For instance, a meta-analysis on the comprehension of visualization showed that compared with novices, experts had higher proportion of fixations on task-relevant areas (Gegenfurtner, Lehtinen, & Säljö, 2011). A recent study showed that participants with higher other-regarding concern, indexed by the ring measure of social value orientation as a personality trait (Liebrand & McClintock, 1988), paid more attention to the other's payoffs when deciding on between different monetary distributions between themselves and another person (Fiedler, et al., 2013). This finding provides an important link between social preference and attention-based information searching. Other gaze measures, such as the distribution of the first- and last-fixation towards the specific information, might also help to reveal changes of attention during the decision-making process (Krajbich, Armel, & Rangel, 2010; Krajbich & Rangel, 2011).

Enlightened by the study by Fiedler *et al* (2013), we argued that the empathy-dependent shift of altruistic choice preference might be induced by the attention distribution towards different aspects of the norm violation situation, which might serve as the potential underlying cognitive mechanism. Given the nature of empathic concern, participants as third-party deciders with higher levels of empathic concern usually pay more attention to the victim (i.e., consider and understand his/her feelings, especially if the victim has been mistreated by others), which is in turn more likely to activate the goal of restorative justice (see Study 3) and finally drives helping rather than punishment behavior. Accordingly, bystanders with lower levels of empathic concern punish more often, as they pay less attention to the victim but more to the offender, which might highlight the goal of retributive justice (see Study 3).

Additionally, it is also possible, with the above proposal, to directly manipulate the attention focus of third-party deciders towards different aspects of the unfair situation, namely either to consider the unfairness of the offender (i.e., offender-focus block, OB) or to think about the feelings of the victim (i.e., victim-focus block, VB), similar to Study 3. The proposal will be further confirmed if we find either a main effect of attention focus and/or the an interaction with the empathic concern level, in terms of both choice behavior and measures of eye-movement during the altruistic decision-making process; such findings could provide a direct link between the external attentional modulator (i.e., attention focus), behavior, and the underlying attention-based decision process.

In summary, Study 4 aims to answer the following three specific research questions, with the help of the eye-tracking technique: first, whether and how empathic concern level (measured by the empathic concern subscale of the IRI; see

also Studies 1-3) influences the eye-movements (in addition to altruistic choice), as a measure of attention, during the decision-making process in a third-party context in which participants decide naturally (i.e., baseline block, BB); second, whether and how external attention focus (i.e., OB or VB) influences both altruistic choice and gaze measures; third, whether and how their interaction (i.e., attention focus × empathic concern) influences these measures. To address these questions, we used a similar design to that of Study 3, except that participants always completed the decision task in BB first and then were informed about the other two conditions to rule out a potential confounding influence of the different focus conditions (see corresponding empirical section for details).

# 2 Study 1: Neural Correlates of Third-Party Altruistic Decision-Making and Its Link with Empathic Concern[5]

## 2.1 Hypotheses

Based on previous findings and our research questions, we have the following hypotheses:

Hypothesis 1 (H1): We expect that participants as third-party deciders will show stronger activation in regions involved in reward process, especially in (both dorsal and ventral) striatum, either when they choose to help the victim or punish the offender, compared with to control trials.

H2a: We expect that empathic concern will influence the participant's choice preference. In particular, participants with higher empathic concern will prefer to help the victim, whereas lower empathic participants will punish the offender more frequently.

H2b: We expect that empathic concern will also affect the neural correlates during the help versus punishment choice. However, we do not have a clear prediction on which region will be affected, as there is no previous study which provides sufficient hints for this hypothesis.

## 2.2 Methods

### 2.2.1 Participants

Thirty-six healthy participants (mean age = 22.72 ± 2.85; 24 females) were recruited from the Online Recruitment System for Economic Experiments (ORSEE) for the present fMRI study. All of them reported no history of neurological or psychiatric disorders. To collect the real decisions used for the fMRI study, we recruited another 84 participants (mean age = 23.58 ± 6.13; 54 females) from the same subject pool for an independent behavioral experiment. The study was approved by the ethics committee of the University of Bonn. All participants signed

[5] The study based on this chapter has been published during the PhD study period of the author with permission. The full citation is here: Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience, 9*, 24.

the written consent form based on the Declaration of Helsinki (BMJ 1991; 302: 1194).

### 2.2.2 Decision Collection and Behavioral Task

Following the "no deception" rule of experimental economics (Glimcher & Fehr, 2013), we first collected real decisions from another group of participants (i.e., behavioral participants) before we ran the fMRI study. In particular, the recruited behavioral participants were invited to the Bonn EconLab and asked to play a Dictator Game, which has two roles, namely an offender (labeled as "proposer") and a victim (labeled as "recipient"). Participants were randomly assigned to one of the two roles and kept that role for 10 rounds of the game. In each round, one offender was paired with an anonymous victim and was endowed with 100 monetary units (MUs; 1 MU = 20 Cents). His or her task was to determine how to split this amount with the victim, by choosing from one of the five options listed as follows: 100/0, 90/10, 80/20, 70/30, 60/40, 50/50 (i.e., payoff for the offender/victim). The presentation of the stimuli and response collection was conducted via Z-tree, the most popular software for behavioural economic experiments (Fischbacher, 2007) .

It is necessary to note the following aspects of the behavioral task. First, participants in the role of the offender were informed before the behavioral task that a certain proportion of their choices, together with the name initials, would be presented to third parties (e.g., fMRI participants) who would complete another part of the study later. They could further affect the final payoff for both offenders and victims denoted in those decisions. Hence, behavioral participants only received a € 4 show-up fee at the end and would receive the choice-dependent payoff (M = € 10.1, SD = € 7.3) a month later (i.e., at the time when the fMRI study was completed). Second, in each round the victim matched with a certain offender who was never the same person, as confirmed by the perfect stranger matching strategy.

From the total of 420 choices made by all offenders (i.e., 142, 82, 57, 33, 43, 63 choices for the split 100/0, 90/10, 80/20, 70/30, 60/40, 50/50 respectively), we finally selected 160 choices in response to the unfair split (i.e., 100/0, 90/10, 80/20, 70/30, 60/40) as stimuli used in the later fMRI study. These choices (i.e., stimuli) were evenly distributed over each of the unfair split in either the decision condition (i.e., 24 choices for each split) or the control condition (i.e., 4 for choices each split).

### 2.2.3 fMRI Paradigm

Two functional scanning runs were divided by a self-paced break. Each run consisted of 80 incentivized trials. In 75% of these trials (i.e., the choice trials), participants were presented with the choice made by the specific offender together with the name initials of both parties (i.e., the offender and the victim) and were asked whether they would like to help the victim by increasing his/her payoff or punish the offender by reducing his/her payoff, each time with their own endowment (i.e., 50 MU per trial; 1 MU = 20 Cents). A bar in magenta was shown below the option once participants made the decision by pressing one of the two buttons with the left or the right finger, which was recorded via response grips (Nordic NeuroLab, Bergen, Norway). Independent of the decision time, the decision phase lasted 4s followed by an inter-stimulus fixation point (1-3s). For trials in which participants failed to respond in 4s, the endowment was deprived for that trial. Then came the next screen in which participants were asked to further indicate the exact amount they would like to transfer by moving a cursor in a scale ranges from 0 to 50 (with the step of 5) in 4 seconds (i.e., the transfer phase). The trial ended with a jittered fixation cross (3-7s). For the remaining 25% trials (i.e., the control trials), which were indicated by a white frame, the procedure was the same, except that participants only needed to observe the decisions and transfers already made by the computer (see Figure 5). To balance the decision, half of these trials were set to the help choice, while others were set to the punishment choice, which was consistent across all five monetary splits. All trials were presented in a pseudo-random order, fixed across participants. Participants saw the stimuli via video goggles (Nordic NeuroLab, Bergen, Norway). The stimuli presentation during the experiment was performed with Presentation 14.9 (Neurobehavioral System, Albana, Canada).

Apart from that, we considered the following details to make the paradigm stand against the potential confounds. First, the cost ratio was set to 1:3, meaning that 1 MU transferred from participants could either reduce 3 MU from the offender or add 3 MU to the victim. This was in line with previous literature (Fehr & Fischbacher, 2004; Leliveld et al., 2012). Second, to avoid demand characteristics, the key words such as “help” and “punish” were never used in either the instructions or the fMRI screen; instead, words with neutral emotion (e.g., “increase” and “subtract”) were adopted respectively. Third, to avoid the association between position and specific option, we counterbalanced the position of the two options (i.e., “increase” and “subtract”) in the decision phase across trials. Fourth, to ensure that all costly altruistic choices were made voluntarily, participants were explicitly told that in the transfer phase they could transfer 0 MU. Fifth, the starting position of the cursor was randomly determined to be located between 0 and

50 in the transfer phase of each trial. Finally, the offender could not lose money (i.e., the minimum payoff was 0).

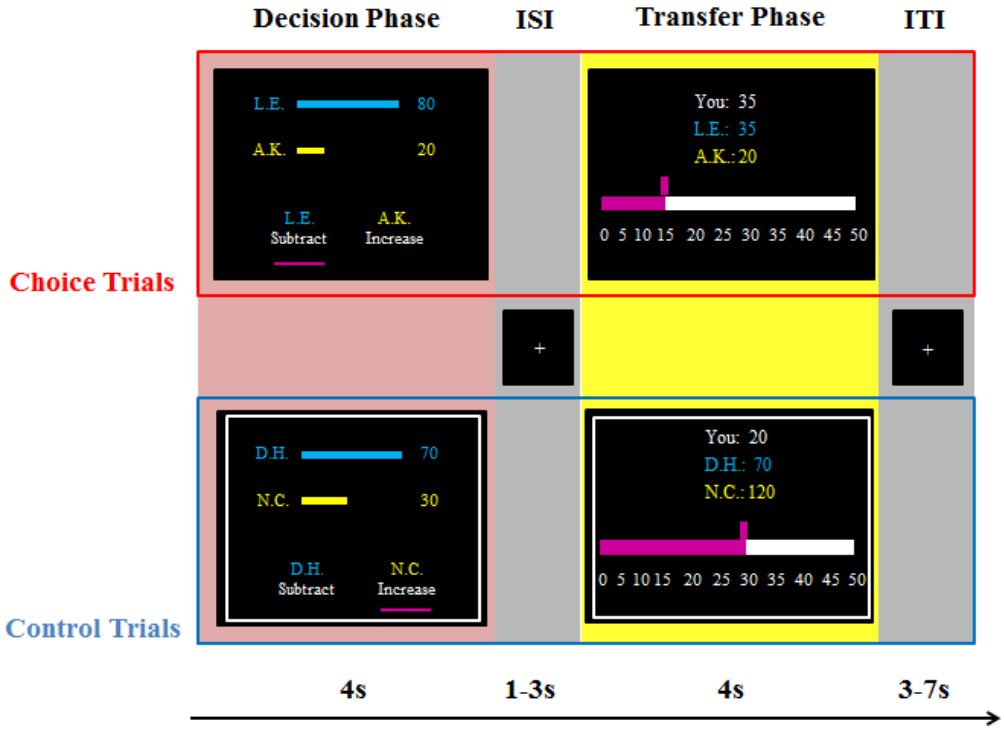


Figure 5. Example of the procedure for the choice trials as well as the control trials. In the example of the choice trial, the participant subtracted 15 MUs from the offender (i.e., L.E.); in the example of the control trial, the participant observed the computer to add 30 MUs to the victim (i.e., N.C.). Abbreviations: MU = monetary unit; ISI = inter-stimulus interval; ITI = inter-trial interval.

### 2.2.4 Procedure

Upon arrival, participants were informed about the behavioral experiment and about the upcoming task in the scanner. They confirmed to have understood the task by passing a short comprehension test directly after the instructions. Before they were sent to the scanner, they were also familiarized with the task using practice trials. A structural scanning was performed following the functional scanning.

After coming out of the scanner, participants filled out a series of online questions including in the empathic concern subscale of Interpersonal Reactivity Index (IRI) which measures empathic concern as a stable personality trait (see Table 1 for items of empathic concern subscale; see Appendix Table 1 for all items of IRI

scale). They also rated the perceived fairness of the five unfair monetary splits used in the study, together with the 50/50 fair split as control, on a 8-point Likert scale (i.e., 1 = very fair, 8 = very unfair). At the end of the study, participants were paid 10 € for their attendance. Additionally, one out of the 160 trials were randomly selected to pay the fMRI participants as well (M = € 7.0, SD = € 2.5). This payoff further determined the choice-dependent extra payoff for corresponding offenders and victims.

Table 1. Items of the empathic concern subscale of IRI

| Content | Answer Scale |
|---|---|
| I often have tender, concerned feelings for people less fortunate than me. | 0---1---2---3---4 |
| Sometimes I don't feel very sorry for other people when they are having problems.* | 0---1---2---3---4 |
| When I see someone being taken advantage of, I feel kind of protective towards them. | 0---1---2---3---4 |
| Other people's misfortunes do not usually disturb me a great deal.* | 0---1---2---3---4 |
| When I see someone being treated unfairly, I sometimes don't feel very much pity for them.* | 0---1---2---3---4 |
| I am often quite touched by things that I see happen. | 0---1---2---3---4 |
| I would describe myself as a pretty soft-hearted person. | 0---1---2---3---4 |

Note: 0 refers to "does not describe me well", 4 refers to "describes me very well". * refers to reverse-scored items. IRI refers to interpersonal reactivity index.

### 2.2.5 Data Collection

All imaging data was collected via the 3-Tesla Siemens Trio platform at the Life & Brain Imaging Center, located at the University Hospital Bonn. For images of the fMRI task, 37 slices of the axial plane (in-plane resolution = 2 ×2 $mm^2$, matrix = 96×96, slice thickness = 3 mm, FOV = 192×192 $mm^2$) covering the whole brain were acquired via a T2*-weighted echo planar imaging (EPI) sequences with blood-oxygenation-level dependent (BOLD) contrast (TR = 2500 ms, TE = 30 ms, flip angle = 90°). We also obtained a high-resolution anatomical scanning with 3D MPRage sequences for anatomical co-registration and normalization (TR = 1660 ms, TE = 2.75 ms, flip angle = 9°, matrix = 320 × 320, slice thickness = 0.8 mm, FOV = 256 × 256 $mm^2$).

### 2.2.6 Data Quality Check and Analyses

Given the goal of the current study, we excluded 10 participants as they failed to show enough decisions (i.e., with a lenient criterion: at least 5 decisions per run) to help (n = 1), punish (n = 7), or both altruistic choices (n = 2) in both of the two functional runs, since few trials might lead to unstable estimation for the target effect according to the low signal-to-ratio feature for fMRI data analyses. Besides, we also excluded one participant due to the incomplete data. Henceforth, data from the remaining 25 participants was adopted for further analyses.

#### 2.2.6.1 Behavioral Data

For the behavioral data, the mean proportion of choice, the mean decision time as well as the mean transfer amount were calculated for help and punishment choice respectively for each participant. Statistical analyses were performed via SPSS 22 (SPSS Inc.). Paired t-test, repeated-measure analysis of variance (ANOVA) as well as Pearson correlation were used to test hypotheses and to perform exploratory analyses. All reported p values were two-tailed and $p < 0.05$ was considered significant.

#### 2.2.6.2 fMRI Data

##### *2.2.6.2.1 Preprocessing*

For fMRI data, SPM8 (Wellcome Trust Department of Cognitive Neurology, London, UK) was used for analysis. For raw EPI images within each run of each participant, we started with the preprocessing, including the following steps. To begin with, we discarded the first three volumes to make sure of a stable BOLD signal in the remaining images. Next, EPI images were realigned to the first volume in order to correct for head motions (< 2.5 mm). After the head motion correction, the images were corrected at the temporal domain via slice timing, which aimed to ensure all slices within one volume were adjusted to the same time point. Then, the mean EPI image within this run was computed and co-registered to the anatomical image, which was followed by the segmentation. With the parameters of the normalization to the Montreal Neurological Institute (MNI) space generated after the segmentation, all EPI images were projected onto MNI space with a $2 \times 2 \times 2$ $mm^3$ resolution. In the next step we applied the spatial smoothing on all images with an 8-mm FWHM (full width half maximum) isotropic Gaussian kernel. To further remove low-frequency drifts, we also performed a high-pass temporal filtering with a cut-off of 128 s.

##### *2.2.6.2.2 General linear model (GLM) analyses*

The GLM mass-univariate regression approach was adopted for the individual-level fixed-effect analyses. This GLM focused on the decision-phase and included

four regressors of interest within in each run, given the decisions participants or the computer made, namely onsets of stimuli presentation during help choice, punish choice, as well as the corresponding control trials (i.e., help_control, and punish_control). All onsets of other events were pooled to one regressor (i.e., other), including onsets of stimuli presentation during keep choice (i.e., participants in these trials kept all the endowment) or no response, and onset of all transfer phase. For the choice less than 5 trials in some runs, onsets of stimuli presentation during that choice were also treated as other regressor and not modeled as an independent event. To control for motion, we additionally included the six estimated head movement parameters in the GLM design matrix. Individual contrasts between regressors of interested pooling the effect across two runs were built, namely the contrast help vs. help_control, punish vs. punish_control, as well as help vs. punish.

For the group-level random-effect analyses, we first performed a one-sample t-test on the contrast help vs. help_control as well as punish vs. punish_control respectively to check for decision-relevant activity. In order to know the common activation pattern in association with the help and punishment choices, we ran a conjunction analysis between the two contrasts mentioned above within the flexible factorial model. Apart from that, we also applied a regression analysis to test whether empathic concern could modulate the altruistic-relevant neural activation. To this end, the contrast help vs. punish was used as the dependent variable with the scores measured by empathic concern subscale of IRI as the predictor in the regression.

*2.2.6.2.3 Explorative functional connectivity analysis*

To explore the neural network involved in help and punishment decisions, we performed a standard psycho-physiological interaction (PPI) analysis (K Friston, et al., 1997; Gitelman, Penny, Ashburner, & Friston, 2003). In principle, this analysis aims to address the question of how a given target region changes its functional connectivity with other part of the brain, measured by the correlation between the time series of BOLD signals in both regions, dependent on different experimental conditions. It has become one of the most popular approaches within the field of cognitive neuroscience to test the context-dependent functional network in fMRI studies (O'Reilly, Woolrich, Behrens, Smith, & Johansen-Berg, 2012).

To do this, we first defined the source regions, namely the striatum, in terms of the conjunction activation of the contrast help vs. help_control and punish vs. punish_control at the group level. To refine and make sure that the joint activation was located in the anatomical region of striatum, we defined two spheres (i.e., the left and the right side) centered at the peak voxel of the joint activation, using the radium of 8 mm and then intersecting these spheres with the bilateral striatum

mask of the AAL template. Taking the individual difference of neural activity into account, we drew the volume of interest (VOI; a 6mm sphere) from the individual contrast help vs. help_control and punish vs. punish_control, respectively for each participant, within the two group-level source masks. To build the interaction term, we then extracted the time series of each VOI (i.e., the physiological term), deconvolved and mulplied them with the psychological term, namely the onsets vectors of either help vs. help_control (i.e., weight: 1, -1) or punish vs. punish_control (i.e., weight: 1, -1), according to the recommended procedure by Gitelman *et al.* (2003). Then we ran four GLM regression analyses separate for help and punishment choice with either the left or the right striatum as the seed VOI at the individual level, each including three regressors of interest (i.e., the PPI term, the physiological term, the psychological term) within each run controlling for head motion. Next, the individual contrast image pooling the effect across two runs was built while focusing on the PPI term vs. the implicit baseline. These images were then forwarded to the group-level random-effect analyses one-sample t-tests, which identified the other regions displaying increased functional connectivity with seed VOI (i.e., either the left or the right striatum) during either help or punishment choices.

For whole-brain analyses mentioned above, we adopted the uncorrected voxel-level $p < 0.001$ with the extent threshold at $k = 50$. For display reason, we also extracted and plotted the parameter estimates (i.e., contrast values) together with time course of percent signal change of the peak voxel in above analyses by MarsBar (http://marsbar.sourceforge.net).

## 2.3 Results

### 2.3.1 Behavioral Results

We first investigated whether empathic concern (mean ± S.D. = 17.24 ± 3.70; Range: from 7 to 23) is correlated to the proportion of help or punishment choice via the Pearson correlation analysis. Consistent with our initial hypothesis (see H2a), we found that participants with higher empathic concern helped the victim more often ($r = 0.441$, $p = 0.027$), whereas those less empathic third parties more preferred to punish the offender ($r = -0.461$, $p = 0.02$, Figure 6A). A similar exploratory analyses further showed that empathic concern could also modulated the decision process (i.e., the mean difference of decision time between help and punishment choice), showing that participants helped faster but prolonged punishment choice with increasing empathic concern level ($r = -0.406$, $p = 0.044$, Figure 6B).

Moreover, we also examined whether participants differed in the following behavior measures between help and punishment choice, namely choice proportion (%), decision time (ms) and transfer amount (MU), by using paired-samples T-test (see Table 2 for summary of descriptive statistics). Although we detected the significant difference in neither proportion ($t(24) = 0.632$, $p = 0.533$) nor decision time ($t(24) = -0.326$, $p = 0.747$) between trials with help and punishment choices, we observed that participants invested more MUs to punish the offender than to compensate the victim ($t(24) = 3.266$, $p = 0.003$).

In addition we checked whether objective inequality affected participants' subjective rating on unfairness. A one-way repeated measure ANOVA revealed a main effect of inequity on unfairness rating ($F(5, 120) = 225.967$, $p < 0.001$, partial $\eta^2 = 0.904$), which was further confirmed by the post-hoc analyses that participants perceived stronger unfairness with increasing inequality of the monetary split ($ps < 0.05$, Bonferroni corrected; see Table 3 for summary of descriptive statistics).

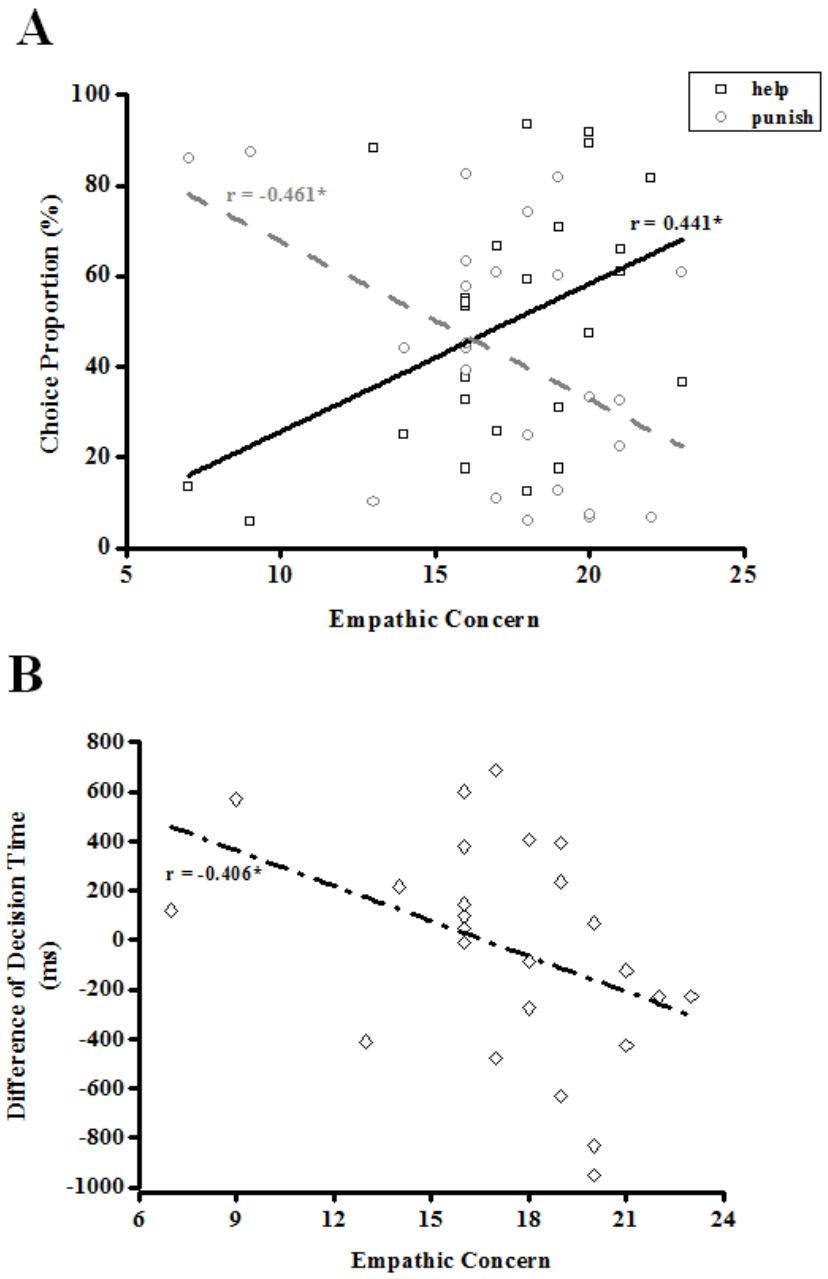


Figure 6. (A) Correlation between empathic concern level and proportion of either help or punishment choice; (B) Correlation between empathic concern level and the difference in decision time between help and punishment choice. Significance level: *p < 0.05.

Table 2. Descriptive summary of behavioral measures during the fMRI task

| | help<br>Mean (S.D.) | punishment<br>Mean (S.D.) |
|---|---|---|
| Choice Proportion (%) | 49.30 | 42.40 |
| | (27.28) | (27.90) |
| Decision Time (ms) | 1583.15 | 1611.45 |
| | (431.63) | (402.22) |
| Transfer Amount (MU) | 11.07 | 16.15 |
| | (5.07) | (6.86) |

Note: S.D. refers to standard deviation; MU refers to monetary unit.

Table 3. Descriptive summary of post-scanning rating

| | 50/50 | 60/40 | 70/30 | 80/20 | 90/10 | 100/0 |
|---|---|---|---|---|---|---|
| Unfairness Rating | 1.48 | 3.52 | 5.24 | 6.24 | 7.32 | 8.00 |
| | (1.12) | (1.30) | (0.93) | (0.93) | (0.48) | (0.00) |

Note: Values refer to the mean; standard deviations are provided in parentheses; unfairness ratings range from 0 (not at all) to 8 (very much).

## 2.3.2 Imaging Findings

### 2.3.2.1 Neural correlates of third-party help and punishment

To test H1, we compared the neural correlates during decisions to help (vs. help_control) and to punish (vs. punish_control) respectively. Consistent with our initial prediction, we found an increased response in bilateral striatum in association with either help or punishment choices. Additionally, both contrasts revealed activation in other regions including inferior/superior parietal lobule (BA 39/40) and mid-cingulate cortex extending to supplementary motor area (BA 4/6). To further confirm the common neural substrates underlying both altruistic choices in such context, we ran a conjunction analysis for the contrast of help vs. help_control and punish vs. punish_control, which showed again the involvement of the bilateral striatum (see Table 4 and Figure 7)[6]. We consequently asked our-

[6] To rule out the effect of button press differed between the choice and the control trials, we ran another GLM with the same regressors as the main GLM except that we modeled the onset of the

selves, whether there was difference in neural correlates between two altruistic choices. We henceforth compared the contrast of help vs. punish via a one-sample t-test, finding no significance under the pre-defined threshold.

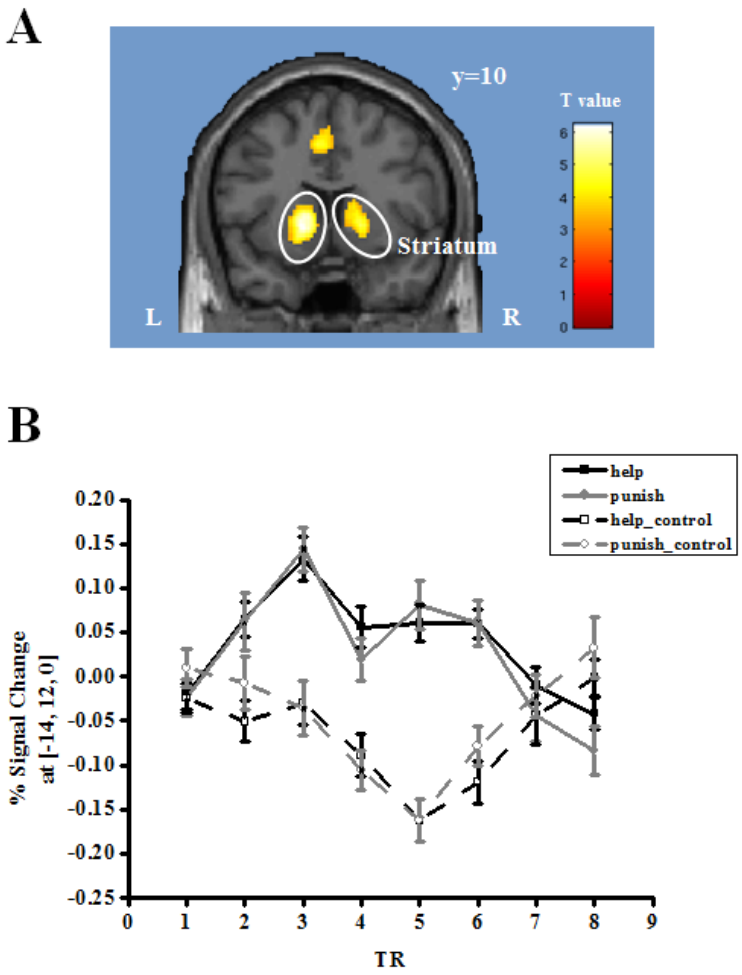


Figure 7. (A) Conjunction activations of both contrasts of help (vs. help_control) and punishment (vs. punish_control); (B) Timecourse of percent (%) signal change in the local peak voxel of left striatum in all conditions. Display threshold: $p < 0.001$ at voxel-level, uncorrected; Error bars: SEM. Abbreviations: L = left, R = right.

button press separately. The main findings in striatum during altruistic decisions remained significant (see Table 5).

Table 4. Neural activations in response to third-party altruistic decisions (vs. control conditions)

| Brain Region | Hemi-sphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***help > help_control*** | | | | | | | |
| MFG | L | 93 | -46 | 36 | 22 | 46 | 4.16 |
| MFG | R | 147 | 40 | 48 | 8 | 46 | 5.20 |
| ACC/SMA | B | 937 | -4 | 12 | 42 | 6/24/32 | 6.92* |
| Insula/STG | R | 254 | 46 | -18 | 10 | 13/22 | 6.90* |
| PCG/PoCG/IPL/SPL | L | 2877 | -46 | -2 | 58 | 1/2/3/6/ 7/39/40 | 8.93* |
| PCG/PoCG/IPL/SPL | R | 2301 | 26 | -64 | 58 | 1/2/3/4/ 7/39/40 | 7.31* |
| IOG/MOG | L | 1438 | -38 | -78 | 0 | 17/18/19 | 7.56* |
| IOG/MOG | R | 1760 | 34 | -84 | 2 | 17/18/19 | 8.85* |
| Caudate/Putamen | L | 574 | -14 | 14 | 4 | | 8.35* |
| Caudate/Putamen | R | 264 | 16 | 14 | -2 | | 7.71* |
| ***punish > punish_control*** | | | | | | | |
| SMA/MCC/ACC | B | 1167 | 10 | -8 | 50 | 6/24/31/32 | 6.71* |
| PCG/PoCG/IPL/SPL | L | 1870 | -40 | -38 | 44 | 2/3/4/ 7/39/40 | 7.22* |
| PCG/PoCG | R | 1047 | 48 | -18 | 50 | 2/3/4 | 6.74* |
| STG/ Insula | L | 206 | -50 | -34 | 8 | 13/41 | 4.68* |
| IOG/MOG/MTG | L | 573 | -44 | -72 | 6 | 17/18/ 19/37 | 5.55* |
| IOG/MOG/MTG | R | 629 | 46 | -66 | 2 | 17/18/ 19/37 | 7.08* |
| Caudate/Putamen | L | 599 | -16 | 10 | -2 | | 7.48* |
| Caudate/Putamen | R | 255 | 24 | -12 | 2 | | 7.26* |
| ***Conjunction*** | | | | | | | |
| Caudate/Putamen | L | 382 | -16 | 12 | 0 | | 6.26* |
| Caudate/Putamen | R | 250 | 16 | -20 | 6 | | 6.08* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| PCG/PoCG/IPL/SPL | L | 1493 | -38 | -38 | 38 | 2/3/4/6/40 | 5.13* |
| PCG/PoCG | R | 922 | 40 | -12 | 58 | 2/3/4/6 | 5.69* |
| MTG/MOG | L | 125 | -44 | -70 | 6 | 19/37 | 5.05 |
| ITG/ MTG | R | 236 | 46 | -66 | 4 | 19/37 | 5.57* |
| SMA/MCC/ACC | M | 583 | -4 | 14 | 46 | 6/9/24/32 | 4.97* |
| STG/Insula/PoCG | L | 324 | -38 | -34 | 16 | 13/41/42 | 4.16* |
| STG/Insula | R | 232 | 50 | -14 | 10 | 13/22/41 | 4.56* |
| IOG/MOG | L | 96 | -26 | -92 | -4 | 18 | 4.10 |
| IOG/MOG | R | 130 | 30 | -86 | -2 | 18/19 | 4.28 |

Note: Threshold is set to $p < 0.001$, $k = 50$, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): L = left, R = right, B = bilateral; ACC = Anterior Cingulate Gyrus, IFG = Inferior Frontal Gyrus, IOG = Inferior Occipital Gyrus, IPL=Inferior Parietal Lobule, ITG = Inferior Temporal Gyrus, MCC = Mid-cingulate Cortex, MFG=Middle Frontal Gyrus, MOG = Middle Occipital Gyrus, MTG=Middle Temporal Gyrus, PCG=Precentral Gyrus, PoCG=Postcentral Gyrus, SMA=Supplementary Motor Area, SMG=Supramarginal Gyrus, SPL = Superior Parietal Lobule, STG = Superior Temporal Gyrus.

Table 5. Neural activations in response to third-party altruistic decisions (vs. control conditions) controlling for button pressing

| Brain Region | Hemi-sphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***help > help_control*** | | | | | | | |
| PoCG/PCG | L | 355 | -46 | -28 | 64 | 1/3/4/6 | 6.03* |
| IPL/PoCG/PCG | R | 602 | 58 | -30 | 54 | 1/2/3/4/6/40 | 5.56* |
| Caudate/Putamen/Pallidum | L | 274 | -12 | 14 | 2 | | 5.87* |
| Caudate/Putamen/Pallidum | R | 225 | 16 | 12 | 0 | | 6.98* |
| Thalumus | L | 158 | 16 | -16 | 4 | | 5.15 |
| Thalumus | R | 162 | -18 | -12 | 6 | | 4.73 |
| ***punish > punish_control*** | | | | | | | |
| SMA | R | 90 | 16 | -8 | 52 | 24 | 5.17 |
| PoCG/PCG | L | 240 | -36 | -22 | 48 | 3/4 | 4.79* |
| PoCG/PCG | R | 353 | 38 | -12 | 54 | 3/4/6 | 5.83* |
| Cau- | B | 1766 | 6 | -28 | -10 | | 7.37* |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| date/Putamen/Pallidum/ | | | | | | | |
| Thalamus/Brainstem | | | | | | | |
| ***Conjunction*** | | | | | | | |
| Caudate/Putamen/Pallidum | L | 205 | -14 | 10 | 0 | | 4.81 |
| Caudate | R | 126 | 18 | 6 | 0 | | 4.42 |
| /Pallidum/Putamen | | | | | | | |
| PoCG/PCG | L | 195 | -50 | -20 | 54 | 3/4 | 3.99 |
| PoCG/PCG | R | 196 | 38 | -14 | 60 | 3/4/6 | 4.50 |
| Thalamus/Brainstem | B | 657 | -2 | -24 | -12 | | 5.27* |
| ***Button Press*** | | | | | | | |
| SFG/MFG | R | 185 | 32 | -4 | 66 | 6 | 5.14 |
| PCG | R | 72 | 36 | -10 | 42 | 6 | 4.25 |
| IOG/MOG/SOG/ | B | 66552 | 22 | -78 | -14 | 3/4/5/6/ | 15.64* |
| FG/Precuneus/Cuneus/ | | | | | | 7/8/9/10/ | |
| PoCG/PCG/ | | | | | | 13/18/19/ | |
| IPL/SPL/ITG/STG/ | | | | | | 20/22/23/ | |
| SMA/ACC/PCC/ | | | | | | 24/30/32/ | |
| MFG/Insula/PHG/ | | | | | | 37/38/39/ | |
| Caudate/Putamen/ | | | | | | 40/42/43/ | |
| Cerebellum | | | | | | 44/45/ | |
| | | | | | | 46/47 | |

Note: In this GLM, we added the onset of button presses to control for motor related activity, with the other regressors being the same as the main GLM. Threshold is set to $p < 0.001$, $k = 50$, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): L = left, R = right, B = bilateral; ACC = Anterior Cingulate Gyrus, FG = Fusiform Gyrus, IFG = Inferior Frontal Gyrus, IOG = Inferior Occipital Gyrus, IPL = Inferior Parietal Lobule, ITG = Inferior Temporal Gyrus, MFG = Middle Frontal Gyrus, MOG = Middle Occipital Gyrus, MTG = Middle Temporal Gyrus, PCC = Posterior Cingulate Cortex, PCG=Precentral Gyrus, PoCG=Postcentral Gyrus, PHG = Parahippocampal Gyrus, SFG = Superior Frontal Gyrus, SMA = Supplementary Motor Area, SOG = Superior Occipital Gyrus, SPL = Superior Parietal Lobule, STG = Superior Temporal Gyrus.

#### 2.3.2.2 Empathic Concern Modulates Neural Correlates During Third-Party Altruistic Choices

The regression analyses on the contrast help vs. punish with the empathic concern scores as the predictor showed that the activity in a frontal-parietal network, mainly including the left part of the lateral prefrontal cortex (LPFC, BA 9) as well as inferior parietal lobule (IPL, BA 7/40; see Table 6 and Figure 8). These findings supported H2b and also explained to some degrees the lack of the main effect mentioned above.

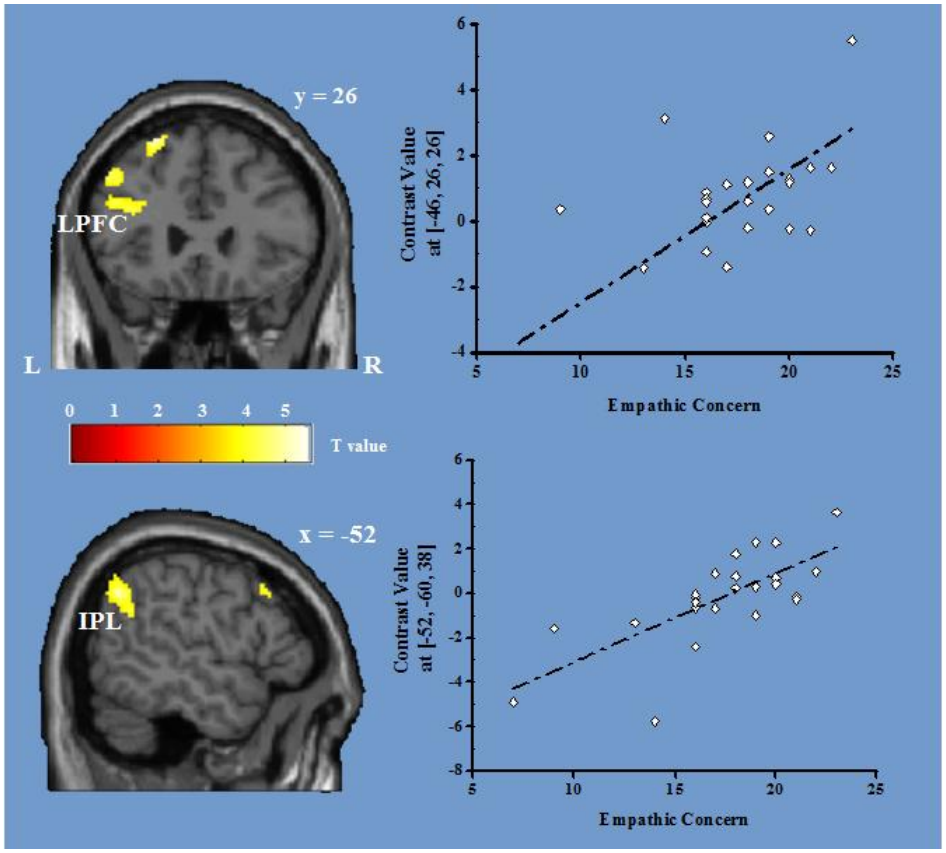


Figure 8. Regions reflecting the correlation between the contrast of help vs. punishment and empathic concern level. Scatter plots showed the relationship between contrast values of peak voxel and empathic concern, only with the goal of illustration. Display threshold: $p < 0.001$ at voxel-level, uncorrected. Abbreviations: L = left, R = right, LPFC = lateral prefrontal cortex; IPL = inferior parietal lobule.

Table 6. Correlation between brain activation of the contrast help vs. punishment and empathic concern scores

| Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| IFG/MFG | L | 84 | -34 | 24 | 22 | 45/46 | 4.62 |
| MFG | L | 150 | -46 | 20 | 40 | 8/9 | 4.79 |
| MFG/FP | L | 79 | -38 | 54 | 6 | 10 | 4.34 |
| SFG/MFG | L | 312 | -24 | 26 | 60 | 6/8/9 | 5.43* |
| SFG | R | 112 | 20 | 66 | 10 | 10 | 5.60 |
| IPL/SPL/AG/SMG | L | 620 | -32 | -74 | 50 | 7/39/40 | 5.47* |
| MTG | R | 73 | 66 | -2 | -24 | 21 | 5.54 |
| ITG | R | 58 | 60 | -20 | -18 | 20/21 | 4.36 |

Note: Threshold is set to $p < 0.001$, $k = 50$, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): L = left, R = right, B = bilateral; AG = Angular Gyrus, FP = Frontal Pole, IFG = Inferior Frontal Gyrus, IPL = Inferior Parietal Lobule, ITG = Inferior Temporal Gyrus, MFG=Middle Frontal Gyrus, MTG=Middle Temporal Gyrus, SFG = Superior Frontal Gyrus, SMG=Supramarginal Gyrus, SPL = Superior Parietal Lobule.

#### 2.3.2.3 PPI Results

The explorative functional connectivity analyses via PPI showed that the bilateral striatum, as our seed regions, increased the connection with the right LPFC (BA 45/46) during help decision (vs. help_control) (see Figure 9), whereas they enhanced the connectivity with left LPFC (BA 44/45) during punishment choices (vs. punish_control) (see Figure 10; see Table 7 for other PPI results).

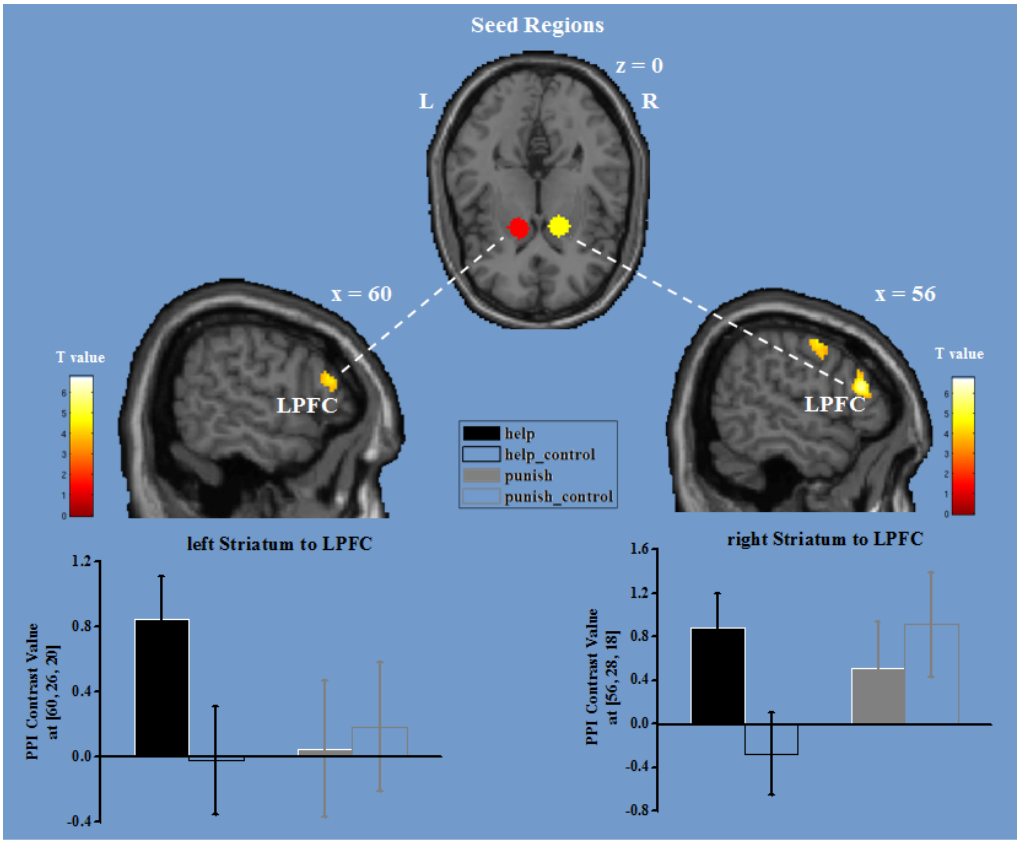


Figure 9. Regions reflecting enhanced functional connectivity with bilateral striatum during help (vs. help_control). Bar plots showed the contrast value of PPI in the peak voxel of LPFC with bilateral striatum in all conditions (vs. implicit baseline respectively), only with the goal of illustration. Display threshold: $p < 0.001$ at voxel-level, uncorrected; Error bars: SEM. Abbreviations: PPI = psycho-physiological interaction, L = left, R = right, LPFC = lateral prefrontal cortex.

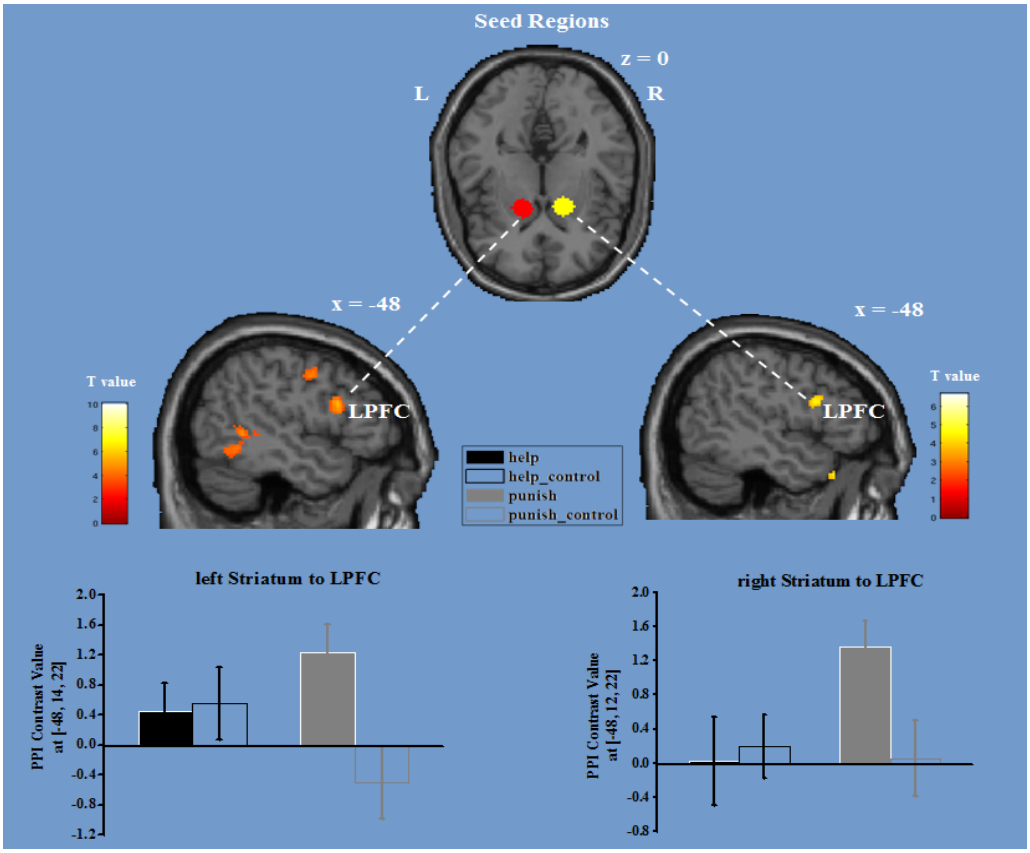


Figure 10. Regions reflecting enhanced functional connectivity with bilateral striatum during punishment (vs. punish_control). Bar plots showed the contrast value of PPI in the peak voxel of LPFC with bilateral striatum in all conditions (vs. implicit baseline respectively), only with the goal of illustration. Display threshold: $p < 0.001$ at voxel-level, uncorrected; Error bars: SEM. Abbreviations: PPI = psycho-physiological interaction, L = left, R = right, LPFC = lateral pre-frontal cortex.

Table 7. Regions reflecting enhanced functional connectivity with striatum during third-party altruistic decisions (vs. control conditions)

| Seed Region | Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|---|
| | | | | x | y | z | | |
| Left Striatum | ***help > help_control*** | | | | | | | |
| | MFG/IFG | R | 50 | 60 | 26 | 20 | 45/46 | 4.47 |
| | SMA | B | 78 | -8 | 4 | 64 | 6 | 4.71 |
| | MCC/PCC | B | 319 | -8 | -26 | 46 | 24/31 | 4.73* |
| | STG/TP | L | 104 | -36 | 12 | -22 | 38 | 4.92 |
| | Precuneus/Cuneus | R | 110 | 24 | -82 | 26 | 7/18/31 | 4.06 |
| | LG/FG/Cuneus | B | 3300 | 18 | -72 | -4 | 17/18/19/37 | 6.79* |
| | Thalamus | B | 81 | -4 | -4 | 10 | | 5.27 |
| | ***punish > punish_control*** | | | | | | | |
| | IFG | L | 169 | -48 | 14 | 22 | 44/45 | 5.50* |
| | SFG | L | 76 | -22 | -4 | 52 | 6 | 4.45 |
| | PCG | L | 528 | -44 | -4 | 44 | 6 | 6.44* |
| | PCG | R | 270 | 40 | -4 | 40 | 6 | 5.80* |
| | STG/MTG | L | 323 | -58 | -34 | 2 | 21/22 | 5.33* |
| | MTG/TP | R | 105 | 58 | 10 | -18 | 21/38 | 6.63 |
| | LG/FG/Cuneus/Precuneus/PHG | B | 12443 | 20 | -68 | -2 | 7/17/18/19/31 | 10.10* |
| | Putamen/Amygdala | L | 101 | -18 | 8 | -6 | | 4.54 |
| | Putamen | R | 58 | 22 | 10 | -4 | | 4.33 |
| Right Striatum | ***help > help_control*** | | | | | | | |
| | MFG/IFG | R | 172 | 56 | 28 | 18 | 45/46 | 5.88 |
| | SFG/MFG | R | 71 | 28 | 40 | 42 | 8/9 | 4.77 |
| | SMA | L | 60 | -12 | 4 | 62 | 6 | 4.00 |
| | PCG | R | 196 | 54 | -2 | 52 | 6 | 4.73* |
| | STG/MTG | L | 234 | -58 | -32 | 0 | 22 | 4.76* |
| | MTG/TP | R | 76 | 48 | 4 | -18 | 21/38 | 4.94 |
| | ITG/FG | L | 277 | -36 | -36 | -20 | 20/36 | 5.18* |
| | PHG/FG | R | 55 | 32 | -28 | -24 | 36 | 4.07 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Cuneus | R | 50 | 18 | -80 | 30 | 7/31 | 4.13 |
| LG/FG | B | 5026 | -6 | -74 | 2 | 17/18/19/37 | 6.80* |
| Puta-men/Insula/PHG | L | 352 | -28 | -20 | 0 | | 6.23* |
| Caudate/Putamen | R | 57 | 22 | 14 | 8 | | 4.70 |
| ***punish > punish_control*** | | | | | | | |
| MeOFG/ACC | B | 315 | 0 | 42 | -6 | 10/11/32 | 6.37* |
| PCG | L | 278 | -40 | -6 | 38 | 6 | 4.73* |
| PCG | R | 135 | 34 | -2 | 34 | 6 | 6.07 |
| MTG | R | 55 | 40 | -78 | 18 | 19 | 4.54 |
| TP | L | 55 | -44 | 24 | -32 | 38 | 5.67 |
| TP | R | 86 | 58 | 10 | -16 | 38 | 5.65 |
| SPL | L | 163 | -20 | -72 | 56 | 7 | 4.58 |
| Cuneus | R | 124 | 16 | -88 | 22 | 18 | 4.35 |
| LG/ FG/Precuneus/ Cuneus/PHG | B | 4781 | -18 | -84 | 10 | 17/18/19/23/30/31 | 6.69* |
| Putamen | R | 59 | 26 | 10 | -4 | | 4.64 |

Note: Threshold is set to $p < 0.001$, k=50, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): L = left, R = right, B = bilateral; ACC = Anterior Cingulate Gyrus, AG = Angular Gyrus, FG = Fusiform Gyrus, IFG = Inferior Frontal Gyrus, IOG = Inferior Occipital Gyrus, IPL=Inferior Parietal Lobule, ITG = Inferior Temporal Gyrus, LG = Lingual Gyrus, MCC = Mid-cingulate Cortex, MFG=Middle Frontal Gyrus, MeOFG = Medial Orbital Frontal Gyrus, MTG=Middle Temporal Gyrus, PCC = Posterior Cingulate Cortex, PCG=Precentral Gyrus, PoCG=Postcentral Gyrus, PHG = Parahippocampal Gyrus, SFG = Superior Frontal Gyrus, SMA=Supplementary Motor Area, SPL = Superior Parietal Lobule, STG = Superior Temporal Gyrus, TP = Temporal Pole.

## 2.4 Discussion

### 2.4.1 Shared Representation for Third-Party Help and Punishment Decision in Striatum

In line with H1 and previous fMRI studies focusing on help (Genevsky, et al., 2013) as well as altruistic punishment choice (de Quervain et al., 2004), we observed for the first time that striatum (esp. the ventral part) was activated during

both help and punishment choice in a third-party context where the third-party decider had both altruistic options to restore the justice.

The striatum is well known for processing the reward for more than half a century. The first direct evidence linking striatum and reward were from a neurophysiological study on rats (Olds & Milner, 1954). With electrodes permanently implanted in the brain, rats could get the electrical stimulation towards several specific regions while pressing a lever. It was important to note that rats received no other reward (e.g., water, food) during the experiment. Results showed that rats frequently pressed the lever which led the electrical stimulation on the striatal area, suggesting the strong relationship between striatum and reward. In non-primate electrophysiological studies, striatal neurons (e.g., in putamen, caudate and nucleus accumbens) were found to fire when animals were presented with reward itself or reward-predicting stimuli (Apicella, Scarnati, Ljungberg, & Schultz, 1992; Schultz, 2015). With the proliferation of the fMRI studies, multiple evidence that human reward processing and the relevant decision-making relied on the function of striatum was accumulated (K. S. Wang, Smith, & Delgado, 2016). The most common paradigm for human reward processing research is a simple guessing paradigm and relevant modified tasks. In such paradigms, people are always asked to make a simple guess and to win the money if their response is correct, such as guessing whether the next number is larger than the current one or which out of one to four boxes (from 1 to 4) contains a randomly hidden ball. Results have consistently showed the engagement of striatum at the moment of winning money (Fliessbach et al., 2010). Moreover, several studies further extend the effect of the social reward on the striatum (Bhanji & Delgado, 2014). An earlier study showed that the striatum was strongly activated when participants viewed attractive (vs. unattractive) faces (Aharon et al., 2001). The similar phenomenon was also observed when people gained the attention and the potential approval from others, in comparison to making decisions alone, while deciding whether to donate to the charity (Izuma, Saito, & Sadato, 2010). Furthermore, a study directly compared the neural correlates of receiving either monetary or social reward (i.e., obtaining the positive evaluation on their personality from other strangers), finding that both types of reward robustly activated striatum (Izuma, Saito, & Sadato, 2008). Given the above evidence, our results suggest that people might gain reward experience via either compensating the victim or punishing the offender, even with the cost of their own money.

A supplementary evidence which partially supports the shared neural representation between the two altruistic choices is that the LPFC, despite being in different hemispheres, increased the functional connectivity with the bilateral striatum during help or punishment decisions. This result partly consisted with the previous finding that third-party punishment elicited stronger activity in left LPFC in com-

parison with direct punishment (Strobel, et al., 2011). From the anatomical perspective, the connection between the LPFC and the striatum (Haber & Knutson, 2009) provide the basis for the task-dependent functional connectivity. From the functional perspective, the LPFC has long been regarded as a key area activated during goal-directed decision-making as well as cognitive control (Miller & Cohen, 2001; Tanji & Hoshi, 2008). Several studies adopting the brain stimulation technique revealed the causal relationship between the LPFC (esp. the right side) and decisions in either the social or non-social domain. For instance, participants in the recipient role during the Ultimatum Game were more likely to accept an unfair offer after the right LPFC was inhibited by low-frequency repetitive TMS compared with the sham control group. Recent evidence with tDCS further confirmed the crucial function of right LPFC in norm compliance. In particular, enhancing the excitability of the right LPFC via anodal tDCS reduced the voluntary sharing percentage of a proposer in a standard Dictator Game, but enhanced the sanction-induced sharing percentage in a context where the recipient can costly punish the unfair proposer. Interestingly, the opposite effect was observed if the right LPFC was inhibited by the cathodal tDCS (Ruff, et al., 2013). The similar effect was replicated in a later TMS study (Strang, et al., 2014). Besides, with an inter-temporal choice task, participants chose the option with immediate monetary reward more often after the inhibitory TMS on the left, but not the right, LPFC, indicating that left LPFC also engages in cognitive control (Figner et al., 2010). Given the above-mentioned evidence from previous literature, our PPI results indicate that third-party deciders, despite experiencing positive emotion and rewarding, still need more cognitive control to inhibit selfish impulses during the altruistic but costly decisions.

Given the common underlying neural substrates, our PPI results also hinted that there might still be some difference in neural processing during these two altruistic choices from a functional network perspective. Particularly, we found that the vmPFC was more closely associated with striatum during punishment choices. Based on previous literature, we know that the vmPFC is crucial for value computation during decision-making (Clithero & Rangel, 2013; Ruff & Fehr, 2014), engaged in integrating affective information (Naqvi, Shiv, & Bechara, 2006) and is also sensitive to reward processing together with the striatum (Bartra, McGuire, & Kable, 2013). Given the multiple functions that the vmPFC might be involved in, it becomes difficult to find a reasonable explanation to this explorative finding.

### 2.4.2 The Role of Empathic Concern in Affecting Choice Preference and Its Neural Correlates

As predicted in H2a, we found that empathic concern correlated positively with the proportion of help choices and negatively correlated the punishment proportion of third-party deciders. Our behavioral results replicate previous finding with the one-shot third-party paradigm (Leliveld et al., 2012) and extend the similar effect into a multi-shot game. Surprisingly, we also showed the modulatory effect of empathic concern on the decision process, namely that people with higher empathic concern were faster in making help choice but slower to punish on average. In the theoretical framework of dual system, reduced decision time is usually regarded as a sign for the automatic process. For instance, Rand and colleagues showed in a series of behavioral studies that participants are more cooperative if they made the decision faster than if they made the decision under time constraint, thus suggesting the heuristic and spontaneous nature of human altruism (Rand, Greene, & Nowak, 2012). Based on these findings it seems that help choice is the automatic option for higher empathic participants but with more controlled processes for lower empathic participants. If this explanation is true, we would predict that regions relevant to cognitive control, such as LPFC, were less activated during help vs. punishment choice in higher (vs. lower) empathic participants. However, our imaging findings conflicted with this prediction. Instead, participants with higher empathic concern displayed higher activity in fronto-polar regions (i.e., left part of LPFC and IPL) during help (vs. punishment) choice. Given the role of frontopolar region in attention (Corbetta & Shulman, 2002; Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008), we proposed an alternative possibility that high empathic third-party deciders prefer to help the victim as they have allocated more attention to the victim, which is in turn driven by the personality trait. However, this unsolved conflict in the explanation motivates future studies to test the above hypothesis.

### 2.4.3 Limitations

The current study bears several limitations. To begin with, we had to exclude nearly one thirds of participants for the fMRI analyses due to the huge individual difference of choice preference across participants. As previously mentioned, most participants were excluded mainly because that they failed to show sufficient altruistic choice in either or both types. Even in the remaining 25 participants, some of them showed strong preference and stick to one of the altruistic choice, which might lead to the unstable estimation of BOLD signal on the effect of the less preferred choice. As it is a common problem for fMRI studies related with decision-making focusing on a certain type of choice, it is not easy to find a good solution.

Perhaps the easiest way is to increase the sample size so that we can guarantee enough participants who fit the aim of the study even after exclusion, which, on the other hand, causes other difficulty in practice (e.g., increasing the research budget).

Another limitation is that since participants in the control trials did not need to respond (i.e., only observing the decisions made by the computer), we could not completely rule out the confounding difference in motor-relevant activity between the decision condition and the control condition. The current design has its advantages, namely that additional affective (e.g., anger) or cognitive (e.g., conflict) processes can be avoided due to the forced response (esp. if the indicated response was against with the voluntary response under a certain context). Given that striatum is part of the motor network (Witt, Laird, & Meyerand, 2008) and also engages in motivated action during decision-making (Guitart-Masip, Duzel, Dolan, & Dayan, 2014), however, the activation difference between choice and control conditions in striatum might also be partly due to the difference in motor requirement. Although later analysis which explicitly modeled the button press still confirmed our main results (i.e., stronger activities in bilateral striatum were associated with both altruistic choice after controlling for the motor effect), further study should take this problem into account and make a better control (e.g., using the condition by asking participants to perform a simple comparison between the payoffs of the offender and the victim as a high-level control).

### 2.4.4 Summary

In a nutshell, the current fMRI study reveal, for the first time, the neural correlates of costly help and punishment choices from third-party deciders by adopting a modified third-party paradigm. The common representation in the striatum during both choice types in such context suggests a reward experience of the human altruism during costly restoring the social norm. Moreover, we again confirm the role of empathic concern in modulating third party's altruistic choice preference, and further show the accompanying neural correlates in frontopolar regions, indicating the mechanism underlying such empathy-dependent choice modulation. These results extend our horizon and knowledge in understanding third-party altruistic decision-making, a special form of human altruism (Fehr & Fischbacher, 2003; Nowak & Sigmund, 2005).

# 3 Studies 2A and 2B: The Effect of Oxytocin on Third-Party Decision-Making and Its Neural Correlates[7]

## 3.1 Hypotheses: Study 2A

According to previous studies, we pose the following hypotheses to address our research questions:

H1: At the behavioral level, we expect that third-party deciders will show more altruistic choices of either type after they are treated with intranasal OXT, compared with the PLC control group.

H2: At the neural level, we expect that intranasal OXT will modulate the reward-relevant processes (esp. in NAcc) as well as the mentalizing processes (esp. in ToM network, mainly TPJ and MPFC) during altruistic decision-making and the accompanying perception (i.e., observing other's being helped or punished).

## 3.2 Methods: Study 2A

### 3.2.1 Participants

We recruited 41 healthy males (mean age: 25.1 ± 3.9 yrs) to attend the present pharmacological-fMRI study. To make sure all participants fit the strict healthy criterion, we performed a clinical screen with the Mini-International Neuropsychiatric Interview (Sheehan et al., 1998) for each participant separately before the MRI session. As a consequence, we ensured that all participants were without any current or past psychiatric or neurological disorders, were free of dependence and addiction to cigarettes, drug or alcohol abuse. Besides, we also guaranteed that all participants were in good health condition (i.e., no caffeine or alcohol intake, with regular sleep, without cold) on the day of fMRI study. This study was approved by the ethics committee of the Medical Faculty of the University of Bonn. All participants signed the written consent based on the Declaration of Helsinki (BMJ 1991; 302: 1194).

[7] The study based on this chapter (Study 2A) has been published during the PhD study period of the author with permission. The full citation is here: Hu, Y., Scheele, D., Becker, B., Voos, G., David, B., Hurlemann, R., & Weber, B. (2016). The Effect of Oxytocin on Third-Party Altruistic Decisions in Unfair Situations: An fMRI Study. *Scientific Reports, 6*.

### 3.2.2 Design

A within-subject, double-blind, placebo-controlled design was adopted. The key independent variable was the drug treatment, which means that each participant attended the fMRI study twice, with one time getting OXT (24 Internet Unit; Sigma Tau; 3 puffs per nostril alternately, each with ~4 IU) and the other time getting PLC with self-administered intranasal spray. Both the experimenters and participants did not know the real treatment on the scanning day.

### 3.2.3 fMRI Paradigm

The fMRI task paradigm was basically the same as we used in Study 1, with the following exceptions. First, we shortened the length of the task by reducing the number of decision trials to 80 so that we finally only made one scanning run (i.e., 80 decision trials with 40 control trials) for one session. Second, the stimuli were presented not via video goggles; instead they were projected on a 32-inch MRI compatible TFT LCD monitor (NordicNeuroLab, Bergen, Norway) positioned at the rear of the magnet bore and participants saw the stimuli via an MRI compatible mirror during the whole experiment. Last but not least, the deception rule was used for the current study. In specific, we did not collect the real choices from another independent group before the fMRI study so that participants' decisions would not make real monetary consequence on others, which was unknown to them.

### 3.2.4 Procedure

As mentioned above, participants were assessed with MINI on a separate day before the MRI session. On the same day, they also filled out the empathic concern subscale of Interpersonal Reactivity Index (IRI) scale for the measurement of empathic concern as an individual trait (Davis, 1983).

On the day of each scanning session, participants first filled out the a series of questionnaires including the State-Trait Anxiety Inventory (STAI), commonly used for measuring the state anxiety (Spielberger, Gorsuch, Lushene, & Vagg, 1970), and the Positive and Negative Affective Schedule (PANAS), commonly used for measuring the state emotion (Watson, Clark, & Tellegen, 1988). Next, participants were provided with the nasal spray and asked to administer a dose of 24 IU of either OXT or PLC by themselves. After that, participants were informed about the third-party task together with other tasks via reading instructions, which was followed by the practice in a separate behavioral testing room.

The scanning started around 30 min after the intranasal administration[8]. Following a 6-min resting-state scanning (i.e., unrelated with the current study and will be reported in another study), the third-party task began and lasted about 30 min. Besides, participants also completed another task that is irrelevant for this study. In total participants stayed in the scanner with functional scanning for about 60 min.

With a short break after scanning, participants finished a rating task for the monetary split they saw just now in the scanner. In particular, they were asked the following three questions, namely 1) "How unfair is this monetary splits offered by Player A to Player B?", 2) "To what degree do you think the proposer deserves punishment?", and 3) How much empathy do you feel for the recipient?" with the fixed order within each participant while the counterbalanced order across different participants, by indicating their evaluation on a 9-point Likert scale (0 = Not at all, 8 = very much). Next participants' anxiety and emotion state were measured again via STAI as well as PANAS. To further control the side effect of OXT on general cognitive ability, we also measured the attention performance via the d2 task (Brickenkamp, 1995). In the very end of each session, participants were also asked to report whether they received OXT or PLC treatment in this session (see Figure 11 for iilustration of whole procedure). Participants were paid 60 € for their attendance together with the task-dependent extra payment (~ € 25) after the 2nd session of the MRI measurement.

[8] We acquired a T1 anatomical scan for each participant before the drug administration when they did the task for the first time.

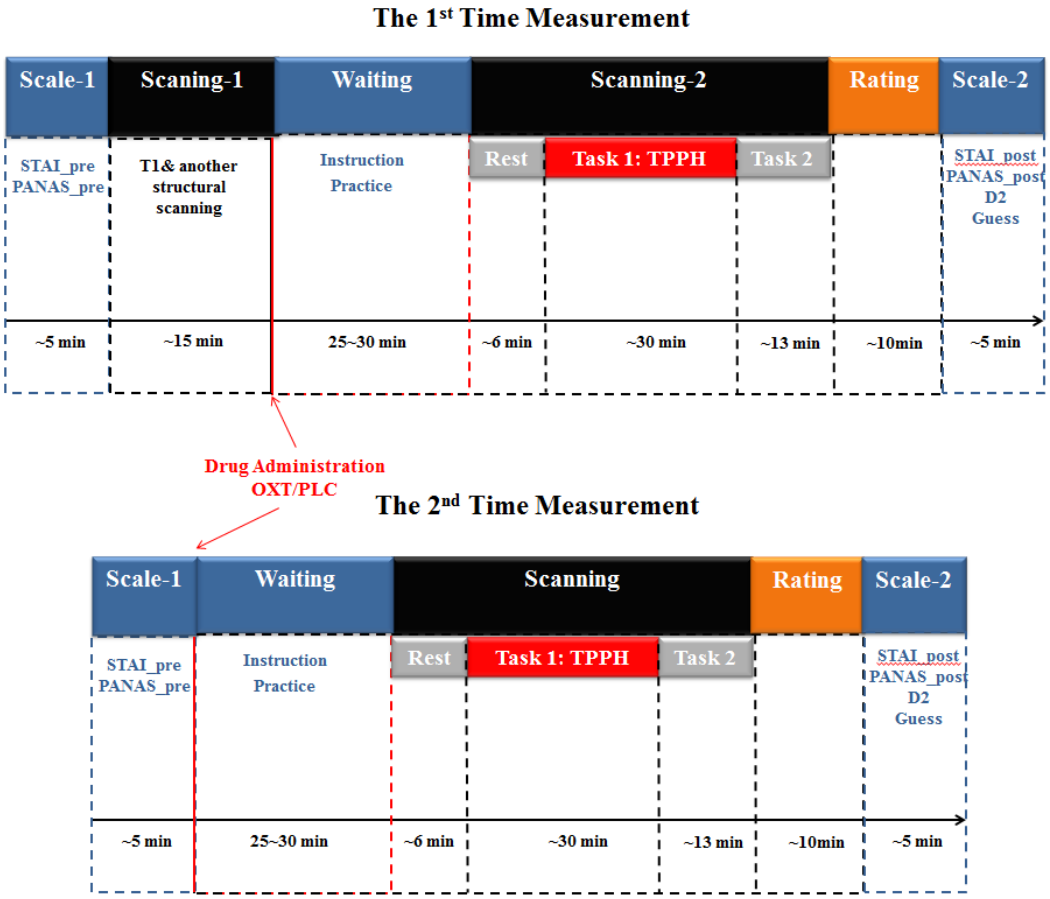


Figure 11. Experimental procedure for both measurements. D2 is a cognitive test used to check attention ability. Abbreviations: OXT = oxytocin, PLC = placebo, STAI = the state-trait anxiety inventory, PANAS = the positive and negative affective schedule, Rest = resting-state scanning, TPPH = third-party punishment and help, pre = before drug treatment, post = after drug treatment, min = minute.

### 3.2.5 Data Collection

All imaging data were collected via the 3-Tesla Siemens Trio platform at the Imaging Center of Life & Brain, University Hospital Bonn. The sequence used for both functional and structural images were the same as Study 1.

### 3.2.6 Data Quality Check and Analyses

Twenty-two (out of 41) participants were kept for the further data analyses. We excluded 13 participants as they failed to show enough decisions (i.e., with a lenient criterion: at least 5 decisions per run) to help (n = 1), punish (n = 9), or both choices (n = 3) in either one or both sessions. Besides, we also excluded one par-

ticipant who quitted during the scanning, 1 participants who received the same drug treatment for both sessions, 1 participants with extreme low IRI empathic concern level (i.e., out of 3 standard deviation of the whole sample) as well as 3 participants with extreme headmotion (i.e., > 3mm).

#### 3.2.6.1 Behavioral Data

Similar to Study 1, we calculated the mean proportion of choice behavior, the mean decision time as well as the mean transfer amount for help and punishment choice respectively for each participant in both sessions (see Table 8). The statistical approach with SPSS 22 (SPSS Inc.) was also similar to Study 1, except that we also used $\chi^2$ test to rule out the side effect of belief.

#### 3.2.6.2 fMRI Data

##### *3.2.6.2.1 Preprocessing*

We used SPM 8 (Wellcome Trust Department of Cognitive Neurology, London, UK) to analyze the fMRI data. The data was preprocessed following the same procedure in Study 1.

##### *3.2.6.2.2 General linear model (GLM) Analyses*

The GLM mass-univariate regression approach was adopted for the individual-level fixed-effect analyses. For each session, a separate GLM was built, which focused on the decision-phase and included four regressors of interest within in each run according to the decisions participants or the computer made, namely onsets of stimuli presentation during help, punish, help_computer, as well as punish_computer conditions. We pooled other uninterested events to a single regressor (i.e., other; same with Study 1) and also added the 6 parameters of head motion in the design matrix. Individual contrasts were built, including the contrast help vs. help_control, punish vs. punish_control, help vs. punish as well as regressors of interest vs. implicit baseline (i.e., help, punish, help_computer, punish_computer vs. implicit baseline respectively).

For the group-level random-effect analyses, we performed a 2 (i.e., treatment: OXT/PLC) × 2 (i.e., self-decision vs. computer: help vs. help_computer / punish vs. punish_computer) repeated measure flexible ANOVA to further test the three-way interaction between three factors, namely treatment (OXT/PLC), agency (self-decision/computer), and decision (help/punish). To further explore whether OXT can modulate the effect of empathic concern on altruistic decision-making, we also performed an additional regression analysis with the contrast [PLC_(help vs punish) vs OXT_(help vs punish)] as the dependent variable and the empathic concern scores as the predictor. For the whole-brain analysis, we adopted the

threshold of $p < 0.001$ uncorrected at peak voxel level with an extent threshold of $k = 50$.

*3.2.6.2.3 Region of interest (ROI) analyses*

Based on our hypotheses, we defined the following ROIs for the three-way interaction analyses mentioned above. Concerning the reward-relevant region, we focused on the bilateral NAcc, which were created based on the masks from the AAL template. Concerning the mentalizing process, we focused on two regions, namely bilateral TPJ and MPFC. Since these regions were not defined given the traditional anatomical template, we used the coordinate-based approach to draw the masks based on a recent meta-analysis literature on the neural correlates of mentalizing (Schurz, et al., 2014). Specifically, a sphere with the radium of 5mm centering on the following coordinates respectively (MNI space, x/y/z, with unit of mm): [-53/-59/20] for the left TPJ, [56/-56/18] for the right TPJ, and [-1/56/24] for the MPFC. All ROI masks were created via the Wake Forest University Pickatlas toolbox (WFU; http://fmri.wfubmc.edu/software/pickatlas ). For statistical analysis, we took the threshold of voxel-wise $p < 0.05$ and familywise error (FWE) corrected for multiple comparisons within the searching volume (i.e., the ROI). To further reveal the interaction, we extracted the parameter estimates (i.e., contrast values) of the peak voxel survived the FWE correction via MarsBar (http://marsbar.sourceforge.net/ ).

## 3.3 Results: Study 2A

### 3.3.1 Behavioral Results

#### 3.3.1.1 Proportion of Altruistic Choice

To test our first hypothesis (H1), we performed paired sample T-tests to compare the choice proportion between OXT and PLC treatment for the help and punishment respectively. However, the results failed to support H1 by showing no difference of drug treatment in either choice (both $ps > 0.7$). An exploratory correlation analyses showed OXT could also not influence the relationship between the difference of choice proportion (i.e., help vs. punish) and empathic concern ($r = -0.083$, $p = 0.714$).

Table 8. Descriptive summary of behavioral measures during the fMRI task

| | help Mean (S.D.) | | punish Mean (S.D.) | |
|---|---|---|---|---|
| | OXT | PLC | OXT | PLC |
| Choice Proportion (%) | 52.67 | 53.64 | 38.30 | 37.16 |
| | (20.74) | (20.65) | (19.28) | (20.97) |
| Decision Time (ms) | 1630.76 | 1732.08 | 1680.41 | 1801.77 |
| | (229.60) | (399.76) | (220.97) | (427.63) |
| Transfer Amount (MU) | 11.99 | 12.70 | 15.13 | 15.50 |
| | (5.92) | (5.86) | (5.57) | (5.82) |

Note: S.D.refers to standard deviation; MU refers to monetary unit.

#### 3.3.1.2 Other Measures

To further test whether intranasal OXT affects the decision process and transfer amount, we focused on these trials in which participants made altruistic choices (i.e., at least transferred 5 MU) and performed a repeated measurement 2 (treatment: OXT/PLC) × 2 (decision: help/punish) on the individual mean decision time (in ms) as well as mean transfer amount (in MU) respectively. We found a trend-to-significant main effect of treatment on decision time ($F(1, 21) = 3.051$, $p = 0.095$, partial $\eta^2 = 0.093$), namely that participants treated with intranasal OXT responded a bit faster in comparison to their choices in the PLC condition, regardless of whether they helped or punished. Besides, we also observed a main effect of decision on transfer amount ($F(1, 21) = 6.295$, $p = 0.02$, partial $\eta^2 = 0.231$), namely that participants punished the offender stronger than helped the victim.

We also checked the effect of OXT on post-scanning subjective rating as well as other controlled measures (before and after the scanning), including state anxiety, positive and negative state emotion and attention performance. With paired samples t-test, none of these above measures showed significant difference (see Table 9 and Table 10 for details). Apart from that, we ruled out the association between participant's belief and real treatment (correct estimates: OXT, n = 10; PLC, n =14; $\chi^2(1) = 0.376$, $p = 0.54$).

Table 9. Descriptive summary of control measures

| | OXT<br>Mean (S.D.) | PLC<br>Mean (S.D.) | Paired t-test<br>t (p) |
|---|---|---|---|
| Positive affect pre | 29.50 (5.91) | 30.73 (5.49) | -1.144 (0.266) |
| Positive affect post | 25.45 (6.44) | 25.32 (7.45) | 0.168 (0.868) |
| Negative affect pre | 11.18 (1.01) | 11.55 (1.79) | -1.250 (0.225) |
| Negative affect post | 11.91 (2.54) | 11.59 (2.13) | 0.718 (0.481) |
| State anxiety pre | 44.09 (1.44) | 44.18 (2.34) | -0.153 (0.880) |
| State anxiety post | 44.05 (2.77) | 43.68 (1.89) | 0.584 (0.565) |
| Attention | 191.05 (78.00) | 189.50 (86.96) | 0.140 (0.890) |

Note: The PANAS was used for assessing positive/negative mood and STAI_state for state anxiety. Both mood and anxiety were measured before and after the treatment; the D2 test was used for assessing attention. S.D. refers to standard deviation; OXT refers to oxytocin, PLC refers to placebo, pre refers to before treatment, post refers to after the scanning task.

Table 10. Descriptive summary of post-scanning rating

| | OXT<br>Mean (S.D.) | PLC<br>Mean (S.D.) | Paired t-test<br>t (p) |
|---|---|---|---|
| Perceived unfairness of offer | 4.73 (1.13) | 4.95 (0.90) | -1.755 (0.094) |
| Deservedness for punishing the offender | 4.26 (1.39) | 4.52 (1.00) | -1.161 (0.259) |
| Empathic concern for the victim | 4.43 (1.12) | 4.62 (0.86) | -0.824 (0.419) |

Note: All the post-scanning ratings range from 0 (not at all) to 8 (very much). S.D. refers to standard deviation; OXT refers to oxytocin, PLC refers to placebo.

## 3.3.2 Imaging Results

### 3.3.2.1 ROI Findings

Partially supporting H2, the ROI-based three-way interaction between treatment (OXT/PLC), agency (self- /computer-decision), and decision (help/punish), defined by the contrast "PLC_[(help vs help_computer) vs (punish vs punish_computer)] vs OXT_[(help vs help_computer) vs (punish vs punish_computer)]", showed the significant activation only in the left TPJ (peak MNI coordinates: -54/-54/22; t(63) = 3.54, p(FWE) = 0.005; see Figure 12) and trend-to-significant activation in but not in the right TPJ (peak MNI coordinates: 50/-58/20; t(63) = 2.35, p(FWE) = 0.079) as well as the MPFC (peak MNI coordinates: -2/56/20; t(63) = 2.25, p(FWE) = 0.095; see Table 11 for other activations

at the whole-brain level), whereas we failed to observe any significant activation in NAcc with the same threshold. Given the results of statistical significance, we only did post-hoc analyses on the left TPJ. Post-hoc analyses on the parameter estimates extracted from the peak voxel in left TPJ showed a treatment (OXT/PLC) × decision (help/punish) interaction for both the computer-decision condition ($F(1,21) = 10.536$, $p = 0.004$, partial $\eta^2 = 0.334$) and the self-decision condition ($F(1,21) = 4.901$, $p = 0.038$, partial $\eta^2 = 0.189$) with different direction. However, the post-hoc paired T-test only showed the OXT-relevant increased activity in left TPJ during trials in help_computer (vs. punish_computer) conditions ($t(21) = 2.348$, $p = 0.029$) but not in other contrasts (all ps > 0.16).

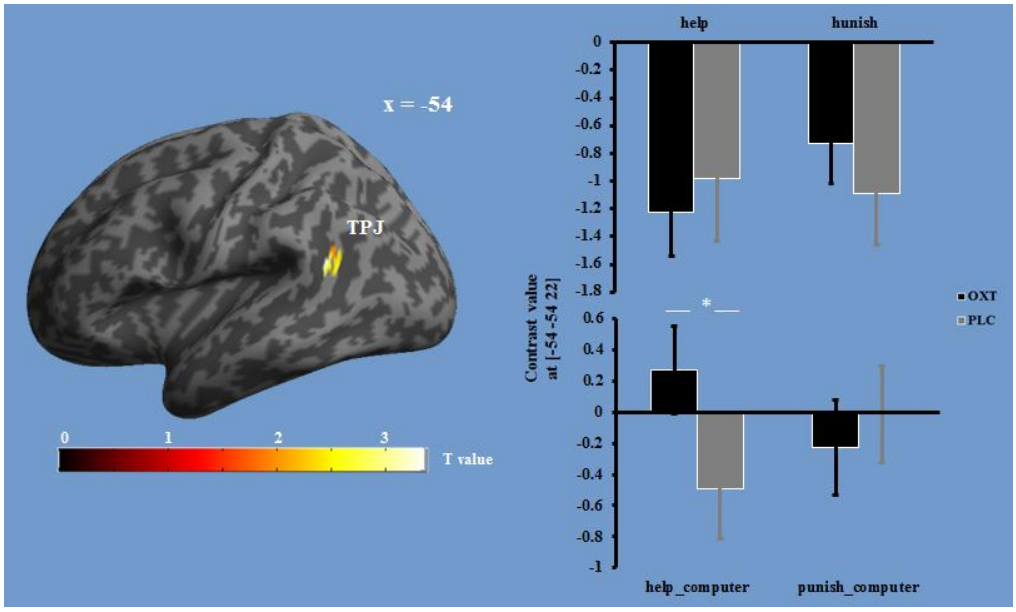


Figure 12. Left TPJ reflecting three-way interaction between drug treatment, agency, and decision (i.e., [PLC_(help – help_computer) – (punish – punish_computer)] vs. [OXT_(help – help_computer) – (punish – punish_computer)]). Bar plots showed the contrast value in the peak voxel of the left TPJ in all conditions. Display threshold: $p < 0.05$ at voxel-level within the mask, uncorrected. Significance level: *p < 0.05; Error bars: SEM. Abbreviations: OXT=oxytocin, PLC=placebo, TPJ=temporo-parietal junction.

Table 11. Regions reflecting the three-way interaction between drug treatment (OXT/PLC), agency (self-decision/computer), and decision (help/punish)

| Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***[PLC_(help - help_computer) - (punish - punish_computer)] - [OXT_(help - help_computer) - (punish - punish_computer)]*** | | | | | | | |
| SFG/MFG | R | 234 | 32 | 16 | 54 | 6/8 | 4.07* |
| IFG/MFG | R | 98 | 56 | 22 | 22 | 45/46 | 3.89 |
| TPJ/SMG/STG | L | 58 | -52 | -50 | 20 | 40 | 3.90 |
| IPL | R | 341 | 42 | -46 | 46 | 40 | 4.18* |
| SMA/PaCL | B | 337 | -2 | -12 | 70 | 6 | 4.67* |
| MTG | L | 59 | -44 | -44 | -8 | 37 | 3.85 |
| PCG/PoCG | L | 74 | -54 | -10 | 10 | 43 | 3.75 |
| PCG/PoCG | R | 141 | 52 | -14 | 30 | 3/4 | 4.00 |
| Thalamus | L | 91 | -18 | -22 | 14 | | 4.52 |
| ***[OXT_(help - help_computer) - (punish - punish_computer)] - [PLC_(help - help_computer) - (punish - punish_computer)]*** | | | | | | | |
| No cluster | | | | | | | |

Note: Threshold is set to $p < 0.001$, k=50, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): OXT=oxytocin, PLC=placebo; L=left, R=right, B=bilateral, BA=Brodmann Area; IFG=Inferior Frontal Gyrus, IPL=Inferior Parietal Lobule, MFG=Middle Frontal Gyrus, MTG=Middle Temporal Gyrus, PCG=Precentral Gyrus, PoCG=Postcentral Gyrus, PaCL=Paracentral Lobule, SFG=Superior Frontal Gyrus, SMA=Supplementary Motor Area, SMG=Supramarginal Gyrus, TPJ=Temporo-parietal Junction.

#### 3.3.2.2 Other Whole-Brain Level Findings

Besides, regions engaged in decision-making as well as action preparation, including the left middle frontal gyrus (BA 9/46), precentral gyrus (BA 3/4/7) as well as supplementary motor areas (BA 6/8) were observed with higher activity during trials in which participants made altruistic decisions (either help or punishment) themselves as opposed to observing the computer's decision. In the reverse contrast, we found higher activation in mentalizing network, such as bilateral TPJ (BA 39/40) and MPFC (BA 9/10), responding to trials with computer's decisions (see Figure 13; also see Table 12 for other activations). With the same threshold,

however, the whole-brain analyses did not detect other significant clusters in other main effect (i.e., treatment, decision) or interaction (i.e., the two way interaction: treatment × agency, treatment × decision, agency × decision; the three-way interaction).

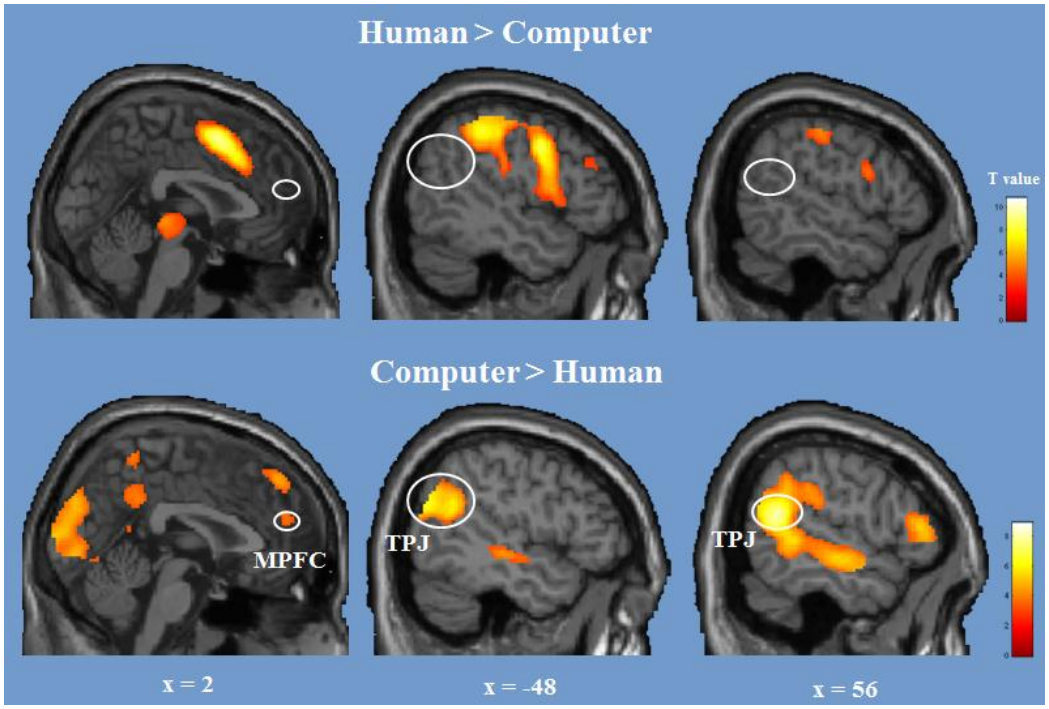


Figure 13. Regions reflecting the main effect of agency (upper panel contrast: self-decision vs. computer; lower panel contrast: computer vs. self-decision). Display threshold: $p < 0.001$ at the voxel-level, uncorrected. Abbreviations: MPFC = Medial Prefrontal Cortex; TPJ = Temporo-parietal Junction.

Table 12. Regions reflecting the effect of agency

| Brain Region | Hemi-sphere | Cluster Size | MNI Coordi-nates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***Self-decision - Computer-decision*** | | | | | | | |
| MFG/IFG | L | 146 | -38 | 48 | 8 | 10/46 | 4.05 |
| MFG | L | 410 | -38 | 32 | 26 | 9/46 | 6.31* |
| IFG | R | 216 | 60 | 10 | 24 | 9/45 | 4.50 |
| Insula | R | 240 | 34 | 20 | 8 | 13 | 5.61* |
| SMA/ACC/IPL/SPL/ Precuneus/PCG/PoCG/ Insula/ SFG/MFG | B | 13095 | -6 | 10 | 50 | 2/3/4/ 6/7/8/9/1 3/ 24/32/40 | 10.86* |
| MOG/IOG | L | 377 | -28 | -90 | -4 | 18/19 | 7.39* |
| MOG/IOG/MTG | R | 556 | 32 | -90 | -4 | 18/19 | 7.08* |
| Thalamus/Brainstem | B | 1052 | -4 | -26 | -2 | | 6.01* |
| ***Computer-decision - Self-decision*** | | | | | | | |
| SFG | B | 361 | 4 | 46 | 48 | 8/9 | 4.74* |
| MPFC/SFG | B | 588 | -12 | 58 | 22 | 9/10 | 4.27* |
| IFG/MFG | R | 678 | 50 | 42 | 2 | 45/46 | 6.33* |
| SFG/MFG | R | 389 | 24 | 26 | 46 | 8 | 5.52* |
| TPJ/IPL/SMG/AG/ MTG/STG | L | 1027 | -48 | -70 | 26 | 39/40 | 6.36* |
| TPJ/IPL/SMG/AG/ MTG/STG | R | 4051 | 56 | -52 | 18 | 21/22/ 39/40 | 7.97* |
| Precuneus/PCC/MCC | B | 1036 | -12 | -52 | 32 | 7/31 | 4.90* |
| MTG/STG | L | 1173 | -56 | -16 | -8 | 21/22 | 5.63* |
| PoCG | L | 122 | -24 | -40 | 60 | 3 | 3.88 |
| Precuneus/PoCG | R | 526 | 12 | -50 | 62 | 3/5/7 | 4.81* |
| PHG/FG | L | 319 | -26 | -50 | -6 | 19 | 5.31* |
| PHG/FG | R | 474 | 24 | -44 | -10 | 19 | 5.83* |
| Cuneus/LG/SOG/MOG | B | 3692 | -8 | -94 | 12 | 7/17/ 18/19/31 | 8.92* |
| Hippocampus/ PHG/Amygdala | L | 110 | -24 | -4 | -14 | | 4.20 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Hippocampus/ PHG/Amygdala | R | | 124 | 20 | -4 | -14 | 4.69 |

Note: Threshold is set to $p < 0.001$, $k = 50$, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): L = left, R = right, B = bilateral, BA = Brodmann Area; ACC = Anterior Cingulate Cortex, FG = Fusiform Gyrus, IFG = Inferior Frontal Gyrus, IPL = Inferior Parietal Lobule, IOG = Inferior Occipital Gyrus, LG = Ligual Gyrus, MCC = Mid-Cingulate Cortex, MFG = Middle Frontal Gyrus, MOG = Middle Occipital Gyrus, MPFC = Medial Prefrontal Cortex, MTG = Middle Temporal Gyrus, PCG = Precentral Gyrus, PoCG = Postcentral Gyrus, PHG = Parahippocampa Gyrus, SFG = Superior Frontal Gyrus, SMA = Supplementary Motor Area, SMG = Supramarginal Gyrus, SOG = Superior Occipital Gyrus, SPL = Superior Parietal Lobule, STG = Superior Temporal Gyrus; TPJ = Temporo-parietal Junction.

#### 3.3.2.3 Regression Findings

An explorative regression analysis showed that OXT, in comparison to PLC, reduced the positive correlation between individual empathic concern score (Mean ± S.D. = 17.95 ± 2.72; Range: from 14 to 24) and neural activity in bilateral inferior parietal lobules (IPL) during help (vs. punishment) choice (i.e., the contrast [PLC_(help - punish) - OXT_(help - punish)]). We extracted the parameter estimates of the peak voxel in bilateral IPL from the two contrasts respectively (i.e., contrasts help vs punish in PLC and OXT) and ran post-hoc correlation analyses with empathic concern scores to further reveal the moderation effect. Under the PLC treatment, help-dominated IPL activity (i.e., the contrast help vs. punish) positively correlated with empathic concern score, which was disappeared under the OXT treatment (see Figure 14; also see Table 13 for other activations).

Table 13. Regions reflecting the influence of empathic concern on the OXT effect on third-party altruistic decisions

| Brain Region | Hemi-sphere | Cluster Size | MNI Coordi-nates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***[PLC_(help - punish) - OXT_(help - punish)] & Empathic Concern*** | | | | | | | |
| IPL | L | 122 | -54 | -40 | 46 | 40 | 4.72 |
| IPL | R | 276 | 44 | -48 | 50 | 7/40 | 5.77* |
| PCG/PoCG | L | 92 | -34 | -26 | 52 | 3/4 | 4.37 |
| ***[OXT_(help - punish) - PLC_(help - punish)] & Empathic Concern*** | | | | | | | |
| - | | | | | | | |

Note: Threshold is set to $p < 0.001$, k=50, uncorrected; * Significant at $p < 0.05$ family wise error corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling toolbox for SPM8): OXT = Oxytocin, PLC = Placebo; L = left, R = right, BA = Brodmann Area; IPL = Inferior Parietal Lobule, PCG = Precentral Gyrus, PoCG = Postcentral Gyrus.

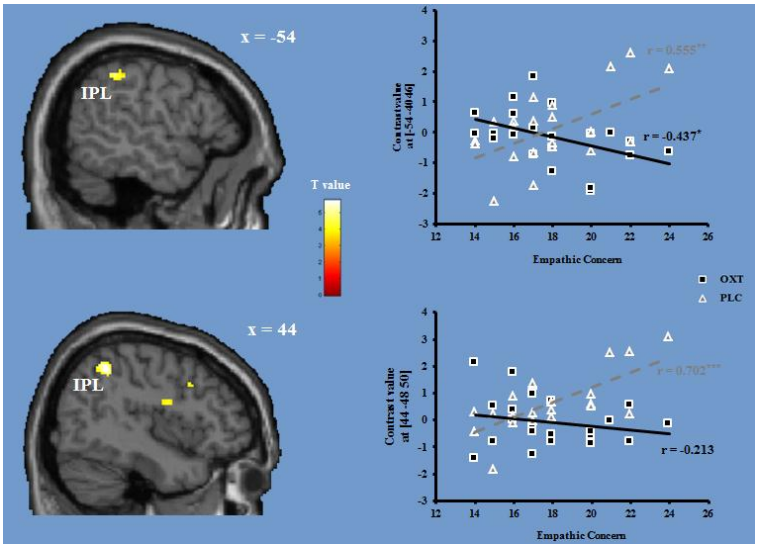


Figure 14. Bilateral IPL reflecting the modulatory influence of empathic concern on the effect of OXT on altruistic decisions (i.e., PLC_(help – punish) vs. OXT_(help – punish)). Display threshold: $p < 0.001$ at voxel-level, uncorrected. Scatter plot of showed the relationship between empathic concern and contrast values in peak voxel of bilateral IPL of the contrast help vs. punish in each drug condition respectively. Significance level: *$p < 0.05$, ***$p < 0.01$, ***$p < 0.001$. Abbreviations: OXT = oxytocin, PLC = placebo; IPL = inferior parietal lobule.

## 3.4 Hypotheses: Study 2B

According to Study 2A and other previous studies, we pose the following hypotheses in response to our research questions:

H1: We expect that third-party deciders will be more likely to either help the victim or punish the offender after they are treated with intranasal OXT, compared with the PLC control group.

H2: We expect that third-party deciders will be faster in making the altruistic choices (i.e., help or punishment) after they are treated with intranasal OXT, compared with the PLC control group.

## 3.5 Methods: Study 2B

### 3.5.1 Participants

We recruited 132 healthy males to the current study via ORSEE (for similar procedure, see Study 1). All participants had to fit the attendance criterion (for similar procedure, see Study 2A) and signed the written consent based on the Declaration of Helsinki (BMJ 1991; 302: 1194). Additionally, we recruited 121 female participants from the same subject pool for an independent behavioral study. This study was approved by the ethics committee of the Medical Faculty of the University of Bonn.

### 3.5.2 Design

A between-subject, double-blind, placebo-controlled design was adopted. Participants were randomly assigned to one of the drug treatments, namely receiving either OXT (24 Internet Unit; Sigma Tau; 3 puffs per nostril alternately, each with ~4 IU) or PLC intranasal spray. As usual, neither the experimenters nor the participants knew the real treatment on the day of experiment.

### 3.5.3 Decision Collection and Behavioral Paradigm

One week before the current experiment, we collected the real choices from online participants, which were used for the later third-party task via Qualtrics (https://www.qualtrics.com/). Specifically, online participants (i.e., the role of Player A; offender) were endowed with 100 MU (1 MU = 0.05 €) and asked to choose one of the three splits (i.e., self/other payoff: 50/50, 60/40, 90/10) between

themselves and an anonymous Player B (i.e., victim). They were also informed that their decisions might be selected and forwarded to a group of third parties in a later study, whose decisions could affect their final payoff as well as that of their matched partners. In the end, we randomly selected 61 decisions (of Player A; 15, 24, 22 choices for the split 50/50, 60/40, 90/10 respectively) from 121 participants and matched the rest of 60 participants (as the Player B) each with a different Player A.

The behavioral paradigm of the present study differed from the paradigm of the Study 2A in the following aspects. First, it was a between-subject design, which meant that participants only received one of the drug treatments and finished the task. Second, it was a "one-shot" game, namely participants made one decision in response to different possible monetary splits respectively, unlike the Study 2A in which participants made several decisions for each of the monetary splits. Third, the strategy method was used. Particularly, each participant, as third-party decider, was not informed the real decision of the offender and needed to made one decision in terms of each possible choice of the offender. Fourth, participants were endowed 100 MU which was always higher than what Player A kept for themselves. In this way we could rule out the possibility of disadvantageous inequality aversion (Jordan, McAuliffe, & Rand, 2014), another potential motivation for driving punishment behavior. Fifth, the option "keep" was presented together with the other two options (i.e., "increase", "subtract") during the decision phase. Sixth, the unfairness rating (i.e., "How unfair do you think of the split offered by the Player A to the Player B?"; a 9-point Likert scale: 0 = fair, 8 = very unfair) was placed immediately after the decision task (i.e., decision and transfer phase; see Figure 15 for details). Last, The stimuli were presented via z-Tree 3.5.1 (Fischbacher, 2007).

### 3.5.4 Procedure

Participants (i.e., third party) were assessed with clinical interview and completed a series online questionnaires (including IRI) in the morning of the experiment days. In the afternoon of the same day, participants in a group of approximately 15 people (i.e., 1 session; 10 sessions in total) arrived at the BonnEconLab. They were randomly assigned to independent cabins and self-administered a dose of 24 IU of either OXT or PLC nasal spray. Next, participants were provided with instructions of the tasks (i.e., including the current task together with other three tasks). The current task lasted around 6 min, which started around 75 min after the intranasal administration (~ 30 min) and another 4 irrelevant tasks (~ 45 min;

would be reported in other studies)[9]. At the end of the experiment, participants were paid by cash accordingly (the payment dependent on the current task: ~ € 4.6). Notably, we did not measure the state affect, anxiety and the attention performance as what we did in Study 2A, due to 1) that no evidence (i.e., Study 2A and other previous studies) has shown that OXT shows the side-effect on these measures and 2) practical reasons (e.g., the duration of the whole experiment; see Figure 16 for details).

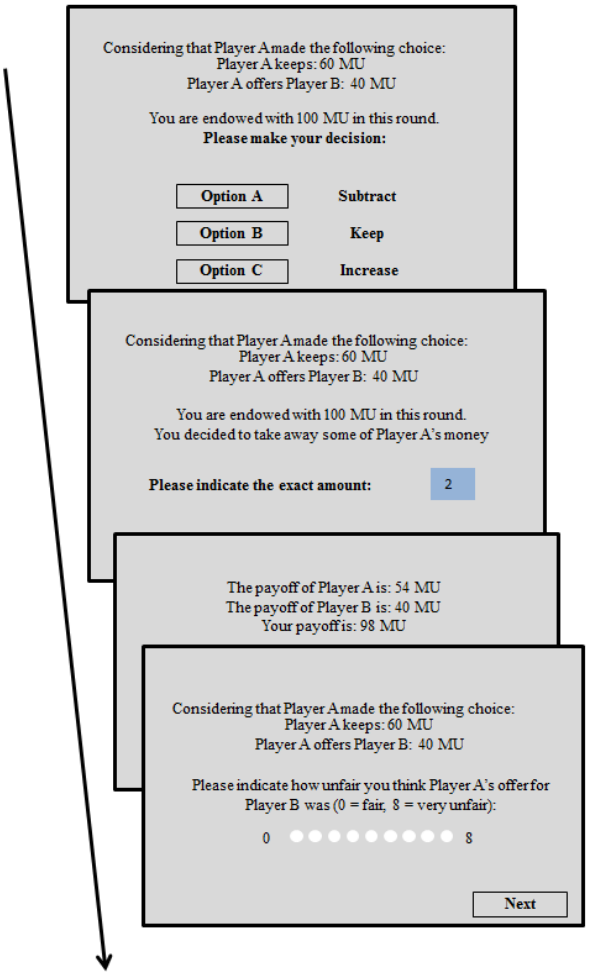


Figure 15. Example of the trial procedure. In this example, the participant subtracted 2 MUs from Player A. Abbreviations: MU = monetary unit.

[9] Note that the order of the first three tasks was always fixed. The order of the task reported here (task 4) and the rest task (task) was counterbalanced across participants (see Figure 16 for details).

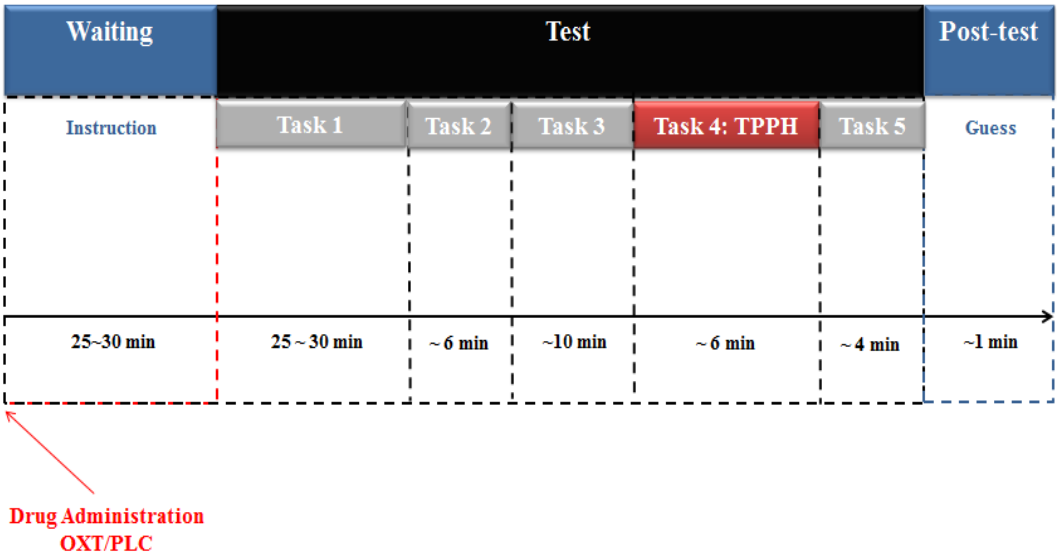


Figure 16. Experimental procedure. Note that the order of Task 4 and Task 5 was counterbalanced across participants. Abbreviations: OXT = oxytocin, PLC = placebo, TPPH = third-party punishment and help, min = minute.

### 3.5.5 Data Collection & Analyses

All behavioral data was collected via the Z-tree 3.5.1. Raw data was re-organized to the long format for later analyses via R 3.3.0 (https://www.r-project.org/) and all plots were created via the ggplot2 R package (http://ggplot2.org/). Statistical analyses were performed in STATA 13 (College Station, TX: StataCorp LP; http://www.stata.com/). We adopted the regression as our main statistical approach and used the robust standard errors clustered on subject to account for the non-independence of repeated measurement of the same participant (Hayes & Cai, 2007). For each dependent measures, we ran three regressions in total. To test our main hypotheses (H1 & H2), the main regression was performed only with the drug treatment (i.e., dummy variable; PLC as the reference group) as the predictor, the controlled regression was similar but additionally taking the monetary split (i.e., dummy variable; 50/50 as the reference group) into account, and the exploratory regression was with both variables and their interaction as the predictors. The descriptive summary for all measures was also listed (see Table 14).

Table 14. Descriptive summary of behavioral measures

| | | help Mean (S.D.) | | punish Mean (S.D.) | | keep Mean (S.D.) | |
|---|---|---|---|---|---|---|---|
| | | OXT | PLC | OXT | PLC | OXT | PLC |
| Choice Proportion (%) | 50/50 | 2.60 | 3.43 | 1.56 | 0 | 29.17 | 29.90 |
| | 60/40 | 7.29 | 8.82 | 5.73 | 4.90 | 20.31 | 19.61 |
| | 90/10 | 9.90 | 8.33 | 14.58 | 11.27 | 8.85 | 13.73 |
| Response Time (s) | 50/50 | 14.01 (4.81) | 16.97 (10.76) | 11.92 (4.87) | NA | 12.67 (9.47) | 12.27 (9.09) |
| | 60/40 | 15.73 (7.46) | 22.66 (18.27) | 19.79 (12.01) | 27.53 (18.84) | 14.11 (12.16) | 12.90 (14.22) |
| | 90/10 | 20.64 (13.16) | 23.15 (15.64) | 17.45 (9.05) | 21.25 (15.62) | 13.35 (9.31) | 14.96 (16.71) |
| Transfer Amount (MU) | 50/50 | 22.20 (21.43) | 10.43 (9.93) | 4.00 (5.20) | NA | 0 | 0 |
| | 60/40 | 12.86 (10.12) | 9.44 (4.64) | 7.73 (4.36) | 5.10 (3.84) | 0 | 0 |
| | 90/10 | 19.63 (10.41) | 22.53 (22.37) | 30.65 (24.02) | 26.91 (17.12) | 0 | 0 |

Note: OXT refers to oxytocin, PLC refers to placebo, S.D. refers to standard deviation, MU refers to monetary unit, NA refers to not applicable.

## 3.6 Results: Study 2B

### 3.6.1 Behavioral Results

#### 3.6.1.1 Side Effect

Similar to Study 2A, participant could not correctly guess the real drug treatment (correct estimates: OXT, n = 54; PLC, n =14; $\chi^2(1) = 1.456$, $p = 0.23$).

#### 3.6.1.2 Choice

To test our first hypothesis (H1), we performed three logistic regression on each of the possible choice (i.e., help vs. non-help; punishment vs. non-punishment; keep vs. non-keep) respectively. Contrary to our hypothesis, the main regression showed that OXT did not make third-party deciders more altruistic in either helping the victim (Odds ratio = 0.952, z = -0.17, p = 0.866) or punishing the offender (Odds ratio = 0.874, z = -0.42, p = 0.678), nor did it make participants more selfish (i.e., keep; Odds ratio = 1.036, z = 0.15, p = 0.883; see Table 15 for regression details). The results hold if we controlled for the effect of split; besides, participants were more (less) likely to help or punish (keep) with the increasing inequality of the monetary split. No interaction effect was observed between drug treatment and monetary split on the choice behavior (see Figure 17; also see Table 16 for details).

Table 15. Results of repeated-measure logistic regression predicting help, punishment, and keep choice by drug treatment

| | help | punish[a] | keep |
|---|---|---|---|
| PLC (ref.) | | | |
| OXT | 0.952 | 0.874 | 1.036 |
| | (-0.17) | (-0.42) | (0.15) |
| Constant | 0.259*** | 0.320*** | 1.72*** |
| | (-6.92) | (-5.26) | (3.20) |
| Pseudo-$R^2$ | 0.0001 | 0.0007 | 0.0001 |
| Observations | 396 | 264 | 396 |

Note: Values refer to odds ratio. The z statistics are provided in parentheses. Robust standard errors clustered on subject for each independent variable are used. OXT refers to oxytocin, PLC refers to placebo, ref. refers to reference.

Significance level: *** p < .001.

[a]For punishment choice, data in the 50/50 case are not used due to inaccurate estimation because of sparse observation.

Table 16. Results of repeated-measure logistic regression predicting help, punishment, and keep choice by drug treatment, offer, and their interaction

| | help | help | punish[a] | punish[a] | keep | keep |
|---|---|---|---|---|---|---|
| PLC (ref.) | | | | | | |
| OXT | 0.950 | 0.738 | 0.870 | 1.204 | 1.043 | 0.803 |
| | (-0.17) | (-0.49) | (-0.42) | (0.39) | (0.15) | (-0.40) |
| 50/50 (ref.) | | | | | | |
| 60/40 | 3.200*** | 3.137** | NA | NA | 0.191*** | 0.164*** |
| | (3.93) | (2.86) | | | (-5.42) | (-4.26) |
| 90/10 | 3.750*** | 2.905* | 2.300** | 2.964** | 0.094*** | 0.080*** |
| | (3.73) | (2.21) | (3.14) | (2.87) | (-7.62) | (-5.86) |
| OXT × 60/40 | | 1.053 | | NA | | 1.359 |
| | | (0.09) | | | | (0.50) |
| OXT × 90/10 | | 1.715 | | 0.588 | | 1.383 |
| | | (0.76) | | (-1.00) | | (0.52) |
| Constant | 0.102*** | 0.115*** | 0.202*** | 0.172*** | 7.642*** | 8.714*** |
| | (-6.71) | (-5.40) | (-5.85) | (-5.11) | (6.54) | (5.40) |
| Pseudo-$R^2$ | 0.043 | 0.046 | 0.028 | 0.031 | 0.131 | 0.131 |
| Observations | 396 | 396 | 264 | 264 | 396 | 396 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Robust standard errors clustered on subject for each independent variable are used. OXT refers to oxytocin, PLC refers to placebo, ref. refers to reference, NA refers to not applicable.

Significance level: * $p < .05$, ** $p < .01$, *** $p < .001$.

[a]For punishment choice, data in the 50/50 case are not used due to inaccurate estimation because of sparse observation.

### 3.6.1.3 Decision Time

To test our second hypothesis (H2), we performed two linear regressions on the decision time for each of the two altruistic choices (i.e., help and punishment) respectively. Since the decision times were not normally distributed (Jarque-Bera (S-K) test: $\chi^2(2) = 225.77$, $p < 0.001$), we adopted the log-transformed decision time instead. Contrary to our hypothesis, OXT did not facilitate either of the altruistic choice (help: b = -0.090, t = -0.59, p = 0.559; punishment: b = -0.177, t = -1.13, p = 0.263) in the main regression (see Table 17 for regression details). The results hold if we controlled for the effect of split; besides, decision time for either choice did not vary in terms of the monetary split or its interaction with drug treatment (see Table 18 for regression details).

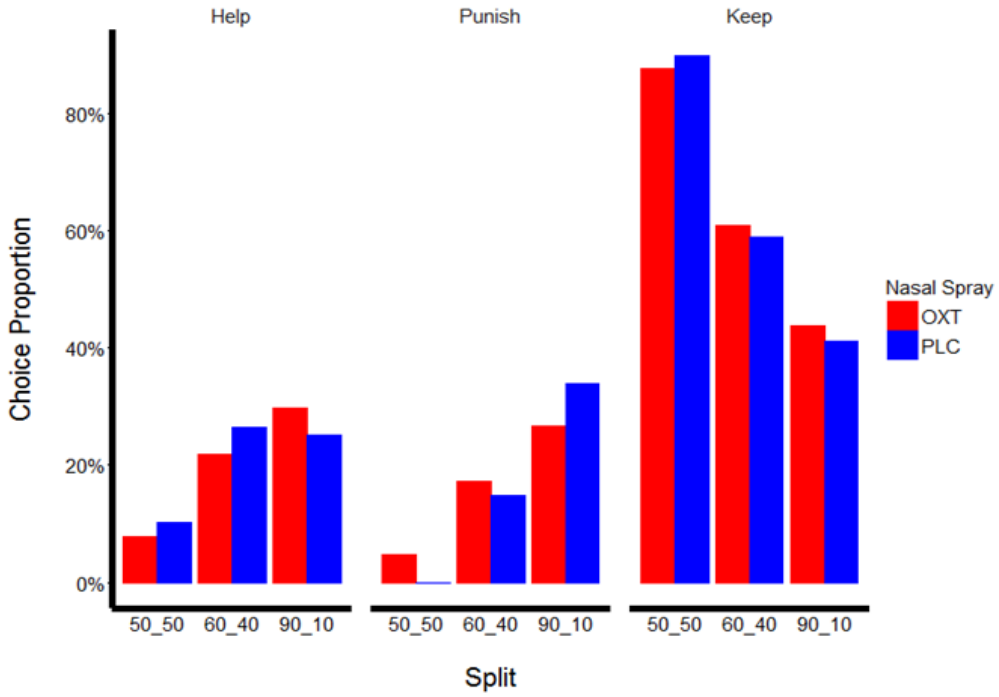


Figure 17. Proportion of each type of choices. Abbreviations: OXT = oxytocin, PLC = placebo.

#### 3.6.1.4 Explorative Analyses on Other Measures

With the same procedure, we also found that intranasal OXT did not affect the transfer amount of either choice (help: b = 2.569, t = 0.75, p = 0.458; punishment: b = 1.340, t = 0.46, p = 0.793) in the main regression. The results hold if we controlled for the effect of split. However, we found a differential effect of monetary split on transfer mount of different choice. Specifically, participants did not increase the transfer amount to help the victim with the increasing inequality of monetary split (both $p > 0.3$), whereas they punished more for the most unequal case (i.e., 90/10: b = 22.355, t = 6.52, $p < 0.001$). No interaction effect was detected in both cases (see Figure 18).

The main regression on the perceived unfairness also yielded non-significant effect of OXT (b = -0.047, t = -0.21, p = 0.832), which hold after controlling the effect of split. The controlled analyses also revealed that participants felt more unfair while the split went unequal (60/40: b = 2.303, t = 11.15, $p < 0.001$; 90/10: b = 5.780, t = 22.02, $p < 0.001$). Moreover, we observed an unexpected interaction effect in the exploratory analyses. In particular, participants with the OXT (vs. PLC) felt more unfair for the 50/50 (b = 0.724, t = 2.89, p = 0.005) split but less unfair for the unequal split (60/40: b = -1.074, t = -2.65, p = 0.009; 90/10: b = -1.242, t = -2.39, p = 0.018; also see Table 17 and Table 18 for regression details).

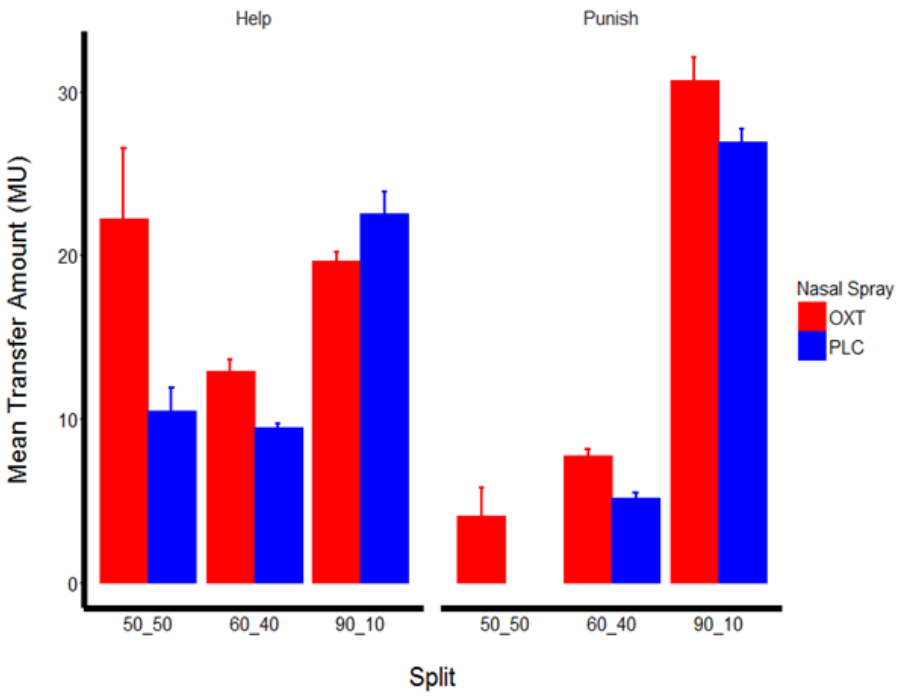


Figure 18. Mean transfer amount of either help or punishment choice. Error bars: SEM. Abbreviations: MU = monetary unit, OXT=oxytocin, PLC=placebo.

Table 17. Results of repeated-measure of linear regression predicting the other dependent variables by drug treatment

| | Log decision time | | | Transfer Amount | | Rating |
|---|---|---|---|---|---|---|
| | help | punish[a] | keep | help | punish[a] | |
| PLC (ref.) | | | | | | |
| OXT | -0.089 | -0.177 | 0.041 | 2.569 | 1.340 | -0.048 |
| | (-0.59) | (-1.13) | (0.49) | (0.75) | (0.26) | (-0.21) |
| Constant | 2.832*** | 2.933*** | 2.342*** | 14.90*** | 20.30*** | 3.157*** |
| | (25.13) | (27.03) | (40.51) | (6.01) | (6.80) | (21.05) |
| $R^2$ | 0.005 | 0.021 | 0.001 | 0.008 | 0.001 | 0.0001 |
| Observations | 80 | 61 | 252 | 80 | 61 | 396 |

Note: Values refer to unstandardized coefficient. The *z* statistics are provided in parentheses. Robust standard errors clustered on subject for each independent variable are used. OXT refers to oxytocin, PLC refers to placebo, ref. refers to reference.

Significance level: *** $p < .001$.

[a]For punishment choice, data in the 50/50 case are not used due to inaccurate estimation because of sparse observation.

Table 18. Results of repeated-measure linear regression predicting other dependent variables by drug treatment

| | Log Decision Time (s) | | | | | | Transfer Amount (MU) | | | | Rating | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | help | help | punish[a] | punish[a] | keep | keep | help | help | punish[a] | punish[a] | | |
| PLC (ref.) | | | | | | | | | | | | |
| OXT | -0.104 | -0.051 | -0.193 | -0.289 | 0.040 | 0.010 | 1.781 | 11.77 | 3.348 | 2.627 | -0.0475 | 0.724** |
| | (-0.69) | (-0.18) | (-1.22) | (-1.03) | (0.48) | (0.10) | (0.49) | (1.22) | (0.74) | (1.49) | (-0.21) | (2.89) |
| 50/50 (ref.) | | | | | | | | | | | | |
| 60/40 | 0.146 | 0.194 | NA | NA | 0.008 | -0.039 | -4.433 | -0.984 | NA | NA | 2.303*** | 2.824*** |
| | (0.78) | (0.64) | | | (0.09) | (-0.32) | (-1.02) | (-0.27) | | | (11.15) | (10.73) |
| 90/10 | 0.250 | 0.254 | -0.177 | -0.250 | 0.079 | 0.081 | 5.469 | 12.10+ | 22.35*** | 21.81*** | 5.780*** | 6.382*** |
| | (1.30) | (0.79) | (-1.06) | (-1.01) | (0.82) | (0.55) | (1.02) | (1.94) | (6.52) | (5.74) | (22.02) | (21.27) |
| OXT × 60/40 | | -0.113 | | NA | | 0.096 | | -8.359 | | NA | | -1.074** |
| | | (-0.32) | | | | (0.54) | | (-0.88) | | | | (-2.65) |
| OXT × 90/10 | | -0.019 | | 0.148 | | -0.002 | | -14.67 | | 1.107 | | -1.242* |
| | | (-0.05) | | (0.44) | | (-0.01) | | (-1.32) | | (0.16) | | (-2.39) |
| Constant | 2.668*** | 2.646*** | 3.057*** | 3.107*** | 2.323*** | 2.337*** | 14.59** | 10.43** | 4.723+ | 5.100*** | 0.462** | 0.0882 |
| | (15.31) | (10.60) | (18.59) | (14.70) | (36.87) | (33.17) | (3.41) | (2.88) | (1.93) | (4.27) | (3.29) | (1.18) |
| $R^2$ | 0.023 | 0.024 | 0.040 | 0.044 | 0.004 | 0.005 | 0.109 | 0.140 | 0.300 | 0.300 | 0.586 | 0.593 |
| Observations | 80 | 80 | 61 | 61 | 252 | 252 | 80 | 80 | 61 | 61 | 396 | 396 |

Note: Values refer to unstandardized coefficient. The *z* statistics are provided in parentheses. Robust standard errors clustered on subject for each independent variable are used. OXT refers to oxytocin, PLC refers to placebo, ref. refers to reference, NA refers to not applicable.

Significance level: $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

[a]For punishment choice, data in the 50/50 case are not used due to inaccurate estimation because of sparse observation.

## 3.7 Discussion: Studies 2A and 2B

### 3.7.1 The Effect of Intranasal OXT on Altruistic Decisions in Third-Party Context

Contrary to our original hypotheses (Study 2A: H1a; Study 2B: H1), intranasal OXT did not improve the proportion of either help or punishment choices, compared with the PLC condition, in both studies with similar but slightly different design and paradigm. There was mixed evidence for the association between OXT and pro-social behaviors. Earlier studies explored that OXT could improve the

human altruism especially in the male sample. The most well-known example is the one on trust behavior (Kosfeld, et al., 2005), namely that healthy male with intranasal OXT treatment increased their investment for the anonymous trustee but not increased their risky behavior, in comparison with the PLC condition. However, the later studies showed the null effect of OXT in improving human altruism, indicating that the prosocial effect of OXT might be dependent on other factors. Concerning the altruistic giving behavior, Zak and colleagues (2007) found that participants receiving OXT only increased the amount giving to the other unknown recipients when they had the chance to reject the offer rather than in a standard dictator game, indicating that the intranasal OXT might only change the sanction-induced altruism (Zak, et al., 2007). A recent example is that people with intranasal OXT treatment increased their donating behavior only when the monetary donation aimed to benefit the people (i.e., prosocial frame) instead of to protect the environment (i.e., pro-environment frame) of the rainforest area in Africa (Marsh et al., 2015). Besides, participants, under the OXT condition, were found to cooperate more often only with their in-group partner but not the out-group partner, which was replicated in a series of experiments with different modified prisoner-dilemma paradigms (De Dreu, 2012; De Dreu, et al., 2010; De Dreu & Kret, 2016). Given the above evidence, we argue that the null effect of OXT in increasing third-party altruistic choices might be due to the following reasons: either all other parties involved in our paradigm were anonymous for the third-party participants or they were not fear of any potential negative consequence for not being altruistic (i.e., all decisions they made were voluntary).

Although intranasal OXT did not affect the third-party altruistic behavior in both studies, we observed an unexpected but interesting trend-to-significant effect of OXT in facilitating altruistic decision processes, indexed by reducing the average decision time for both help and punishment choices (vs. PLC) in Study 2A. This result is consistent with the social salience hypothesis (Shamay-Tsoory, 2010; Shamay-Tsoory & Abu-Akel, 2016), which addresses the key role of OXT in enhancing the social cues in different contexts. In a recent review paper, Ma and colleagues (2016) proposed the social adaptation model of OXT, a more comprehensive theoretical framework in explaining the effect of OXT in social behavior, which also covered the OXT-dependent social salience enhancement as one of the crucial underlying mechanisms (Ma, Shamay-Tsoory, Han, & Zink, 2016). However, this result was not replicated in the Study 2B. One possible reason could be that participants started the task in around 75 min after the intranasal administration, which would definitely reduce the effect of OXT. Thus whether OXT could affect the altruistic decision process should be cautiously treated and needs further replication.

### 3.7.2 Intranasal OXT Modulates Neural Correlates of Different Altruistic Decisions and Accompanying Perception Process

Contrary to our original hypothesis (Study 2A: H2) based on previous literature as well as our finding in Study 1, intranasal OXT did not affect the neural processing during either altruistic decisions or perception process in the NAcc, the key region of the reward circuit (Haber & Knutson, 2009). Likewise, the evidence of OXT in enhancing reward-relevant activation is not robustly reported, indicating the involvement of other potential modulators during this process. For instance, male participants with intranasal OXT treatment did not show higher neural activity in the striatum (including NAcc) in response to attractive women's faces who were not familiar to them, although the intranasal OXT significantly increased the NAcc activity when they viewed their romantic partner's faces (Scheele, et al., 2013). In another study, OXT even exerts a reversed effect by reducing the activation in reward neuro-circuits when fathers viewed their own kids' faces in comparison with faces of other unknown children (Wittfoth-Schardt et al., 2012). These evidence together suggests that the effect of OXT on reward system might be modulated by the social context, especially the social distance. Thus, the null effect of OXT in our case might be also due to the far social distance between participants and other players involved. This explanation, however, should be tested by future studies with a similar paradigm in which the social distance should be explicitly manipulated (e.g., either the offender or the victim is a friend, an in-group member or a stranger of the third-party participant).

On the other hand, the result supports part of the H2 of Study 2A that intranasal OXT modulated the mentalizing network (esp. left TPJ) during either the decision or perception process in such context, which specifically increased the activity of left TPJ when participants observed others being helped. Consistently, we also found the involvement of these regions during observation (computer) trials in comparison to decision trials, which is in favor of the explanation of these regions as mentalizing-relevant process based on previous fMRI studies (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Schaafsma, et al., 2014; Schurz, et al., 2014) and a lesion study highlighting the crucial contribution of left TPJ in such process (Samson, Apperly, Chiavarino, & Humphreys, 2004).

As mentioned before (see introduction), the ability of meantalizing and a relevant process (e.g., empathy) is regarded as the precursor for the human altruism, especially when making altruistic decisions in such a complex situation (De Waal, 2008). There is evidence from an earlier behavioral study finding that intranasal OXT could improve the mentalizing ability measured by the Reading the Mind in the Eyes Test (RMET) (Domes, et al., 2007) (but also see (Radke & de Bruijn, 2015)), a paradigm in which participants were asked to judge the emotion of the

target in the stimuli only in terms of their eyes (Liebrand, Jansen, Rijken, & Suhre, 1986). Moreover, a recent study with a hypothetical third-party context revealed that participants with intranasal OXT had stronger harmful feelings for the victim but they rated higher for the deservedness to punish the offender whom were involved in the criminal scenarios (Krueger, et al., 2013), which further suggested the asymmetry of OXT-dependent enhancement for pro-social perception that are more sensitive to the victim. Besides, participants were also found showing more empathy for other's pain only when taking the perspective of others instead of themselves (Abu-Akel, Palgi, Klein, Decety, & Shamay-Tsoory, 2015). Intriguingly, a recent study focusing on the female sample revealed that OXT promotes participant's spontaneous anthropomorphism, measured by the way participants used to interpret the geometric movement (Scheele et al., 2015)

Regarding the neural processing, there is also some, but not much, evidence detecting the OXT-dependent change of TPJ activation in different paradigms requiring the ability of social cognition in close relationship with mentalizing, from both the healthy and the clinical samples. Based on the healthy male sample, Lancaster and colleagues (2015) showed the link between the left TPJ (as well as other regions) activity in response to the biological motion (e.g., the geometric shapes move in a regular way, compared with random movement), measured via the blood, and OXT plasma levels (Lancaster et al., 2015). Based on a sample of high-function autistic children and adolescents, a recent fMRI study detected the increased TPJ activity during the mentalizing-relevant task with the RMET paradigm (Gordon et al., 2013). On the basis of above evidence, our results indicate that OXT might also induce anthropomorphic tendency in males via modulating the activity in TPJ in such context, which may strengthen the social salience cue and then facilitate altruistic decisions as well as relevant prosocial processes. However, such interpretation should be treated with caution as the link between the OXT reception gene and the TPJ is still unknown (Haas, Anderson, & Smith, 2013).

### 3.7.3 Intranasal OXT Modulates Empathy-Dependent Neural Correlates of Different Altruistic Decisions

The explorative analyses in Study 2A further revealed the modulatory role of OXT in altering the empathy-dependent neural activity in bilateral IPL during help (vs. punishment) choices. In particular, different activation between help and punishment in IPL was positively correlated with individual empathic concern scores, as what we observed in Study 1. However, such relationship was reduced under the intranasal OXT treatment. Since the IPL is a crucial part of the attention and control network (Corbetta, Patel, & Shulman, 2008; Corbetta & Shulman,

2002), our results indicate that OXT might change the salience of social cues relying on the attention system which is also modulated by the endogenous personality trait of empathic concern, in consistent with the social salience theory (Shamay-Tsoory & Abu-Akel, 2016). Again, this effect needs to be replicated by future studies.

### 3.7.4 Limitations

There are some limitations which might affect the generalizability of Study 2A and 2B. First and foremost, as in Study 1, we excluded several participants (~ 46%) of Study 2A in the later analyses. Most of them were excluded because they failed to show sufficient help and punishment choice in either or both of the sessions. The only solution might still be, within the scope of the fund and the time, to increase the sample size which can benefit the statistical power. An additional advantage of the large samples is that investigators can then divide the participants into different groups in terms of their social preference so that they can compare the difference between different groups both at the behavioral and the neural level. Another problem is that we only recruited male participants in both Study 2A and 2B. As more and more evidence indicated the gender difference in OXT-induced effect in social cognition and decisions (Chen et al., 2015; Feng et al., 2015; Gao et al., 2016; Rilling et al., 2014; Scheele et al., 2014), future studies should also recruit the healthy females and compare the effect of OXT on the same measures between genders. Last but not least, participants of Study 2B started the current task after three other non-relevant tasks due to practical reasons (i.e., ~75 min after intranasal administration), which might lead to the reduced effect of OXT and other potential confounding problems such as fatigue as well as proactive interference.

### 3.7.5 Summary

Studies 2A and 2B provide, to our knowledge, the first evidence of how intranasal OXT affects the altruistic decision-making of healthy males in the third-party social context. Moreover, Study 2A, adopting the fMRI technique, further reveals the effect of OXT on the neural correlates of decision and accompanying perception processes. In Study 2A, we showed that in the subsample of altruistic participants, OXT slightly facilitates altruistic choices (i.e., either help or punishment) by reducing the decision time, despite that it did not improve either the proportion or the intensity (i.e., transfer amount) for both altruistic choices. We replicated the null effect of OXT on the altruistic choice but did not observe the OXT-dependent change in decision time in Study 2B. At the neural level (i.e., only in Study 2A), OXT selectively increased the activity in left TPJ when participants viewed the

victim being helped, which indicated the plausible role of OXT in promoting the anthropomorphic tendency during mentalizing process. Besides, OXT also modulated the empathy-dependent activity in IPL for help (vs. punishment) choices, suggesting its role in influencing the salience endogenously dependent on the empathic concern via the attention system during altruistic decision-making. In sum, the current results extend our knowledge of the linkage between OXT and a unique form of human altruism in a more complex social context and the potential underlying neural mechanism.

# 4 Study 3: The Effect of Other-Regarding Focus on Third-Party Altruism and Its Neural Correlates

## 4.1 Hypotheses

According to previous findings and our research questions, we have the following hypotheses:

H1: Compared with deciding naturally (i.e., baseline block, BB) we expect that participants as third-party deciders will choose to punish the offender more often when instructed to consider the (un)fairness of offender's proposal (i.e., offender-focus block, OB). Alternatively, participants are expected to increase the frequency to help the victim once they focus on the victim's feelings after receiving the offer (i.e., victim-focus block, VB).

H2: At the neural level, we expect that activation in TPJ will be higher during decision-making in both OB and VB, compared with BB.

H3: At the neural level, we also expect that the control network (e.g., anterior cingulate cortex, ACC; inferior frontal gyrus, IFG) will show stronger activation while making the decisions that cause conflicts with the focus (e.g., the choice of help conflicts with OB), compared with making the same decision in BB.

## 4.2 Methods

### 4.2.1 Participants

We recruited 50 healthy participants to attend the current fMRI study (23 male; mean age = 24.6 ± 3.5; 4 left handedness) via online flyers at the University of Bonn and social media. The study was approved by the ethics committee of the University of bBonn and written informed consent was given by all participants according to the Declaration of Helsinki (BMJ 1991; 302: 1194).

### 4.2.2 Paradigm and Stimuli

The current study utilized a mixed fMRI design and comprised one functional scanning, which consisted of 18 blocks equally distributed to three conditions of other-regarding focus conditions (i.e., BB, OB, and VB; see Figure 19A). To minimize the potential confounding effect of proactive inference caused by experimental manipulation, we fully randomized the order of blocks for each subject,

however we ensured no more than three consecutive blocks belonging to the same focus condition. Each block started after a 5-s instruction, which asked participants, as third-party decider, to either focus on the (un)fairness of Player A's (i.e., OB) offer, focus on the feeling of Player B (i.e., VB), or decide naturally (i.e., without a specific focus; BB; see Figure 19B) before making decisions. Eight trials began after the instruction, which consisted of seven target trials with an unfair offer (i.e., the payoff of the offender was at least twice as that of the victim) and one filler trial with a fair offer. The order of these trials was also randomized across participants. Within each trial, participants saw a monetary allocation between a specific Player A and Player B (the total amount ranged from € 9 to € 11), identified only by the initials, and then were asked to decide whether to decrease the payoff of Player A or increase the payoff of Player B by using their own endowment (i.e., € 10) in 4 s (i.e., the decision phase). After a jittered ISI fixation (3 ~ 5 s), they were further asked to indicate the exact amount on a VAS ranging from 0 to 10 with the changing step of € 0.5, within 4 s (i.e., the transfer phase; see Figure 19C). The cost ratio was also set to 1:3. Any fast response (i.e., responding less than 200 ms) or missing response (i.e., not resonding in 4 s) during the decision phase was warned with a message and the endowment in that trial was deprived. For other details about the paradigm, see Study 1 and Study 2A.

Notably, the current study adopted different stimuli to make the context closer to a real-life situation and reduce degree of losing attention from the participants due to repetition of exact the same stimuli (see Appendix Table 2 for all stimuli). First, Euro was used as the currency unit instead of an arbitrary monetary unit (MU). Second, we refined and increased the variation of payoff by keeping two digits round to the same integer (i.e., a random number within ± 0.2). Last but never the least, we ensured the average of total payoffs of all unfair trials (i.e., € 10) to be the same across all blocks to rule out the confounding effect of monetary amount.

### 4.2.3 Procedure

Upon arrival, participants were informed about the context and given the first part of the instructions, which contained general information about the third-party paradigm without mentioning the focus manipulation (i.e., BB). Next, participants passed an comprehension quiz and performed some practice trials to be familiarized with the task. After that, we provided them the second part of the instructions which explicitly indicated the other two focus conditions (i.e., OB and VB). Critically, participants were also told that they should always make the decisions they preferred, which aimed to reduce the demand characteristics as previous studies did (Hare, Malmaud, & Rangel, 2011). Then, they did another round of practice

that also included the instruction phase while in the scanner. The functional scanning lasted around 40 min and was followed by a 6-min structural scanning. Afterwards, participants completed a rating task to indicate their unfairness feeling to offers that appeared during the fMRI experiment on a 9-point Likert (0 = not at all, 8 = very much). Participants were paid at the end of the experiment (up to € 25).

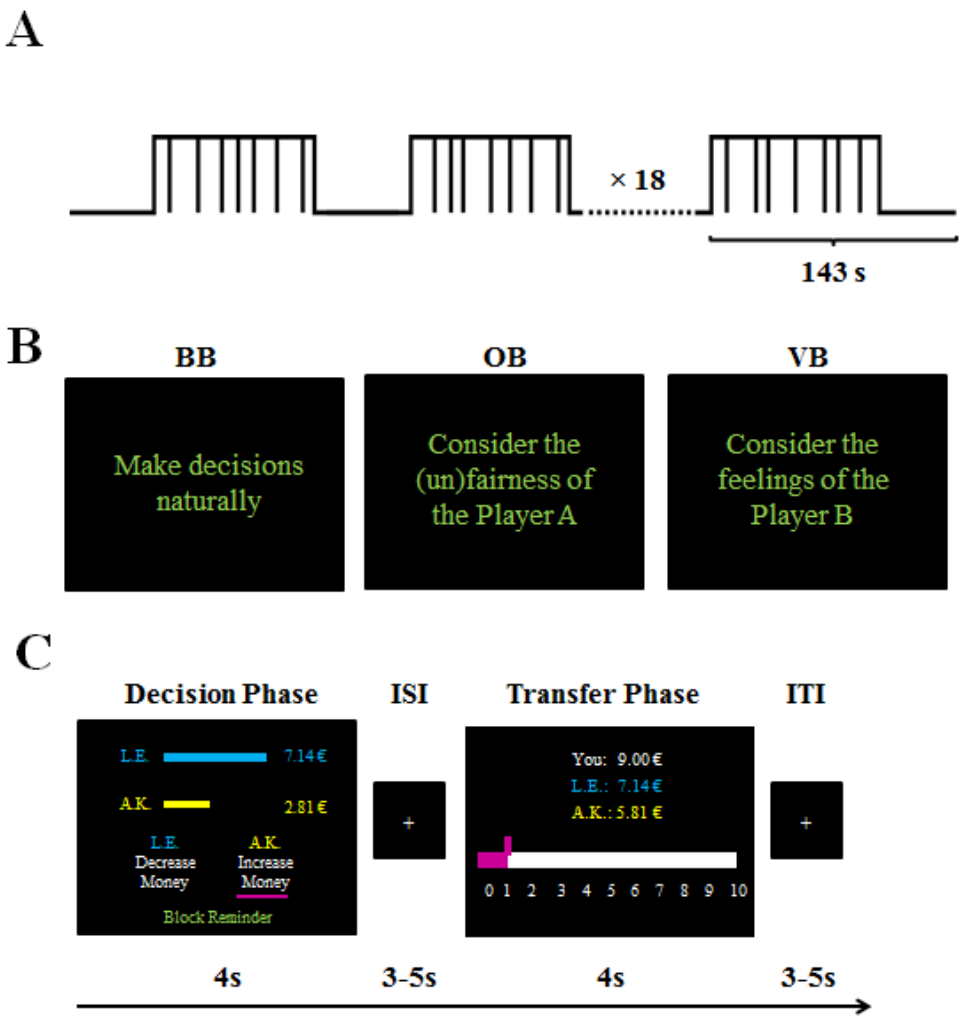


Figure 19. (A) Illustration for the mixed design; (B) Instructions screen presented before each block; (C) Example for the trial procedure. The offender was labeled as Player A, the victim was labeled as Player B in the whole experiment. In this example, the the participant added € 1 to the victim (i.e., A.K.). Abbreviations: BB = baseline block, OB = offender-focused block, VB = victim-focused block, ISI = inter-stimulus interval, ITI = inter-trial interval.

### 4.2.4 Data Collection

All imaging data were collected via the 3-Tesla Siemens Trio platform at the Imaging Center of Life & Brain, University Hospital Bonn. The sequence used for both functional and structural images were the same as Study 1 and Study 2A.

### 4.2.5 Data Quality Check and Analyses

We excluded four out of 50 participants from the analyses because of excessive head movement (N = 3) and quit of the scanning (N = 1). Given our research questions (also hypotheses) and the power of statistical analyses, we based our later analyses on the following three (sub)samples: i.e., the MAIN sample (N = 46), the HELP subsample (N = 42; participants chose at least five times to help in all three focus conditions), the PUNISH subsample (N = 22; participants chose at least five times to punish in all three focus conditions). Additionally, we did an explorative analysis of the interaction effect between attention focus (i.e., BB, OB, and VB) and altruistic decision type (i.e., help and punish) on the decision-relevant neural activities, which was performed on the HELPUN subsample (N = 20; ; participants chose at least five times to both help and punish in all three focus conditions).

#### 4.2.5.1 Behavioral Data

To test H1, we calculated the proportion of help and punishment choices in each focus condition in the MAIN sample and performed a repeated measure one-way ANOVA on each choice respectively. To check the robustness of this result, we did the same analyses on remaining three subsamples. Besides, we also checked the decision time and transfer amount of help choices in the HELP subsample and that of punishment choices in the PUNISH subsample, via the same analyses. Moreover, we adopted a 3 (attention focus: BB, OB, and VB) × 2 (altruistic choice type: help and punish) repeated-measure ANOVA to check the main effects and their interaction on the decision time and transfer amount in the HELPUN subsample. All of the above analyses were performed via SPSS 22 (SPSS Inc.). Mauchly's sphericity test was used to check the assumption of sphericity for ANOVA and a Greenhouse-Geisser correction was applied if this assumption was violated. Bonferroni correction was adopted to control for multiple comparisons in post-hoc analyses. All reported p-values were two-tailed and $p < 0.05$ was considered significant.

#### 4.2.5.2 fMRI Data

##### *4.2.5.2.1 Preprocess*

We used SPM 8 (Wellcome Trust Department of Cognitive Neurology, London, UK) to analyze the fMRI data. The data was preprocessed following a similar procedure as used in Study 1 and Study 2A, except that we adopted a cut-off value of 286 s (i.e., twice the block duration; instead of the default 128 s) to model the block effect in the high-pass temporal filtering.

##### *4.2.5.2.2 General linear model (GLM) analyses*

To test H2, GLM1 was built on the MAIN sample, which included three regressors of interest, namely onsets of decision phase during all choices for unfair trials in BB, OB, and VB (i.e., BBdec OBdec VBdec). To test H3, we also built GLM2 and GLM3 based on the HELP and PUNISH subsample respectively. GLM2 consisted of three regressors of interest, namely onsets of decision phase during help choices in BB, OB and VB (i.e., BBhelp, OBhelp, VBhelp). Similarly, GLM3 included the same regressors, but with punishment choice instead (i.e., BBpunish, OBpunish, VBpunish). To exploratively test the interaction effect on choice-relevant activation, we also built GLM4 on the HELPUN subsample with six regressors of interested included: i.e., BBhelp, BBpunish, OBhelp, OBpunish, VBhelp, VBpunish (similar with previous GLMs). For all GLMs, the duration of these regressors was considered and set equivalent to the real decision time. For regressors of non-interested in all GLMs, see Table 19 for details.

For each GLM, we creased the individual contrasts of regressors of interest and forwarded them to a one-way flexible factorial ANOVA model in which pairwise (and the reverse) comparisons were performed in terms of the corresponding samples respectively at the group level (see Table 19).

Table 19. Information of GLMs

| GLM | Regressors of non-interested | Target Contrast |
|---|---|---|
| GLM1 | 1-3): onsets of BB, OB, and VB blocks | ***Individual Level:*** |
| (MAIN sample; | (duration equals 143 s; the period from | BBdec vs implicit baseline |
| N = 46) | the offset of the instruction to the onset | OBdec vs implicit baseline |
| | of the instruction of the next block); | VBdec vs implicit baseline |
| | 4) onsets of all transfer phases (duration | ***Group Level:*** |
| | equals 4 s); | OBdec vs BBdec |
| | 5) onsets of all instructions (duration | VBdec vs BBdec |
| | equals 5 s); | OBdec vs VBdec |
| | 6) onsets of stimuli presentation during | |
| | invalid decision phases (i.e., no response | |
| | trials, duration equals the 4 s; trials with | |
| | the decision time less than 200 ms or fair | |
| | offers, duration equals the decision | |
| | time); | |
| | 7-12) headmotion parameters | |
| GLM2 | 1-5): same as GLM1; | ***Individual Level:*** |
| (HELP subsample; | 6) onsets of stimuli presentation during | BBhelp vs implicit baseline |
| N = 42) | invalid decision phases (i.e., keep and | OBhelp vs implicit baseline |
| | punishment choice, duration equals the | VBhelp vs implicit baseline |
| | decision time; no response trials, dura- | ***Group Level:*** |
| | tion equals the 4 s; trials with the deci- | OBhelp vs BBhelp |
| | sion time less than 200 ms or fair offers, | VBhelp vs BBhelp |
| | duration equals the decision time); | OBhelp vs VBhelp |
| | 7-12) headmotion parameters | |
| GLM3 | 1-5): same as GLM1; | ***Individual Level:*** |
| (PUNISH subsample; | 6) onsets of stimuli presentation during | BBpunish vs |
| N = 22) | invalid decision phases (i.e., keep and | implicit baseline |
| | help choice, duration equals the decision | OBpunish vs |
| | time; no response trials, duration equals | implicit baseline |
| | the 4 s; trials with the decision time less | VBpunish vs |
| | than 200 ms or fair offers, duration | implicit baseline |
| | equals the decision time); | ***Group Level:*** |
| | 7-12) headmotion parameters | OBpunish vs BBpunish |

| | | |
|---|---|---|
| | | VBpunish vs BBpunish<br>OBpunish vs VBpunish |
| GLM4<br>(HEPUN subsample;<br>N = 20) | 1-5): same as GLM1;<br>6) onsets of stimuli presentation during invalid decision phases (i.e., keep choice, duration equals the decision time; no response trials, duration equals the 4 s; trials with the decision time less than 200 ms or fair offers, duration equals the decision time);<br>7-12) headmotion parameters | ***Individual Level:***<br>BBhelp vs. BBpunish<br>OBhelp vs. OBpunish<br>VBhelp vs. VBpunish<br>***Group Level:***<br>OB(help-punish) vs BB(help-punish)<br>VB(help-punish) vs BB(help-punish)<br>OB(help-punish) vs VB(help-punish) |

Note: GLM refers to general linear model, dec refers to decision, BB refers to baseline block, OB refers to offender-focused block, VB refers to victim-focused block.

#### *4.2.5.2.3 Explorative functional connectivity analysis*

To further address how attention focus influences the functional connectivity between the bilateral TPJ and other brain regions, we performed exploratory analyses by using the generalized form of context-dependent psycho-physiological interactions analysis (gPPI toolbox: https://www.nitrc.org/projects/gppi). Compared to the standard PPI approach (K Friston, et al., 1997), gPPI spans the whole experimental space which allows modelling of more than two task conditions and furthermore improves the model fit by increasing the specificity for true negative findings and sensitivity for true positive findings (Cisler, Bush, & Steele, 2014; McLaren, Ries, Xu, & Johnson, 2012). To ensure a reasonable interpretation and to maintain sufficient statistical power, we only performed PPI analyses on GLM1 with the left TPJ as the seed region as it was jointly activated in OBdec and VBdec (both compared with BBdec) at the group level. Specifically, the source mask was defined as a sphere with a radius of 4 mm centered at the peak voxel of the corresponding group-level contrasts within the left TPJ mask, which was applied to all participants in the MAIN sample. Afterwards we extracted its time series (physiological terms), which were deconvolved, multiplied by each regressor in that GLM (psychological terms) and then reconvolved with the HRF to generate the PPI terms (Gitelman, et al., 2003). Next, all terms including the 6 head motion parameters were forwarded to a new GLM. The individual contrasts were built, based on parameter estimates for the PPI terms. Finally, a group-level one-sample t-test analysis was performed to identify the brain regions displaying

increased functional connectivity with the seed regions during either OBdec or VBdec (both compared with BBdec).

We reported our results in GLM 1-3 with a cluster-level whole-brain corrected (WBC) threshold of $p < 0.05$ while controlling for family-wise error (FWE) rate with an uncorrected voxel-level threshold of $p < 0.001$ (Eklund, Nichols, & Knutsson, 2016). An a priori TPJ mask (Hutcherson, Bushong, & Rangel, 2015) was used for small volume correction (SVC) given the initial hypotheses. Besides, we used a lenient uncorrected voxel-level threshold of $p < 0.005$ with the extent threshold of 100 for the results of GLM4 (due to the smaller sample size) as well as the explorative PPI analyses. Region labelling and data visualization followed the same procedure as Study 1. In addition, we extracted the beta values of the peak voxels in above-mentioned contrasts for display using the MarsBaR toolbox (http://marsbar.sourceforge.net/).

## 4.3 Results

### 4.3.1 Behavioral Results

As predicted in H1, participants showed higher (lower) help proportion (main effect of attention focus: $F(2,90) = 21.10$, $p < 0.001$, partial $\eta^2 = 0.32$; Post-hoc ps $< 0.01$) but lower (higher) punishment proportion (main effect of attention focus: $F(2,90) = 17.91$, $p < 0.001$, partial $\eta^2 = 0.29$; Post-hoc ps $< 0.01$) in VB (OB), compared with BB, in the MAIN sample (see Figure 20). A similar behavioral pattern was also observed in the rest three subsamples (All $Fs > 9$, ps $< 0.001$, partial $\eta^2$s $> 0.3$; see Table 20 for descriptive summary of choice proportion).

Regarding specific types of choice, it was found that participants in the HELP subsample took longer in deciding to help the victim in OB, compared with either BB or VB (main effect of attention focus: $F(2,82) = 17.23$, $p < 0.001$, partial $\eta^2 = 0.30$; Post-hoc ps $< 0.001$). No other effect was detected neither in transfer amount during help choices in the HELP subsample ($p > 0.06$) nor in both measures during punishment choices in the PUNISH subsample (both ps $> 0.06$). By analyzing both altruistic choices on each measure respectively in the HELPUN sample, we found that participants in general responded slower in OB (main effect of attention focus: $F(2, 38) = 3.75$, $p = 0.047$, partial $\eta^2 = 0.17$; Post-hoc $p = 0.002$, compared with BB) as well as during punishment (main effect of altruistic choice type: $F(1, 19) = 5.84$, $p = 0.026$, partial $\eta^2 = 0.23$; Post-hoc $p = 0.026$). Apart from that we did not observe any other effect on both measures (all $p > 0.06$; see Table 21 for descriptive summary of decision time and transfer amount).

In addition, participants in all (sub)samples reported significantly higher feelings of unfairness to target unequal offers than to filler equal offers (all ts > 23, p < 0.001; see Table 22 for descriptive summary of rating).

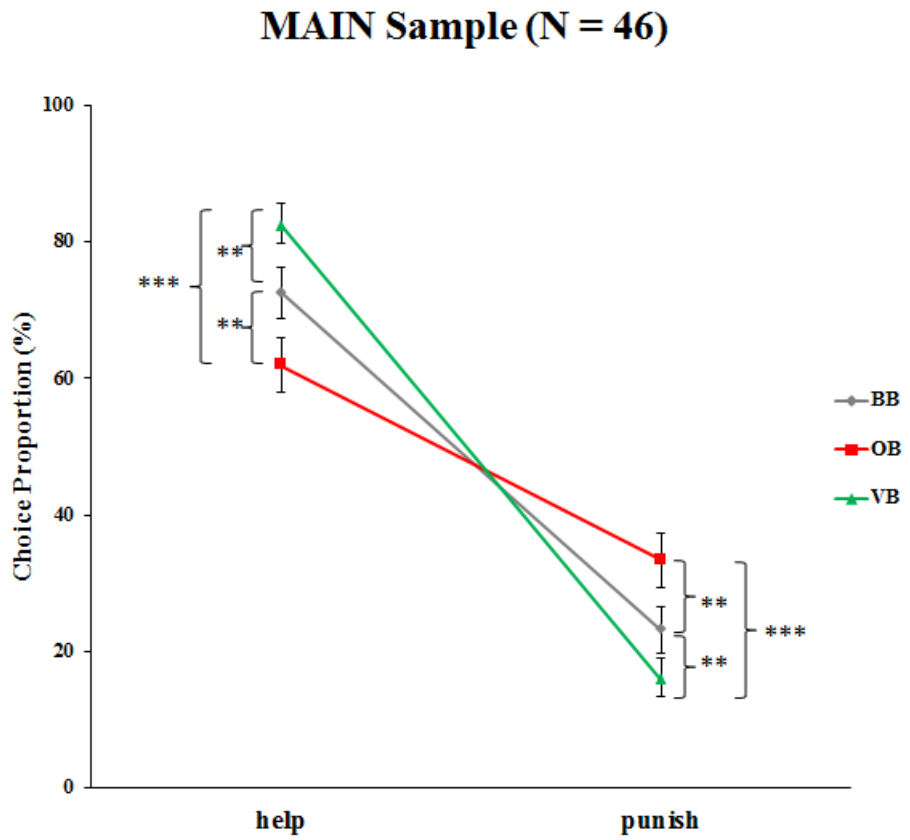


Figure 20. Proportion of altruistic choices in different focus conditions in the MAIN sample. Significance level: *p <0.05, **p < 0.01, ***p < 0.001, Bonferroni correction; Error bars: SEM. Abbreviations: BB = baseline block, OB = offender-focused block, VB = victim-focused block.

Table 20. Descriptive summary of altruistic choice proportion (%) during the fMRI task

| | help | | | punish | | |
|---|---|---|---|---|---|---|
| | Mean (S.D.) | | | Mean (S.D.) | | |
| | BB | OB | VB | BB | OB | VB |
| ***MAIN sample*** | 72.57 | 61.69 | 82.45 | 23.19 | 33.44 | 15.99 |
| (N = 46; GLM1) | (25.99) | (26.86) | (21.22) | (23.33) | (26.73) | (20.44) |
| ***HELP subsample*** | 76.36 | 67.35 | 83.44 | 21.43 | 28.00 | 14.85 |
| (N = 42; GLM2) | (20.23) | (20.99) | (17.65) | (19.45) | (20.25) | (16.49) |
| ***PUNISH subsample*** | 55.75 | 46.86 | 67.21 | 41.88 | 48.91 | 30.85 |
| (N = 22; GLM3) | (19.47) | (18.84) | (21.88) | (19.11) | (20.27) | (20.96) |
| ***HELPUN subsample*** | 59.77 | 51.55 | 69.53 | 37.74 | 43.81 | 28.34 |
| (N = 20; GLM4) | (14.36) | (11.74) | (16.21) | (13.24) | (12.32) | (14.43) |

Note: S.D. refers to standard deviation, BB refers to baseline block, OB refers to offender-focused block, VB refers to victim-focused block.

Table 21. Descriptive of decision time and transfer amount during the fMRI task

| | Target offer with unequal monetary allocation Mean (S.D.) | Filter offer with equal monetary allocation Mean (S.D.) |
|---|---|---|
| ***MAIN sample*** | 5.95 (0.60) | 0.54 (0.80) |
| (N = 46; GLM1) | | |
| ***HELP subsample*** | 5.97 (0.62) | 0.57 (0.82) |
| (N = 42; GLM2) | | |
| ***PUNISH subsample*** | 5.92 (0.58) | 0.54 (0.62) |
| (N = 22; GLM3) | | |
| ***HELPUN subsample*** | 5.91 (0.59) | 0.54 (0.63) |
| (N = 20; GLM4) | | |

Note: S.D. refers to standard deviation, BB refers to baseline block, OB refers to offender-focused block, VB refers to victim-focused block.

Table 22. Descriptive summary of post-scanning rating

| | | BB | OB | VB |
|---|---|---|---|---|
| | | Mean (S.D.) | Mean (S.D.) | Mean (S.D.) |
| ***All valid choices of MAIN sample (N = 46; GLM1)*** | | | | |
| Decision Time (ms) | | 1562.12 (386.98) | 1736.47 (398.72) | 1563.49 (402.20) |
| Transfer Amount (€) | | 2.28 (1.28) | 2.30 (1.27) | 2.50 (1.35) |
| ***Help choices of HELP subsample (N = 42; GLM2)*** | | | | |
| Decision Time (ms) | | 1571.22 (399.38) | 1731.18 (438.56) | 1569.66 (416.95) |
| Transfer Amount (€) | | 2.28 (1.28) | 2.30 (1.27) | 2.50 (1.35) |
| ***Punishment choices of PUNISH subsample (N = 22; GLM3)*** | | | | |
| Decision Time (ms) | | 1814.82 (364.88) | 1901.08 (368.29) | 1945.22 (363.91) |
| Transfer Amount (€) | | 2.09 (0.89) | 2.12 (0.62) | 2.26 (1.05) |
| ***Help and punishment choices of HELP UN subsample (N = 20; GLM4)*** | | | | |
| Decision Time (ms) | *Help* | 1800.85 (375.72) | 1913.37 (418.93) | 1778.45 (420.36) |
| | *Punishment* | 1844.99 (366.26) | 1934.31 (360.70) | 1958.63 (379.40) |
| Transfer Amount (€) | *Help* | 2.13 (1.00) | 2.18 (1.22) | 2.44 (1.39) |
| | *Punishment* | 2.16 (0.91) | 2.15 (0.63) | 2.30 (1.07) |

Note: Unfairness ratings range from 0 (not at all) to 8 (very much); S.D. refers to standard deviation, BB refers to baseline block, OB refers to offender-focused block, VB refers to victim-focused block.

## 4.3.2 Imaging Findings

### 4.3.2.1 General Effect of Attention Focus on Decision-Relevant Activities

As predicted in H2, participants in the MAIN sample (GLM1) showed higher activation in bilateral TPJ during decision-making while considering the unfairness of the offender's behavior (vs. BB). Similarly, we also observed increased decision-relevant activities in the left TPJ while participants took the victim's feeling

into account (vs. BB; see Figure 21). For other activations yielded from above-mentioned and remaining contrasts, see Table 23 for details.

Table 23. Decision-relevant activities reflecting the effect of different attention focus in the MAIN sample (N = 46; GLM1)

| Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***OBdec vs. BBdec*** | | | | | | | |
| TPJ | L | 508 | -54 | -50 | 22 | 40 | 4.71* |
| TPJ | R | 126 | 62 | -46 | 30 | 40 | 3.93† |
| IFG/AI | L | 114 | -46 | 30 | -6 | 47 | 4.14 |
| PCG | L | 274 | -42 | 2 | 58 | 6/8 | 4.52* |
| ***BBdec vs. OBdec*** | | | | | | | |
| No cluster | | | | | | | |
| ***VBdec vs. BBdec*** | | | | | | | |
| TPJ | L | 165 | -50 | -48 | 22 | 40 | 4.07† |
| ***BBdec vs. VBdec*** | | | | | | | |
| No cluster | | | | | | | |
| ***OBdec vs. VBdec*** | | | | | | | |
| ACC/MCC/SMA | B | 626 | 6 | 22 | 46 | 6/8/32 | 5.00* |
| Thalamus/Caudate/ | B | 194 | -2 | -2 | 16 | | 4.95* |
| Lateral Ventricle | | | | | | | |
| ***VBdec vs. OBdec*** | | | | | | | |
| No cluster | | | | | | | |

Note: Threshold is set to $p < 0.001$, $k = 100$, uncorrected; * Significant at $p < 0.05$ family wise error corrected at the cluster level; †Significant at $p < 0.1$ family wise error (FWE) rate corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling template toolbox for SPM8): dec = decision, BB = baseline block, OB = offender-focused block, VB = victim-focused block; L = left, R = right, B = bilateral, BA = Brodmann Area; ACC = Anterior Cingulate Cortex, AI = Anterior Insula, IFG = Inferior Frontal Gyrus, MCC = Mid-Cingulate Cortex, PCG = Precentral Gyrus, SMA = Supplementary Motor Area, TPJ = Temporo-parietal Junction.

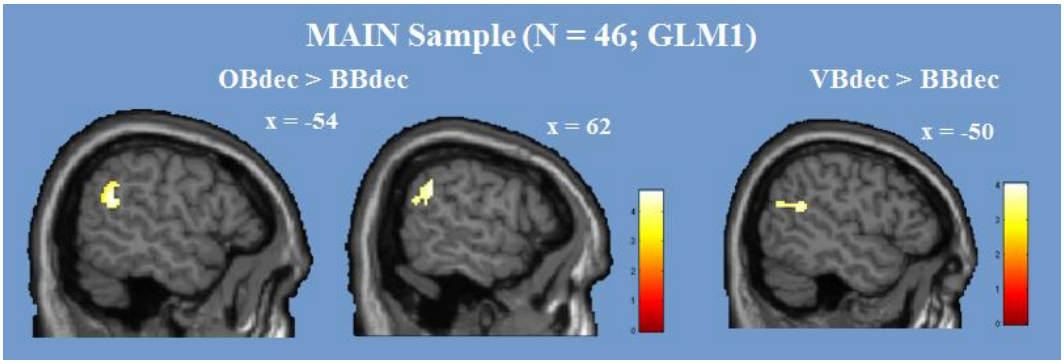


Figure 21. Choice-relevant activities in TPJ reflecting the effect of attention focus. Display threshold: $p < 0.001$ at the voxel-level, uncorrected. Abbreviations: BB = baseline block, OB = offender-focused block, VB = victim-focused block; dec=decision, TPJ = temporo-parietal junction.

#### 4.3.2.2 The Effect of Attention Focus on Activities of Specific Decision

Regarding the effect of attention on help-relevant activities, we found in the HELP subsample (GLM2) that higher activation in the dorsal part of ACC extending to the supplementary motor area (SMA) and bilateral IFG extending to the anterior insula (AI) while participants made help choice in OB, compared to either VB or BB (i.e., OBhelp vs. VBhelp, and OBhelp vs. VBhelp; see Figure 22; also see Table 24 for other activations). For punishment-relevant neural activation, participants in the PUNISH subsample (GLM3) exhibited reduced activation only in the right IFG in OB compared with BB (i.e., OBpunish vs. BBpunish; see Figure 22). No other significant activations were observed in remaining contrasts. In sum, these results were consistent with H3.

Table 24. Help-relevant activities reflecting the effect of attention focus in the HELP subsample (N = 42; GLM2)

| Brain Region | Hemi-sphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***OBhelp vs. BBhelp*** | | | | | | | |
| IFG | L | 217 | -54 | 16 | 6 | 45/47 | 4.46* |
| AI | L | 141 | -28 | 18 | -6 | 13 | 4.54† |
| IFG/AI | R | 420 | 48 | 24 | 4 | 13/45/47 | 5.26* |
| PCG/MFG | L | 291 | -44 | 12 | 46 | 6/8 | 4.40* |
| MFG | R | 128 | 38 | 26 | 38 | 9 | 4.26 |
| ACC/MCC /SMA | B | 173 | 0 | 30 | 44 | 6/8/9 | 4.11* |
| MeFG/SFG | R | 115 | 12 | 6 | 64 | 6/8/9 | 4.04 |
| TPJ | L | 191 | -50 | -48 | 22 | 40 | 4.58* |
| TPJ/IPL | R | 323 | 58 | -46 | 34 | 40 | 4.24* |
| ***BBhelp vs. OBhelp*** | | | | | | | |
| No cluster | | | | | | | |
| ***VBhelp vs. BBhelp*** | | | | | | | |
| No cluster | | | | | | | |
| ***BBhelp vs. VBhelp*** | | | | | | | |
| No cluster | | | | | | | |
| ***OBhelp vs. VBhelp*** | | | | | | | |
| IFG/AI | R | 161 | 42 | 20 | -8 | 13/45/47 | 4.28† |
| IFG/MFG | R | 118 | 38 | 46 | 6 | 10 | 4.50 |
| ACC/MCC/SMA | B | 1104 | 6 | 22 | 46 | 6/8/9/32 | 5.13* |
| IPL | R | 214 | 54 | -50 | 42 | 40 | 4.42* |
| Caudate/Lateral Ventricle | B | 191 | -4 | -2 | 16 | | 4.38* |
| ***VBhelp vs. OBhelp*** | | | | | | | |
| No cluster | | | | | | | |

Note: Threshold is set to $p < 0.001$, $k = 100$, uncorrected; * Significant at $p < 0.05$ family wise error corrected at the cluster level; †Significant at $p < 0.1$ family wise error (FWE) rate corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling template toolbox for SPM8): BB=baseline block, OB=offender-focused block, VB=victim-focused block, L=left, R=right, B=bilateral, BA=Brodmann Area; ACC = Anterior Cingulate Cortex, AI = Anterior Insula, IFG = Inferior Frontal Gyrus, IPL = Inferior Parietal Lobule, MCC = Mid-Cingulate Cortex, MFG = Middle Frontal Gyrus, MeFG = Medial Frontal Gyrus, PCG = Precentral Gyrus, SFG = Superior Frontal Gyrus, SMA = Supplementary Motor Area, TPJ = Temporo-parietal Junction.

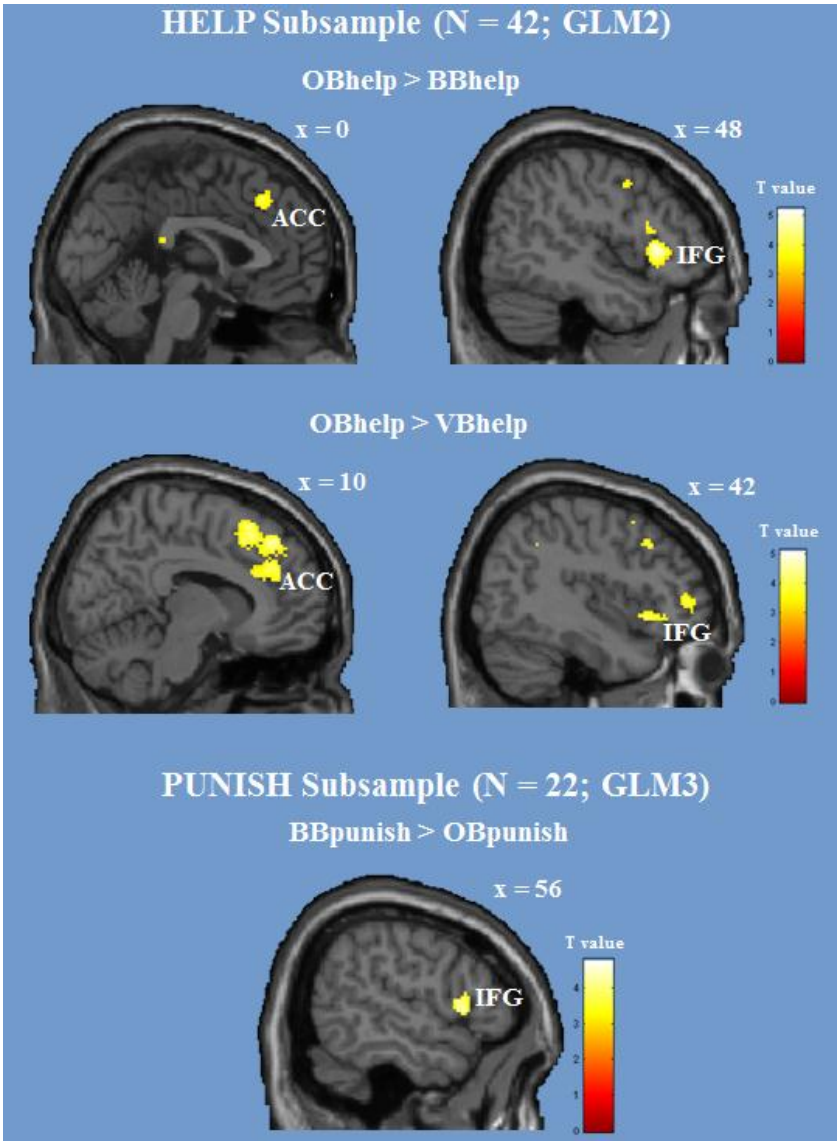


Figure 22. Regions reflecting the conflict between the effect of attention focus and the choice. Display threshold: $p < 0.001$ at the voxel-level, uncorrected. Abbreviations: BB = baseline block, OB = offender-focused block, VB = victim-focused block, ACC = anterior cingulate cortex, IFG = inferior frontal gyrus.

#### 4.3.2.3 Interaction Effect on Decision-Relevant Activities

Focusing on the HELPUN subsample (GLM 4), we found that the differential activation in the right IFG extending to AI between help and punishment was higher when participants considered the offender's unfairness,compared to deciding naturally (i.e., [OBhelp − OBpunish] vs [BBhelp − BBpunish]; see Figure 23). Besides, helping in OB resulted in stronger activitions in the dACC/SMA as well

as the right inferior parietal lobule in OB than that in VB (i.e., [OBhelp − OBpunish] vs [VBhelp − VBpunish]; see Table 25 for other activations).

Table 25. Differential activities between help vs. punishment reflecting the effect of attention focus in the HELPUN subsample (N = 20; GLM4)

| Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***OB(help - punish)* vs. *BB(help - punish)*** | | | | | | | |
| IFG/AI | R | 493 | 48 | 20 | 14 | 13/44/45/47 | 4.42† |
| IPL | R | 129 | 52 | -44 | 34 | 40 | 3.22 |
| ***BB(help - punish)* vs. *OB(help - punish)*** | | | | | | | |
| No cluster | | | | | | | |
| ***VB(help - punish)* vs. *BB(help - punish)*** | | | | | | | |
| MTG/MOG/SOG | R | 101 | 28 | -68 | 20 | 31 | 3.35 |
| ***BB(help - punish)* vs. *VB(help - punish)*** | | | | | | | |
| No cluster | | | | | | | |
| ***OB(help - punish)* vs. *VB(help - punish)*** | | | | | | | |
| IFG/AI | R | 136 | 48 | 24 | 2 | 13/45/47 | 3.43 |
| ACC/OFC | B | 167 | -4 | 28 | -4 | 10/24/32 | 4.08 |
| SFG/SMA/MCC | B | 1134 | 18 | 14 | 48 | 6/8 | 4.83* |
| IPL | R | 356 | 54 | -56 | 48 | 40 | 4.22† |
| ***VB(help - punish)* vs. *OB(help - punish)*** | | | | | | | |
| No cluster | | | | | | | |

Note: Threshold is set to $p < 0.005$, $k = 100$, uncorrected; * Significant at $p < 0.05$ family wise error (FWE) corrected at the cluster level; †Significant at $p < 0.1$ family wise error (FWE) rate corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling template toolbox for SPM8): BB = baseline block, OB = offender-focused block, VB = victim-focused block; R = right, BA = Brodmann Area; AI = Anterior Insula, IFG = Inferior Frontal Gyrus, MCC = Mid-Cingulate Cortex, MOG = Middle Occipital Gyrus, MTG = Middle Temporal Gyrus, SFG = Superior Frontal Gyrus, SMA = Supplementary Motor Area, SOG = Superior Occipital Gyrus.

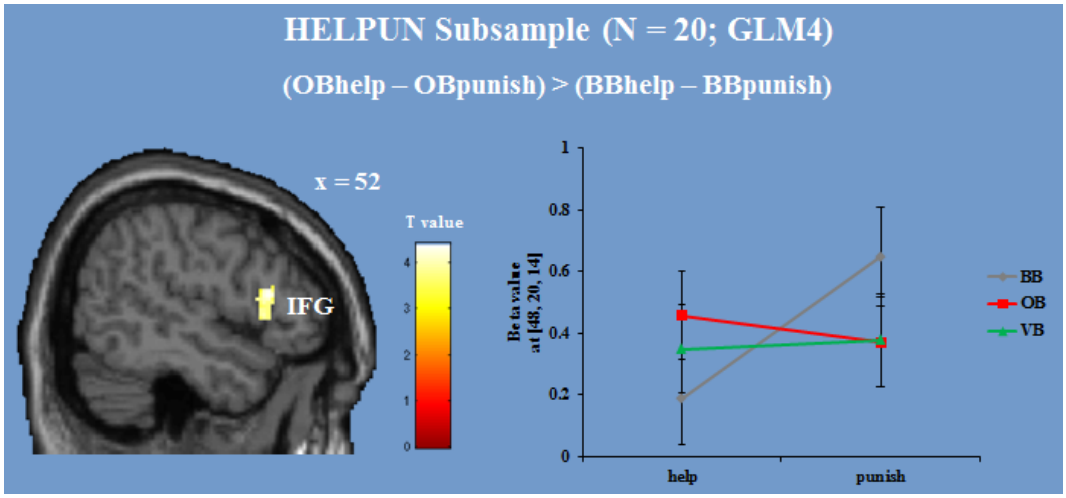


Figure 23. IFG reflecting the interaction between the effect of attention focus (in OB) and the choice in the HELPUN subsample. The line plot showed the beta value in the peak voxel of the right IFG in all conditions, only with the goal of illustration. Display threshold: $p < 0.005$ at the voxel-level, uncorrected; Error bars: SEM. Abbreviations: BB = baseline block, OB = offender-focused block, VB = victim-focused block, IFG = inferior frontal gyrus.

#### 4.3.2.4 Focus-Dependent Functional Connectivity During Decision-Making

Given the results of GLM1, the conjunction analyses showed that only the left TPJ was more activated during decision-making in OB and VB compared with BB (i.e., GLM1: conjunction between OBdec vs. BBdec and VBdec vs. BBdec, MNI coordinates of peak voxel: -50/-48/22). Hence, we only performed the exploratory PPI analyses in GLM1 with the left TPJ as seed regions. We found that the left AI/IFG exhibited an enhanced connectivity with the left TPJ during the decisions making in OB compared with BB (i.e., OBdec > BBdec), which also held true in the contrast of VBdec vs. BBdec with a more lenient threshold ($p < 0.005$ uncorrected with $k = 100$; see Figure 24; also see Table 26 for other activations).

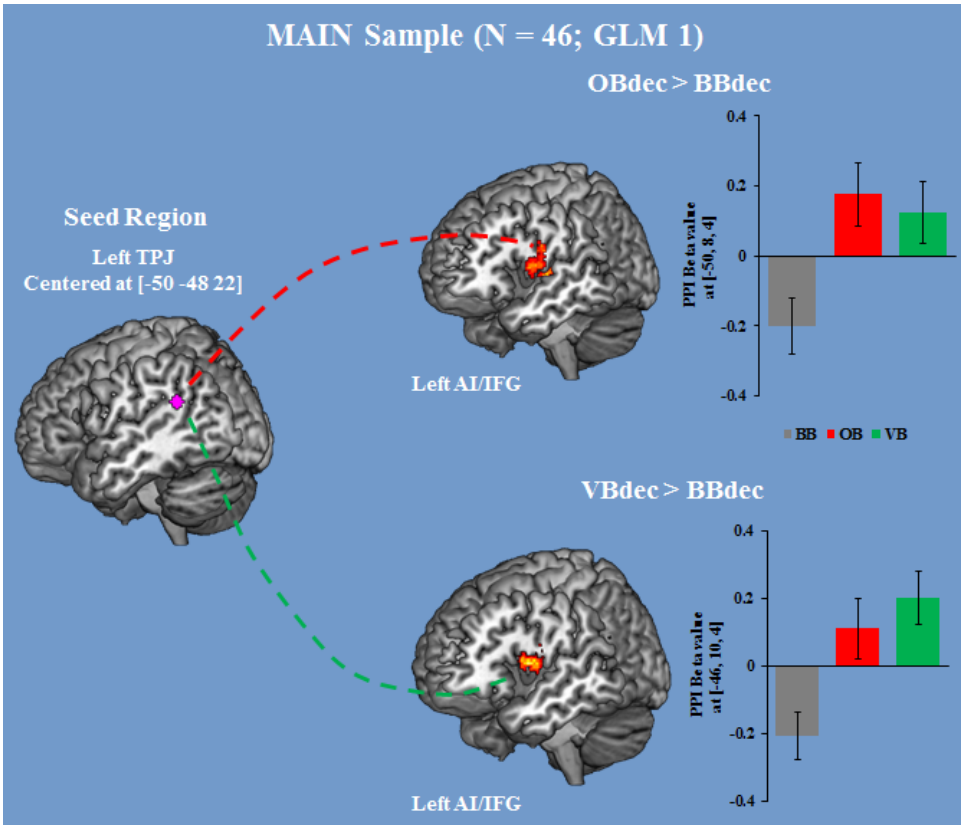


Figure 24. Regions reflecting enhanced functional connectivity with left TPJ during decisions in OB and VB (vs. BB respectively) in the MAIN sample. Bar plots showed the beta value of PPI in the peak voxel of left AI/IFG with left TPJ in all conditions, only with the goal of illustration. Display threshold: $p < 0.005$ at voxel-level, uncorrected; Error bars: SEM. Abbreviations: PPI = psycho-physiological interaction, BB = baseline block, OB = offender-focused block, VB = victim-focused block, dec = decision, AI = anterior insula, IFG = inferior frontal gyrus, TPJ = temporo-parietal junction.

Table 26. Regions reflecting enhanced functional connectivity with the left TPJ during decision-making in OB or VB (both vs. BB) in the MAIN sample (N = 46; GLM1)

| Brain Region | Hemisphere | Cluster Size | MNI Coordinates | | | BA | T-value |
|---|---|---|---|---|---|---|---|
| | | | x | y | z | | |
| ***OBdec vs. BBdec*** | | | | | | | |
| IFG/AI/Thalamus/ Putamen/Lateral Ventricle | L | 1847 | -50 | 8 | 4 | 13 | 3.47* |
| Thalamus | R | 329 | 16 | -24 | 10 | | 3.72† |
| Caudate/Putamen/Insula | R | 153 | 32 | 20 | 18 | 13 | 3.83 |
| MFG/SMA | R | 124 | 18 | -8 | 66 | 6 | 3.76 |
| ***BBdec vs. OBdec*** | | | | | | | |
| No cluster | | | | | | | |
| ***VBdec vs. BBdec*** | | | | | | | |
| IFG/AI/PCG | L | 196 | -46 | 10 | 4 | 13/44 | 3.77 |
| PoCG/PCG | R | 178 | 64 | -14 | 38 | 2/3/4/6 | 3.56 |
| ***BBdec vs. VBdec*** | | | | | | | |
| No cluster | | | | | | | |

Note: Threshold is set to $p < 0.005$, $k = 100$, uncorrected; * Significant at $p < 0.05$ family wise error corrected at the cluster level; †Significant at $p < 0.1$ family wise error (FWE) rate corrected at the cluster level.

Abbreviations (brain regions are labeled according to the automated anatomic labeling template toolbox for SPM8): dec=decision, BB=baseline block, OB=offender-focused block, VB=victim-focused block, L=left, B=bilateral, BA=Brodmann Area; AI = Anterior Insula, IFG = Inferior Frontal Gyrus, MFG = Middle Frontal Gyrus, PCG = Precentral Gyrus, PoCG = Postcentral Gyrus, SMA = Supplementary Motor Area, TPJ = Temporo-parietal Junction.

## 4.4 Discussion

### 4.4.1 The Effect of Attention Focus on (Altruistic) Choice Preference in a Third-Party Context

Behavioral results on choices showed that manipulating the focus of third-party deciders is an effective way to reshape their choice pattern, namely that they increased (reduced) the frequency to punish (help) while considering the unfairness of offenders but were more (less) likely to help (punish) when they thought about the feelings of the victim. These findings are consistent with our prediction in H1 and previous studies testing the causal relationship between attention focus and choice in other domains. For instance, hungry but non-dieting participants were

more likely to choose healthy but less delicious food when they considered the healthinessof the food during decision-making, compared with their choices when deciding naturally (Hare, et al., 2011). In a recent unpublished study, participants, as the role of proposer in the dictator game, became more generous in sharing the money with the unknown recipients when they thought more on either the right thing to do or the recipient's feeling (Hutcherson & Rangel, 2014). Our current study thus extends such attention-induced effect on decision-making to an incentivized third-party context. Moreover, we observed the similar results across different subsamples, indicating that this effect could hardly be affected by individual difference.

### 4.4.2 TPJ: A Key Region Reflecting the Effect of Other-regarding Focus during Decision-making

As predicted in H2, decision-relevant TPJ activation was stronger in both conditions where participants were asked to consider aspects of other parties, either the offender's unfair behavior or the victim's feeling. TPJ has been shown to be in close link with theory-of-mind (ToM)/mentalizing ability in a large amount of literature (Schaafsma, et al., 2014; Schurz, et al., 2014). Since that either OB or VB reuired more perspective-taking process, highly relevant with mentalizing ability, compared with deciding naturally, it is plausible to label the activation of TPJ during decision-making (esp. help choice) as the mark of mentalizing in these two conditions. However, this explanation should be treated cautiously due to the reverse inference problem (Poldrack, 2006) and the multifunction of TPJ, which is not limited to ToM but extend to other cognitive ability like attention reorientation (Corbetta, et al., 2008; Lee & McCarthy, 2014; Mars et al., 2012; Mitchell, 2008).

### 4.4.3 Engagement of Control Network in Modulating the Decision Process Influenced by Attention Focus

In line with our prediction in H3, we found that the control neural network, especially the IFG/AI and the dACC/SMA, was strongly involved in resolving the conflict between the goal of specific other-regarding focus and specific choice made. Particularly, both regions were strongly activated when third parties decided to help the victim in OB, compared with either VB or BB. The right IFG/AI also displayed lower activation during punishment choice when participants considered the offender's unfair behavior than when they decided without any specific focus. The reverse activation pattern of IFG/AI during different altruistic choices in OB and BB reasonably yielded its significant interaction in the HELPUN

subsample. Previous studies have already shown that the IFG/AI contributed crucially to cognitive control, such as response inhibition, or task switching (Aron, Robbins, & Poldrack, 2004; Nelson et al., 2010); but also see (Aron, Robbins, & Poldrack, 2014) and was considered as one of the key regions in the ventral attention network (Dosenbach, et al., 2008; Vossel, Geng, & Fink, 2014). More relevantly, Hare and colleagues (2011) observed the stronger activation in the IFG/AI when participants took into account the healthiness of food in a food choice task, indicating its potential role in reflecting the attention-induced change during decision-making. Together with the results of attention-induced change of choice (i.e., higher punishment proportion but lower help proportion in OB), these findings suggest that this region could be the neural hub for modulating choices depending on different attention focus.

Notably, the dACC/SMA, a region tightly associated with cognitive control (Shackman et al., 2011), especially the monitor and resolution of conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Botvinick, Cohen, & Carter, 2004), also participated in the modulation of attention focus during help choices. Given the fact that the process of conflict resolution was usually reflected by the change of decision time, such as the classical "Stroop" task (MacDonald, Cohen, Stenger, & Carter, 2000; Pardo, Pardo, Janer, & Raichle, 1990), we could probably explain the engagement of dACC in our results in modulating the decision-making process, not the choice per se, given different focus condition. Consistent with this interpretation, we also observed the significant delay during help choice in OB, compared with either VB or BB.

### 4.4.4 Cross-Talk between TPJ and Control Network during Decision Process Dependent on Attention Focus

Our explorative PPI analyses showed that the left TPJ enhanced its task-dependent functional connectivity with the left AI/IFG during decision-making in either OB or VB (vs. BB). Consistently, two recent studies also showed a strong functional coupling between TPJ (esp. the anterior and right part) and bilateral AI extending to IFG via connectivity-based data-driven analyses (Bzdok et al., 2013; Mars, et al., 2012). Given the role of the AI/IFG in attention and cognitive control, our PPI results indicate that ToM processes are connected to inhibitory control reflected by the enhanced functional link between TPJ and AI/IFG, which influences third parties' (altruistic) choices. This finding also sheds light on the mechanism underlying the shift of attention-dependent altruism preference from a network perspective. However, we could not rule out alternative explanations as the AI is also involved in equally relevant psychological functions such as empathy (Bernhardt & Singer, 2012; Fan, Duncan, de Greck, & Northoff, 2011; Gu et al., 2012).

### 4.4.5 Limitations

To begin with, we had comparatively few observations of punishment choice especially when participants were instructed to think about the feeling of the victim, which thus led to the unreliable estimation of the BOLD signal for this condition due to insufficient statistical power. Since participants could voluntarily make the decisions in the current design, it is not possible to control their choice. Nevertheless, previous studies also showed that participants preferred helping to punishing if both options were available (Jordan, Hoffman, Bloom, & Rand, 2016; Lotz, Okimoto, et al., 2011). Future studies could use cases with other types of more severe norm violations (e.g., criminal scenes), instead of the fairness norm violation, which might help to elicit more punishment behaviors (Buckholtz, et al., 2008; Krueger, et al., 2014). A relevant disadvantage is that we could investigate the effect of attention focus on specific altruistic choice only based on (sub)samples with different participants (either the number or the identity), which sets limitations to the consistency as well as the generalization of our findings.

### 4.4.6 Summary

In sum, Study 3 provides the first empirical evidence of how simple manipulation in changing the focus on different aspects of a third-party incentivized context affects the way of third-party observers adopt to costly intervene in a situation of injustice, which helps to maintain the justice and social norm. Furthermore, it tries to clarify the neural basis underlying this modulation, with the help of the fMRI technique, and indicates the role of TPJ and control network (esp. AI/IFG, dACC/SMA) in this attention-dependent decision process. Regarding the practical implications, we hope to shed light on our understanding of the process behind legal judgment, as it is a highly complex decision-making process in a social context that is easily influenced and reshaped by different attention foci.

# 5 Study 4: The Cognitive Basis Underlying Third-Party Altruistic Decision-Making

## 5.1 Hypotheses

Based on previous findings and our research questions, we derived the following three sets of hypotheses. For each set of hypotheses, three sub-hypotheses were provided basedb on different dependent measures in which we were interested (i.e., choice proportion, processing speed/general processing depth, and fixation-based attention proportion).

$H_1$: Empathic concern will influence third-party altruistic decisions and information search patterns in the baseline block (BB).

*$H_{1a}$: Third parties will choose to help more often with an increasing empathic concern level and choose to punish more often with a decreasing empathic concern level in the BB.*

*$H_{1b}$: Decision time (DT) and the number of fixations when deciding to help will be smallest for those participants with a high empathic concern level, whereas they will be largest for punishment decisions in the BB.*

*$H_{1c}$: Higher empathic concern level will result in a higher proportion of fixation towards victim-relevant payoff information in the BB.*

$H_2$: Directing attention towards different focuses will influence third-party altruistic choices and information search patterns.

*$H_{2a}$: Third parties will choose to help more often in the victim-focused block (VB) and will choose to punish more often in the offender-focused block (OB), compared to the BB.*

*$H_{2b}$: The DT and number of fixations in help (punishment) decisions in the VB will be the smallest (largest) and those in the OB will be the largest (smallest).*

*$H_{2c}$: Third parties will show an increase in the proportion of fixation directed towards victim-related payoff information in the VB and a decrease in the OB (compared with the BB).*

$H_3$: The empathic concern level will modulate the attention effect on third-party altruistic choices and information search patterns.

*$H_{3a}$: The empathic concern level will positively correlate with the help proportion in the VB and the BB; the slope in the VB will thereby be larger than in the BB. The empathic concern level will negatively correlate with the punishment proportion in the OB and BB; the slope in the OB will be larger than that of the BB.*

*$H_{3b}$: An increase in empathic concern level will lead to a decrease in DT and the number of fixations in decisions to help. This effect will be strong in the BB and even stronger in the VB. A higher empathic concern level will result in a longer DT and a larger number of fixations in punishment decisions. This effect will be particularly strong in the BB and even stronger in the OB.*

*$H_{3c}$: A higher empathic concern level will result in a higher proportion of fixation towards victim-relevant payoff information. This effect will be accentuated in the VB.*

## 5.2 Methods

### 5.2.1 Participants

Forty-seven healthy German participants were recruited as third parties for the current eye-tracking study (17 males: mean age = 24.26 ± 6.02 yrs). They were recruited via the Online Recruitment System for Economic Experiments (Greiner, 2004). Following the rule of "no deception" in experimental economics, we recruited additional 47 pairs of first (i.e., offender) and second party (i.e., victim) from the same subject pool for an online study to collect the real choices and used them as the stimuli for the eye-tracking study.

### 5.2.2 Online Decision Collection

Online choices from the first-party (i.e., offender) were collected 5 days before the eye-tracking study via Unipark (http://www.unipark.com/de/). The online task basically followed the procedure used in the Study 1, but with the following exceptions. First, all 94 participants played the role of dictators (i.e., offender) and they were told that some of their choices (i.e., half of them) would be chosen for the other experiment. In particular, each of the selected choice was matched with a real person as the recipient (i.e., victim) and then forwarded to a third person (i.e., third-party decider, namely participants in the eye-tracking part), together with the initials of both parties, who could affect their final payoffs. Second, each participant was presented with 99 binary decision tasks. In each task, participants needed

to choose one of the two given options which characterized the different money split between themselves and another anonymous recipient. Third, we intentionally paired the target option (i.e., the unequal monetary split used in the eye-tracking part as stimulus) with an unattractive option (i.e., an equal split but with very low joint payoff) in each binary decision task. In this way participants' choices were biased so that we obtained all target choices.

### 5.2.3 Eye-Tracking Stimuli

In terms of the real decisions from the offenders in the online part, we first created the template including 28 unequal splits with different money allocations as the target trials plus 5 equal splits as the filler trial for the eye-tracking study. Unequal template splits were selected using the following rules: (a) the sum payoff ranges from €6 to €14; (b) the payoff for both offender and victim should be less than €10 to avoid the difference in eye-movements due to the different digits in payoff (i.e., splits like (10,0), (11,2), (12,1) were not used) as well as the proportion (i.e., splits like (6,0), (7,0), (8,0), (9,0) were not used; see Table 27 for details).

Each template split occurred once in each of the three block but with slight different form, to increase the complexity of the stimuli and reduce the repetition. In specific, we further modified the template stimuli by adding a random fluctuation within the range of +/- 0.2 on the integer so that it is with two more digits. As a consequence, three differential sets of stimuli were generated with the same average payoff. We randomly assigned these three sets of stimuli to each of the focus blocks (i.e., conditions) respectively across participants in the eye-tracking study (see Appendix Table 3 for all stimuli).

Table 27. The template stimuli for the eye-tracking study

| Total Payoff (Offender, Victim; €) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| (3,3)* | | | | | | | | |
| (4,2) | (4,3) | (4,4)* | | | | | | |
| (5,1) | (5,2) | (5,3) | (5,4) | (5,5)* | | | | |
| | (6,1) | (6,2) | (6,3) | (6,4) | (6,5) | (6,6)* | | |
| | | (7,1) | (7,2) | (7,3) | (7,4) | (7,5), | (7,6) | (7,7)* |
| | | | (8,1) | (8,2) | (8,3) | (8,4) | (8,5) | (8,6) |
| | | | | (9,1) | (9,2) | (9,3) | (9,4) | (9,5) |

Note: * Equal split is used in the filler trial.

### 5.2.4 Eye-Tracking Paradigm

The eye-tracking session included three blocks differing in instructed focuses, namely the baseline block (i.e., deciding naturally; BB), the offender-focused block (i.e., thinking about the (un)fairness of the offender's behavior and its link with social norm before deciding; OB), or the victim-focused block (i.e., thinking about the feeling of victim affected by the offender's behavior; VB). In order to remove the proactive inhibition from the previous condition, we deliberately put the BB as the 1st block and introduce another two blocks with the order counter-balanced across participants to them only when they completed the BB. Thus, participants were given the instruction of BB, including the general information about the previous online task and the upcoming eye-tracking third-party task, once they arrived. For instructions of another blocks (i.e., OB, VB), participants were only informed at the beginning of each block after BB.

Within each block, participants started with five practice trials to familiarize themselves with the display and response after reading the instruction, and then performed 33 incentive trials. The trial procedure was similar to that used in Study 3, with the following exceptions (see Figure 25). First, all texts were in white which avoided the confounding factor of lower-level features driving different fixations. Second, the relative payoff information (%) was added together with the absolute payoff information and they were put in a white ellipse equally divided into four parts. In this way we could increase the amount of information so that we gained more fixations and meanwhile the fixations towards the specific piece of information were easily separated. Third, we kept the same display within each participant but balanced the location of the offender as well as the victim together with their payoff information (i.e., both absolute and relative) across participants to rule out the eye-movement effect led by specific display of information (see Figure 26). Fourth, neither did the decision phase nor the transfer phase has a time limitation. Fifth, a 2s black phase was adopted at the beginning of each trial to refresh the eye-movement pattern. Stimuli were displayed on either a 17" or a 19" color monitor with a native resolution of 1280 × 1024 and presented via Presentation 14.9 (Neurobehavioral System, Albana, Canada). The pixel size of the information presented was kept constant with all three eye-trackers.

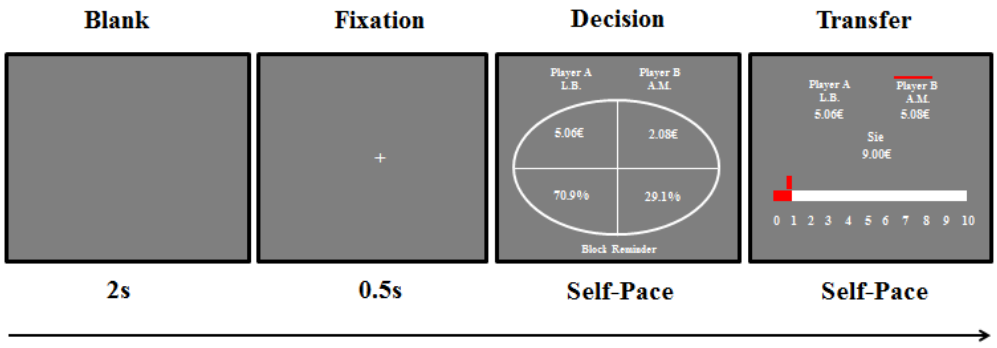


Figure 25. Example of the trial procedure. In this example, the participant added € 1 to the victim (i.e., A.M., labeled as Player B) instead of subtract the money from the offender (i.e., L.B., labeled as Player A).

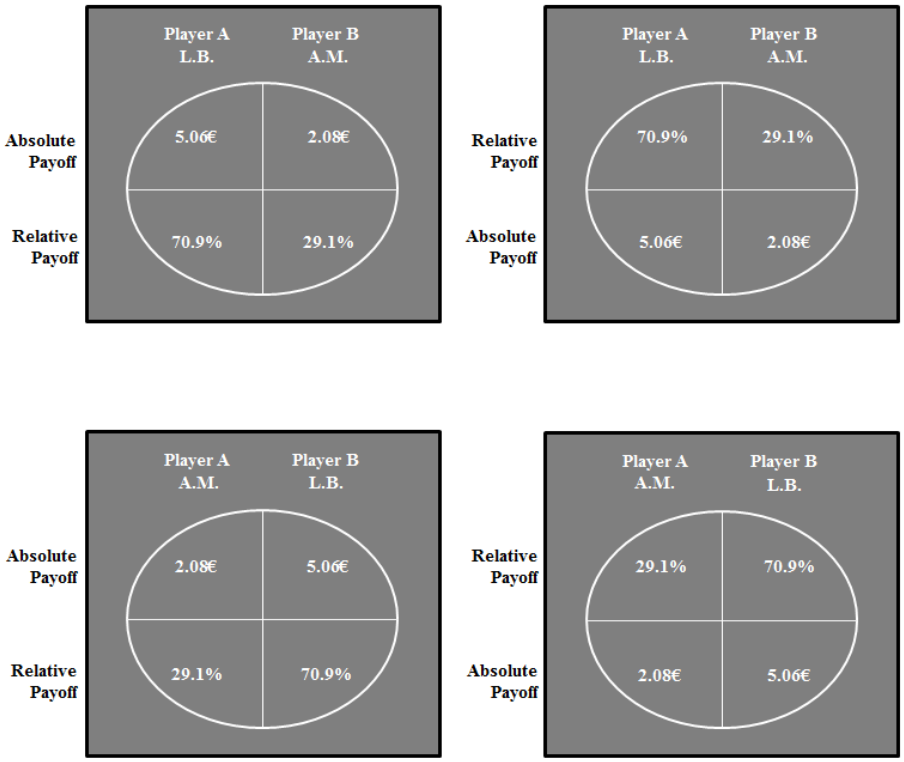


Figure 26. Four types of display used in the current study. Upper-left: the offender (victim) labeled as Player A (B) with the absolute (relative) payoff listed in the upper (lower) half of the ellipse. Upper-right: the offender (victim) labeled as Player A (B) with the absolute (relative) payoff listed in the lower (upper) half of the ellipse. Lower-left: the offender (victim) labeled as Player B (A) with the absolute (relative) payoff listed in the upper (lower) half of the ellipse. Lower-right: the offender (victim) labeled as Player B (A) with the absolute (relative) payoff listed in the lower (upper) half of the ellipse.

### 5.2.5 Procedure

At least 12 hours prior to the eye-tracking experiment, all participants (i.e., third parties in the eye-tracking session) completed the online measurement of empathic concern (Mean ± S.D. = 18.39 ± 3.72; Range: 9 to 26) by IRI (same as used in previous studies) as well as demographical questions via Unipark. Upon arrival, they were given the instruction about the BB and then completed the task in the BB with both their behaviors and eye-movement recorded. Unknown to them, they were asked to do the same task in another two blocks (i.e., OB and VB). After the decision task, participants finished another task which would be reported separately. Finally, participants received a € 5 participation fee and one trial was randomly selected to pay all three parties correspondingly.

### 5.2.6 Data Collection

Behavioral data were collected via Presentation 14.9. Data of gaze behavior were recorded by the eye gaze binocular system (LC Technologies; see Figure 27) with a remote binocular sampling rate of 120 Hz and an accuracy of about 0.45˚.

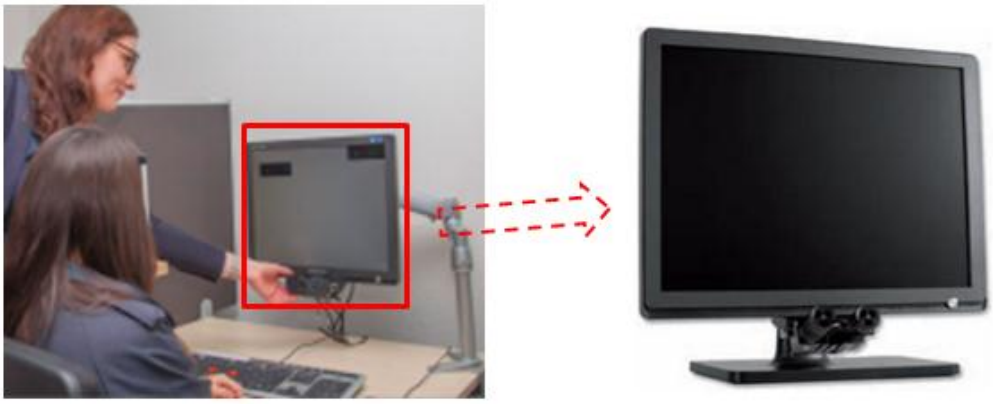

Figure 27. Illustration for the LC eye gaze binocular system in the Decision Lab, Max Planck Institute for Research on Collective Goods, Bonn. Source for the left figure (with small adaptation): https://www.coll.mpg.de/node/7417; source for the right figure: http://eyegaze.com/wp-content/uploads/EAS%20Binocular%20Technical%20Specifications.pdf.

### 5.2.7 Data Analyses

#### 5.2.7.1 Pre-processing & Areas of Interested (AOI) of Eye-tracking Data

We pre-processed the raw data with an in-house algorithm to define valid fixations, namely periods of relatively stable gazes (located within a radius of 30 pix-

els) between two saccades lasting at least 50 ms. To ensure the fixations of later analyses locating within the information we were most interested, we limited all fixations within the following four non-overlapping areas of interest (AOIs), namely one square area (i.e., size: 100×100 pixels) covering the information of either absolute/relative payoff of the offender or the victim respectively (see Figure 28). Besides, we also created three text AOIs (i.e., two AOI covering either the initial of the offender or the victim with the size of 100×100 pixels; one AOI covering the reminder of each block with the size of 390×100 pixels) which were used as the additional criteria to check the quality of the eye-tracking measurement. Moreover, we combined the absolute and relative payoff AOI for the offender and the victim respectively, to simplify the later analyses.

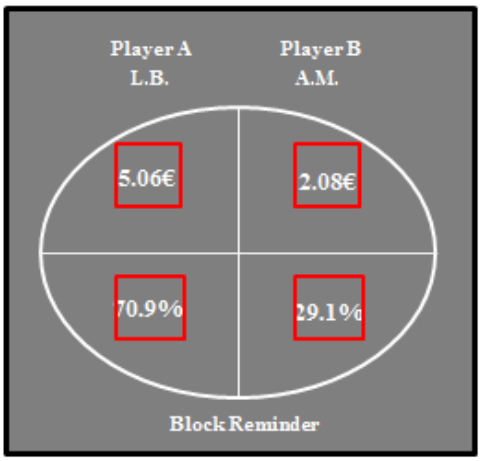


Figure 28. Illustration for the payoff-relevant AOIs (marked with red frame).

### 5.2.7.2 Exclusion Criteria for Different Analyses

For choice behavior, we adopted the 3,864 out of 4,653 trials in total for analyses by excluding trials with the missing response of empathic concern scale (99 trials, appro. 2.1%) together with filler trials (i.e., trials with equal splits; 690 trials, appro. 14.8%).

On top of that, we applied additional rules for excluding trials in analyses of process measures (i.e., fixation numbers/decision time, fixation proportion), namely 1) trials with extreme decision time (i.e., < 200 ms or > mean + 3SD of certain participant; 95 trials, appro. 2.0%); 2) trials with poor recording of eye-movement (i.e., trials with 50% fixations falling outside of all AOIs or those with all fixations within text AOIs; 472 trials, appro. 10.14%); 3) trials with keep choices (i.e., trials in which participants transferred € 0; 802 trials, appro. 17.2%).

#### 5.2.7.3 General Statistical Approach

We performed all statistical analyses via STATA 13 (College Station, TX: StataCorp LP). Given the repeated measure panel data (i.e., multiple observation for each condition per participant) and unbalanced observations within each condition (i.e., different numbers of each choice per participant), we adopted the mixed-effect regression (i.e., linear or logistic regression; main STATA command: xtlogit or xtreg) which treated the effect of participant as the random effect.

Our analyses were performed in a way to directly test the proposed hypotheses. In particular, we highlighted the following measures as dependent variables, which included choice behavior (H1a, H2a, H3a), fixation number in 4 payoff-relevant AOIs along with decision time as the measure of the general information search behavior (H1b, H2b, H3b), and fixation proportion towards the victim payoff-relevant AOIs as the measure of attention distribution towards the specific piece of information (H1c, H2c, H3c). In addition we also checked other measures (i.e., behavior: transfer amount of altruistic choices; eye-movement: the distribution of the first- and the last fixation) as the explorative analyses. Notably, the fixation numbers and decision times were log-transformed before the analyses since they were not normally distributed (Jarque-Bera (S-K) test: fixation number: $\chi^2(2) = 1237.52$, $p < 0.001$; decision time: $\chi^2(2) = 1461.52$, $p < 0.001$). For descriptive summary of all measures mentioned above, see Table 28 for details.

Table 28. Descriptive summary of all measures

| | BB | | OB | | VB | |
|---|---|---|---|---|---|---|
| | Mean (S.D.) | | Mean (S.D.) | | Mean (S.D.) | |
| | help | punish | help | punish | help | punish |
| Choice Proportion (%)[a] | 58.2 | 13.5 | 58.5 | 16.4 | 62.7 | 11.1 |
| | (34.9) | (25.2) | (37.5) | (27.9) | (37.1) | (23.8) |
| Transfer Amount (€)[a] | 1.67 | 2.01 | 1.63 | 1.58 | 1.67 | 1.73 |
| | (0.87) | (1.51) | (0.92) | (0.75) | (1.04) | (1.43) |
| Decision Time (s)[b] | 3.05 | 3.41 | 2.12 | 2.95 | 2.02 | 2.85 |
| | (1.36) | (1.46) | (1.16) | (2.08) | (1.13) | (1.71) |
| Fixation Number[b] | 9.86 | 11.06 | 7.84 | 10.01 | 7.38 | 9.94 |
| | (4.27) | (5.70) | (4.15) | (6.27) | (4.06) | (6.59) |
| Fixation Proportion towards victim-payoff AOI (%)[b] | 61.05 | 44.21 | 60.45 | 42.24 | 59.01 | 39.84 |
| | (6.49) | (9.85) | (10.97) | (16.07) | (13.75) | (16.64) |
| First Fixation Proportion towards victim-payoff AOI (%)[b] | 72.01 | 57.13 | 74.73 | 50.18 | 75.04 | 60.80 |
| | (36.01) | (42.70) | (34.25) | (43.11) | (35.97) | (44.90) |
| Last Fixation Proportion towards victim-payoff AOI (%)[b] | 77.68 | 19.18 | 66.30 | 24.03 | 66.63 | 16.17 |
| | (13.96) | (26.52) | (21.47) | (32.74) | (21.83) | (25.74) |

Note: [a] The total observation equals 3864 (trials). [b] The total observation equals 2945 (trials) due to additional criteria for process measures (see Supplementary Information for details). S.D. refers to standard deviation.

## 5.3 Results

### 5.3.1 Baseline Block (BB)

The goal of this part analyses was to investigate the effect of empathic concern on altruistic choices and eye-movement during decision-making in the BB (H1a-c). Thus the only main predictor for the following analyses was the empathic concern level. Besides, the trial as an index of time was also added to these regression analyses to rule out the effect of practice.

#### 5.3.1.1 Choice Behavior

We found that third parties with a higher level of empathic concern were more likely to help the victim (Odds ratio = 1.20, z = 1.93, p = 0.053) but less likely to punish the offender (Odds ratio = 0.77, z = 1.92, p = 0.055), as we predicted in H1a. Besides, we did not observe the effect of empathic concern on the behavior of keep (i.e., whether to intervene or not; Odds ratio = 0.96, z = 0.46, p = 0.650; see Figure 29A; also see Table 29 for regression details).

Table 29. Results of repeated-measure mixed-effect logistic regression predicting the help, punishment or keep choice by empathic concern with the time effect (i.e., trials) controlled in the baseline block (BB)

| | help | punish | keep |
|---|---|---|---|
| Empathic concern | $1.215^{+}$ | $0.769^{+}$ | 0.959 |
| | (1.93) | (-1.92) | (-0.46) |
| Trial | $1.017^{*}$ | $0.952^{***}$ | 1.005 |
| | (2.09) | (-3.85) | (0.56) |
| Constant | $0.024^{+}$ | 5.638 | 0.532 |
| | (-1.96) | (0.68) | (-0.37) |
| McKelvey & Zavoina's $R^2$ | 0.085 | 0.212 | 0.006 |
| Observations | 1288 | 1288 | 1288 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as random effect.

Significance level: $^{+}\ p < .10$, $^{*}\ p < .05$, $^{***}\ p < .001$.

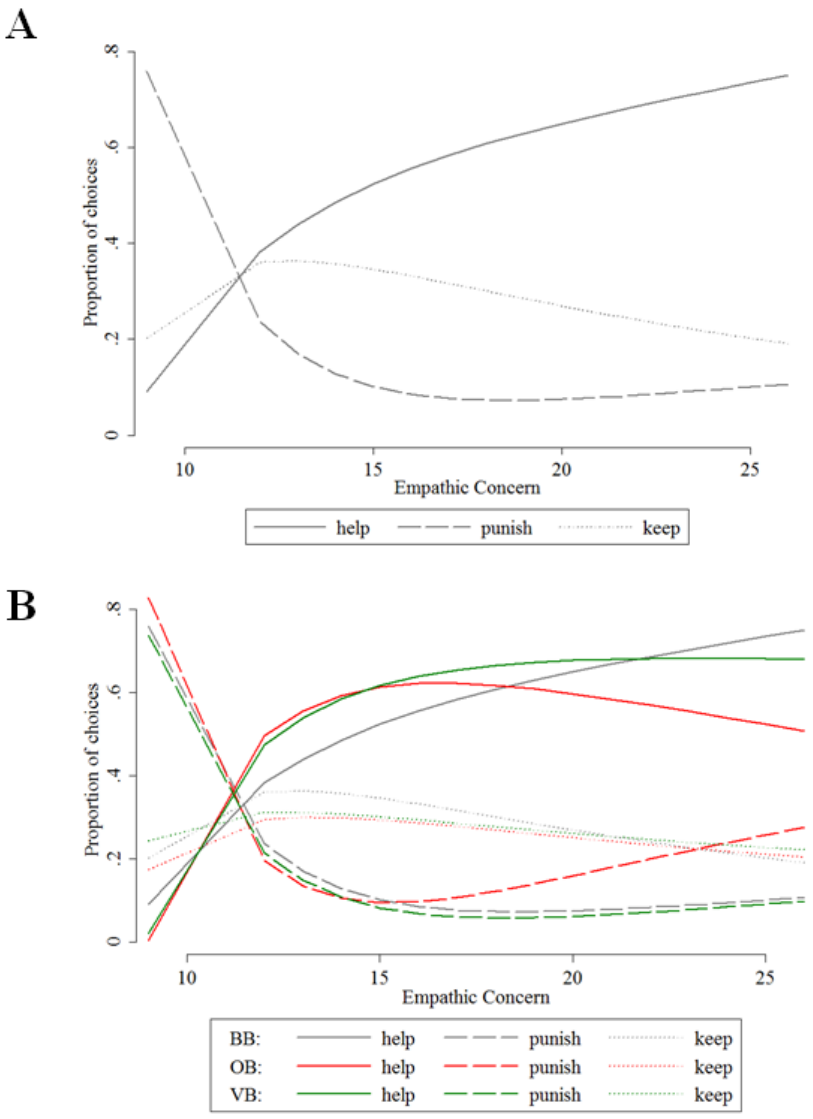


Figure 29. (A) Choice proportion predicted by empathic concern level in BB; (B) Choice proportion predicted by empathic concern level in all conditions. The curve plot showed the fractional polynomial relationship between choice proportion and empathic concern level for each type of choices respectively.

#### 5.3.1.2 Transfer Amount

Empathic concern could predict the amount of third parties to neither help the victim (b = -0.027, z = -0.67, p = 0.500) nor punish the offender (b = 0.012, z = 0.14, p = 0.891).

#### 5.3.1.3 Indices of Information Searching

In contrast to H1b, we found that neither (log-transformed) fixation number (help: b = 0.002, z = 0.12, p = 0.907; punishment: b = -0.004, z = -0.12, p = 0.901) nor (log-transformed) decision time (help: b = -0.014, z = -0.70, p = 0.487; punishment: b = -0.011, z = -0.44, p = 0.662) of either altruistic choices could be predicted by empathic concern of third parties.

#### 5.3.1.4 Fixation Proportion towards Victim-payoff AOIs

In consistency with H1c, we found that third parties with a higher empathic concern level distributed a higher fixation proportion towards victim-relevant information (i.e., victim-payoff AOIs; b = 1.046, z = 2.83, p = 0.005). Post-hoc analyses showed that this effect existed in both help (b = 0.679, z = 2.34, p = 0.020) and punishment choices (b = 1.021, z = 2.50, p = 0.013; see Figure 30A; also see Table 30 for regression details of all above measures).

Table 30. Results of repeated-measure mixed-effect linear regression predicting the transfer amount, log-transformed fixation number, log-transformed decision time and fixation proportion by empathic concern for help and punishment choices in the BB

| | Transfer Amount[a] (in €) | | Fixation Number (Log) | | Decision Time (Log; in ms) | | Fixation Proportion | | |
|---|---|---|---|---|---|---|---|---|---|
| | help | punish | help | punish | help | punish | help+punish | help | punish |
| Empathic concern | -0.027 | 0.012 | 0.002 | -0.004 | -0.014 | -0.011 | 1.046** | 0.679* | 1.021* |
| | (-0.67) | (0.14) | (0.12) | (-0.12) | (-0.70) | (-0.44) | (2.83) | (2.34) | (2.50) |
| Trial | -0.006 | -0.028* | -0.015*** | -0.013** | -0.022*** | -0.012** | 0.129* | 0.072 | 0.098 |
| | (-1.53) | (-2.31) | (-7.97) | (-2.98) | (-11.52) | (-3.24) | (2.35) | (1.20) | (0.79) |
| Constant | 2.312** | 2.377 | 2.395*** | 2.475*** | 8.564*** | 8.375*** | 35.35*** | 46.66*** | 21.90** |
| | (2.97) | (1.45) | (6.31) | (4.40) | (21.96) | (18.57) | (5.03) | (8.20) | (2.93) |
| $R^2$(overall) | 0.008 | 0.035 | 0.052 | 0.056 | 0.103 | 0.092 | 0.052 | 0.018 | 0.061 |
| Observations | 750 | 174 | 673 | 148 | 673 | 148 | 821 | 673 | 148 |

Note: Values refer to unstandardized coefficients. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as random effect. Time effect (as trials) was controlled in the analysis. BB refers to the baseline block.

Significance level: * $p < .05$, ** $p < .01$, *** $p < .001$.

[a]Analyses on transfer amount keeps the same dataset used for choice behavior.

#### 5.3.1.5 Distribution of the First and Last Fixation

Besides the hypothesized measures, we also took a look at the fixation at certain time point during decision-making as explorative analyses. The most representative fixations were the first and the last fixation. We found that the empathic concern level could even bias the third party's attention towards the victim-relevant information at the very beginning (Odds ratio = 1.305, z = 1.79, p = 0.074), which also showed a similar trend on the last fixation (Odds ratio = 1.080, z = 1.50, p = 0.134; see Table 31 for regression details).

Table 31. Results of repeated-measure mixed-effect logistic regression predicting the distribution of the first and the last fixation (towards victim payoff-relevant AOIs) by empathic concern for help and punishment choices respectively in the BB.

| | First Fixation | | | Last Fixation | | |
|---|---|---|---|---|---|---|
| | help+punish | help | punish | help+punish | help | punish |
| Empathic Concern | 1.305$^{+}$ | 1.312 | 1.353 | 1.080 | 1.048 | 1.016 |
| | (1.79) | (1.55) | (1.64) | (1.50) | (1.26) | (0.20) |
| Trial | 1.032$^{*}$ | 1.043$^{**}$ | 0.988 | 1.005 | 0.983$^{+}$ | 1.042$^{+}$ |
| | (2.49) | (2.74) | (-0.44) | (0.54) | (-1.68) | (1.70) |
| Constant | 0.028 | 0.033 | 0.010 | 0.505 | 2.278 | 0.072$^{+}$ |
| | (-1.28) | (-1.03) | (-1.41) | (-0.69) | (1.12) | (-1.81) |
| McKelvey & Zavoina's $R^2$ | 0.117 | 0.149 | 0.11 | 0.019 | 0.089 | 0.337 |
| Observations | 821 | 673 | 148 | 821 | 673 | 148 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. BB refers to the baseline block.

Significance level: $^{+}$ $p < .10$, $^{*}$ $p < .05$, $^{**}$ $p < .01$.

### 5.3.2 All Blocks

The goal of this part analyses was to investigate the effect of attention focus (H2a-c) and its interaction with empathic concern (H3a-c) on altruistic choice and eye-movement during decision-making in all three blocks. Thus the main predictors for the analyses relevant to H2a-c included attention focus (reference condition: BB; two dummy variables coding for OB and VB) and empathic concern level. The main predictors for the analyses relevant to H3a-c were the same except the attention focus × empathic concern interaction terms (dummy variables; two dummy variables coding for OB × empathic concern, and VB × empathic con-

cern). To minimize the effect of collinearity between the regressors in the analyses relevant to H3a-c, the empathic concern score was always mean-centered which was then used to create dummy variables coding for the interaction. Similarly, the trial as an index of time was also added to these regression analyses to rule out the effect of practice.

#### 5.3.2.1 Choice Behavior

Regarding the effect of attention focus (H2a), we found that third parties increased the possibility to punish (Odds ratio = 2.213, z = 3.17, p = 0.002) and reduced the likelihood to keep (Odds ratio = 0.661, z = -2.07, p = 0.039) when they considered the unfairness of the offender. Although the results did not reach statistical significance, there was a trend showing that third parties were more likely to help the victim in the VB (Odds ratio = 1.302, z = 1.43, p = 0.153). These findings were basically consistent with our predictions in H2a.

Regarding the interaction effect (H3a), we found that participants with higher level of empathic concern were more likely to punish in the OB (Odds ratio = 1.125, z = 3.09, p = 0.002) but withdraw to help in either the OB (Odds ratio = 0.885, z = -3.88, p < 0.001) or the VB (Odds ratio = 0.925, z = -2.44, p = 0.015). These results were not in line with our expectations in H3a (see Figure 29B; also see Table 32 for regression details).

Table 32. Results of repeated-measure mixed-effect logistic regression predicting the help, punishment or keep choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions

| | help | help | punish | punish | keep | keep |
|---|---|---|---|---|---|---|
| BB (ref) | | | | | | |
| OB | 0.916 | 0.980 | 2.213** | 1.954** | 0.661* | 0.660* |
| | (-0.49) | (-0.11) | (3.17) | (2.58) | (-2.07) | (-1.98) |
| VB | 1.302 | 1.377+ | 0.941 | 0.793 | 0.729 | 0.726 |
| | (1.43) | (1.70) | (-0.23) | (-0.83) | (-1.56) | (-1.53) |
| Empathic Concern (EC) | 1.195 | 1.279* | 0.862 | 0.817+ | 0.983 | 0.948 |
| | (1.55) | (2.11) | (-1.26) | (-1.68) | (-0.17) | (-0.52) |
| BB (ref) × EC | | | | | | |
| OB × EC | | 0.885*** | | 1.125** | | 1.055+ |
| | | (-3.88) | | (3.09) | | (1.65) |
| VB × EC | | 0.925* | | 1.007 | | 1.062+ |
| | | (-2.44) | | (0.17) | | (1.86) |
| Trial | 1.002 | 1.001 | 0.993+ | 0.996 | 1.002 | 1.002 |
| | (0.73) | (0.40) | (-1.83) | (-0.99) | (0.74) | (0.62) |
| Constant | 0.033 | 0.871 | 0.472 | 0.028*** | 0.423 | 0.312** |
| | (-1.58) | (-0.32) | (-0.34) | (-7.58) | (-0.45) | (-3.00) |
| McKelvey & Zavoina's $R^2$ | 0.039 | 0.042 | 0.037 | 0.044 | 0.002 | 0.003 |
| Observations | 3864 | 3864 | 3864 | 3864 | 3864 | 3864 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

### 5.3.2.2 Transfer Amount

Unlike choice behavior, transfer amount of neither help nor punishment could be predicted by attention focus, empathic concern or their interaction (all ps > 0.15; see Table 33 for regression details).

Table 33. Results of repeated-measure mixed-effect linear regression predicting the transfer amount of help and punishment choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions.

| | help | help | punish | punish |
|---|---|---|---|---|
| BB (ref) | | | | |
| OB | -0.009 | -0.017 | -0.0005 | 0.112 |
| | (-0.11) | (-0.20) | (-0.00) | (0.42) |
| VB | -0.004 | -0.012 | 0.137 | 0.252 |
| | (-0.05) | (-0.15) | (0.56) | (0.97) |
| Empathic Concern (EC) | -0.031 | -0.028 | -0.017 | 0.007 |
| | (-0.79) | (-0.76) | (-0.33) | (0.13) |
| BB (ref) × EC | | | | |
| OB × EC | | 0.0006 | | -0.053 |
| | | (0.04) | | (-1.43) |
| VB × EC | | -0.010 | | 0.003 |
| | | (-0.68) | | (0.07) |
| Trial | -0.00002 | 0.0001 | -0.005 | -0.0075* |
| | (-0.02) | (0.11) | (-1.61) | (-2.07) |
| Constant | 2.243** | 1.655*** | 2.360* | 2.135*** |
| | (2.98) | (12.66) | (2.52) | (9.40) |
| $R^2$(overall) | 0.006 | 0.005 | 0.002 | 0.005 |
| Observations | 2311 | 2311 | 528 | 528 |

Note: Values refer to unstandardized coefficients. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

#### 5.3.2.3 Indices of Information Searching

Partially supporting our predictions in H2b, we observed that third parties increased the either (log-transformed) fixation numbers (help: b = 0.104, z = 2.40, p = 0.017; punishment: b = 0.189, z = 1.78, p = 0.075) or (log-transformed) decision times (help: b = 0.122, z = 2.83, p = 0.005; punishment: b = 0.158, z = 1.65, p = 0.099) during either altruistic choices only in OB (for above analyses in VB: all ps > 0.3).

Table 34. Results of repeated-measure mixed-effect linear regression predicting the fixation number and decision time of help and punishment choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions.

| | Fixation Number (Log) | | | | Decision Time (Log; in ms) | | | |
|---|---|---|---|---|---|---|---|---|
| | help | help | punish | punish | help | help | punish | punish |
| BB (ref.) | | | | | | | | |
| OB | 0.104* | 0.111* | 0.189+ | 0.235* | 0.122** | 0.119** | 0.158+ | 0.192+ |
| | (2.40) | (2.49) | (1.78) | (2.14) | (2.83) | (2.69) | (1.65) | (1.93) |
| VB | 0.029 | 0.036 | -0.012 | 0.033 | 0.036 | 0.035 | 0.013 | 0.044 |
| | (0.68) | (0.84) | (-0.12) | (0.30) | (0.87) | (0.80) | (0.14) | (0.46) |
| Empathic Concern (EC) | 0.001 | 0.006 | 0.009 | -0.008 | -0.017 | -0.010 | 0.009 | -0.010 |
| | (0.05) | (0.30) | (0.32) | (-0.25) | (-0.82) | (-0.50) | (0.37) | (-0.37) |
| BB (ref) × EC | | | | | | | | |
| OB × EC | | -0.012 | | 0.014 | | -0.010 | | 0.021 |
| | | (-1.40) | | (0.72) | | (-1.18) | | (1.20) |
| VB × EC | | -0.004 | | 0.046* | | -0.011 | | 0.045* |
| | | (-0.54) | | (2.14) | | (-1.39) | | (2.30) |
| Trial | -0.007*** | -0.007*** | -0.007*** | -0.008*** | -0.009*** | -0.009*** | -0.007*** | -0.008*** |
| | (-11.28) | (-11.01) | (-5.32) | (-5.33) | (-14.97) | (-14.23) | (-5.92) | (-5.75) |
| Constant | 2.224*** | 2.244*** | 2.304*** | 2.468*** | 8.312*** | 7.994*** | 8.042*** | 8.202*** |
| | (5.75) | (32.66) | (4.58) | (21.29) | (21.24) | (120.76) | (17.78) | (78.48) |
| $R^2$(overall) | 0.087 | 0.087 | 0.095 | 0.101 | 0.151 | 0.152 | 0.117 | 0.121 |
| Observations | 2063 | 2063 | 432 | 432 | 2063 | 2063 | 432 | 432 |

Note: Values refer to unstandardized coefficients. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Concerning the interaction effect (H3b), we found that participants with a higher empathic concern level increased the effort of information searching (log-transformed fixation number: b = 0.046, z = 2.14, p = 0.032; log-transformed decision time: b = 0.045, z = 2.30, p = 0.021) during punishment choice when they focused on the feeling of the victim. These findings were not consistent with our expectations in H3b (see Table 34 for regression details).

#### 5.3.2.4 Fixation Proportion towards Victim-Payoff AOIs

Against our predictions in H2c, we did not observe the attention-induced change in fixation proportion towards victim-relevant information in either the OB (b = -0.930, z = -0.40, p = 0.486) or the VB (b = 0.083, z = 0.06, p = 0.949). Analyzing the interaction effect revealed that third parties with a higher level of empathic concern paid less attention to the victim-relevant information in either the OB (b = -0.511, z = -2.04, p = 0.042) or the VB (b = -0.552, z = -2.15, p = 0.032), which was not expected in H3c (see Figure 30B; see Table 35 for regression details).

Table 35. Results of repeated-measure mixed-effect linear regression predicting the fixation proportion of attention towards victim-relevant information for help, punishment and both choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions.

| | help+punish | help+punish | help | help | punish | punish |
|---|---|---|---|---|---|---|
| BB (ref.) | | | | | | |
| OB | -0.930 | -0.928 | 0.902 | 0.536 | -5.827$^{+}$ | -6.726$^{+}$ |
| | (-0.70) | (-0.67) | (0.63) | (0.37) | (-1.75) | (-1.93) |
| VB | 0.083 | 0.090 | 0.491 | 0.168 | -6.876$^{*}$ | -7.694$^{*}$ |
| | (0.06) | (0.07) | (0.36) | (0.12) | (-2.08) | (-2.23) |
| Empathic Concern (EC) | 0.685 | 1.018$^{*}$ | 0.454 | 0.725$^{+}$ | 0.796$^{+}$ | 0.403 |
| | (1.45) | (2.05) | (1.15) | (1.75) | (1.87) | (0.69) |
| BB (ref) × EC | | | | | | |
| OB × EC | | -0.511$^{*}$ | | -0.272 | | 0.716 |
| | | (-2.04) | | (-0.97) | | (1.21) |
| VB × EC | | -0.552$^{*}$ | | -0.568$^{*}$ | | 0.171 |
| | | (-2.15) | | (-2.09) | | (0.25) |
| Trial | -0.010 | -0.009 | -0.021 | -0.013 | 0.033 | 0.051 |
| | (-0.53) | (-0.49) | (-1.05) | (-0.66) | (0.76) | (1.10) |
| Constant | 44.29$^{***}$ | 56.89$^{***}$ | 52.76$^{***}$ | 60.96$^{***}$ | 27.18$^{***}$ | 40.76$^{***}$ |
| | (5.00) | (32.08) | (7.07) | (42.35) | (3.40) | (18.01) |
| $R^2$(overall) | 0.015 | 0.017 | 0.003 | 0.005 | 0.071 | 0.064 |
| Observations | 2495 | 2495 | 2063 | 2063 | 432 | 432 |

Note: Values refer to unstandardized coefficients. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}$ $p < .10$, $^{*}$ $p < .05$, $^{***}$ $p < .001$.

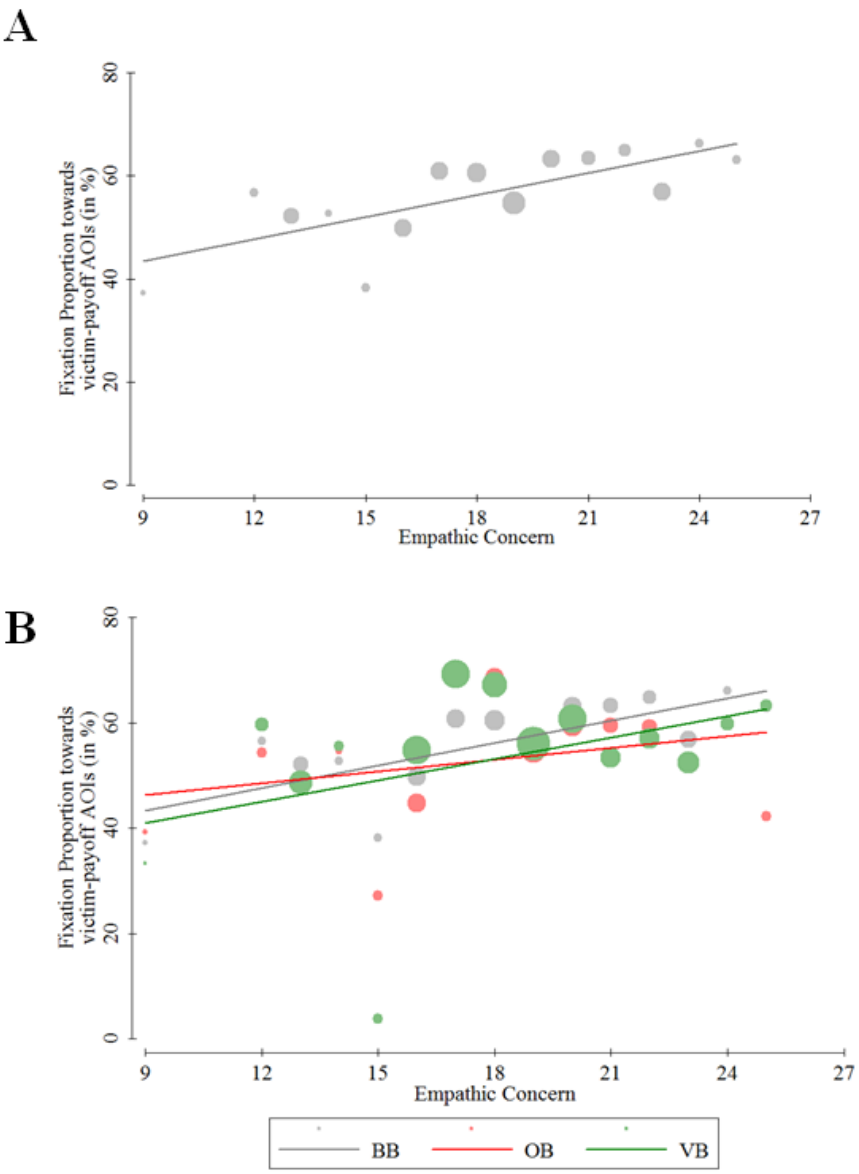


Figure 30. (A) Fixation proportion towards the victim-payoff AOIs predicted by empathic concern level in BB; (B) Fixation proportion towards the victim-payoff AOIs predicted by empathic concern level in all conditions. The line plot showed the linear relationship between fixation proportion and empathic concern level for altruistic choices (i.e., help and punishment).

#### 5.3.2.5 Distribution of the First and Last Fixation

We also ran similar regressions as explorative analyses on the first and the last fixation. Participants were less likely to pay attention to the victim-relevant information at the first glance while considering the unfairness of the offender (Odds ratio = 0.550, z = -2.18, p = 0.030). Besides we found that third parties with

higher levels of empathic concern were less likely to look at the victim-relevant information in either the OB (Odds ratio = 0.809, z = -4.22, p < 0.001) or the VB (Odds ratio = 0.828, z = -3.52, p < 0.001; see Table 36 for regression details). Analyses of the last fixation mirrored the findings of the first fixation (see Table 37 for regression details).

Table 36. Results of repeated-measure mixed-effect logistic regression predicting the distribution of the first fixation towards victim-relevant information for help, punishment and both choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions.

| | help+punish | help+punish | help | help | punish | punish |
|---|---|---|---|---|---|---|
| BB (ref.) | | | | | | |
| OB | 0.550$^{*}$ | 0.539$^{*}$ | 0.712 | 0.591 | 0.363 | 0.337 |
| | (-2.18) | (-2.19) | (-1.06) | (-1.60) | (-1.25) | (-1.35) |
| VB | 0.998 | 0.964 | 0.863 | 0.689 | 0.778 | 0.697 |
| | (-0.01) | (-0.13) | (-0.47) | (-1.10) | (-0.35) | (-0.50) |
| Empathic Concern (EC) | 1.109 | 1.261$^{+}$ | 1.136 | 1.266 | 1.071 | 1.271 |
| | (0.84) | (1.83) | (0.90) | (1.62) | (0.46) | (1.36) |
| BB (ref.) × EC | | | | | | |
| OB × EC | | 0.809$^{***}$ | | 0.844$^{**}$ | | 0.804$^{+}$ |
| | | (-4.22) | | (-2.72) | | (-1.70) |
| VB × EC | | 0.828$^{***}$ | | 0.819$^{**}$ | | 0.726$^{*}$ |
| | | (-3.52) | | (-3.15) | | (-1.99) |
| Trial | 1.012$^{**}$ | 1.010$^{**}$ | 1.015$^{**}$ | 1.016$^{**}$ | 1.005 | 1.002 |
| | (3.13) | (2.60) | (3.22) | (3.26) | (0.54) | (0.17) |
| Constant | 0.476 | 3.569$^{**}$ | 0.448 | 5.107$^{***}$ | 0.450 | 2.062 |
| | (-0.32) | (2.80) | (-0.30) | (3.32) | (-0.28) | (1.08) |
| McKelvey & Zavoina's $R^2$ | 0.023 | 0.028 | 0.029 | 0.029 | 0.017 | 0.033 |
| Observations | 2495 | 2495 | 2063 | 2063 | 432 | 432 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 37. Results of repeated-measure mixed-effect logistic regression predicting the distribution of the last fixation towards victim-relevant information for help, punishment and both choice by attention focus, empathic concern (odd columns) and their interaction (even columns) in all conditions.

| | help+punish | help+punish | help | help | punish | punish |
|---|---|---|---|---|---|---|
| BB (ref.) | | | | | | |
| OB | 0.727$^{+}$ | 0.751 | 0.701$^{+}$ | 0.687$^{+}$ | 1.834 | 1.917 |
| | (-1.79) | (-1.55) | (-1.73) | (-1.79) | (1.10) | (1.12) |
| VB | 0.860 | 0.890 | 0.772 | 0.758 | 1.035 | 1.075 |
| | (-0.84) | (-0.63) | (-1.26) | (-1.30) | (0.06) | (0.12) |
| Empathic Concern (EC) | 1.031 | 1.082$^{+}$ | 1.033 | 1.059 | 0.997 | 1.011 |
| | (0.72) | (1.68) | (0.82) | (1.23) | (-0.06) | (0.13) |
| BB (ref.) × EC | | | | | | |
| OB × EC | | 0.920$^{*}$ | | 0.970 | | 0.975 |
| | | (-2.55) | | (-0.77) | | (-0.29) |
| VB × EC | | 0.940$^{+}$ | | 0.960 | | 0.993 |
| | | (-1.86) | | (-1.05) | | (-0.06) |
| Trial | 0.996 | 0.996 | 0.995$^{*}$ | 0.995$^{+}$ | 0.998 | 0.997 |
| | (-1.45) | (-1.62) | (-1.98) | (-1.75) | (-0.27) | (-0.35) |
| Constant | 1.376 | 2.470$^{***}$ | 2.437 | 4.452$^{***}$ | 0.188 | 0.182$^{***}$ |
| | (0.39) | (5.20) | (1.16) | (8.72) | (-1.48) | (-4.58) |
| McKelvey & Zavoina's $R^2$ | 0.015 | 0.019 | 0.028 | 0.029 | 0.016 | 0.017 |
| Observations | 2495 | 2495 | 2063 | 2063 | 432 | 432 |

Note: Values refer to odds ratio. The *z* statistics are provided in parentheses. Data clusters specific to subject were treated as the random effect. Time effect (as trials) was controlled in the analysis. ref refers to reference, BB refers to baseline block, OB refers to offender-focus block, VB refers to victim-focus block.

Significance level: $^{+}\ p < .10$, $^{*}\ p < .05$, $^{***}\ p < .001$.

## 5.4 Discussion

### 5.4.1 Empathic Concern Can Not Only Predict Third-Party Altruistic Choice But Also Gaze Searching

Testing the link between empathic concern and choice behavior, as well as the attentional-gaze distribution during decision-making in the third-party context, is the main aim of Study 4 (H1a). Regarding choice behavior, we replicated previous findings showing that the level of empathic concern could positively (negatively) predict the costly helping (punishing) behavior of a third-party bystander in response to the violation of a social norm (i.e., unfairness in this case) (Leliveld, et al., 2012).

More importantly, empathic concern does not only influence altruistic choices, but also exerts an impact on eye-movement, albeit, in a specific manner. Despite not observing the expected relationship between empathic concern and the extent of information searching behavior (indexed by decision time and fixation number respectively; H1b), our results revealed an association between empathic concern and attention distribution to victim-relevant information (i.e., AOIs covering both the absolute and relative monetary payoff made to the victim), reflected by the fixation proportion (H1c). Particularly, the fixation proportion towards victim payoff-relevant information increased with empathic concern level when participants decided naturally, regardless of the altruistic choice participants made. These findings demonstrate that the proportion of fixations to a specific piece of information recorded by eye-tracking equipment provides a direct and accurate measure of underlying empathic concerns for investigating cognitive processing and information searching, which can thus be used as a reliable index for measuring how people's attention is allocated during the decision-making process (Fiedler & Glöckner, 2015; Orquin & Loose, 2013). Notably, the fixation proportion to a specific piece of information seems to be crucial as it can clearly disentangle different sources contributing to the attentional effect. Hence, the contingency between the fixation proportion on the victim's payoff and empathic concern indicates that more empathic third parties allocate more attention specifically to the victim rather than to the offender, instead of enhancing general processing depth during decision-making. As a result, highly empathic people might consider the feelings of the victim more and ultimately be more likely to help the victim.

We also noticed that the empathy-driven attentional effect was more salient at the very begianning. Visual perceptual studies have shown that first fixation is usually driven by the properties of the stimuli themselves, such as their saliency (Parkhurst, Law, & Niebur, 2002). Nevertheless, this cannot explain our findings since the basic visual properties of the stimuli (esp. the number of digits in all payoffs) for both victim- and offender-relevant information were the same for all participants. Given that the location of the payoff-relevant content is fixed for each participant across trials in the current paradigm, participants might be aware of the target information beforehand. Therefore, this finding suggest that empathic concern could direct the attention of third parties paid towards the victim's payoff information, even on first glance. Interestingly, such an effect was dampened on the last fixation. This is probably due to the fact last fixation is more closely related to final choice, as hinted at previous evidence. For instance, in a value-based binary food choice task, the final choices (left or right item) of participants were predicted by the location of the last fixation unless that item was much worse than

the alternative one (Krajbich, et al., 2010). Nevertheless, such explanations should be treated cautiously and replication is required by future studies.

### 5.4.2 The Effect of Attention Focus and Its interaction with Empathic Concern on Altruistic Choice

As predicted in H2a, third parties were more likely to punish the offender when taking the unfairness of the offender into account (OB), which was consistent with previous findings in food choice (Hare, et al., 2011) and the social decision (Gromet & Darley, 2009; Hutcherson & Rangel, 2014; Makwana, et al., 2014) paradigms. Notably, participants only showed a trend-to-significant increase in helping behavior when focusing on the feeling of the victim (VB). Such findings seem to indicate that the help choice is by nature driven by empathy for the victim, which is set as default (i.e., BB) and already in consistency with the manipulation used in the VB.

Moreover, compared with lower empathic participants, people with higher empathic concern increased the possibility of punishing while reduced the likelihood of helping in the OB (vs. the BB).Against H3a, our findings further clarified a main effect of attention focus on the altruistic choice mentioned above. A recent behavioral study provides a clue for understanding the current results. Gummerum and colleagues (2016) showed that third parties helped the victim more often in a third-party compensation game (i.e., whereby only help was possible) when empathic anger (towards the offender who harmed the victim) rather than self-focused anger (towards the offender who harmed the participants themselves) was elicited (Gummerum, et al., 2016). Given the nature of empathic concern, the highly empathic people felt more empathic anger in the OB. Intriguingly, they switched their approach from compensating the victim to punishing the offender, and justified the punishment as a better way to achieve justice when both altruistic options were available. Besides, the result that they were also less likely to help the victim in the VB might be driven by the fact that they unexpectedly behaved more selfishly, since they might be averse to being “forced” to help. Taken together, the above-mentioned results suggest that the attention-induced effect on decision-making could be extended to a more complex social context, and such an effect could be further modulated by empathic concern.

One additional point that might be worthwhile to discuss is the fact that transfer amount, as another important measure of third parties’ altruistic behaviors, did not vary much across the different focus conditions, different levels of empathic concern, or their interactions. These results, compared with the findings on choice behavior, suggest that the altruistic preference, instead of the altruistic intensity, is

more sensitive to either personality trait variance or the experimental conditions, which provides a better measure for altruism.

### 5.4.3 The Effect of Attention Focus and Its Interaction with Empathic Concern on the Eye-movements of Third parties during Altruistic Decision-making

Concerning the information searching behavior (H2b), third parties took more time to decide to costly intervene, accompanied by more fixations in search of information, when they considered the offender's behavior in the context highlighting social norms (OB). This finding indicated the need for more in-depth cognitive processing in the OB, as participants must additionally consider and reevaluate the situation from the offender's viewpoint, which differs from how third parties naturally think and respond to norm violations. However, no such change in information searching was observed in the VB, which further showed the weak effect of the VB in influencing choice behaviors. Apart from that, the unexpected interaction (H3b) between attention focus (i.e., the VB) and empathic concern in cognitive processing during punishment choices suggested that the third party might struggle more when the final choice they made (i.e., to punish the offender) contradicted both intrinsic (i.e., higher level of empathic concern) and external consideration (i.e., perceiving the feelings of the victim).

Unlike general processing depth, we did not observe the expected difference in attentional distribution varied across different focus conditions, as measured by the fixation proportion, towards victim-relevant information (H2c). Interestingly, we detected a surprising interaction effect (H3c) whereby third parties with a higher level of empathic concern attended less to victim-relevant information in the OB. Such a finding indicates that the instruction of considering the unfairness of offender drove highly empathic bystanders to attend more to the offender, rather than the victim, since the search space only included the offender and the victim-relevant information in this case. In addition, there seemed to be a perfect match between the behavior finding (see above section) and fixation proportion in terms of the interaction effect. In either the OB or the VB, highly empathic participants paid less attention to victim-relevant information, which might be another explanation for the decreased possibility of helping seen in both conditions.

In addition to the above, we also found that third parties in the OB were less (more) likely to attend to victim-relevant (offender-relevant) information during either the first or last fixation, which was in consistency with the results on fixation number. Thus, this suggests that we could induce bias in third party attention even at the specific point when they either started, or finished making decisions by manipulating the focus.

### 5.4.4 Limitations

One limitation of the current study was that the analysis on punishment choice might be underpowered, because participants systematically chose to punish less frequently (only ~13% of all trials), similar to what was found in Studies 1-3 (see General Discussion for more details). Given the mild degree of norm violation seen in the current setting (i.e., merely unfair money allocation), future studies could increase the frequency of punishment by increasing the severity of the norm violation, such as introducing the intention of the offender (Buckholtz et al., 2015). Furthermore, a context in which only punishment (and keeping the money) is allowed (e.g., third-party punishment game) might also help to increase punishment behaviors and gaze behaviors during punishment-related decision-making.

Another possible limitation was the unbalanced order between the BB and the other two focus manipulation blocks (the OB and VB) among participants. As mentioned before, we designed the experiment in this way by assigning the BB as the first experimental condition, to guarantee that the participants' behaviors and information search patterns were not biased by the attention focus. However, such a design could also be disadvantageous in that the results of the OB and VB are always systematically confounded by fatigue, as well as familiarity, due to the fixed order, although we mitigated such confounding by counterbalancing the order of the OB and VB across participants, and by controlling for the time effect (i.e., trial) in the regression analyses. Besides, it is also possible that participants had already established decision strategies that could not be easily influenced by the introduction of further instructions. For these reasons, the extent of information searching required could be reduced in either the OB or the VB, which might result in insensitivity of fixation proportion to the manipulation of attention focus. In order to address this problem, future studies should modify the design. For instance, it is plausible to employ a full between-subjects design. Alternatively, it might also be possible to increase the number of blocks, with fewer trials, and to fully randomize the different attention focus blocks within participants (as we did in Study 3).

### 5.4.5 Summary

With a modified third-party paradigm, the current study captures for the first time how empathic concern affects the underlying cognitive processes during altruistic decision-making by recording the gaze behavior. Moreover, we shed some light on the influence of the attention focus on different contextual cues highlighting particular aspects of an unfair situation (i.e., focusing on either the suffering of the victim or the conduct of the offender) and its interaction with empathic concern on

altruistic choice and the accompanying eye-movement pattern. Taken together, these findings provide direct empirical evidence for the proposal which explains the empathy-dependent altruistic preference shift via attention, implied by the previous study (Leliveld, et al., 2012). More broadly, these findings could encourage future studies on investigating the cognitive mechanism underlying moral judgments and decision-making by employing implicit attention measures such as eye-tracking techniques (Fiedler & Glöckner, 2015).

# 6 General Discussion

Although it is not a secret that human beings will voluntarily help a victim and punish an offender in response to a social norm violation, at cost to their own time, money, and energy even if their own well-beings are not affected directly, it is still unknown as to how our brain makes these altruistic decisions, and whether other factors could also influence such decisions and their underlying neural correlates as well as the cognitive processes. In the series of experiments introduced in the current PhD dissertation, we adopted the modified third-party paradigm commonly used in behavioral economics, in combination with popular cognitive neuroscience methods (esp. fMRI and eye-tracking), to address the above research questions.

Specifically, Study 1 mainly investigated the neural correlates of third-party altruistic decision-making and revealed that help and punishment choices shared common neural substrates in the bilateral striatum (esp. ventral part), indicating that a reward component accompanies third-party altruism. Moreover, Study 1 showed the link between empathic concern and third-party altruism at the behavioral and neural level. Study 2 mainly tested the effect of intranasal oxytocin on third-party altruistic choice behavior (Studies 2A and 2B) as well as its neural correlates (Study 2A). Albeit that oxytocin did not influence the third-party altruistic choice behavior, Study 2A showed that oxytocin (vs. placebo) enhanced the activation in the left TPJ while participants observed the victim being helped by the computer, suggesting its role in improving mentalizing ability during social interactions. Study 3 explored the role of other-regarding attention focus in modulating third-party altruistic decisions, as well as their neural correlates. The induced attention focus not only changed the behavior of third parties, but also affected the accompanying decision-relevant activation in the TPJ and control network, providing new empirical evidence for attention-decision coupling. To further clarify the cognitive process underlying third-party altruistic decision-making, Study 4 adopted eye-tracking methods and showed that the attention distribution of third parties towards the victim's payoffs, measured by the fixation proportion, was affected by individual empathic concern levels as well as its interaction with instructed attention focus.

Given that the results, as well as the limitations, of each study have been discussed in detail directly in each corresponding empirical chapter, in the remaining part of this section I will discuss issues of more general interest (i.e., some common features of our findings across studies, debatable results, and implications), describe future directions of research on this topic, and provide a short conclusion to end the main part of the dissertation.

## 6.1 Third-Party Deciders Prefer Helping the Victim to Punishing the Offender

Among all four studies, we noticed a common and interesting phenomenon, namely that participants on average were at least two-fold more likely to help the victim rather than punish the offender[10]. Intriguingly, this finding, despite not being the focus of the current dissertation, is consistent with previous studies showing that third parties transferred more money to help than to punish (Chavez & Bicchieri, 2013; Lotz, Okimoto, et al., 2011)[11]. Recent studies have attempted to explain such behavioral bias of third-party deciders from the evolutionary perspective of social signaling (Jordan, et al., 2016; Raihani & Bshary, 2015b). For instance, Raihani and Bshary (2015) found that, compared with the third-party punishers who costly punished the selfish offender, third-party helpers who costly compensated the recipient were more rewarded by a fourth party (i.e., in a modified dictator game where the fourth party could increase the bonus of the third party with a cost ratio of 1:5) when the third party could either help, punish or keep their endowment. Additionally, the fourth party was more likely to reward third-party helpers when the third party could only help or keep, rather than only punish or keep. These results suggested that another motivation which might cause negative impact on the reputation of punishers may refrain participants from choosing punishment, especially when they have other options to be altruistic (Raihani & Bshary, 2015a). Consistent with these ideas, another recent study showed that third-party punishers were less trusted, if they also had the chance to help, by the fourth party (i.e., in a trust game where the fourth party played as the investor, while the third-party punisher or helper played as the trustee); this supported their game-theoretical model, which proposes that third-party punishment, although considered as an important signal for trustworthiness, is less salient and informative than costly helping (Jordan, et al., 2016)[12]. Taken together, these findings suggest a way to increase your reputation and impression for others, namely by punishing the offender when you are not directly affected, this signals that you are trustworthy; but you will be considered even more trustworthy and kind by others if you try your best to help the victim.

[10] It seems that this does not hold true for Study 1 at the first glance. However we should not forget that we excluded 10 participants and seven of them chose to always help the victim (i.e., for the remaining three participants, 2 of them always punished the offender and 1 was always selfish).

[11] Notably, participants in these studies could perform both help and punishment at the same time, unlike the context where participants could either help or punish once in all studies included in the present dissertation.

[12] Unlike previous studies, participants in this study helped the victim only as second parties (in a context similar to that of the traditional dictator game), and not as third parties.

## 6.2 Other Potential Motivations That Drive Third-Party Help and Punishment

Third-party help and punishment, in the studies described herein, are framed as altruistic behaviors/decisions based on a consequence-oriented definition, namely, benefiting others (i.e., recipients) at cost to the actors (see Introduction). However, altruism can also be defined in a more strict sense, in terms of the motivation underlying the behaviors, whereby only behaviors with the goal of benefiting others without benefiting the actor (either immediately or in the long run) are purely altruism. In our studies, both the offenders and victims are anonymous (i.e., only with name initials) and are strangers to the third party participants; all contexts are framed as the one-shot game. Therefore, the participants never know who they helped and only meet the other parties once[13], which ensures that they cannot receive any payback from other parties in the future.

Despite such altruistic motivation, it is still possible that these behaviors could be driven by other motivations. For instance, each third party in Studies 1 and 2A was endowed with 50 MU, which was always lower than the initial payoff of the offender (i.e., from 60 to 100 MU). Thus third-party participants might punish the offender simply due to envy or aversion to disadvantageous inequality, such that they reduced the payoff inequality between themselves and the offender via punishment (Fehr & Schmidt, 1999; Pedersen, et al., 2013). In trying to address this potential confound, participants were always endowed with more than the offender (and, of course, the victim) in the other three studies (i.e., Study 2B: offender maximal payoff: 90 MU, third party initial endowment: 100 MU; Studies 3 and 4: offender maximal payoff: ~ € 9, third party initial endowment: € 10). As a result, we still observed that third parties costly punished unfair offenders rather than selfishly keeping all of their endowment. These results were also in consistent with a previous study showing that third-party punishment intensity did not vary between envy (i.e., the maximal payoff of the offender was 100 MU, whereas the third party was endowed with 50 MU) and neutral (i.e., both the maximal payoff of the offender and initial endowment of the third party were 50 MU or 100 MU) conditions in a similar paradigm (Jordan, McAuliffe, & Rand, 2014).

Besides the motivation of envy, another potential motivation that might drive helping behavior is efficiency (Engelmann & Strobel, 2004), which refers to con-

[13] In practice, we might use the different choices of the same offender, matched with different victims, given the limitations in time and budget (e.g., Study 1 included 160 trials per participant; to ensure different offenders and victims in each trial we have to recruit 320 participants, which is much more difficult and less efficient). Since third-party participants completed many trials (i.e., ~100), especially during the fMRI measurement, and were not asked to memorize any initials, we assumed that they did realize that it was sometimes the same offender and thus treated the people in each trial as different individuals.

cern for maximizing the sum for all individuals in the group, since we applied the same cost ratio (i.e., 1:3) on the helping behaviors such that third-party participants could always increase the total payoff to all three parties (i.e., by producing money) via helping the victim in our cases. To fully rule out the motivation of efficiency, future studies should use a cost ratio of 1:1 (as in the normal dictator game) in the third-party decision task (see also Future Directions).

## 6.3 Empathic Concern Can Predict the Choice Preference, But Not Always

Previous evidence showed that participants as third-party deciders were biased towards an altruistic choice preference depending on the stable personality trait of empathic concern (Leliveld, et al., 2012). In Study 1, we also showed that empathic concern positively (negatively) correlated with the proportion of helping (punishment) behaviors, and also influenced the decision-making process (i.e., decision time of making altruistic choices) in a multi-shot game. The predictive effect of empathic concern on altruistic choice preference was replicated in Study 3, where participants first took part in the third-party task and then made similar decisions by considering either the offender's social norm violation or the victim's feelings. However, we did not observe the same significant results in the other two studies.

One possible areason for such an inconsistency could be contextual influences during the task. A common feature of Studies 1 and 3 (esp. in the baseline condition) is that participants in both studies were only informed about the third-party task and nothing else besides. Therefore, any decisions participants made in these two tasks can be regarded as the "natural" decisions, so that the only factor that might have influenced their choice preference was empathic concern. On the contrary, participants in two other studies, while receiving third-party task information, were also informed about the other experimental conditions at the same time, which accompanied the decision-making process along the whole task and might reduce the effect of empathic concern on altruistic choices. In particular, participants self-administered the OXT spray intra-nasally in Study 2. In Study 4, participants were given an additional instruction regarding the attention focus before they started the task in the scanner. From the results, we know that these other experimental conditions exerted an influence either on the choice behavior or its neural correlates, which in the end affects the modulatory effect of empathic concern on choice preference in the same context.

## 6.4 Distributed Neural Representation of Third-Party Altruistic Decision-making

### 6.4.1 Reward Network

In Study 1, we showed that both help and punishment choices from a third-party decider caused more activation in the striatum compared with the control condition. Since we mentioned many times in the previous section that the striatum (esp. the ventral part) is closely associated with either basic reward processing (e.g., food, water; see Haber & Knutson, 2009; Wang, Smith & Delgado, 2016) or social rewarad processing (e.g., money, positive feedback; see Bhanji & Delgado, 2014), our results suggest a hedonic component to the third-party altruistic decision-making, which is similar to the imaging findings for other forms of human altruism, such as second-party punishment (De Quervain, et al., 2004) and charitable donation (Genevsky, et al., 2013; Harbaugh, Mayr, & Burghart, 2007). In fact, such explanation is also consistent with the behavioral finding that spending the money on someone else makes people happier than keeping it for oneself (Dunn, et al., 2008), which held true even across different cultures (Aknin, et al., 2013). From a theoretical perspective, this finding might provide a potential proximate explanation for the origin of reputation-based indirect reciprocity, which complements the ultimate explanation (i.e., third-party punishment is an important mechanism that helps to enforce the development and maintenance of the social norm) mentioned in previous studies (Bendor & Swistak, 2001; Fehr & Fischbacher, 2004b).

### 6.4.2 Control Network

We also observed the involvement of the control network, especially the latera prefrontal cortex (LPFC; including the dorsal and ventral part) and the anterior cingulate cortex (ACC), during third-party altruistic decisions in our studies. To be specific, participants showed enhanced functional coupling between the LPFC and striatum during either help or punishment choices, compared with the control condition, respectively, in Study 1. Also in this study, we found that more empathic participants showed a stronger neural response during help (vs. punishment) choices. In Study 4, we showed that the ACC, as well as LPFC, was strongly activated for choices in conflict with the attention focus (i.e., helping the victim under the condition whereby participants were asked to focus on the offender's behavior, which violated a social norm, in comparison with the same choice in the baseline condition). As mentioned in the previous section, the LPFC, as well as ACC, is a crucial part of the attention network (Dosenbach, et al., 2008; Vossel, et al., 2014). More importantly, additional evidence has shown that the LPFC strongly was

strongly related to norm-related behavior in a social context. For instance, manipulating the excitement of the LPFC (esp. the right part) via non-invasive brain stimulation techniques can affect either sanction-induced sharing behavior (Ruff, et al., 2013; Strang, et al., 2014) or the acceptance of an unfair offer (Knoch et al., 2006), which involves the competition between the selfish motive and other-regarding concern. Our results further extend the role of the control network (esp. the LPFC) to altruistic decisions made in a more complex social context.

### 6.4.3 Mentalizing Network

Apart from the regions mentioned above, we also found that the mentalizing network (esp. TPJ) is involved in the third-party altruistic decision-making and the accompanying perceptual process. In Study 2, male participants who had the intranasal OXT treatment showed selectively higher activation in the left TPJ when they observed the victim being helped (vs. the placebo condition). The effect on the TPJ was stronger when participants considered either the offender's violation or the victim's feelings during decision-making, compared with when they arrived at a decision naturally in the baseline condition of Study 4. Despite there being no direct evidence from previous studies of the link between the mentalizing network and altruistic decisions made per se in social context (Buckholtz, et al., 2008; De Quervain, et al., 2004; Spitzer, Fischbacher, Herrnberger, Grön, & Fehr, 2007), our results showed that, actually mentalization might always involves in the social decision process and appears sensitive to other factors influencing such process. These findings also provide support for a theoretical framework in which mentalizing ability is regarded as a fundamental ability for the evolution of human altruism (De Waal, 2008).

### 6.4.4 Relationship with the Third-Party Punishment Neural Network

Buckholtz and Marois (2012) proposed a neural network in support of third-party (punishment) decision-making in the context of a legal judgment. In particular, the mentalizing-relevant region (esp. the TPJ) and the affect-relevant region (esp. the amygdala) of the brain encode the intention of the offender and the harmful consequence of the crime, respectively, during the scene evaluation phase. Signals from both types of information are then integrated in the MPFC, another key region closely associated with social cognition, which then sends the information on to the DLPFC for selection and implementation of the final decisions (Buckholtz & Marois, 2012). On the basis of this framework, Krueger & Hoffman (2006) refined the model by highlighting the role of the dorsal—along with the ventral—part of MPFC during information integration, and in supplementing the role of the

anterior insula in processing affective consequences (Krueger & Hoffman, 2016). Consistent with the neural circuitry mentioned above, our studies also revealed involvement of the LPFC in altruistic decision-making biased by empathic concern (the dorsal part; Study 1), and showed how the LPFC involvement reflected in cases of conflict between choice and the attention focus (the ventral part; Study 3). We also confirmed an important role of TPJ within this context, which could be further modulated by other factors (i.e., intranasal oxytocin, Study 2A; attention focus, Study 3). Furthermore, we pinpointed for the first time the hedonic component (i.e., striatum, Study 1) of third-party altruistic decision-making, which was not taken into considerations in previous work. Given the key difference between tasks (i.e., our studies adopted an unfairness-based economic decision paradigm in which both the punishment and helping options were available; the studies mentioned above adopted the criminal justice judgment paradigm in which participants could only punish), our studies basically replicated the previous findings and further extended the neural network underlying altruistic decision in the third-party context.

## 6.5 Implications for Applied Research

In companies or organizbations, a very common phenomenon is employee mistreatment (e.g., the employee is paid much less than what he/she deserves, or is demoted or even replaced by another colleague who is less competent). Several applied studies focused on developing a theoretical model to explain and predict how the third parties would respond in real life (Skarlicki & Kulik, 2004; Skarlicki, et al., 2015; Zhu, Martens, & Aquino, 2012). Usually, these kinds of models tried to characterize all of the cognitive stages, from perception of the violation, evaluation and attribution, blame to final decisions to act, together with several factors that modulated the cognitive processing in each stage. From a very general perspective, our studies could inform such cognitive models by providing more details from measures at different levels of analysis (i.e., behavior, cognition, brain activation), and even suggested other possible cognitive or affective processes during this procedure. Moreover, our studies also highlight additional factors modulating a third party's reaction that were not included in the models.

## 6.6 Future Directions

### 6.6.1 Content-Based Concerns

Although the present series of studies already investigated factors (i.e., empathic concern, oxytocin, and other-regarding attention) that are considered most likely to influence the third-party altruistic decision-making, there remain several variables that might affect people's choice in this context.

First, the cost ratio of the altruistic choice might influence a participant's (as a third party) altruistic decision-making and its neural correlates. In the current series of studies, we inherited a cost ratio of 1:3 from the original study on third-party punishment by Fehr & Fischbacher (2004), in which participants could either take 3 MU off of the payoff of the offender, or increase by 3 MU the payoff to the victim by transferring 1 MU. The purpose of setting this cost ratio is to motivate more punishment behavior. However, this gives rise to another potential explanation for helping behavior, namely efficiency. Particularly, the motivation behind helping behaviors, with this cost ratio, might merely be to create more money for the victim or even for both sides (i.e., the victim as well as the third-party decider him-/herself). There is already evidence showing that participants are less likely to punish (and even stopped punishing) the free-rider in a public game if the price of punishment is sufficiently expensive (i.e., from the condition of paying out 1 MU to decrease by 4 MU, to the condition of paying out 4 MU to decrease by 1 MU) (Carpenter, 2007). Thus, future studies might need to compare different cost ratios (e.g., cheap/equal/expensive cost ratio: 1:3/1:1/3:1) to further assess whether this influences both the choice behaviors and its neural correlates within the same paradigm.

Second, the social link between the third-party deciders and the other two parties might influence the decision and its neural correlates. In the current series of studies, all three parties are anonymous. We deliberately used such design to rule out other confounding factors in the original study. Recent studies have already shown that social relationship do affect third-party punishment behavior. For instance, participants punished the out-group offender more harshly compared with the in-group offender (Schiller, et al., 2014). Such in-group bias in third-party punishment can even emerge at the age of 6 (Jordan, McAuliffe, & Warneken, 2014). These evidences suggest that third-party helping choices might also be influenced by in-group bias or other related factors (e.g., ethnic group and degree of social distance, such as family members, friends, and strangers).

Third, the intention behind the offender's behavior might influence a third-party decider's choice and its neural correlates. In practice, a judge always passes the sentences involving different degrees of punishment to criminals, depending

on whether they committed the offense (e.g., causing a death) on purpose or by accident. Laboratory experiments also confirmed this commonsense finding, namely that participants rated offenders fully responsible for a crime as being both more blameworthy and deserving of greater punishment compared to those with diminished responsibility (Buckholtz, et al., 2015). With the third-party punishment paradigm, another study replicated this finding by showing that participants as third-party deciders meted out stronger punishments to unfair dictators if their decisions were made by themselves in comparison with the non-intention condition, in which those decisions were randomly determined by the computer (Zhong, Chark, Hsu, & Chew, 2016). Thus it also might be interesting for future studies using the third-party task to take intention into account.

Fourth, a post-hoc literature search showed that other personality traits besides empathic concern could also influence the altruistic decisions of bystanders. For example, justice sensitivity, a trait capturing the subjective readiness and strength in response to an injustice viewed from different perspectives (e.g., offender, victim, bystander, beneficiary) (Schmitt, Gollwitzer, Maes, & Arbach, 2005), has also been shown to consistently predict third-party altruistic choices in similar unequal situations (Baumert, Schlösser, & Schmitt, 2014; Baumert & Schmitt, 2016; Lotz, Baumert, Schlösser, Gresser, & Fetchenhauer, 2011). Thus, future studies on this topic should also consider other personality traits as predictive measures.

Last but not least, it would be valuable for future studies to investigate the dynamic learning procedure underlying a third-party decider's choice, within a repeated game paradigm in which participants can alter their behavior and strategy based on the offender's behavior. In the ultimatum game, a previous study adopted a norm-training paradigm showed how participant's behavior changed in response to an unpredictable shift in norm (e.g., from advantageous inequality to equality), and also how the brain encodes such a learning process (Xiang, Lohrenz, & Montague, 2013). Although there might be difficulties in directly applying a similar procedure to the third-party paradigm, this should still be investigated in the future.

## 6.6.2 Methods-Based Concerns

### 6.6.2.1 The Approach of Computational Modelling

As we mentioned in our previous studies, we always had to exclude 30% of, or even more, participants as they did not make enough altruistic choices (i.e., either help, punishment or both) to be used in the later fMRI analyses. To address this limitation, the easiest and the most straightforward solution is to recruit more participants to maintain a big sample. However, such a solution is not always feasible

in practice, as it would entail investing more money in, and to prolonging, the project.

Alternatively, this limitation could be partially solved or mitigated by taking a new approach to analyze the data, namely computational modelling. Simply speaking, computation modelling characterizes human cognition and information processing with the help of formal mathematical equations. The most important feature of this approach is that it can be used to generate more precise predictions, which can be further used to compare different hypotheses (Busemeyer & Diederich, 2010; Glimcher & Fehr, 2013). In general, this approach consists of the following steps: designing a task, coming up with assumptions, building the computational model based on those hypotheses, and estimating model parameters; if the goal of the study is to compare several competing models, then investigators also need to quantitatively compare these models (Ahn, Haines, & Zhang, 2016; Busemeyer & Diederich, 2010). Notably, all the response (or choice) data of a participant will be used to estimate the individual model parameters. This is quite different from the traditional approach, which is to categorize the data into different conditions based on the participant's choice. In this regard, we suggest that computational modelling can take fuller advantage of the data than the traditional approach.

Nowadays, the computational modelling approach is becoming more and more popular in combination with neuroscience techniques (esp. fMRI and EEG) to fill the knowledge gap regarding how, instead of what, our brain processes the information and make decisions (Forstmann & Wagenmakers, 2015; O'Doherty, Hampton, & Kim, 2007). Applying this combination approach to patients with psychiatric disorders even resulted in the emerging field of computational psychiatry (Montague, Dolan, Friston, & Dayan, 2012; Stephan & Mathys, 2014; X.-J. Wang & Krystal, 2014).

Concerning third-party altruistic behavior, a recent fMRI study adopted this approach to investigate the neural and computational mechanism underlying third-party punishment (Zhong, et al., 2016). Similar to the standard procedure, participants as third-party deciders were presented with an allocation choice between an offender and a victim (i.e., payoff of the offender/victim in MU: 50/50, 80/20, 90/10, 100/0) and then decided how much they would like to spend, from their own endowment (i.e., 160 MU per round), to punish the offender at a high cost ratio (i.e., 1:5; investing 1 MU could decrease the offender's endowment by 5 MU). To further investigate how the brain computes the subjective utility of the punishment behavior, researchers adopted a modified inequality aversion model that incorporated the parameter to capture the aversion of the third-party decider for the inequality between the offender and the victim, which thus extended the traditional egoistic model of inequality aversion (Fehr & Schmidt, 1999). Using

the individual estimated parameter, they computed the subjective utility given the participant's punishment amount in each round, and further found that it correlated with ventral medial prefrontal cortex (vmPFC) and right TPJ activation during decision-making in such a context, which provides insights into the origin of third-party punishment.

Besides static utility maximization models, more recent studies have started to apply the dynamic sequential sampling model (SSM) or attention diffusion drift model (DDM), which not only consider the choice but also take the reaction time, as well as other process measures (e.g., eye-movements), into account (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Krajbich, et al., 2010; Ratcliff, Smith, Brown, & McKoon, 2016), with respect to the field of social decision-making. For example, Krajbich and colleagues (2015) in a recent study, showed that the attentional DDM with model parameters based on a previous food choice task can accurately predict the choice and decision time in a series of social decision tasks (e.g., dictator game) for a totally different sample (Krajbich, Hare, Bartling, Morishima, & Fehr, 2015), suggesting a common cognitive mechanism between social and non-social decision-making. Another fMRI study also confirmed that the DDM can nicely fits with altruistic choice and the related decision-making process in a binary dictator game. Moreover, this study first linked the model-predicted choice with the neural correlates (i.e., vmPFC and TPJ), which generated new insights into the nature of human altruism (Hutcherson, et al., 2015). In sum, these findings showed the great potential to extend either static or dynamic models to the third-party context in the future.

#### 6.6.2.2 From GLM Analyses to Representational Analyses

Traditional GLM analysis provides information describing how neural signals correlate with different cognitive states or experimental conditions. However, it is not very successful in directly revealing how psychological functions are represented in the brain. This limitation can be addressed by more recent representational analyses, which usually includes the following two types: multi-voxel pattern analysis (MVPA) and representational similarity analysis (RSA) (Poldrack & Farah, 2015). To put it simply, MVPA, based on the principle of machine learning, is used to decode and categorize different psychological states from the activation patterns of different voxels (Haxby, 2012; Norman, et al., 2006). RSA aims to compare how the brain activation patterns differs from different stimuli or psychological states (Kriegeskorte, et al., 2008). These methods have been applied to several domains of cognitive neuroscience, such as visual perception (Bracci, Caramazza, & Peelen, 2015; Haxby et al., 2001) and memory (Lewis-Peacock & Norman, 2014; Xue et al., 2010).

Notably, a recent fMRI study, adopting a modified third-party legal justice paradigm and combining traditional general linear model (GLM) analysis and MVPA, showed that the right dorso-lateral prefrontal cortex (DLPFC) could accurately predict the punishment level of third parties, rather than the evaluation on degrees of either harmful consequence or intention of offender committing the crime, at the time of arriving at a punishment decision (Ginther, et al., 2016). These MVPA-based findings clarified the unique role of the right DLPFC in third-party punishment decisions, which avoids the problematic issue of interpretation, such as reverse inference (Poldrack, 2006), associated with GLM analyses. This study also paves the way for more future studies to apply the representational methods to the topic of third-party altruistic decision-making.

#### 6.6.2.3 Other Notes

Other possible analysis approaches by using (functional) MRI data might benefit future studies on third-party altruistic decision-making. For example, effective connectivity methods, such as dynamic causal modeling (Karl Friston, et al., 2003), can be used to investigate how information is processed by different regions sensitive to different types of decision (e.g., help or punishment) or experimental conditions, from a network perspective. Moreover, structural imaging methods, such as voxel-based morphometry (VBM) (Ashburner & Friston, 2000) and diffusion tensor imaging (DTI) (Le Bihan et al., 2001), could be adopted, together with computational modelling, to reveal the link between the anatomical basis of individual difference in choice preferences in the third-party context.

Beyond (functional) MRI, it is also possible to use other techniques to investigate the same research question from different perspectives. An interesting question would be as follows: when does the brain show the first sign of making an altruistic decision in the third-party context? This could be answered via the time-sensitive EEG technique. Last but by no means the least, future studies could also try to find the genetic basis of third-party altruistic decision-making and further reveal its link to the neural correlates measured via the above techniques.

## 6.7 Conclusion

Let us return to our original research question: Why do (some) third parties intervene at self-cost when they face a situation in which the social norm is violated and their own interests are not even affected? What factors might influence their choices in such situations? By adopting a behavioral economics paradigm in combination with neuroscience techniques (esp. fMRI and eye-tracking), the studies included in the current dissertation try to provide potential answers (or at least

some helpful insights) to these questions by integrating multiple levels of analysis (i.e., behavior, cognitive and neural levels). Together with the existing literature, we hope that the findings of the above-mentioned studies could shed some light on the underlying cognitive and neural mechanisms of third-party altruistic decision-making. However, it is always necessary to bear in mind the limitations of these studies, with respect to the design and analysis. Although there is still a long way to go to unveil the mysteries of third-party altruistic decision-making and human altruism, future studies will be promising with better designs and advanced methodologies.

# Bibliography

Abu-Akel, A., Palgi, S., Klein, E., Decety, J., & Shamay-Tsoory, S. (2015). Oxytocin increases empathy to pain when adopting the other-but not the self-perspective. *Social Neuroscience, 10*(1), 7-15.

Aharon, I., Etcoff, N., Ariely, D., Chabris, C. F., O'Connor, E., & Breiter, H. C. (2001). Beautiful faces have variable reward value: fMRI and behavioral evidence. *Neuron, 32*(3), 537-551.

Ahn, W.-Y., Haines, N., & Zhang, L. (2016). Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *bioRxiv*, 064287.

Aknin, L. B., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Burns, J., Biswas-Diener, R., . . . Norton, M. I. (2013). Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology, 104*(4), 635.

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal, 103*(418), 570-585.

Apicella, P., Scarnati, E., Ljungberg, T., & Schultz, W. (1992). Neuronal activity in monkey striatum related to the expectation of predictable environmental events. *Journal of neurophysiology, 68*(3), 945-960.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences, 8*(4), 170-177.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: one decade on. *Trends in Cognitive Sciences, 18*(4), 177-185.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage, 11*(6), 805-821.

Bandettini, P. A. (2012). Twenty years of functional MRI: the science and the stories. *Neuroimage, 62*(2), 575-588.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., & Hyde, J. S. (1992). Time course EPI of human brain function during task activation. *Magnetic resonance in medicine, 25*(2), 390-397.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37-46.

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76*, 412-427.

Batson, C. D. (2014). *The altruism question: Toward a social-psychological answer*: Psychology Press.

Batson, C. D., & Powell, A. A. (2003). Altruism and prosocial behavior. In T. Millon & M. J. Lerner (Eds.), *Handbook of Psychology* (Vol. 5, pp. 463-484).

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic Games: A Performance-based Assessment of Fairness and Altruism. *European Journal of Psychological Assessment, 30*(3), 178-192.

Baumert, A., & Schmitt, M. (2016). Justice sensitivity *Handbook of social justice theory and research* (pp. 161-180): Springer.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*(6), 1452-1469.

Bazemore, G. (1998). Restorative justice and earned redemption communities, victims, and offender reintegration. *American Behavioral Scientist, 41*(6), 768-813.

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K. M., . . . Krueger, F. (2016). Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence. *Social neuroscience*, 1-11.

Bendor, J., & Mookherjee, D. (1990). Norms, third-party sanctions, and cooperation. *Journal of Law, Economics, & Organization*, 33-63.

Bendor, J., & Swistak, P. (2001). The evolution of norms. *American Journal of Sociology, 106*(6), 1493-1545.

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual review of neuroscience, 35*, 1-23.

Bhanji, J. P., & Delgado, M. R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 61-73.

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. *Cognition, 120*(2), 215-224.

Blakemore, S.-J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing? *Annual Review of Psychology, 65*, 187-207.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences, 33*(1), 10-16.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review, 108*(3), 624.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in cognitive sciences, 8*(12), 539-546.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences, 100*(6), 3531-3535.

Boyd, R., & Richerson, P. J. (1989). The evolution of indirect reciprocity. *Social Networks, 11*(3), 213-236.

Boyd, R., & Richerson, P. J. (2002). Group beneficial norms can spread rapidly in a structured population. *Journal of theoretical biology, 215*(3), 287-296.

Bracci, S., Caramazza, A., & Peelen, M. V. (2015). Representational similarity of body parts in human occipitotemporal cortex. *The Journal of Neuroscience, 35*(38), 12977-12985.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics, 14*(3), 375-398.

Brickenkamp, R. (1995). Aufmerksamkeits-Belastungs-Test 'd2', erweiterte und neu gestaltete Auflage. *Diagnostica, 41*, 291-296.

Brüne, M., Scheele, D., Heinisch, C., Tas, C., Wischniewski, J., & Güntürkün, O. (2012). Empathy Moderates the Effect of Repetitive Transcranial Magnetic Stimulation of the Right Dorsolateral Prefrontal Cortex on Costly Punishment. *PloS one, 7*(9), e44747.

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

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature neuroscience, 15*(5), 655-661.

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From Blame to Punishment: Disrupting Prefrontal Cortex Activity Reveals Norm Enforcement Mechanisms. *Neuron, 87*(6), 1369-1380.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*: Sage.

Bzdok, D., Langner, R., Schilbach, L., Jakobs, O., Roski, C., Caspers, S., . . . Eickhoff, S. B. (2013). Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *Neuroimage, 81*, 381-392.

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

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 43*(1), 9-64.

Camerer, C., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *The Journal of Economic Perspectives, 9*(2), 209-219.

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization, 62*(4), 522-542.

Carter, C. S. (2014). Oxytocin pathways and the evolution of human behavior. *Annual Review of Psychology, 65*, 17-39.

Charvet, L., Cleary, R., Vazquez, K., Belman, A., & Krupp, L. (2014). Social cognition in pediatric-onset multiple sclerosis (MS). *Multiple Sclerosis Journal, 20*(11), 1478-1484.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology, 39*, 268-277.

Chen, X., Hackett, P. D., DeMarco, A. C., Feng, C., Stair, S., Haroon, E., . . . Rilling, J. K. (2015). Effects of oxytocin and vasopressin on the neural

response to unreciprocated cooperation within brain regions involved in stress and anxiety in men and women. *Brain Imaging and Behavior*, 1-13.

Cisler, J. M., Bush, K., & Steele, J. S. (2014). A comparison of statistical methods for detecting context-modulated functional connectivity in fMRI. *Neuroimage, 84*, 1042-1052.

Clithero, J. A., & Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, nst106.

Coke, J. S., Batson, C. D., & McDavis, K. (1978). Empathic mediation of helping: A two-stage model. *Journal of Personality and Social Psychology, 36*(7), 752-766.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron, 58*(3), 306-324.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215.

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition, 50*(1), 41-77.

Darley, J. M., & Pittman, T. S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review, 7*(4), 324-336.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113-126.

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

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., . . . Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328*(5984), 1408-1411.

De Dreu, C. K., & Kret, M. E. (2016). Oxytocin conditions intergroup relations through upregulated in-group empathy, cooperation, conformity, and defense. *Biological Psychiatry, 79*(3), 165-173.

De Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science, 305*, 1254-1258.

De Waal, F. B. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology, 59*, 279-300.

Deshpande, G., LaConte, S., James, G. A., Peltier, S., & Hu, X. (2009). Multivariate Granger causality analysis of fMRI data. *Human Brain Mapping, 30*(4), 1361-1373.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *Neuroimage, 55*(2), 705-712.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological psychiatry, 61*(6), 731-733.

Dosenbach, N. U., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in cognitive sciences, 12*(3), 99-105.

Dunn, E. W., Aknin, L. B., & Norton, M. I. (2008). Spending money on others promotes happiness. *Science, 319*(5870), 1687-1688.

Dunn, E. W., Aknin, L. B., & Norton, M. I. (2014). Prosocial Spending and Happiness Using Money to Benefit Others Pays Off. *Current Directions in Psychological Science, 23*(1), 41-47.

Eisenberg, N., & Miller, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin, 101*(1), 91.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *The American Economic Review, 94*(4), 857-869.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica, 73*(6), 2017-2030.

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews, 35*(3), 903-911.

Fehr, E., & Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences, 11*(10), 419-427.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature, 425*(6960), 785-791.

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185-190.

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

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human nature, 13*(1), 1-25.

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

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 817-868.

Feinstein, A., Magalhaes, S., Richard, J.-F., Audet, B., & Moore, C. (2014). The link between multiple sclerosis and depression. *Nature Reviews Neurology, 10*(9), 507-517.

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

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

Feng, C., Hackett, P. D., DeMarco, A. C., Chen, X., Stair, S., Haroon, E., . . . Rilling, J. K. (2015). Oxytocin and vasopressin effects on the neural response to social cooperation are modulated by sex in humans. *Brain Imaging and Behavior, 9*(4), 754-764.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: an eye-tracking analysis. *Frontiers in Psychology, 3*.

Fiedler, S., & Glöckner, A. (2015). Attention and moral behavior. *Current Opinion in Psychology, 6*, 139-144.

Fiedler, S., Glöckner, A., Nicklisch, A., & Dickert, S. (2013). Social Value Orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational Behavior and Human Decision Processes, 120*(2), 272-284.

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience, 13*(5), 538-539.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics, 10*(2), 171-178.

Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology, 33*(4), 897-913.

Fliessbach, K., Rohe, T., Linder, N. S., Trautner, P., Elger, C. E., & Weber, B. (2010). Retest reliability of reward-related BOLD signals. *Neuroimage, 50*(3), 1168-1176.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-Based Cognitive Neuroscience: A Conceptual Introduction. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 139-156): Springer.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*(3), 347-369.

Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., & Dolan, R. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage, 6*(3), 218-229.

Friston, K., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage, 19*(4), 1273-1302.

Friston, K., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping, 2*(4), 189-210.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531-534.

Gao, S., Becker, B., Luo, L., Geng, Y., Zhao, W., Yin, Y., . . . Hurlemann, R. (2016). Oxytocin, the peptide that bonds the sexes also divides them. *Proceedings of the National Academy of Sciences*, 201602620.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*(4), 523-552.

Genevsky, A., Västfjäll, D., Slovic, P., & Knutson, B. (2013). Neural Underpinnings of the Identifiable Victim Effect: Affect Shifts Preferences for Giving. *The Journal of Neuroscience, 33*(43), 17188-17196.

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience, 9*.

Gimpl, G., & Fahrenholz, F. (2001). The oxytocin receptor system: structure, function, and regulation. *Physiological reviews, 81*(2), 629-683.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. *The Journal of Neuroscience, 36*(36), 9420-9434.

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

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and human behavior, 24*(3), 153-172.

Gitelman, D. R., Penny, W. D., Ashburner, J., & Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage, 19*(1), 200-207.

Glass, L., Moody, L., Grafman, J., & Krueger, F. (2015). Neural signatures of third-party punishment: evidence from penetrating traumatic brain injury. *Social cognitive and affective neuroscience*, nsv105.

Glimcher, P. W., & Fehr, E. (2013). *Neuroeconomics: Decision making and the brain*: Academic Press.

Glöckner, A., & Herbold, A. K. (2011). An eye‐tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making, 24*(1), 71-98.

Gordon, I., Vander Wyk, B. C., Bennett, R. H., Cordeaux, C., Lucas, M. V., Eilbott, J. A., . . . Pelphrey, K. A. (2013). Oxytocin enhances brain function in children with autism. *Proceedings of the National Academy of Sciences, 110*(52), 20953-20958.

Gray, K., Ward, A. F., & Norton, M. I. (2014). Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General, 143*(1), 247.

Greiner, B. (2004). The online recruitment system ORSEE 2.0–a guide for the organization of experiments in economics. *University of Cologne, Working Paper Series in Economics, 10*, 2004.

Gromet, D. M., & Darley, J. M. (2009). Punishment and beyond: Achieving justice through the satisfaction of multiple goals. *Law & Society Review, 43*(1), 1-38.

Gu, X., Gao, Z., Wang, X., Liu, X., Knight, R. T., Hof, P. R., & Fan, J. (2012). Anterior insular cortex is necessary for empathetic pain perception. *Brain, 135*(9), 2726-2735.

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends in cognitive sciences, 18*(4), 194-202.

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second-and third-party punishment behavior. *Cognition, 133*(1), 97-103.

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology, 65*, 94-104.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization, 3*(4), 367-388.

Haas, B. W., Anderson, I. W., & Smith, J. M. (2013). Navigating the complex path between the oxytocin receptor gene (OXTR) and cooperation: an endophenotype approach. *Front. Hum. Neurosci, 7*(801), 10.3389.

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

Hallett, M. (2000). Transcranial magnetic stimulation and the human brain. *Nature, 406*(6792), 147-150.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology, 7*(1), 17-52.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316*(5831), 1622-1625.

Hare, T. A., Malmaud, J., & Rangel, A. (2011). Focusing attention on the health aspects of foods changes value signals in vmPFC and improves dietary choice. *The Journal of Neuroscience, 31*(30), 11077-11087.

Hauser, M. D., Chen, M. K., Chen, F., & Chuang, E. (2003). Give unto others: genetically unrelated cotton-top tamarin monkeys preferentially give food to those who altruistically give food back. *Proceedings of the Royal Society of London B: Biological Sciences, 270*(1531), 2363-2370.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage, 62*(2), 852-855.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*(4), 709-722.

Hechter, M., & Opp, K.-D. (2001). *Social norms*: Russell Sage Foundation.

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

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., . . . Henrich, N. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767-1770.

Horne, C. (2001). Sociological perspectives on the emergence of social norms. In M. Hechter & K. D. Opp (Eds.), *Social Norms* (pp. 3-34): Russell Sage Foundation.

Hruschka, D. J. (2010). *Friendship: Development, ecology, and evolution of a relationship* (Vol. 5): Univ of California Press.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1): Sinauer Associates Sunderland.

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., . . . Maier, W. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *The Journal of Neuroscience, 30*(14), 4999-5007.

Hutcherson, C., Bushong, B., & Rangel, A. (2015). A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron, 87*(2), 451-462.

Hutcherson, C., & Rangel, A. (2014). Ethics or empathy? Different appraisals activate distinct social cognitive brain regions during altruistic choice. *Annual Conference of Society for Neuroeconomics, Miami, USA, Sept.26-28, 2014*.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron, 58*(2), 284-294.

Izuma, K., Saito, D. N., & Sadato, N. (2010). Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of cognitive neuroscience, 22*(4), 621-631.

Jordan, J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*, 473-476.

Jordan, J., McAuliffe, K., & Rand, D. (2014). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, 1-23.

Jordan, J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences, 111*(35), 12710-12715.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285-S300.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., & Brugger, P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *The Journal of Neuroscience, 26*(24), 6469-6472.

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.

Konovalov, A., & Krajbich, I. (2016). Over a Decade of Neuroeconomics What Have We Learned? *Organizational Research Methods*, 1094428116644502.

Koopmans, R., & Rebers, S. (2009). Collective action in culturally similar and dissimilar groups: an experiment on parochialism, conditional cooperation, and their linkages. *Evolution and human behavior, 30*(3), 201-211.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*(7042), 673-676.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292-1298.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A Common Mechanism Underlying Food Choice and Social Decisions. *PLoS Comput Biol, 11*(10), e1004371.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences, 108*(33), 13852-13857.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Krueger, F., & Hoffman, M. (2016). The Emerging Neuroscience of Third-Party Punishment. *Trends in neurosciences*.

Krueger, F., Hoffman, M., Walter, H., & Grafman, J. (2014). An fMRI investigation of the effects of belief in free will on third-party punishment. *Social Cognitive and Affective Neuroscience*, nst092.

Krueger, F., Parasuraman, R., Moody, L., Twieg, P., de Visser, E., McCabe, K., . . . Lee, M. R. (2013). Oxytocin selectively increases perceptions of harm for victims but not the desire to punish offenders of criminal offenses. *Social cognitive and affective neuroscience, 8*(5), 494-498.

Kurzban, R., Burton-Chellew, M. N., & West, S. A. (2015). The Evolution of Altruism in Humans. *Annual Review of Psychology, 66*, 575-599.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., . . . Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences, 89*(12), 5675-5679.

Lancaster, K., Carter, C. S., Pournajafi-Nazarloo, H., Karaoli, T., Lillard, T. S., Jack, A., . . . Connelly, J. J. (2015). Plasma oxytocin explains individual differences in neural substrates of social perception. *Frontiers in human neuroscience, 9*.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological bulletin, 89*(2), 308.

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: concepts and applications. *Journal of magnetic resonance imaging, 13*(4), 534-546.

Lee, S. M., & McCarthy, G. (2014). Functional Heterogeneity and Convergence in the Right Temporoparietal Junction. *Cerebral Cortex*, bhu292.

Leliveld, M. C., Dijk, E., & 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, 42*(2), 135-140.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Competition between items in working memory leads to forgetting. *Nature communications, 5*.

Liebrand, W. B., Jansen, R. W., Rijken, V. M., & Suhre, C. J. (1986). Might over morality: Social values and the perception of other players in experimental games. *Journal of Experimental Social Psychology, 22*(3), 203-215.

Liebrand, W. B., & McClintock, C. G. (1988). The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. *European Journal of Personality, 2*(3), 217-230.

Loewenstein, G., Rick, S., & Cohen, J. D. (2008). Neuroeconomics. *Annual Review of Psychology, 59*, 647-672.

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., & Fetchenhauer, D. (2011). Individual differences in third-party interventions: How justice sensitivity shapes altruistic punishment. *Negotiation and Conflict Management Research, 4*(4), 297-313.

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology, 47*(2), 477-480.

Ma, Y., Shamay-Tsoory, S., Han, S., & Zink, C. F. (2016). Oxytocin and Social Adaptation: Insights from Neuroimaging Studies of Healthy and Clinical Populations. *Trends in cognitive sciences, 20*(2), 133-145.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472), 1835-1838.

Makwana, A., Polania, R., & Hare, T. (2014). Behavioral and neural effects of highlighting monetary gain in the Ultimatum Game. *Annual Conference of Society for Neuroeconomics, Miami, USA, Sept.26-28, 2014*.

Mars, R. B., Sallet, J., Schüffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral cortex, 22*(8), 1894-1903.

Marsh, N., Scheele, D., Gerhardt, H., Strang, S., Enax, L., Weber, B., . . . Hurlemann, R. (2015). The Neuropeptide Oxytocin Induces a Social Altruism Bias. *The Journal of Neuroscience, 35*(47), 15696-15701.

McCall, C., Steinbeis, N., Ricard, M., & Singer, T. (2014). Compassion meditators show less anger, less punishment, and more compensation of victims in response to fairness violations. *Frontiers in Behavioral Neuroscience, 8*.

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage, 61*(4), 1277-1286.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167-202.

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral cortex, 18*(2), 262-271.

Mitzkewitz, M., & Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory, 22*(2), 171-198.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences, 16*(1), 72-80.

Naqvi, N., Shiv, B., & Bechara, A. (2006). The role of emotion in decision making a cognitive neuroscience perspective. *Current Directions in Psychological Science, 15*(5), 260-264.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives in Psychological Science, in press*.

Nelissen, R. M., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making, 4*(7), 543-553.

Nelson, S. M., Dosenbach, N. U., Cohen, A. L., Wheeler, M. E., Schlaggar, B. L., & Petersen, S. E. (2010). Role of the anterior insula in task-level control and focal attention. *Brain Structure and Function, 214*(5-6), 669-680.

Nitsche, M., & Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *The Journal of Physiology, 527*(3), 633-639.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences, 10*(9), 424-430.

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature, 393*(6685), 573-577.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291-1298.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences, 1104*(1), 35-53.

O'Reilly, J. X., Woolrich, M. W., Behrens, T. E., Smith, S. M., & Johansen-Berg, H. (2012). Tools of the trade: psychophysiological interactions and functional connectivity. *Social cognitive and affective neuroscience, 7*(5), 604-609.

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences, 87*(24), 9868-9872.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences, 89*(13), 5951-5955.

Olds, J., & Milner, P. (1954). Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology, 47*(6), 419.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta psychologica, 144*(1), 190-206.

Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences, 87*(1), 256-259.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research, 42*(1), 107-123.

Patil, I., Young, L., Sinay, V., & Gleichgerrcht, E. (2016). Elevated moral condemnation of third-party violations in multiple sclerosis patients. *Social Neuroscience*, 1-22.

Paulhus, D. L., & Carey, J. M. (2011). The FAD–Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment, 93*(1), 96-104.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society of London B: Biological Sciences, 280*(1758), 20122723.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review Psychology, 56*, 365-392.

Pfeiffer, T., Rutte, C., Killingback, T., Taborsky, M., & Bonhoeffer, S. (2005). Evolution of cooperation by generalized reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences, 272*(1568), 1115-1120.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59-63.

Poldrack, R. A., & Farah, M. J. (2015). Progress and challenges in probing the human brain. *Nature, 526*(7573), 371-379.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*: Cambridge University Press.

Radke, S., & de Bruijn, E. R. (2015). Does oxytocin affect mind-reading? A replication study. *Psychoneuroendocrinology, 60*, 75-81.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98-103.

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69*(4), 993-1003.

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

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences, 17*(8), 413-425.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in cognitive sciences, 20*(4), 260-281.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences, 109*(37), 14824-14829.

Rilling, J. K., DeMarco, A. C., Hackett, P. D., Chen, X., Gautam, P., Stair, S., . . . Patel, R. (2014). Sex differences in the neural and behavioral response to intranasal oxytocin and vasopressin during human social interaction. *Psychoneuroendocrinology, 39*, 237-248.

Rilling, J. K., DeMarco, A. C., Hackett, P. D., Thompson, R., Ditzen, B., Patel, R., & Pagnoni, G. (2012). Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. *Psychoneuroendocrinology, 37*(4), 447-461.

Rocca, M. A., Amato, M. P., De Stefano, N., Enzinger, C., Geurts, J. J., Penner, I.-K., . . . Filippi, M. (2015). Clinical and imaging assessment of cognitive dysfunction in multiple sclerosis. *The Lancet Neurology, 14*(3), 302-317.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549-562.

Ruff, C. C., & Huettel, S. A. (2013). Experimental methods in cognitive neuroscience *Neuroeconomics: Decision making and the brain* (Vol. 2, pp. 77-108).

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

Sabbagh, C., & Schmitt, M. (2016). *Handbook of social justice theory and research*: Springer.

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience, 7*(5), 499-500.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2014). Deconstructing and reconstructing theory of mind. *Trends in cognitive sciences*.

Scheele, D., Schwering, Elison, Spunt, Maier, & Hurlemann, R. (2015). A human tendency to anthropomorphize is enhanced by oxytocin. *European Neuropsychopharmacology*, in press.

Scheele, D., Striepens, N., Kendrick, K. M., Schwering, C., Noelle, J., Wille, A., . . . Hurlemann, R. (2014). Opposing effects of oxytocin on moral judgment in males and females. *Human brain mapping, 35*(12), 6067-6076.

Scheele, D., Wille, A., Kendrick, K. M., Stoffel-Wagner, B., Becker, B., Güntürkün, O., . . . Hurlemann, R. (2013). Oxytocin enhances brain reward system responses in men viewing the face of their female partner. *Proceedings of the National Academy of Sciences, 110*(50), 20308-20313.

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and human behavior, 35*(3), 169-175.

Schmitt, M., Gollwitzer, M., Maes, J., & Arbach, D. (2005). Justice sensitivity: Assessment and location in the personality space. *European Journal of Psychological Assessment, 21*(3), 202-211.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological review, 84*(1), 1.

Schultz, W. (2015). Reward. In A. W. Toga (Ed.), *Brain mapping: An encyclopedic reference* (Vol. 2, pp. 643-651): Academic Press.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews, 42*, 9-34.

Selten, R. (1965). *Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes*.

Seyfarth, R. M., & Cheney, D. L. (2012). The evolutionary origins of friendship. *Annual review of psychology, 63*, 153-177.

Shackman, A. J., Salomons, T. V., Slagter, H. A., Fox, A. S., Winter, J. J., & Davidson, R. J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nature Reviews Neuroscience, 12*(3), 154-167.

Shamay-Tsoory, S. G. (2010). Oxytocin, social salience, and social approach. *Biological psychiatry, 67*(6), e35.

Shamay-Tsoory, S. G., & Abu-Akel, A. (2016). The social salience hypothesis of oxytocin. *Biological psychiatry, 79*(3), 194-202.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., . . . Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry, 59*, 22-33.

Skarlicki, D. P., & Kulik, C. T. (2004). Third-party reactions to employee (mis) treatment: A justice perspective. *Research in organizational behavior, 26*, 183-229.

Skarlicki, D. P., O'Reilly, J., & Kulik, C. T. (2015). The third-party perspective of (in) justice. In R. Cropanzano & M. Ambrose (Eds.), *Oxford Handbook of Justice in Work Organizations* (pp. 235-255): Oxford University Press.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108-131.

Sommers, S. R., & Ellsworth, P. C. (2000). Race in the courtroom: Perceptions of guilt and dispositional attributions. *Personality and Social Psychology Bulletin, 26*(11), 1367-1379.

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., & Vagg, P. R. (1970). State-trait anxiety inventory (STAI). *BiB 2010*, 180.

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

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current opinion in neurobiology, 25*, 85-92.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. (2014). Be nice if you have to-The neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, nsu114.

Strang, S., Grote, X., Kuss, K., Park, S. Q., & Weber, B. (2016). Generalized Negative Reciprocity in the Dictator Game–How to Interrupt the Chain of Unfairness. *Scientific Reports, 6*.

Striepens, N., Kendrick, K. M., Maier, W., & Hurlemann, R. (2011). Prosocial effects of oxytocin and clinical evidence for its therapeutic potential. *Frontiers in Neuroendocrinology, 32*(4), 426-450.

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

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*(2), 149-178.

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological reviews, 88*(1), 37-57.

Toi, M., & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology, 43*(2), 281.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 35-57.

Turillo, C. J., Folger, R., Lavelle, J. J., Umphress, E. E., & Gee, J. O. (2002). Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes, 89*(1), 839-865.

Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law and Society Review*, 237-265.

Van Lange, P. A., Ouwerkerk, J. W., & Tazelaar, M. J. (2002). How to overcome the detrimental effects of noise in social interaction: the benefits of generosity. *Journal of Personality and Social Psychology, 82*(5), 768.

Vossel, S., Geng, J. J., & Fink, G. R. (2014). Dorsal and ventral attention systems distinct neural circuits but collaborative roles. *The Neuroscientist, 20*(2), 150-159.

Wang, K. S., Smith, D. V., & Delgado, M. R. (2016). Using fMRI to Study Reward Processing in Humans: Past, Present, and Future. *Journal of Neurophysiology*, jn. 00333.02015.

Wang, X.-J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron, 84*(3), 638-654.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063.

Will, G.-J., Crone, E. A., van den Bos, W., & Güroğlu, B. (2013). Acting on observed social exclusion: Developmental perspectives on punishment of excluders and compensation of victims. *Developmental psychology, 49*(12), 2236.

Williams, K. D., Cheung, C. K., & Choi, W. (2000). Cyberostracism: effects of being ignored over the Internet. *Journal of personality and social psychology, 79*(5), 748.

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. *Neuroimage, 42*(1), 343-356.

Wittfoth-Schardt, D., Gründing, J., Wittfoth, M., Lanfermann, H., Heinrichs, M., Domes, G., . . . Waller, C. (2012). Oxytocin modulates neural reactivity to children's faces as a function of social salience. *Neuropsychopharmacology, 37*(8), 1799-1807.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience, 33*(3), 1099-1108.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science, 330*(6000), 97-101.

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive Intergroup Bias in Third-Party Punishment. *Journal of Experimental Psychology: General*.

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PloS one, 2*(11), e1128.

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *Neuroimage, 129*, 95-104.

Zhu, L. L., Martens, J. P., & Aquino, K. (2012). Third party responses to justice failure: An identity-based meaning maintenance model. *Organizational Psychology Review*, 2041386611434655.

# List of Figures

# List of Tables

# Appendix

Appendix Table 1. All items in the IRI

| No. | Subscale | Content | Answer Scale |
|---|---|---|---|
| 1 | FS | I daydream and fantasize, with some regularity, about things that might happen to me. | 0---1---2---3---4 |
| 2 | EC | I often have tender, concerned feelings for people less fortunate than me. | 0---1---2---3---4 |
| 3 | PT | I sometimes find it difficult to see things from the "other guy's" point of view.* | 0---1---2---3---4 |
| 4 | EC | Sometimes I don't feel very sorry for other people when they are having problems.* | 0---1---2---3---4 |
| 5 | FS | I really get involved with the feelings of the characters in a novel. | 0---1---2---3---4 |
| 6 | PD | In emergency situations, I feel apprehensive and ill-at-ease. | 0---1---2---3---4 |
| 7 | FS | I am usually objective when I watch a movie or play, and I don't often get completely caught up in it.* | 0---1---2---3---4 |
| 8 | PT | I try to look at everybody's side of a disagreement before I make a decision. | 0---1---2---3---4 |
| 9 | EC | When I see someone being taken advantage of, I feel kind of protective towards them. | 0---1---2---3---4 |
| 10 | PD | I sometimes feel helpless when I am in the middle of a very emotional situation. | 0---1---2---3---4 |
| 11 | PT | I sometimes try to understand my friends better by imagining how things look from their perspective. | 0---1---2---3---4 |
| 12 | FS | Becoming extremely involved in a good book or movie is somewhat rare for me.* | 0---1---2---3---4 |
| 13 | PD | When I see someone get hurt, I tend to remain calm.* | 0---1---2---3---4 |
| 14 | EC | Other people's misfortunes do not usually disturb me a great deal.* | 0---1---2---3---4 |
| 15 | PT | If I'm sure I'm right about something, I don't waste much time listening to other people's arguments.* | 0---1---2---3---4 |
| 16 | FS | After seeing a play or movie, I have felt as though I were one of the characters. | 0---1---2---3---4 |
| 17 | PD | Being in a tense emotional situation scares me. | 0---1---2---3---4 |

| | | | |
|---|---|---|---|
| 18 | EC | When I see someone being treated unfairly, I sometimes don't feel very much pity for them.* | 0---1---2---3---4 |
| 19 | PD | I am usually pretty effective in dealing with emergencies.* | 0---1---2---3---4 |
| 20 | EC | I am often quite touched by things that I see happen. | 0---1---2---3---4 |
| 21 | PT | I believe that there are two sides to every question and try to look at them both. | 0---1---2---3---4 |
| 22 | EC | I would describe myself as a pretty soft-hearted person. | 0---1---2---3---4 |
| 23 | FS | When I watch a good movie, I can very easily put myself in the place of a leading character. | 0---1---2---3---4 |
| 24 | PD | I tend to lose control during emergencies. | 0---1---2---3---4 |
| 25 | PT | When I'm upset at someone, I usually try to "put myself in his shoes" for a while. | 0---1---2---3---4 |
| 26 | FS | When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me. | 0---1---2---3---4 |
| 27 | PD | When I see someone who badly needs help in an emergency, I go to pieces. | 0---1---2---3---4 |
| 28 | PT | Before criticizing somebody, I try to imagine how I would feel if I were in their place. | 0---1---2---3---4 |

Note: 0 refers to “does not describe me well”, 4 refers to “describes me very well”. * refers to reverse-scored items. IRI includes 28 items in total and is consisted of four subscales with 7 items respectively. IRI refers to interpersonal reactivity index; PT refers to perspective-taking; FS refers to fantasy; EC refers to empathic concern; PD refers to personal distress.

Appendix Table 2. Stimuli used in the fMRI task (Study 3)

| Block No. | Offer Type | Payoff | |
|---|---|---|---|
| | | Offender | Victim |
| 1 | fair | 4.53 | 4.47 |
| 1 | unfair | 6.83 | 3.17 |
| 1 | unfair | 6.98 | 2.02 |
| 1 | unfair | 7.94 | 1.06 |
| 1 | unfair | 8.00 | 2.00 |
| 1 | unfair | 8.04 | 2.96 |
| 1 | unfair | 8.92 | 2.08 |
| 1 | unfair | 9.19 | 0.81 |
| 2 | fair | 5.00 | 5.00 |
| 2 | unfair | 6.98 | 3.03 |
| 2 | unfair | 7.17 | 1.83 |
| 2 | unfair | 7.98 | 1.02 |
| 2 | unfair | 8.02 | 1.98 |
| 2 | unfair | 8.11 | 2.89 |
| 2 | unfair | 8.97 | 2.03 |
| 2 | unfair | 9.02 | 0.98 |
| 3 | fair | 5.55 | 5.45 |
| 3 | unfair | 6.81 | 2.19 |
| 3 | unfair | 6.84 | 3.16 |
| 3 | unfair | 7.96 | 3.04 |
| 3 | unfair | 8.07 | 0.93 |
| 3 | unfair | 8.16 | 1.84 |
| 3 | unfair | 9.01 | 1.99 |
| 3 | unfair | 9.03 | 0.97 |
| 4 | fair | 5.53 | 5.47 |
| 4 | unfair | 6.89 | 3.11 |
| 4 | unfair | 7.04 | 1.96 |
| 4 | unfair | 7.86 | 2.14 |
| 4 | unfair | 8.03 | 0.97 |
| 4 | unfair | 8.09 | 2.91 |
| 4 | unfair | 8.83 | 2.17 |
| 4 | unfair | 8.89 | 1.11 |

| | | | |
|---|---|---|---|
| 5 | fair | 4.54 | 4.46 |
| 5 | unfair | 6.97 | 3.04 |
| 5 | unfair | 7.10 | 1.90 |
| 5 | unfair | 7.91 | 2.09 |
| 5 | unfair | 7.92 | 1.08 |
| 5 | unfair | 8.20 | 2.80 |
| 5 | unfair | 8.99 | 1.01 |
| 5 | unfair | 9.13 | 1.87 |
| 6 | fair | 5.01 | 4.99 |
| 6 | unfair | 7.01 | 2.99 |
| 6 | unfair | 7.14 | 1.86 |
| 6 | unfair | 7.82 | 1.18 |
| 6 | unfair | 8.13 | 1.87 |
| 6 | unfair | 8.18 | 2.82 |
| 6 | unfair | 8.82 | 1.18 |
| 6 | unfair | 9.19 | 1.81 |
| 7 | fair | 5.50 | 5.50 |
| 7 | unfair | 6.85 | 3.15 |
| 7 | unfair | 6.97 | 2.03 |
| 7 | unfair | 7.88 | 3.12 |
| 7 | unfair | 7.94 | 2.06 |
| 7 | unfair | 8.17 | 0.83 |
| 7 | unfair | 8.99 | 2.01 |
| 7 | unfair | 9.13 | 0.87 |
| 8 | fair | 5.03 | 4.97 |
| 8 | unfair | 7.11 | 2.89 |
| 8 | unfair | 7.11 | 1.89 |
| 8 | unfair | 7.89 | 2.11 |
| 8 | unfair | 8.06 | 0.94 |
| 8 | unfair | 8.17 | 2.83 |
| 8 | unfair | 8.87 | 2.13 |
| 8 | unfair | 8.97 | 1.03 |

| | | | |
|---|---|---|---|
| 9 | fair | 5.04 | 4.96 |
| 9 | unfair | 6.89 | 2.11 |
| 9 | unfair | 7.16 | 2.84 |
| 9 | unfair | 7.90 | 3.10 |
| 9 | unfair | 7.98 | 2.02 |
| 9 | unfair | 8.14 | 0.86 |
| 9 | unfair | 8.82 | 2.18 |
| 9 | unfair | 9.01 | 0.99 |
| 10 | fair | 5.51 | 5.49 |
| 10 | unfair | 6.90 | 2.10 |
| 10 | unfair | 7.09 | 2.91 |
| 10 | unfair | 7.83 | 1.17 |
| 10 | unfair | 7.93 | 2.07 |
| 10 | unfair | 8.16 | 2.84 |
| 10 | unfair | 8.84 | 2.16 |
| 10 | unfair | 9.14 | 0.86 |
| 11 | fair | 4.55 | 4.45 |
| 11 | unfair | 7.05 | 1.95 |
| 11 | unfair | 7.05 | 2.95 |
| 11 | unfair | 7.95 | 2.05 |
| 11 | unfair | 8.10 | 2.90 |
| 11 | unfair | 8.12 | 0.88 |
| 11 | unfair | 8.84 | 1.16 |
| 11 | unfair | 9.04 | 1.96 |
| 12 | fair | 5.05 | 4.95 |
| 12 | unfair | 6.93 | 2.07 |
| 12 | unfair | 7.12 | 2.88 |
| 12 | unfair | 7.90 | 1.10 |
| 12 | unfair | 8.03 | 2.97 |
| 12 | unfair | 8.12 | 1.88 |
| 12 | unfair | 9.10 | 1.90 |
| 12 | unfair | 9.17 | 0.83 |

| | | | |
|---|---|---|---|
| 13 | fair | 4.50 | 4.50 |
| 13 | unfair | 6.86 | 2.14 |
| 13 | unfair | 6.94 | 3.07 |
| 13 | unfair | 7.81 | 1.19 |
| 13 | unfair | 7.89 | 3.11 |
| 13 | unfair | 8.05 | 1.95 |
| 13 | unfair | 9.09 | 0.91 |
| 13 | unfair | 9.09 | 1.91 |
| 14 | fair | 5.02 | 4.98 |
| 14 | unfair | 7.00 | 2.00 |
| 14 | unfair | 7.13 | 2.87 |
| 14 | unfair | 7.96 | 1.04 |
| 14 | unfair | 8.05 | 2.95 |
| 14 | unfair | 8.17 | 1.83 |
| 14 | unfair | 8.90 | 1.10 |
| 14 | unfair | 9.12 | 1.88 |
| 15 | fair | 4.52 | 4.48 |
| 15 | unfair | 7.13 | 1.87 |
| 15 | unfair | 7.18 | 2.82 |
| 15 | unfair | 7.86 | 1.14 |
| 15 | unfair | 7.99 | 3.01 |
| 15 | unfair | 8.18 | 1.82 |
| 15 | unfair | 8.87 | 1.13 |
| 15 | unfair | 8.88 | 2.12 |
| 16 | fair | 4.51 | 4.49 |
| 16 | unfair | 6.96 | 2.04 |
| 16 | unfair | 7.19 | 2.81 |
| 16 | unfair | 7.82 | 2.18 |
| 16 | unfair | 8.00 | 3.00 |
| 16 | unfair | 8.02 | 0.98 |
| 16 | unfair | 9.16 | 0.84 |
| 16 | unfair | 9.20 | 1.80 |

| | | | |
|---|---|---|---|
| 17 | fair | 5.54 | 5.46 |
| 17 | unfair | 7.15 | 2.85 |
| 17 | unfair | 7.19 | 1.81 |
| 17 | unfair | 7.84 | 3.16 |
| 17 | unfair | 7.97 | 1.03 |
| 17 | unfair | 8.20 | 1.80 |
| 17 | unfair | 9.05 | 1.95 |
| 17 | unfair | 9.08 | 0.92 |
| 18 | fair | 5.52 | 5.48 |
| 18 | unfair | 6.96 | 3.04 |
| 18 | unfair | 7.09 | 1.91 |
| 18 | unfair | 7.82 | 3.18 |
| 18 | unfair | 7.88 | 2.12 |
| 18 | unfair | 7.95 | 1.05 |
| 18 | unfair | 8.88 | 1.12 |
| 18 | unfair | 9.18 | 1.82 |

Appendix Table 3. Offer combinations shown for the on-line part of the experiment (Study 4)

| Block No. | Options used in the eye-tracking study | | Alternative option | |
|---|---|---|---|---|
| | Offender | Victim | Offender | Victim |
| 1 | 4.14 | 1.99 | 1.03 | 1.03 |
| 1 | 3.88 | 2.87 | 0.88 | 0.88 |
| 1 | 4.98 | 2.89 | 1.11 | 1.11 |
| 1 | 4.94 | 3.94 | 0.97 | 0.97 |
| 1 | 6.13 | 2.84 | 1.20 | 1.20 |
| 1 | 6.10 | 3.97 | 1.03 | 1.03 |
| 1 | 6.18 | 5.10 | 0.90 | 0.90 |
| 1 | 6.82 | 4.15 | 1.02 | 1.02 |
| 1 | 7.18 | 4.93 | 0.96 | 0.96 |
| 1 | 8.18 | 4.16 | 1.12 | 1.12 |
| 1 | 6.87 | 6.14 | 0.97 | 0.97 |
| 1 | 8.02 | 4.82 | 1.10 | 1.10 |
| 1 | 8.04 | 6.00 | 1.16 | 1.16 |
| 1 | 8.91 | 4.92 | 0.85 | 0.85 |
| 1 | 4.92 | 0.83 | 0.88 | 0.88 |
| 1 | 4.88 | 2.11 | 1.09 | 1.09 |
| 1 | 6.13 | 1.00 | 1.08 | 1.08 |
| 1 | 6.02 | 1.93 | 1.08 | 1.08 |
| 1 | 6.88 | 1.08 | 1.14 | 1.14 |
| 1 | 7.05 | 1.95 | 1.08 | 1.08 |
| 1 | 7.95 | 1.05 | 1.09 | 1.09 |
| 1 | 6.97 | 2.93 | 0.85 | 0.85 |
| 1 | 8.11 | 2.03 | 0.91 | 0.91 |
| 1 | 8.83 | 0.86 | 0.98 | 0.98 |
| 1 | 8.16 | 3.18 | 0.92 | 0.92 |
| 1 | 9.12 | 1.87 | 1.18 | 1.18 |
| 1 | 8.90 | 2.86 | 0.85 | 0.85 |
| 1 | 9.11 | 3.88 | 1.01 | 1.01 |
| 1 | 3.18 | 3.18 | 0.84 | 0.84 |
| 1 | 4.17 | 4.17 | 1.09 | 1.09 |
| 1 | 5.00 | 5.00 | 0.90 | 0.90 |
| 1 | 6.07 | 6.07 | 1.05 | 1.05 |
| 1 | 6.87 | 6.87 | 1.04 | 1.04 |

| | | | | |
|---|---|---|---|---|
| 2 | 4.05 | 2.10 | 1.14 | 1.14 |
| 2 | 4.03 | 2.98 | 0.93 | 0.93 |
| 2 | 5.02 | 3.17 | 1.16 | 1.16 |
| 2 | 5.06 | 3.86 | 1.18 | 1.18 |
| 2 | 5.90 | 2.87 | 1.07 | 1.07 |
| 2 | 5.97 | 4.11 | 0.82 | 0.82 |
| 2 | 6.05 | 4.85 | 0.87 | 0.87 |
| 2 | 6.97 | 4.08 | 0.81 | 0.81 |
| 2 | 7.17 | 4.87 | 1.13 | 1.13 |
| 2 | 8.14 | 3.91 | 1.09 | 1.09 |
| 2 | 7.12 | 5.94 | 0.99 | 0.99 |
| 2 | 7.88 | 5.10 | 1.03 | 1.03 |
| 2 | 7.92 | 6.18 | 0.96 | 0.96 |
| 2 | 8.92 | 5.10 | 0.82 | 0.82 |
| 2 | 5.03 | 0.94 | 1.09 | 1.09 |
| 2 | 5.06 | 2.08 | 0.94 | 0.94 |
| 2 | 6.08 | 0.82 | 1.06 | 1.06 |
| 2 | 5.83 | 1.94 | 1.03 | 1.03 |
| 2 | 7.11 | 0.80 | 0.99 | 0.99 |
| 2 | 7.14 | 2.18 | 1.16 | 1.16 |
| 2 | 8.14 | 1.00 | 0.94 | 0.94 |
| 2 | 6.90 | 2.96 | 0.87 | 0.87 |
| 2 | 7.87 | 2.05 | 1.15 | 1.15 |
| 2 | 9.08 | 1.16 | 1.15 | 1.15 |
| 2 | 8.10 | 2.99 | 0.82 | 0.82 |
| 2 | 9.12 | 1.83 | 1.19 | 1.19 |
| 2 | 8.84 | 2.88 | 0.88 | 0.88 |
| 2 | 8.89 | 4.07 | 1.04 | 1.04 |
| 2 | 3.09 | 3.09 | 1.00 | 1.00 |
| 2 | 3.84 | 3.84 | 0.85 | 0.85 |
| 2 | 4.86 | 4.86 | 1.18 | 1.18 |
| 2 | 6.02 | 6.02 | 1.14 | 1.14 |
| 2 | 6.84 | 6.84 | 1.01 | 1.01 |

| | | | | |
|---|---|---|---|---|
| 3 | 3.81 | 1.94 | 1.09 | 1.09 |
| 3 | 4.14 | 2.92 | 0.97 | 0.97 |
| 3 | 5.13 | 3.14 | 0.95 | 0.95 |
| 3 | 5.03 | 3.90 | 0.87 | 0.87 |
| 3 | 5.80 | 2.82 | 0.82 | 0.82 |
| 3 | 5.92 | 3.98 | 1.01 | 1.01 |
| 3 | 5.83 | 5.10 | 1.16 | 1.16 |
| 3 | 6.81 | 3.89 | 1.02 | 1.02 |
| 3 | 7.20 | 4.99 | 0.87 | 0.87 |
| 3 | 7.86 | 4.02 | 0.94 | 0.94 |
| 3 | 7.06 | 6.17 | 0.85 | 0.85 |
| 3 | 7.86 | 5.02 | 1.11 | 1.11 |
| 3 | 7.95 | 5.90 | 0.94 | 0.94 |
| 3 | 8.86 | 4.85 | 1.19 | 1.19 |
| 3 | 4.93 | 0.88 | 1.14 | 1.14 |
| 3 | 4.90 | 2.12 | 0.88 | 0.88 |
| 3 | 5.98 | 1.13 | 0.87 | 0.87 |
| 3 | 6.19 | 2.09 | 1.10 | 1.10 |
| 3 | 7.01 | 0.92 | 0.98 | 0.98 |
| 3 | 7.20 | 1.89 | 1.12 | 1.12 |
| 3 | 7.95 | 1.09 | 1.09 | 1.09 |
| 3 | 7.12 | 2.94 | 1.08 | 1.08 |
| 3 | 8.02 | 2.09 | 0.98 | 0.98 |
| 3 | 8.97 | 1.02 | 1.11 | 1.11 |
| 3 | 7.85 | 3.05 | 1.04 | 1.04 |
| 3 | 8.89 | 1.86 | 1.06 | 1.06 |
| 3 | 8.93 | 3.02 | 1.08 | 1.08 |
| 3 | 8.91 | 3.92 | 1.20 | 1.20 |
| 3 | 2.85 | 2.85 | 0.91 | 0.91 |
| 3 | 4.09 | 4.09 | 1.01 | 1.01 |
| 3 | 5.05 | 5.05 | 0.85 | 0.85 |
| 3 | 5.85 | 5.85 | 0.82 | 0.82 |
| 3 | 7.04 | 7.04 | 1.04 | 1.04 |

Note: Stimuli blocks are randomly assigned to the three attention conditions across participants. Namely, for stimuli sets of block 1, it can either be used as stimuli of BB, OB or VB for different participants. Same logic fits for the other two blocks.