Bachelor Degree Project

Social Punishment:
Evidence from experimental scenarios

Bachelor Degree Project in Cognitive Neuroscience
Basic level 22.5 ECTS
Spring term 2018

Johan Pieslinger

Supervisor: Oskar MacGregor
Examiner: Björn Persson

Abstract

Punishment is the act of penalizing an individual as a response to a transgression. This thesis will deal with punishment in experimental game scenarios and in experimental criminal punishment scenarios, along with their different adaptations. The aim will be to provide an overview of both psychological and neurological underpinnings of punishment by reviewing existing literature. While punishment ought to deter transgressions and promote cooperative behavior, internal neural reward-related systems seem to be a driving factor of the desire to punish wrongdoings. Decisions on whether a transgressor is guilty and deserves punishment is mediated by the medial prefrontal cortex with an emphasis on the ventromedial parts. External influences affect the behavioral output and its underlying neural signatures of punishment. Social context such as peer pressure and in-group bias emphasize the importance of theory of mind related areas when conducting punishment.

Keywords: punishment, transgressor, ultimatum game, public goods game, neuroscience

Table of contents

Social Punishment: Evidence from experimental scenarios

Social punishment is a common aspect of our lives. We might not think about it too much, but we certainly acknowledge that it exists. Most people have at some point stood in a line somewhere, be it in the bank or at a concert, and most people have also come in contact with someone who tries to skip the line in some way and achieve their goal early. Observing this awakes something in us, it might not be too dramatic but a desire to force that person to the back of the line and behave like everyone else. This is in fact a desire to punish someone who has violated the social norm. The norm is that you wait in line behind whoever is before you until it is your turn, when someone violates that norm they are committing a form of transgression. Forcing the line-cutter to the back of the line would be a kind of punishment to reinforce norm-compliance and cooperation. Although this is a rather banal example of how punishment exists in our everyday lives, it shows that punishing cheaters is a common desire in humans. What is probably less known is how punishment works, what motivates us to punish and why it exists. Say that this line-cutter claims that she did not know there was a line, that it was a mistake. I for one would not feel as irritated with the person, since this person did not try to cheat.

The notion that a non-intended transgression makes me less irritated can be seen all across society and one of the most extreme cases of this is in ending another human's life. There is a difference in both crime-definition and punishment depending on if it was intended or not. Manslaughter is less severe than murder and is defined by a lesser intent of the one who committed the crime (The President's Council on Bioethics Staff, 2010). Influences such as intent shows that punishment is a rather complex phenomenon. This thesis will bring forth how researchers are able to explore social punishment in an experimental setting, as well as describing how social punishment functions in both psychological and neural terms.

The aim will be to illustrate the different fundamental functions social punishment in economic games and experimental criminal punishment scenarios draw upon, as well as strengthen these claims with underlying neural activity. Punishment is a vast topic and can be talked about in different contexts. This thesis will overlook the topic of *operant conditioning*, which is controlling behaviour by consequences – such as a child touching a hot stove (behavior) which results in pain (consequence), the child will then be less likely to touch the stove again due to the painful consequence (controlled behaviour; Staddon, & Cerutti, 2003). A main difference between punishment in operant conditioning and social punishment is the emphasis of social interaction. In social punishment there is a high significance of aspects such as intention, fairness and morality, whereas operant conditioning can exist without these aspects, i.e. a hot stove does not intend to hurt (Skinner, 1938). When mentioning punishment in this thesis, punishment will exclusively refer to social punishment unless stated otherwise.

Firstly, two specific games used to exploit punishment in humans will be under the scope - the *ultimatum game* and the *public goods game*. These games, including different versions of them, allows researchers to control different aspects of punishment and figure out how it works in different settings. Both the ultimatum game and the public goods game are founded on economical exchange, and to some extent cooperation. Which means that these games can be employed in a variety of scientific fields, such as economical science, psychology and neuroscience, and punishment is not the only aspect of these games (Barbey, & Grafman, 2011; Jordan, McAuliffe, & Rand, 2016; Yudkin, Rothmund, Twardawski, Thalla, & Van Bavel, 2016). Theories regarding *material opportunism* and *strategy* have been explored using these games and are important aspects influencing the games' outcomes (Güth, & Kocher, 2014; Fehr, & Fischbacher, 2004a; Jordan et al., 2016). This thesis will have an emphasis on punishment, taking other aspects into account when relevant for punishment and the aims of this thesis.

Games are however not the only way to study punishment; *criminal punishment* in an experimental setting, which will be referred to as *criminal scenarios*, gives insight into more severe transgressions. This thesis will summarize contemporary scientific literature on the subject of punishment within these situations. Psychological trends and their underlying neural circuitry will be brought forth to describe how punishment works in a broader sense. There are plenty of psychological studies revolving the public goods game, unfortunately modern scientific imaging studies using the public goods game is rather lacking in comparison to the ultimatum game (Buckholtz, & Marois, 2012). Therefore, the chapter regarding psychological aspects will be including the public goods game, but the chapter about the neural mechanisms of punishment will have a greater focus on criminal scenarios and the ultimatum game.

The instances in which this thesis sets out to explore punishment is confined to an experimental setting, i.e. transgression is often, but not exclusively, a controlled factor by researchers. This means that punishment in this case does not represent all instances in which social punishment is present, for instance between-group retaliation is not covered by the limitations of the studies presented in this thesis.

The parts of this thesis that discuss the neurology of punishment will be based on evidence from both costly and non-costly punishment. By correlating different parts of the brain with different aspects of punishment, a basic overview of the mechanisms involved in punishment will be presented. To alter how punishment works psychologically should also alter its underlying neural signatures and is thus of interest when discussing the neurology behind punishment.

Lastly punishment can be talked about in a number of different aspects, one term that is frequently used is "*altruistic punishment*". This term most often refers to costly *third-party punishment*, where the punishing party is not affected by a transgression and the motivation

for punishment stems from altruistic reasons. However, the term is not used consistently across scientific literature and popular science, some refer to “altruistic punishment” as costly punishment while other refer to it as third-party punishment, regardless if it is costly or not (Buckholtz et al., 2008; Putz, Palotai, Csertő, & Bereczkei, 2016; Riedl, Jensen, Call, & Tomasello, 2015). To avoid confusion, the term “altruistic punishment” will not be used in this thesis, in favor of “third-party punishment” or similar. Instead, when altruism is mentioned, it will refer to a more traditional meaning, i.e. doing something for the good of another at a cost to oneself.

Where can we study punishment?

The study of moral transgressions, or more specifically punishment in response to moral transgressions, requires some kind of human interaction. This chapter will bring forth three primary instances in which researchers are able to explore how punishment works and why it exists - the ultimatum game, the public goods game and criminal scenarios. All of these instances also have some things in common, namely the roles of the players. The distribution of the roles might differ between these but three are always present - *victim*, *transgressor* and *punisher*. These roles might even be combined in different variations in the situations one sets out to test punishment; a combination of the roles of victim and punisher in a participant might test vengeful punishment, for example.

The ultimatum game

The ultimatum game is a game that helps to investigate both economical experiments and social decision-making in humans (Güth, Schmittberger, & Schwarze, 1982; Rilling, King-Casas, & Sanfey, 2008). The game works by involving two players, one proposer and one receiver. The proposer is given a sum of money, out of which she gives - or proposes - a portion to the receiver. The receiver can then either decline or accept the proposal. If the receiver accepts, both participants receive the amount dictated by the proposer, but if the

offer is declined neither of them receives anything. To achieve the most lucrative outcome, the proposer must find an as unfair offer as possible, weighing in their own favour, but fair enough that the receiver is less likely to reject it.

Intuitively, the proposer could offer any amount of money to the receiver if both believe that any money is better than no money (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). This, however, is not typically the case. Studies repeatedly show that if an offer is too unfair, a majority of receivers would rather see that no money is distributed than that the proposer receives too much money (Corradi-Dell'Acqua, Civai, Rumiati, & Fink, 2012; Koenigs, & Tranel, 2007; Oosterbeek, Sloof, Van De Kuilen, 2004; Sanfey et al., 2003). This is where we stumble upon the concept of *costly punishment*, and why games such as the ultimatum game are important for the science behind punishment in humans. When the receiver declines the proposal, she is effectively paying a cost to punish the proposer for conducting what the receiver believes to be unfair behavior.

It might seem highly arbitrary what constitutes an unfair offer. Some receivers might consider anything below 50% to be unfair and that it should be punished, while other receivers might not. Thankfully, follow-up studies and large amounts of data gives a more holistic picture of where the line can be drawn (Sanfey et al., 2003). While the most normal offers are around 50% of the whole sum, what constitutes a low offer is around 20% of the money. If an offer would be considered low, there's usually a 50% chance that the receiver declines the offer (Sanfey et al., 2003). This is because although it might be more self-interestedly rational to choose some money over no money, when presented with an unfair offer, we tend to react emotionally negatively to it, i.e. it evokes emotions such as anger or distrust. If the negative emotional reaction to decline - or punish - outweighs the cognitive motivation to accept the offer, said offer will be rejected (Sanfey et al., 2003).

As offers become more unfair, both emotional responses and rejection rates increase, but this also depends on the nature of the proposer. A typical version of the ultimatum game involves two players - one where there are human participants as the role of both proposer and receiver, and one with a computer in the role of proposer with a human as receiver (typically with the human participant made to believe that there is another human participant, rather than a computer, proposing offers via computer, instead of face-to-face). Interestingly, rejection rates differ between these two instances of the ultimatum game. When there are two humans participating, rejection rates are significantly higher than when there is a computer screen and a human. This could indicate that there is a stronger emotional response when facing a human proposer (Sanfey et al., 2003). A simple explanation of this is that declining the offer is a form of punishing unfair behavior. This, meaning that when faced with a human proposer, one can infer that the proposer had the intention to both willingly and knowingly propose an unfair offer, and that the proposer has morally wronged in some way. In contrast, when faced with the computer opponent, the intention of the proposer becomes unclear. Questions such as: *is this programmed?* or *am I facing a human?* could influence the receiver's response. It becomes more problematic to punish when the moral transgression is less clear, and especially given that computers do not have intentions.

A way of influencing rejection rates in the ultimatum game is manipulating the total size of the sum of money. Should the sum increase, rejection rates would then decrease in response (Oosterbeek et al., 2004). This challenges what constitutes low offers and allows for discussion of low offers on two levels. First there is the relative low offer, mentioned earlier a low offer is somewhere around 20% of the total sum (Sanfey et al., 2003). Secondly, a low relative offer could still be high in monetary gain for the receiver, consider receiving 20% of a 100,000 dollar total. The effects of endowment size can be seen in both second- and third-party versions of the ultimatum game, and thus provides a problem for drawing conclusions

about rejection rates (Jordan et al., 2016; Oosterbeek et al., 2004) To circumvent this problem, looking at averages from multiple studies regarding the ultimatum game points us towards more robust definition of the term *low offer*. If endowment size is set aside, the trend for what constitutes a low offer is around 20% of the sum, and relative offers together with associated rejection rates will be more in focus during this thesis than endowment size (Sanfey et al., 2003).

To develop the field of punishment and third-party punishment, alterations of the ultimatum game have been used (Fehr, & Fischbacher, 2004b). In a version known as the *dictator game*, instead of having a pot that the proposer offers a share of to the receiver, both participants gain a certain amount each (Kahneman, Knetsch, & Thaler, 1986). The proposer in this game is called the dictator, and can then take any amount from the receiver. This is where an independent third party has also been introduced to the game in addition to the original receiver and proposer. The receiver can not decline, as the dictator steals from them, but instead the third party can, for a cost, punish the dictator to lower the amount they receive. Why this version is similar to the ultimatum game is because there is still a pot that is shared, but not in such a clear way. Say the dictator has 50 dollars and the receiver has 50 dollars. The dictator then declares how much of the receivers 50 dollars she will take. If this were the ultimatum game, the proposer would have a 100 dollar pot to share with the receiver. The dictator game is much clearer in the moral transgression, as the wrongfulness in the ultimatum game lies in an unwillingness to share with the receiver, while the wrongdoing in the dictator game constitutes stealing from the receiver. In addition, this motivates the third-party to act upon and punish moral transgressions. The independent third party's role in the dictator game is then to decide whether to punish the dictator or not. This works by having given the third-party player a sum of money as well. The third party can then use his money to lower the monetary gain for the dictator, i.e. if the third-party spends 5 dollars, the

monetary gain for the dictator might be lowered by 15 dollars. This third party is not required to punish the dictator, and would arguably be better off not punishing the dictator, as this would result in not losing any money for the good of someone else (Fehr, & Fischbacher, 2004b).

Studies using the dictator game have explored the way we punish, and a clever use of withholding information about normal punishment rates to the third party as they make the decision have revealed an interesting phenomenon (Fabbri, & Carbonara, 2017). In this case, information about third-party punishment rates among the participants peers was withheld until after the punisher had made the decision. After the decision had been made however, experimenters revealed the average punishment rates to the participant and allowed for revision of the chosen punishment. It seems that even though thoughts and beliefs about what is fair and not is highly subjective, allowing punishers to know how their peers have punished reveals a preference for social conformity. Say that the imaginary people George and John are involved in a dictator game. George wants to punish John, but after George has settled on a punishment it is revealed that all people in George's society punishes much more severely. George will then be, on average, inclined to change his punishment to what fits the social norm. In relation to the dictator game, peer pressure is a powerful factor in how we punish, as we punish individuals who break the social norm we do not ourselves want to break the social norm on how to punish (Fabbri, & Carbonara, 2017). Social conformity will be discussed later in the section - *Social influences on games and punishment*.

The public goods game

Violations of social norms evoke negative feelings toward the transgressor, which in turn motivates people to punish them (Matsumoto, & Hwang, 2015). Exploring social norms and what warrants punishment can be done through the public goods game (Andreoni, 1988). This game involves a group of people, all of whom are given a sum of money. Each

individual then offers a set of this money to a collective pot, which is multiplied by a certain amount. The multiplied collective pot is then distributed equally amongst all participants, and the game continues for multiple rounds. The amount the pot is multiplied by depends on the number of participants, it will always be constructed in a way so that if only one participant contributes, the same individual will end up with less than she started with. This is to make sure that reasons for contributing to the pot stem from altruistic and trusting reasons, rather than egoistic ones (Van Hoorn, Van Dijk, Güroğlu, & Crone, 2016). The success for each participant in this game relies on their willingness to cooperate with each of one's peers. To generate the highest yielding pot, all participants are required to offer all of their money. What makes this game interesting in relation to punishment is that participants are not required to offer any amount of money. Instead, if an individual wants to maximize her own profit, she should not put any money in the pot while at the same time all of her peers offer all of their money.

The public goods game allows for a social norm to be created in an experimental setting. If a group on average offers 10 dollars each to the pot, this becomes the norm. Anything above could be considered as an act of goodness, altruism or even foolishly trustworthy. If one would propose below the norm set by the group, this could be referred to as free-riding, i.e. taking advantage of others cooperation without oneself cooperating. If free-riding would be allowed to run rampant, without any risk of repercussions, cooperation would be deterred as a result. But, if participants in the public goods game are allowed to know what their peers offered and punish *free-riders*, the opposite effect occurs (Fehr, & Fischbacher, 2004a; Fehr, & Gächter, 2002). The punishment used in the public goods game is often a costly one and impacts the monetary gain for the one targeted by the punishment, meaning that when it is revealed who offered what to the pot, all participants are allowed to pay a cost to lower the amount gained by an individual for that round. The group is motivated

to punish people who break the group's norm in order to deter people from taking advantage of cooperators (Fehr, & Gächter, 2002).

Not only is the public goods game interesting because it creates a group dynamic and a social norm, but it also allows for the creation of two norms. The first norm is how much each should offer to the collective pot, the other one is the incentive to punish the ones that violates the first social norm. These norms dictate how much money one should contribute in each situation and to what extent punishment of free-riders should occur. Violating either one or both of these norms yields the highest result for the individual. To put this in perspective of the individual, both norms demands a personal cost but leads to the optimal outcome for the group. This since if nobody is motivated to punish free-riders, free-riding is not deterred. If free-riding is not deterred the motivation for risking money in the collective pot becomes diminished. In this situation, the motivation to punish free-riders becomes a meta-public good, which then allows individuals to free-ride on (Perc, Gómez-Gardeñes, Szolnoki, Floría, & Moreno, 2013). Further studies delving deeper into the subject of the different kinds of public goods state that cooperation alone is not viable, due to the threat of free-riders. However, different factors and public goods can be introduced to incentivise cooperation and create a more prosocial environment. These studies reveal that cooperation can be reinforced not only by punishment but also by rewards (Perc et al., 2013; Szolnoki & Perc, 2010). This begins to show how public goods game and strategy can be a complicated matter, depending on the factors introduced.

Data from public goods games point us in the same direction of social conformity as in the dictator game. If the group's average offer is increased in the public goods game, the individual will conform to the social norm (Fehr, & Fischbacher, 2004a; Fehr, & Fischbacher, 2004b).

A version of the public goods game only involving two players is known as *the prisoner's dilemma* (also known as *the trust game*). The principles behind this game lie in that two players are prisoners who are sentenced to 10 years of imprisonment cumulatively. If both players do not rat the other one out (cooperate), they both receive 2 years imprisonment each. If one does rat the other one out (defect), while the other one doesn't (cooperate), the defector doesn't receive any time in prison and the cooperator receives the full 10 years in imprisonment. If both would defect, they both receive 10 years imprisonment. To result in the most optimal outcome for both participants it is required that both cooperate, but this does not yield the most optimal outcome for the individual. The principles behind this game can be easily translated to involve money or other beneficial circumstances, i.e. if both cooperate they receive 150 dollars to share, but if one defects the defector receives 100 dollars and the cooperator receives nothing, and if both defect they receive nothing at all. There's both an incentive to cooperate and to defect (Fehr, & Fischbacher, 2004b).

When playing a monetary version of the prisoner's dilemma, one study introduced a non-affiliated third party (Fehr, & Fischbacher, 2004b). The third party's role in this version was as a punisher, although the monetary gain for the punisher was not affected by the actions of the two playing the prisoner's dilemma. The punisher could use money to punish the players in the prisoner's dilemma in the same way as in the public goods game, i.e. paying a cost to lower the monetary gain for one of the players. The results from this study indicated three trends (Fehr, & Fischbacher, 2004b). First, punishment was almost always directed towards players who defected, punishment towards cooperators was almost non-existent. Second, punishment was harsher towards defectors when their partner had cooperated, indicating that when there is mutual defection the social norm violation is considered less severe. Third, around 50% of third-party players chose to punish defectors. Here it is important to remember that the third parties monetary gains were not affected by the actions

of the others, rather the third party was affected negatively by paying a cost to punish another player. Therefore, motivation to punish in this case lies in the moral transgression directed towards someone else, rather than when being affected by it such as in the public goods game (Fehr, & Fischbacher, 2004b).

Furthermore, regarding both the public goods game and the ultimatum game, playing the game for multiple rounds have different effects of the results. Such as in the public goods game, iteration allows for contribution based on previous results and other participants willingness to cooperate (Fehr, & Fischbacher, 2004a). This allows for a learning aspect within these games, a kind of strategic development. An interesting phenomenon occurs in the ultimatum game regarding experience and iteration. While the share of the sum in the ultimatum game is higher when the participants are inexperienced, playing multiple rounds of the ultimatum game increases rejection rates of low offers. Playing the ultimatum game long enough leads to a state where equal splits becomes the norm, regardless of reputation among participants (Güth, & Kocher, 2014). While experience increase rejection rates for receivers, the size of the offer from proposers decrease (Güth, & Kocher, 2014; Oosterbeek et al., 2004). This is where a dissociation between experience and iteration is in place. While experience can be gained from iteration, experience can be carried over from multiple game-sessions. In iterated ultimatum games, the norm that develops is dynamic and dependent on that specific situation and participants (Güth, & Kocher, 2014). In line with iteration in the ultimatum game, iteration in economical exchange games overall seem to increase cooperation (Güth, & Kocher, 2014; Oosterbeek et al., 2004; Rand & Nowak, 2013). The increase of cooperation over time can be attributed to the effects such as conditional cooperation, i.e. cooperation if the other one cooperates, and norm-learning, i.e. how to play the game according to others (Güth, & Kocher, 2014; Rand & Nowak, 2013). One of the reasons why iteration of games is different from single rounds is the employment of self-

benefit strategies. Aspects such as reputation and strategy might have the ability to be more expressed than social preference in multiple rounds of the public goods game and might be more beneficial than transgression due to lasting effects of one's choices (Rand & Nowak, 2013).

Experimental criminal scenarios

The public goods game and the prisoner's dilemma both rely on the cooperation of others. If one cooperates, it requires others to cooperate as well. The public goods game shows us a perfect example of trust and how a norm can be formed, as well as the importance of deterring free-riding. However, only looking at games to understand punishment would not be able to paint the whole picture. These games involve direct punishment where the punisher usually, but not necessarily, is in proximity to the transgression itself. Criminal punishment scenarios work as a complement to explore punishment and emotional responses towards more severe moral transgressions, without victimising the punisher by having her be affected by the transgression, such as in the public goods game. These criminal scenarios can work by exposing participants to videos of an individual committing a crime. The participant's task is then to determine whether or not the individual ought to be punished and how severe the punishment ought to be (Buckholtz et al., 2008).

If trends in punishment behavior of participants across games and criminal punishment scenarios would follow the same pattern, conclusions could be drawn on a more fundamental level. To exemplify, if punishment severity in the public goods game relies on how severe the transgression is - i.e. how much lower the contribution to the collective pot is than the group average - as well as in the criminal punishment scenarios - i.e. theft vs. murder -, one can guess that punishment severity does in fact depend on the severity of the moral transgression in other situations as well. This notion on the severity of punishment will be brought up later in both a more psychological context and in a neural context.

Games such as the ultimatum game and the public goods game provide opportunities to explore social norm formation and the violation of said norms. Tweaks to these games allow for different circumstances and alter the role and thought processes of the participants. Yet, understanding what we are able to measure and control in these games are important. In the ultimatum game, what would happen if the proposer were faced with a receiver that has abused her position as a proposer in previous games? How willing are we to punish in the dictator game depending on the size of monetary gain for the dictator (Jordan et al., 2016)? Questions like these illustrate how different variables can be used and manipulated to further explore how punishment in humans works.

Psychology in games and punishment

How much, for what, and whom should we punish, are highly individual questions based on one's moral beliefs, but at the same time questions such as these lie at the heart of how we choose to engage in punishing behavior toward others. While we choose to punish transgressors to promote what we believe to be cooperative behavior, the road from transgression to punishment is dependent on a number of factors - some of which are how severe the transgression was, if there are multiple punishers, if it was an accident or an act of pure malice, etc. This chapter will shed light on some of the most important psychological factors in punishment, as well as influences determining how punishment is conducted.

Why punish?

Cooperation thrives in human societies, even though cooperation is not always the most lucrative solution for the individual. If a person would choose to not spend resources on helping others, but still draw from the benefits of others' cooperation, this would yield in the highest individual profit for this person. This sort of person is often referred to as a "free-rider". Luckily, humans seem to have evolved behavior and emotions targeted towards free-riders, as uncooperative or morally condemnable behavior tends to evoke feelings of disgust,

anger or contempt (Matsumoto, & Hwang, 2015). To be able to explore the psychology behind cooperation and punishment, the public goods game has been extensively used (Fehr, & Fischbacher, 2004a; Fehr, & Gächter, 2002).

As described above, not contributing anything to the collective pot while all the others provide their maximum amount yields the most favorable outcome for the individual. This individual would then be free-riding on the group's effort to achieve the highest yielding outcome for everybody. If punishment of this free-rider is possible, non-cooperative behavior is generally deterred. If punishment of free-riders is costly for the punisher, it gets a bit more complex, as costly punishment then becomes a public good in and of itself (Fehr, & Gächter, 2002). If only one person in a group of eight would punish free-riders at a personal cost, and one of the eight would free-ride, the six remaining individuals would be virtually free-riding on the punishing individual’s effort to deter free-riding. This since the punisher would effectively be attempting to achieve the highest yielding outcome by interacting with the now two different public goods - free-riding deterrence *and* providing money to the collective pot.

This goes to show that cooperation involving resources tend to promote cooperation-promoting behavior - such as punishment for free-riders - as a resource in and of itself.

One important aspect of the public goods game is that everyone is affected by the free-rider, meaning that whoever chooses to punish is also the victim of the free-rider. This gives egotistical motivation to punish free-riders (Fehr, & Gächter, 2002).

When putting cooperation and punishment in the public goods game to the test, one study was conducted having participants playing the public goods game across six rounds in groups of four, one time with and one time without punishment. It was found that 84% of the subjects conducted costly punishment at least once, with punishment being imposed on those who contributed below the group average by the ones who contributed above average. The severity of punishment also increased with higher investment in public goods. Across the two

different conditions, both with and without punishment, average investment was substantially higher (92%) when participants could punish non-cooperative behavior. Furthermore, participants in the public goods game were asked to rate feelings towards a hypothetical free-rider across two scenarios. One scenario where the rest of the group sacrificed a high amount towards the shared pot, and one scenario where the group sacrificed a lower amount. Both scenarios also contained a free-rider. Participants rated higher amounts of negative emotions, such as anger etc., directed towards the free-rider, when the investment from the group was higher (Fehr, & Gächter, 2002).

This can be seen as even though the investment from the free-rider was the same in both scenarios, the transgression relative to the cooperation of others was not, therefore it can be interpreted that the severity of a transgression is dependent on context and not always the act in and of itself. To make it clearer, investing five dollars to the public pot while the other three group members contribute 16 on average is worse than contributing five dollars when the rest of the group invests eight on average (Fehr, & Gächter, 2002).

The data from the public goods game gives good insight as to why we punish. To promote cooperation and deter free-riding in a certain situation can be translated to fit into a more societal setting - to punish negative deviations from the norm promotes norm-compliance, which in turn results in higher cooperation due to lower risk of malicious intent in cooperation-dependent situations (Fehr, & Gächter, 2002). Analyses of iterated public goods games also indicate an importance of transgression deterrence, if cooperation is not reinforced in some way cooperation will almost always fail as a response (Perc et al., 2013; Rand & Nowak, 2013).

Severity of punishment

Evidence suggests that severity of punishment and emotional arousal are highly correlated. As a transgression grows more severe both emotional and punishment-oriented

responses increase in intensity, although severe punishment might not always be dependent on high emotional arousal, as will be discussed in the following section (Buckholtz et al., 2008).

A recurring phenomenon is that the more severe a norm-transgression is deemed to be, the more severe the punishment. This can be seen in the public goods game, criminal punishment scenarios and in the dictator game (Fehr, & Gächter, 2002; Buckholtz et al., 2008; Fehr, & Fischbacher, 2004b). However, severity of those rely on different factors - e.g. is the transgressor responsible, was it intended, how did it affect others? In one criminal punishment study, some of these factors were put to the test (Buckholtz et al., 2008). Participants were tasked with watching the fictional character John in three different criminal scenarios. These scenarios consisted of responsible, diminished-responsibility and non-crime scenarios. The criminal context of the scenarios ranged in severity, ranging from e.g. theft to murder. The participants were tasked with punishing John using different severities of punishment on a 0-9 scale, with 0 indicating no punishment and 9 being the most severe, equating approximately to life imprisonment or a death sentence. As could be expected, punishment severity was positively correlated to the severity of the crime in the responsible scenarios, as well as non-crime scenarios resulting in a 0 punishment rating.

However, the interesting part from this study comes from the diminished-responsibility scenarios. Here John had still committed an act where the outcome was the same as the criminal one, but under mitigating circumstances, e.g. being responsible for an accident. The trend was that under these mitigating circumstances, the intent of John was not the behavioral outcome presented, resulting in a lowered severity of punishment for John’s transgressions. The behavioral notion of punishing intent rather than the result of a specific action is made is further supported by neural data which will be further discussed in the neural chapter of this thesis (Buckholtz et al., 2008).

As previously mentioned, evidence from the public goods game indicates that severity of punishment is in relation to a rather arbitrary norm. The norm itself depends on the cooperators' average investment - deviation from this norm is what could be considered as a transgression or altruism depending on if it is a negative or positive deviation. However, a greater negative deviation leads to greater negative emotional responses from cooperators, which in turns entails an increased risk of punishment for these free-riders, in line with both the emotions evoked as well as the transgression in itself (Fehr, & Gächter, 2002).

To conclude, the greater the transgression, the greater the emotional arousal accompanied by a more severe punishment. It seems that high emotional arousal and severe punishment go hand in hand and might be a driving factor for warranting severe punishment. When diminishing the responsibility, however, it begins to show that it is not the transgression itself that is at focus, but rather that determining a fitting punishment is based in large part on the transgressor's motivation and intent. Emotional arousal, in other words, depends on the contextual situation and the observer herself, and even if theory of mind in certain situations can give rise to a higher emotional arousal, which then motivates more severe punishment (Buckholtz et al., 2008; Fehr, & Gächter, 2002).

Taking some of the context from the criminal-punishment scenarios, when there is a smaller gap between the free-riders and cooperators - both emotional and punishment responses are diminished. It would be interesting to see if the diminished responses in the public goods situations stem from questionable intent and responsibility, similarly to the diminished-responsibility scenarios in criminal punishment. This, since the norm in public goods situations is arbitrary and dependent on the group average - where is the line drawn for common mistakes where it no longer warrants disciplinary actions from cooperators?

It is important to keep in mind that punishment itself needs to be fair. A too gentle punishment would not elicit the disciplinary improvements in a transgressor, whilst a too

harsh punishment becomes a transgression itself. This, since the social norm is that a certain punishment fits a certain crime, if the punishment is too severe, the act of punishing becomes more condemnable than the transgression itself - e.g. punishing theft with a death sentence. A too severe punishment does therefore not fulfill its purpose, such as deterring further uncooperative behavior (Dilts, 2012).

Responsibility

To be able to determine if a transgressor is responsible, or at least how responsible, for her own actions, the observer must be able to infer intent in the individual. A diminished-responsibility scenario should lead the observer to determine that the criminal act was not as intended. To give an example of this, the crime of causing the death of another human being is normally divided into three categories in criminal law: murder, manslaughter and negligent homicide All these crimes involve the same sort of act - causing the death of another human being - but are defined by the intent of the transgressor. Murder, being the most severe and morally despicable in the eyes of the law, is also defined by a high level of intent from the individual (The President's Council on Bioethics Staff, 2010). By looking at studies involving an observer determining punishment for a transgressor committing a criminal offense, mitigating circumstances for the transgression do seem to lead to less severe punishment (Buckholtz et al., 2008). As responsibility becomes a variable of guilt, diminishing the responsibility of the transgression leads to a less guilty transgressor, which in turn lowers the severity of the observer's punishment.

The ultimatum game provides additional, albeit indirect, insights on determining intent. The version of the ultimatum game where one human receiver is pitted against a computer proposer resulted in less severe results. Tendency to punish the transgression was significantly lower when facing a computer proposer rather than when faced with a human one (Sanfey et al., 2003). Considering that declining the offer in the ultimatum game is

essentially costly punishment, willingness to conduct costly punishment is diminished. To put this in perspective - a human might propose an unfair offer due to selfish reasons such as wanting more money to buy more things, whereas a computer can not knowingly or willingly act upon its own selfish goals. One can interpret that the results of not being as prone to punish the computer is due to lowered responsibility and therefore does not evoke the same emotional response typically accompanying punishment (Sanfey et al., 2003; Buckholtz et al., 2008).

In relation with scientific findings and the society, the law does seem in line with the human psyche. Understanding another individual's intent and responsibility is a fundamental part of being able to conduct adequate punishment.

A responsible punisher is not the only thing required to conduct punishment, but the punisher must feel a responsibility to punish (Feng et al., 2016). A modulated version of the dictator game examined what happens when the punisher is the sole punisher or part of a group. The experiment was constructed to test the severity of punishment of participants in an alone condition versus in a group condition. The participants were told that the decisions of the dictator would be revealed to either only the participant or the participant together with four others. The results indicated that when participants were the sole punisher, they punished more severely than in the group condition - a phenomenon referred to as *diffusion of responsibility* (Feng et al., 2016). In the lone condition, the participant was solely responsible for the punishment of the transgressor, whereas in the group condition, the responsibility to punish was diffused among several punishers. This trend - to not as willingly conduct altruistic behavior - can be seen across other circumstances such as in the *bystander effect* (Feng et al., 2016). The bystander effect is when individuals are less likely to help a victim of a crime or accident due to the presence of others, i.e. a number of bystanders apart from oneself (Darley, & Latane, 1968).

On an interesting side-note, empathic response to pain in others is greater when the pain is inflicted by a transgressor rather than self-inflicted by accident. Although empathy is outside the scope of this text, the notion that emotional response is higher when there is a transgressor responsible for the pain of others gives a slight insight to what motivates us to punish (Akitsuki, & Decety, 2009).

Social influences on games and punishment

Instances of social conformity can be found in both cooperation situations as well as in punishment (Fabbri, & Carbonara, 2017; Van Hoorn et al., 2016). Not only working as a motivation and precursor for prosocial behavior, desire for social conformity also shapes how punishment and cooperation looks in different situations (Fabbri, & Carbonara, 2017). One group of individuals especially sensitive to social influences, e.g. peer pressure, are adolescents (Van Hoorn et al., 2016). Although adolescents are infamous for being peer pressured into taking excessive risks, one particular study examined peer pressure effects on prosocial behavior in participants aged 12-16 years (Van Hoorn et al., 2016). The task for the participants was to play the public goods game across three different conditions - one with present observing peers who evaluated the amount contributed to the collective pot, one where present peers were only observing but did not evaluate the participant's decisions, and one where there were no peers present at all. The entire test took place online, presented through a computer screen, the spectators were represented by pictures of similarly aged adolescents, but were in fact youth-actors. Both the participants and the actors met before the online public goods game took place. When in the evaluative condition, participants were presented with positive feedback from five spectators depending on the amount the participant contributed. No feedback was given during the observation-only-condition but participants were still instructed that the spectators would evaluate their decisions. It was found that being observed by similarly aged peers provided an increase in the amount

contributed to the collective pot in the public goods game, as compared to not being observed. The highest increase of contributed amount was found when the observers rated the participant's decisions (Van Hoorn et al., 2016).

This provides an example of how social conformity acts as a motivation to act altruistically in a cooperation-dependent situation. The study just described does not, however, deal with the punishing aspects of social conformity. Third-party punishment has been found to be affected by intergroup bias, i.e. individuals punish out-group members more severely or willingly than in-group members (Yudkin et al., 2016). This begins to show that social status among individuals does form the way we punish. To test punishment in out- and in-group settings, a version of the dictator game was devised (Yudkin et al., 2016). To represent out- or in-group the participants were categorized according to which their favorite sports teams were, as well as by their nationalities. The participants were all the third party in the dictator game, and were watching a transgressor steal money from a victim. The transgressor was either a member of the same group, such as the same nationality or cheering for the same football team, or from another group, such as cheering for another team or being from a different country. The participants were given a sum of money from which they could choose to spend to punish the transgressor in these scenarios. The results showed that swift punishment was significantly harsher than slow punishment of out-group members. This meaning that a low decision time between transgression and punishment increased the severity of punishment of out-group members - a phenomenon referred to as *reflexive intergroup bias* (Yudkin et al., 2016).

Neural correlates of punishment

The neural mechanisms of how we conduct punishment as humans to enforce social norms and compliance within our society depends on the situation, as costly and non-costly punishment are similar in some aspects, while different in others. Taking both of these

instances into account, this chapter will lay down an overview of which processes are responsible for what, and in which scenarios these play a more important role. To begin to understand the biology of how we choose to punish transgressors, some central neural mechanisms responsible for our social lives need to be put forward.

The social brain, cooperation and norm-violations

One important feature in humans as a species is our ability for cooperation and social interaction - an ability often prescribed to the foremost parts of our brains, the prefrontal cortex (PFC). This brain area is bigger in humans than other animals, relative to our brain sizes (Rilling et al., 2008). The PFC is involved in an array of different functions, such as goal-directed behavior, working memory, cognitive control and emotion regulation, all of which can be imagined to influence punishment in some way (Miller, & Cohen, 2001; Barbey, & Grafman, 2011; Koenigs, & Tranel, 2007). The PFC is divided into different main areas, ventromedial, lateral and mid-dorsal, all of which are heavily connected to each other. These areas are also connected to other systems in the brain, including sensory- and emotional systems, leading the PFC to be able to process a wide variety of different information in a focused space (Miller, & Cohen, 2001). In relation to punishment, these mentioned properties of the PFC do all seem to be involved; it is easy to imagine a phenomenon like cognitive control and working memory (typical PFC functions) to be critical not only for conducting fair punishment, but also for punishing fairly in the eyes of one's peers.

Evidence from studies regarding damage to the PFC suggests that the different parts of the PFC are more involved in some functions of punishment than others. For instance, evaluation of the intent of a transgressor and damage done to the victim seems to rely more on the medial parts of the PFC, whereas the dorsolateral parts are more involved in choosing a punishment fitting the crime (Glass, Moody, Grafman, & Krueger, 2015).

In relation to viewing some wrongdoing occur, it seems unlikely that third-party punishment would not involve the emotional processes in our brain. A wrongdoing is not something that people usually view with glee, rather it tends to evoke feelings of anger, disgust or contempt (Matsumoto, & Hwang, 2015). Many of the emotions are often correlated with specific regions in the brain, but looking at some of these areas might show their role in punishment.

Activity in the insula (the insula is usually tied to feelings of disgust; Wicker et al., 2003) seem to help with the identification of norm-violations. When an individual is proposed an unfair offer in the ultimatum game, activity in the insula predicts a preference for rejecting the offer, while inhibition of the insula increases the likelihood of an individual accepting the unfair offer (Seymour, Singer, & Dolan, 2007). Furthermore, evidence from observing participants in a resource-allocation task shows that insular activity corresponds with choosing a fair and equal allocation, as opposed to a more unfair one (Du, & Chang, 2015). This points towards a role for the insula in identifying inequity.

The center for fear, the amygdala, does seem to have its place in punishment as well. Activity in this region corresponds with both emotional arousal, in regard to observing a crime take place, and when choosing the severity of punishment for an individual who has committed a crime (Buckholtz et al., 2008).

Evidence from non-costly punishment

The study by Buckholtz et al. (2008), on the importance of transgressor responsibility for punishment using the criminal scenarios, gives us an insight into where in the brain two fundamental functions of third-party punishment - the ability to determine responsibility and to estimate an appropriate punishment - operate and have in common with other mechanisms (Buckholtz et al., 2008). The ability to determine if a transgressor is responsible for her own actions has been positively correlated with activity in the dorsolateral parts of the PFC in the

right hemisphere (rDLPFC). This means that when an observer deemed a transgressor more responsible for wrongdoing, the rDLPFC showed a higher activity than when the observer experienced the transgressor's responsibility to be lower, regardless of the severity of the crime. Interestingly, no significant difference in rDLPFC activity was found between diminished-responsibility and no-crime scenarios (Buckholtz et al., 2008).

Furthermore, determination of responsibility and severity of the crime does not seem to be processed without determining intent of the transgressor, an ability included in the concept of theory of mind (ToM; Buckholtz et al., 2008). One region especially crucial to ToM is the temporo-parietal junction (TPJ), and in these criminal scenarios just discussed, evidence for involvement of the TPJ in third-party punishment was found (Gallagher, & Frith, 2003; Buckholtz et al., 2008). More specifically, the TPJ showed activation in all of the conditions but had a significantly greater activation in the diminished-responsibility conditions than both the responsible and no-crime conditions. Interestingly, the TPJ had its greatest activation earlier, during low activity in the rDLPFC, which was followed by low TPJ activity and high rDLPFC activity (Buckholtz et al., 2008). To give a more concrete perspective of the phenomenon: the observer first infers the intent of the transgressor which is represented by high activity in the TPJ, to then determine if the transgressor is responsible and should be punished for her actions, which is represented by high rDLPFC activity (Buckholtz et al., 2008).

Continuing on the subject of third-party punishment in a non-cost setting, the study conducted by Buckholtz et al. (2008) gives us further insights. As mentioned previously, rDLPFC activity do not seem to correspond to the severity of a chosen punishment. Instead, activity in affective regions, with an emphasis on the amygdala, seem to predict the severity of the punishment as chosen by an observer, with higher peak activity predicting a more severe punishment.

In this example, emotional arousal also correlated with punishment magnitude, which does seem intuitively logical, with a more severe crime inducing more negative emotions resulting in a more severe punishment. Although affective processes can be a predictor for severity of punishment, imaging and lesion studies have found that both affective- and social decision-making processes are involved in choosing the severity of punishment in a transgression, including amygdala, medial prefrontal cortex (MPFC) and posterior cingulate cortex. Since these regions are active during the assessment of a punishment, and the inhibition of these regions alter punishment, mechanisms such as cognitive control, working memory and emotional control are believed to influence and alter the punishment magnitude together with affective regions (Buckholtz et al., 2008; Glass et al., 2015; Koenigs, & Tranel, 2007).

Converging with evidence from brain injury, it does seem that DLPFC activity is highly involved in the act of punishing a third party, with the rDLPFC being more involved in the act of choosing to punish, as it does not correlate with severity of punishment. This allows for other regions such as affective regions, to be more involved in the assessment of punishment severity (Buckholtz et al., 2008; Glass et al., 2015).

Evidence from costly punishment

The previous section described certain neural aspects of punishment underlying how it is actually conducted, but with one important aspect missing - personal cost for punishment. Studies regarding the ultimatum game provide a perfect opportunity to explore the underlying mechanisms of punishment behavior, including the decision whether to punish or not, and the need to include cost-benefit considerations in addition to responsibility and severity of a transgression (Du, & Chang, 2015).

In most cases, reactions toward unfairness or wrongdoings are negative emotions, but when playing the ultimatum game as the receiver one needs to consider the costs of punishing

versus receiving a reward. Emotional control in this case becomes highly important. Keeping the negative emotions to the side in order to receive a reward, independent of fairness, does seem more beneficial in some cases than not receiving anything. The ventromedial prefrontal cortex (VMPFC) has been attributed this sort of emotional control, and an inhibition of this area should lead to unfair offers being more likely to be rejected (Koenigs, & Tranel, 2007). This has also proved to be the case: diminishing emotional control by VMPFC-inhibition does tend to lead to higher rejection rates from receivers, as negative emotions motivating punishment are not controlled (Koenigs, & Tranel, 2007).

As emotional control is inhibited, one can say that the individual is experiencing high emotional arousal in response to an unfair offer. While VMPFC is highly involved in controlling emotional arousal, emotional arousal is also correlated with activity in the anterior insula (Corradi-Dell'Acqua et al., 2012). In both the ultimatum game and the dictator game, both receiving and observing someone else receive an unfair offer evokes insular activity. This is in line with insula's role in identifying inequity, and insula activity being a predictor for offer-rejection (Seymour et al., 2007; Du, & Chang, 2015).

When playing the ultimatum game as the receiver, feelings of unfairness do seem to drive the decision to refuse an offer, but are these two phenomena the result of the same neural mechanisms, or can they be disassociated? If the right rDLPFC is inhibited by the use of transcranial magnetic stimulation (TMS) it leads to unfair offers being more likely to be accepted without affecting feelings of unfairness (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Seymour et al., 2007). This, together with the role of the rDLPFC in the criminal scenarios previously discussed, a conceptual overlap for both third-party and second-party punishment can be found. Since the decision to reject an offer is essentially a decision to punish unfair or wrongful behavior, one role of the rDLPFC seems to be deciding to punish regardless of the perspective of the punisher. One objection to this derivation might be the

difference of cost and roles between the two scenarios. In the criminal punishment scenarios there was a non-victimized punisher who paid no cost to punishing the wrongdoer, while in the ultimatum game the punisher was also the victim of the transgression as well paying a cost to punish the transgressor. Even though the nature of these two scenarios might affect the role of the rDLPFC, the results of different studies does seem to converge towards the role of the rDLPFC to be a driving area for the decision to punish.

Neural signatures of motivation of punishment

Although modern imaging studies on punishment in the public goods game are rather lacking, the alteration involving only two players – the prisoner's dilemma - can provide data describing what motivates punishment and how deterrence of norm-violations works (De Quervain, Fischbacher, Treyer, & Schellhammer, 2004). In this study, two players were given 10 dollars each, from which player A could trust player B by giving B all of their money. If A would trust B, the money would be quadrupled and then given to B, this resulting in B having 10 + 40 = 50 dollars, and A having nothing. B was then given the option of giving half of the money back to A. Should A not trust B, each player would receive 10 dollars. One notion to keep in mind in this game is that norm-formation should work in the same way as in the public goods game, i.e. by having A trust B, B should reciprocate or a norm-violation will occur (De Quervain et al., 2004). Player A was also able to punish B across four conditions by attributing punishment points, the impact of these punishment points was determined by the condition. First, B's actions are intentional and punishment for A is costly, with every dollar spent from A resulting in a two-dollar reduction from B. In the second condition, B's actions are intentional but without costly punishment for A, with each point of punishment resulting in a two-dollar deduction for B. The third condition still had B responsible for her own actions, but punishment was not costly for either A or B, but only had a symbolic value. In the fourth condition, B's actions were determined randomly by a computer, and

punishment of B was costly for A, with the same ratios (one to two) as in previous conditions (De Quervain et al., 2004). The first and second conditions had the ability to both induce a will to punish the transgressor, and an ability to do so for the participant. In the third and fourth conditions, punishment was either not impactful or not warranted. The experiment tested whether these implications across the conditions affected desire to punish and underlying reward-related neural mechanisms, namely the caudate nucleus. The caudate nucleus has been thought to be involved in decision making motivated by reward (De Quervain et al., 2004).

Behavioral data gathered expressed that across all intentional conditions, participants viewed B's decision to keep the money as highly unfair. The non-intentional condition did not induce the same sense of unfairness, possibly due to diminished responsibility (Buckholtz et al., 2008). Player B's decision to not reciprocate across all intentional conditions were viewed as unfair, but the neural activity in the caudate nucleus differed. Compared to average brain activity, satisfying the desire to punish resulted in an increased activation in the caudate nucleus, i.e. when player A had the ability to exert meaningful punishment upon player B. The opposite effect was found when player A could not punish player B meaningfully, in the symbolic condition, or punishment was not warranted, the non-intentional condition (De Quervain et al., 2004). A correlation between caudate activation and amount invested in costly punishment was also found, with higher activation correlating with a higher investment in punishment.

This data suggests that the desire to punish is satisfied if punishment is impactful on the transgressor and that punishment is motivated by internal reward systems, thus providing insights as to why humans engage in costly punishment. Further evidence points towards the role of internal reward systems, such as the ventral striatum and nucleus accumbens, being a motivation for the punishment of uncooperative or unfair individuals (Singer et al., 2006).

Neural activity behind influences on punishment

Punishing people belonging to one's own group versus punishing out-group individuals fundamentally changes the way we punish. Recalling the importance of ToM and its neural correlates to determine responsibility and intent of another individual, the use of this mechanism and its brain regions relies on whether or not we belong to the same group as the transgressor (Buckholtz et al., 2008; Morese et al., 2016). When punishing in-group members, one can see increased brain activity in regions linked to ToM, such as the VMPFC and TPJ in relation to punishing out-group members (Morese et al., 2016), indicating that one tries to understand the intent and reasons of the transgression more in in-group situations. This provides bias for severity in third-party punishment, as ToM can provide mitigating circumstances for the transgressor, leading to less severe punishments within our own group (Morese et al., 2016; Buckholtz et al., 2008).

Another version of altering intent, or at least having a different perspective on it, is the belief in free will (Krueger, Hoffman, Walter, & Grafman, 2014). An extreme version of belief in free will is that the reasons for why people do what they do stems ultimately from the will of the individual itself without any outside influences whatsoever. The other extreme is the opposite, which claims that free will does not exist and everything can always be traced back to preceding events and that our actions are dependent on these events. Taking these two extremes as two ends of a spectrum results in a scale for belief in free will. Functional magnetic resonance imaging studies have explored the effect of this belief on third-party punishment. In line with what has been discussed in relation to determining intent, an individual who believes in free will should be able to infer a higher intent in the transgressor. This, since free will implies that the transgressor committed the transgression more willingly than according to a more deterministic stance. Neural activity and behavioral data support this notion, to a certain degree (Krueger et al., 2014). In a situation using criminal

punishment scenarios, participants were tasked with punishing a transgressor committing various crimes. Emotional response and functional magnetic resonance imaging data was acquired in addition to the behavioral outputs, i.e. severity of punishment. The results showed that in scenarios which resulted in a lesser emotional response, a neural and behavioral difference was found. Believers in free will punished transgressors more severely than their deterministic peers. This was also accompanied by a greater increase in TPJ activity for the believers of free will. The differences between the two groups of participants diminished in scenarios that induced a high emotional response (Krueger et al., 2014).

How the presence of one's peers affect the public goods game has been discussed earlier, but there is a difference in brain activity in areas related to ToM and social processes (Van Hoorn et al., 2016). To recall - participants were placed in three different conditions: alone, peer observation and peer evaluation. There were significant behavioral changes between these conditions, with increased cooperation with peer observation and peer evaluation (Van Hoorn et al., 2016). The areas TPJ, dorsomedial PFC (DMPFC) and superior temporal sulcus did all differ across the three conditions. When participants were deciding what to contribute to the collective pot, peer presence increased activity in the TPJ and had a positive correlation with the amount the participant contributed. This since peer presence increased the amount participants contributed, as well as TPJ activity. While both the evaluation and observation conditions provided significant effects in TPJ, DMPFC and superior temporal sulcus the alone condition, the brain activity differences between evaluation and observation conditions were not significant, even though behavioral data would suggest increased activity across these regions with a significantly higher amount of contribution in the evaluation condition than the observation condition (Van Hoorn et al., 2016). Activity in the DMPFC is thought to be involved in actively thinking about the intentions and thoughts of others, and together with other ToM areas, the increased activity in

the observation and evaluation conditions could be an indicator that participants thought or worried about the mental states of their observers (Feng et al., 2016).

Furthermore, diffusion of responsibility provides insights into underlying neural networks in punishment. As stated earlier, activity in the insula helps us with the identification of norm violations and can work as a predictor for rejection rates in the ultimatum game (Seymour et al., 2007). In the modified version of the dictator game where third-party punisher participants were faced with a dictator's decision either alone or together with peers, differences in activity in the insula could be found (Feng et al., 2016). As the responsibility was diffused amongst several individuals, activity in the insula was lowered as compared to being tasked with punishing the dictator in the alone condition. The opposite was found for DMPFC activity, with higher activity in the group-condition than in the alone condition. The differences in activity in both insula and DMPFC are also noted in what seems to be the driving underlying neural region, i.e. which of the two areas that is expressed as the region which sends outputs to other regions in the network. In the alone condition it was found that the left anterior insula (left-AI) was the driving neural region. Taking the roles of insula and DMPFC into context, there is first an identification of norm violation, which then sends information to ToM-areas and other brain regions. In contrast, brain activity in the group-condition was driven by DMPFC activity. This resulted in DMPFC, a region linked to ToM, first being activated and sending outputs to identification of norm violation-areas and other brain regions (Feng et al., 2016).

The behavioral data that participants punished less severely in the group condition might be explained by the neural data. Participants in the group condition try to imagine the mental states of their peers, i.e. participants might think "these individuals ought to punish the dictator too, so I don't need to do it as much". Whereas in the alone condition, ToM of

one's peers is obsolete and there is only norm-violation to mediate further brain activity (Feng et al., 2016).

From this section, one can conclude that punishment is prone to external and internal influences. With both beliefs and social context manipulating the way we punish, as well as affecting the underlying neural correlates, objective punishment does seem a distant phenomenon (Morese et al., 2016; Krueger et al., 2014).

Discussion

The aim has been to illustrate how punishment works in different settings, as well as to provide both psychological and neural data behind the phenomenon that is punishment. By using different games, such as the ultimatum game or the public goods game, or criminal scenarios which incorporates punishment, researchers have been able to explore this phenomenon. To enforce social norm compliance and deter free-riding are some of the reasons as to why punishment exists on a functional level (Fehr, & Fischbacher, 2004a; Fehr, & Gächter, 2002). On the individual level on the other hand, observing a transgression take place evokes negative feelings and helps motivate action directed towards the transgressor (Sanfey et al., 2003; Fehr, & Gächter, 2002). A trend for diminishing the severity of the punishment can be seen in both experimental settings and in criminal law, should the transgression be non-intended (Buckholtz et al., 2008; The President's Council on Bioethics Staff, 2010). Punishment is also shown to be prone to social influences, as information about mean punishment severity amongst one's peers and if the target of this punishment is an out-group member both influence the assessment of how harsh a punishment should be (Fabbri, & Carbonara, 2017; Yudkin et al., 2016).

To know what goes on in the brain during punishment, neuroimaging studies provide an opportunity to understand the phenomenon. Moral transgression evokes insula activity in those observing or affected by the transgression (Seymour et al., 2007; Du, & Chang, 2015).

This activity is also correlated with emotional arousal, which gives some indication as to why moral transgression triggers negative feelings in humans (Corradi-Dell'Acqua et al., 2012; Matsumoto, & Hwang, 2015). Although negative emotions alone do not trigger the will to punish, internal reward systems such as the caudate nucleus, nucleus accumbens and ventral striatum work by motivating and rewarding the punishment of moral transgressors (De Quervain et al., 2004; Singer et al., 2006). Even if there is a transgression that usually warrants punishment in some way, it does not always mean that it is justified. The decision to punish, mediated by the rDLPFC, together with ToM areas, such as the TPJ, helps us to decide whether or not the transgressor is responsible for her actions and if the transgressor should be penalized (Buckholtz et al., 2008). To assess a fitting punishment, affective and social decision-making areas involving VMPFC, amygdala and posterior cingulate cortex allows for functions such as emotional arousal, emotional control and working memory to not only punish according to our emotional response but also punish according to social norms (Buckholtz et al., 2008; Glass et al., 2015; Koenigs, & Tranel, 2007).

This brief overview is a compressed description of some of the underlying neural networks responsible for how punishment works. As seen above, punishment is rather sensitive to contextual factors, such as diffusion of responsibility, and these factors have the ability to significantly alter how punishment works (Feng et al., 2016). Thus, certain situations would change how an overview like this would look like.

This thesis has set out to discuss and explore social punishment in a neuroscientific setting, and what we can note is that the extensive use of games and criminal scenarios have enabled researchers to uncover some neurobiological aspects of punishment. A conclusion drawn from this thesis is that punishment occurs as a general phenomenon in humans to promote cooperation and reinforce social norms and draws upon specific neural networks to function. Furthermore, to be able to expand on this subject in the future, more neuroimaging

studies involving the public goods game might provide insights to undercover neural signatures on social norm formation in both short- and long-term cooperation among strangers.

Since punishment is a social phenomenon, insights on the individual level described in this thesis can have implications for the society as a whole. What would be interesting is to see if convictions regarding how to punish transgressors on the individual level corresponds with convictions on how to deal with transgressors on the societal level. If an offense is treated the same on the societal level as the individual, personal proximity becomes disregarded and in-group context remains somewhat intact. Questions such as this helps to create a more robust foundation for real-world implications on the subject of punishment. Punishment is a very real part of most of our lives and should therefore, hopefully, be based on as solid scientific foundations as possible.

To conclude this thesis, the reason for why most people stand in line at the bank is because people pertain to the norm of the act. The norm is enforced by the line-standers who react negatively to, and punish, the people who try to cut in line. Line-cutting is therefore deterred.

References

Akitsuki, Y., & Decety, J. (2009). Social context and perceived agency affects empathy for pain: an event-related fMRI investigation. *Neuroimage*, *47*(2), 722-734. doi:10.1016/j.neuroimage.2009.04.091

Andreoni, J. (1988). Why free ride?: Strategies and learning in public goods experiments. *Journal of public Economics*, *37*(3), 291-304. doi:10.1016/0047-2727(88)90043-6

Barbey, A. K., & Grafman, J. (2011). An integrative cognitive neuroscience theory of social reasoning and moral judgment. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1), 55–67. doi:10.1002/wcs.84

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. doi:10.1038/nn.3087

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. doi:10.1016/j.neuron.2008.10.016

Corradi-Dell'Acqua, C., Civai, C., Rumiati, R. I., & Fink, G. R. (2012). Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: an fMRI study. *Social cognitive and affective neuroscience*, *8*(4), 424-431. doi:10.1093/scan/nss014

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of personality and social psychology*, *8*(4p1), 377. doi:10.1037/h0025589

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The

neural basis of altruistic punishment. *Science*, *305*(5688), 1254. doi:10.1126/science.1100735

Dilts, A. (2012). To kill a thief: Punishment, proportionality, and criminal subjectivity in Locke's Second Treatise. *Political Theory*, *40*(1), 58-83. doi:10.1177/0090591711427000

Du, E., & Chang, S. W. C. (2015). Neural components of altruistic punishment. *Frontiers in Neuroscience*, 9, 26. doi:10.3389/fnins.2015.00026

Fabbri, M., & Carbonara, E. (2017). Social influence on third-party punishment: An experiment. *Journal of Economic Psychology*, *62*, 204-230. doi:10.1016/j.joep.2017.07.003

Fehr, E., & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4), 185–190. doi:10.1016/j.tics.2004.02.007

Fehr, E., & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and human behavior*, *25*(2), 63-87. doi:10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137. doi:10.1038/415137a

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. doi:10.1002/hbm.23057

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, *7*(2), 77-83. doi:10.1016/S1364-6613(02)00025-6

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*, 11(2), 253–262. doi:10.1093/scan/nsv105

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, *108*, 396-409. doi:10.1016/j.jebo.2014.06.006

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of economic behavior & organization*, *3*(4), 367-388. doi:10.1016/0167-2681(82)90011-7

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics, 19*(4), 741-763. doi:10.1007/s10683-015-9466-8

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

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. doi: 10.1126/science.1129156

Koenigs, M., & Tranel, D. (2007). Irrational economic decision-making after ventromedial prefrontal damage: evidence from the Ultimatum Game. *Journal of Neuroscience*, *27*(4), 951-956. doi:10.1523/JNEUROSCI.4606-06.2007

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*, 9(8), 1143–1149. doi:10.1093/scan/nst092

Matsumoto, D., & Hwang, H. C. (2015). Emotional reactions to crime across cultures. *International journal of psychology*, *50*(5), 327-335. doi:10.1002/ijop.12103

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202. doi:10.1146/annurev.neuro.24.1.167

Morese, R., Rabellino, D., Sambataro, F., Perussia, F., Valentini, M. C., Bara, B. G.,

& Bosco, F. M. (2016). Group membership modulates the neural circuitry underlying third party punishment. *PloS one*, *11*(11), e0166357 doi:10.1371/journal.pone.0166357

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental economics*, *7*(2), 171-188. doi:10.1023/B:EXEC.0000026978.14316.74

Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M., & Moreno, Y. (2013). Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface*, *10*(80), 20120997. doi**:**10.1098/rsif.2012.0997

The President's Council on Bioethics Staff, (2010). An Overview of the Impact of Neuroscience Evidence in Criminal Law. In M.J. Farah (Ed.), *Neuroethics, an introduction with readings* (pp. 220-231). Cambridge: The MIT Press.

Putz, Á., Palotai, R., Csertő, I., & Bereczkei, T. (2016). Beauty stereotypes in social norm enforcement: The effect of attractiveness on third-party punishment and reward. *Personality and Individual Differences*, 88, 230-235. doi:10.1016/j.paid.2015.09.025

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, *17*(8), 413-425. https://doi.org/10.1016/j.tics.2013.06.003

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current biology*, *25*(13), 1731-1735. doi:10.1016/j.cub.2015.05.014

Rilling, J. K., King-Casas, B., & Sanfey, A. G. (2008). The neurobiology of social decision-making. *Current Opinion in Neurobiology*, 18(2), 159–165. doi:10.1016/j.conb.2008.06.003

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003).

The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755-1758. doi: 10.1126/science.1082976

Seymour, B., Singer, T., & Dolan, R. (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*, 8(4), 300–311. doi:10.1038/nrn2119

Singer, T., Seymour, B., O'doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7075), 466. doi:10.1038/nature04271

Skinner, B. F. (1938). The behavior of organisms: an experimental analysis. Appleton-Century. *New York*. Retrieved from http://s-f-walker.org.uk/pubsebooks/pdfs/The%20Behavior%20of%20Organisms%20-%20BF%20Skinner.pdf

Staddon, J. E., & Cerutti, D. T. (2003). Operant conditioning. *Annual review of psychology*, *54*(1), 115-144. doi:10.1146/annurev.psych.54.101601.145124

Szolnoki, A., & Perc, M. (2010). Reward and cooperation in the spatial public goods game. *EPL (Europhysics Letters)*, *92*(3), 38003. doi:10.1209/0295-5075/92/38003

Van Hoorn, J., Van Dijk, E., Güroğlu, B., & Crone, E. A. (2016). Neural correlates of prosocial peer influence on public goods game donations during adolescence. *Social cognitive and affective neuroscience*, *11*(6), 923-933. doi:10.1093/scan/nsw013

Wicker, B., Keysers, C., Plailly, J., Royet, J. P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron*, *40*(3), 655-664. doi:10.1016/S0896-6273(03)00679-2

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*, *145*(11), 1448. http://dx.doi:10.1037/xge0000190