Moralistic Punishment Signaling as a Function of Proportionality

Presentado por:

Juan Camilo Salcedo García

DISERTACIÓN

Presentado como requisito para obtener el título de
Doctor en Psicología

Departamento de Psicología
Facultad de Ciencias Sociales

Universidad de los Andes
Bogotá, Colombia

2020

Comité evaluador:

Jurado Interno
Santiago Amaya PhD
Profesor asociado, Departamento de Filosofía, Universidad de los Andes

Jurado externo
David Pizarro PhD
Profesor asociado, Departamento de Psicología, Cornell University

Asesor de disertación
William Jiménez-Leal, PhD
Profesor asociado, Departamento de Psicología, Universidad de los Andes

Table of Contents

Table of Tables

Table of Figures

1. Abstract

Studies on moralistic punishment, that is punishment not in response to direct antagonism, have shown that it promotes cooperation and prosocial behavior, albeit at a cost to the punisher. Contrary to the view that such punishment is entirely altruistic, recent research suggests that punishers gain reputational benefits from the seemingly selfless act of punishing. This implies that moralistic punishment constitutes a signaling mechanism, and that punishment carries key information about the punisher. Nevertheless, both how the signal is determined as well as the prevailing content and honesty of the signaling function of moralistic punishment have not been fully examined.

I argue that the signaling function of punishing depends on an essential component of moralistic punishment, its proportionality. In four experiments I test how proportionality determines the prevailing content and honesty of the signal. In experiments 1A and 1B, I test how deservedness and severity, two fundamental constituents of proportionality, interact to determine the trustworthiness signaled by the punisher. Experiment 2 expands on this, focusing only on how signaled trustworthiness depends on different levels of severity. Experiment 3 continues to focus on severity, by testing how the prevailing content of the signal changes as a function of the punishment's proportionality, whether the attributed motives for the punishment influence the content and how proportionality determines the social role that the punisher is considered for.

2. Introduction

On June 20th, 2018, a few Colombian fans smuggled alcohol inside a pair of binoculars into the Saransk stadium in Russia for the world cup match between Colombia and Japan. They were filmed by fellow supporters toasting and drinking inside the stadium, and the video went viral very quickly. The ensuing moral condemnation on social media was swift and plentiful, and the

incident was even featured in the local news. A few days after, Avianca, the largest airline in Colombia announced that they were firing Luis Felipe Gomez, one of the men seen toasting and drinking in the video, who also happened to be a high-ranking executive at the company. In a public statement, the airline mentioned that they terminated Gomez's contract because his actions were contrary to the principles and values of the company. A couple of days later, Gomez himself issued a statement explaining that he had not smuggled the alcohol. He had simply been invited to have a drink by some friends he had met by chance inside the stadium and that was what the video captured. In the video he is not seen smuggling the alcohol, but he can be seen drinking from the binoculars.

Some reactions to the "punishment" on social media vindicated Avianca's decision arguing that they had to set an example and send a message, while others claimed it was too harsh, especially since Gomez was on vacation and was in no way performing work-related duties. The opinions people voiced regarding Avianca ranged from total condemnation for excessive and unnecessary use of power, to complete exaltation for setting an example and ensuring the values of the company and even the country. In this case, there are arguably two particular issues that swayed people's opinions regarding the airline's decision. First, there is the issue of severity. Some said the airline went too far, arguing a less severe penalty might have sufficed, while others mentioned it was commensurate with the offense because, as an executive, he should have upheld the values of the company in every possible instance. Second, there is the issue of deservedness. Since he didn't smuggle the alcohol himself, some argued he didn't deserve the punishment, while others claimed that drinking from clearly smuggled alcohol deemed him sufficiently guilty and thus deserving of the sanction. The image that laypeople have regarding a punisher is determined by the particularities of the punishment exacted by said punisher. In this case the opinions and

perceptions of the punisher (i.e. Avianca) became a trending topic in social media and inspired a number of opinion pieces both for and against the airline's decision in major news outlets in Colombia. In virtually all of the opinions and commentaries, people based their judgments about the punisher on the perceived severity and deservedness of the punishment.

How does proportionality play out in the more general assessment of both punishment and punisher? Surprisingly, this is a question that has not been explored in psychology, probably because the focus of empirical studies has been so far on the motivations for punishment centered around a limited set of justifications for it (e.g. Carlsmith & Darley, 2008; Osgood, 2017). The impact of proportionality on punishment has mostly been debated as part of a larger philosophical discussion on the justification of punishment (e.g. Sloan & Miller, 1990). In order to track the main ideas on proportionality, in the next section I will provide an overview of the central punishment justifications postulated in the literature. Furthermore, I'll describe how the perception of the punisher has been evaluated within one of those theories (i.e. the expressivist theory of punishment), albeit without a focus on how proportionality determines said perception, and I will give an account of the empirical evidence for the expressivist theory of punishment. Next, I will provide a detailed characterization of punishment, which allows for a wide variety of punishments to communicate specific traits about the punisher (i.e. moralistic punishment) and will then present the extant research which supports the notion of moralistic punishment as a signaling mechanism. Then I will proceed to explain how severity and deservedness can determine what is communicated by moralistic punishment. The document then specifies the research questions, details the four experiments intended to elucidate them and finally offers a general discussion on the implications of the findings, as well as the scope and limitations of the studies.

3. Background

Punishment is a ubiquitous phenomenon in all human societies. It is fundamental in enforcing social norms, laws and justice systems around the world and is also a common practice in private institutions (Bedau & Kelly, 2017). Intuitively, punishment can be understood as a penalty inflicted as a consequence of an offense. The justifications for punishment have been voluminously discussed in the fields of philosophy and theoretical law (e.g. Bentham, 1780; Durkheim & Simpson, 1933; Hart, 1968; Kant & Hastie, 2002; Moore, 2010). On the one hand, punishment has been argued as primarily an instrument of deterrence (Van den Haag, 1975; Wilson, 1983). This perspective posits that punishment is a way to avoid repeated transgressions by the offender or to prevent potential transgressions by other members of society (Sloan & Miller, 1990). On the other hand, it has been argued that the goal of punishment is to provide retribution (Singer, 1979; von Hirsch, 1992). According to this argument, punishing is justified on moral grounds and as such, transgressors justly deserve their punishment even if it does not act as a deterrent for further transgressions (Sloan & Miller, 1990). Other justifications have also been formulated as complimentary to deterrence and retribution, including inequity aversion (Bone, McAuliffe, & Raihani, 2016), incapacitation (Levitt, 2004), restitution (Wietekamp, 1989) and rehabilitation (Moore, 2010).

However, a different theory argues that the goal of punishment is to communicate moral condemnation to the transgressor, thereby reaffirming social values and norms to the public at large (Feinberg, 1965; Hampton, 1984; Morris, 1981; Nozick, 1981). This so-called expressivist view of punishment argues that punishing has a primary symbolic function and serves as a social signaling mechanism (Kahan, 1996). The expressivist theory argues that punishment is in essence a language used to convey moral condemnation (Kahan, 1996). In his defense of the expressivist

function, Feinberg (1965) writes: “punishment is a conventional device for the expression of attitudes of resentment and indignation, and of the judgments of disapproval and reprobation on the part either of the punishing authority or of those ‘in whose name’ the punishment is inflicted.” (p. 400).

3.1. Moralistic punishment as a signaling mechanism: evidence for the expressivist theory

In the psychological arena, studies have focused on the motivations for punishment. In this area, the theory of retribution as a motivator of punishment appears to have more empirical support (Carlsmith, 2006; Carlsmith & Darley, 2008; Carlsmith, Darley, & Robinson, 2002; Darley & Pittman, 2003; Ellsworth & Mauro, 1998; Lerner, Goldberg, & Tetlock, 1998) than the deterrence motivation theory (Vidmar & Ellsworth, 1974; Ellsworth & Ross, 1983; Osgood & Muraven, 2015). By focusing on the deterrence-retribution dichotomy, these researchers have privileged experimental strategies whereby third parties evaluate the transgression and / or transgressor and then assign appropriate punishments in one-shot instances (without any further interactions or consequences for the punisher). Such methodology makes it impossible to determine whether punishing has an expressivist function. Moreover, psychological research on motivation related to the expressivist value of punishment has focused on revenge, with vengeance seekers reporting an attempt to communicate relative strength, disapproval of the transgression, or a demand for respect (Crombag, Rassin, & Horselenberg, 2003; French, 2001; Frey, Pearson, & Cohen, 2015; Gollwitzer, 2009; Vidmar, 2001).

Nevertheless, there is some empirical evidence backing the expressivist theory of punishment. Although not explicitly, it is supported by research on so-called “moralistic”

punishment. Moralistic punishment within costly signaling models typically refers to the opportunistic – that is, not in response to direct antagonism – punishment of social norm violations (Gordon & Lea, 2016). This includes third party punishment (in which unaffected observers punish unfair or selfish behavior) as well as punishment within public goods games. While moralistic punishment has been widely observed in laboratory and field experiments (Balafoutas & Nikiforakis, 2012; Boyd & Mathew, 2015; Fehr & Fischbacher, 2004; Fehr & Gächter, 2000; Jordan, McAuliffe & Rand, 2015; Kurzban, DeScioli, & O'Brien, 2007; Ostrom, Walker, & Gardner, 1992; Sefton, Shupp, & Walker, 2007) it constitutes a puzzle from an evolutionary standpoint. Despite the fact that moralistic punishment has been shown to promote cooperation and prosocial behavior in a variety of contexts (Balliet, Mulder, & Van Lange, 2011; Fehr & Gächter, 2002), from an evolutionary and game theoretic perspective it is not very clear why an individual would choose to incur a cost to punish another when it is not a response to direct antagonism. In particular, the evolution of moralistic punishment as a viable and stable behavior is difficult to explain given that it imposes costs on an individual while the benefits are shared amongst the group as a whole (Dreber, Rand, Fudenberg, & Nowak, 2008). In addition, punishers can be exploited by individuals who cooperate but don't punish, yet reap the cooperative benefits of punishing (Yamagishi, 1988), and can even suffer directly or indirectly from antisocial or retaliatory punishment (Dreber & Rand, 2012). However, the problem of the costs incurred by individuals who engage in moralistic punishment can be solved if punishers can somehow gain indirect benefits from the act of punishing itself. Namely, moralistic punishers can use the act of punishing to signal qualities that make them desirable as partners in other social interactions (Santos, Rankin, & Wedekind, 2010) or to make them seem formidable so that they are not taken

advantage of in future social interactions (Gordon & Lea, 2016). Such reputational benefits can outweigh the costs of punishing and allow it to evolve as a stable behavior

Conceptualized as a signaling mechanism, moralistic punishment serves as a buttress for the expressivist theory of punishment because it allows for the act of punishing to communicate what sort of behavior the punisher disapproves of while at the same time signaling personal character traits. In addition, the expressivist characteristic of moralistic punishment could arguably be subject to a number of contextual variables, including the proportionality of the punishment, which could potentially affect both the honesty, and overall content of what is signaled. Nonetheless, the version of moralistic punishment found in costly signaling research is restricted by the empirical methodology used (Barclay, 2006; Gordon & Lea, 2016; Gordon, Madden, & Lea, 2014; Jordan, Hoffman, Bloom, & Rand, 2016; Nelissen, 2008). Namely, it almost exclusively focuses on monetary punishments with monetary costs on economic games with traditional third-party punishment paradigms. This is problematic because it limits the types of violations worthy of punishment and the type of punishments evaluated to interactions involving money. Likewise, it restricts the possible variability of the punishment's proportionality. However, moralistic punishment is arguably a more common and general phenomenon than has been explored in the research thus far. In order to clarify what moralistic punishment encompasses, I will provide a more detailed characterization of the phenomenon. Such a definition will be necessary afterwards in order to understand how proportionality could possibly be a crucial determinant of the signaling function of moralistic punishment.

3.2. Characterization of moralistic punishment

In order to give a precise description of what I mean by moralistic punishment, I first define what it does not entail. To begin with, moralistic punishment is fundamentally different from the concept of punishment utilized in operant conditioning theory (Chance, 2013). Moralistic punishment is not the presentation of an aversive stimuli (or the removal of a valued stimuli) after a given behavior or response in the hopes of reducing the reoccurrence of said behavior.

Secondly, even though moralistic punishment can be taken as a component of certain justice systems, it is not the same as legal punishment. Most definitions of legal punishment involve the intentional imposition of something burdensome and reprobative on an offender for having committed a crime, by a person or body who claims the authority to do so, such as a judge or jury (Duff & Hoskins, 2018). The difference with moralistic punishment resides in the characterization of said offender and crime, as well as on the nature of the authority. Namely, in legal punishment, the crime and the offender are typically determined by the pre-established legal code of the institution or country in question. The person or body who imposes the punishment is also determined within the institution's legal and/or political system which proclaims it as the pertinent authority. In moralistic punishment, what constitutes a crime (and thus an offender) is not necessarily determined by a law. Instead, a mere perception that a moral norm has been violated is sufficient justification for punishment.

Most normative definitions of punishment expand on legal punishment to include an authority that deliberately inflicts harm upon an offender who has committed a public wrong for which he is morally responsible, and where the harm is crucially proportional to the seriousness of the crime and the moral guilt of the offender (Geeraets, 2018). Moralistic punishment, however,

need not be restricted by the justifications of such normative theories. In fact, moralistic punishment most closely follows the two tenets of Leo Zaibert's (2006) definition of punishment:

(1) Person A should, in order to punish, blame person B for doing X;

(2) In response to X, A should do something to B which A believes painful for B to endure

(p. 31)

As such, moralistic punishment does not require the person imposing the punishment to be a legal authority. More importantly, moralistic punishment does not conceptually require any proportionality between the seriousness of the crime and the hard treatment inflicted in response to be regarded as punishment. However, a key characteristic of moralistic punishment is that it does not take place in response to direct antagonism. In other words, moralistic punishment can only be enacted by individuals that are not directly affected by the transgression in question. This is necessary given that punishment in response to direct antagonism can be greatly driven by retaliatory motives and makes it very difficult to identify the extent to which it is precipitated by moral concerns as opposed to personal interest. In fact, it would be nearly impossible to conceive of a person punishing in response to direct antagonism that does so entirely motivated by moral concerns entirely independent of their personal stake in the matter. Therefore, moralistic punishment is performed solely by third parties. In previous research, moralistic punishment has been commonly referred to as third-party punishment and in the remainder of this text both terms will be used interchangeably to refer to the same concept.[1]

[1] Pedersen, Kurzban and Mccullough (2013) have argued that most third-party punishment elicited in the lab suffers from some key methodological limitations (the most important of which is the "strategy method" employed in a number of published studies) and have therefore concluded that third-party punishment does not exists (or at least is much less prevalent than most researcher have claimed). However, their claim does take into account field studies that have found evidence for third-party punishment in naturalistic settings (e.g. Balafoutas & Nikiforakis, 2012),

Nevertheless, the common conception of third-party punishment observed in the scientific literature can be expanded on to include punishing agents that go beyond single individuals. Within costly signaling theory, virtually all research on third-party punishment has also been conducted from the point of view of the individual. However, third-party punishment can take place in interactions that range from the triadic to those involving potentially millions of individuals thanks to modern social networks. It can potentially be exerted by and within groups of people, clubs, companies, NGOs and governments.

Zaibert's definition of punishment also puts forth a very flexible conceptualization of wrongdoing. Specifically, there is no explicit set of legal, religious or moral norms that prohibit X. Therefore, the wrongdoing in question must not necessarily be a public wrongdoing. Instead, X only needs to be considered wrong against a certain standard to which A is committed, and thus all that is required is for A to consider the person who caused X blameworthy. Moralistic punishment, though, would require that the particular standard to which A is committed be moral in nature. Moralistic punishment can thus be regarded as a response to a moral norm violation which enables the punisher to express his/her own moral outrage, and can therefore be conceived as the action tendency of the moral emotion of outrage (Frijda, 2004). By punishing, the agent is expressing what kind of moral norms she cares about.

Nonetheless, in order for moralistic punishment to have evolved as a stable behavior it must also be able to confer some kind of indirect personal benefit to the punisher that outweighs the cost incurred by punishing. Next, I present two distinct lines of research that point to two different types of indirect benefits punishers can obtain via third-party punishment.

---

nor a study that found that the "strategy method" (and endowment size) do not significantly affect third-party punishment (Jordan, McAuliffe, & Rand, 2016).

3.3. Indirect personal benefits gained with third-party punishment

A first line of research posits that moralistic punishment offers indirect benefits to punishers by allowing them to signal qualities that make them appear as desirable partners in future social interactions. Such reputational benefits can outweigh the costs of punishing and allow it to evolve as a stable behavior (Santos et al., 2010). In support of this signaling quality of moralistic punishment, Barclay (2006) found that punishers in iterated (as opposed to one-shot) public goods games were both rated as more trustworthy, group focused and worthy of respect than non-punishers. In posterior dyadic trust games, those same punishers also benefited monetarily compared to non-punishers. In another study, Raihani and Bshary (2015) found that third-parties that punished selfish dictators in a third-party punishing game, were more likely to be rewarded by bystanders than third-parties who took no action in response to a selfish dictator. Likewise, Jordan, Hoffman, Bloom and Rand (2016) presented evidence which suggests that third-party punishing (TPP) is used as a costly signal of trustworthiness in a traditional trust game paradigm where participants had to decide how much money to entrust a person who had taken part in a previous TPP game as the punisher. Similarly, Nelissen (2008) showed that third-party punishers who incurred a greater cost to punish were rated as friendlier than those who incurred a lower cost, and found that perceptions of fairness mediated the amounts of money entrusted to previous third-party punishers on subsequent trust games. This cumulative evidence shows that the act of punishing says something about the punisher. In particular, this line of research suggests that third party punishment signals some of the underlying character traits which deem the punisher trustworthy, or more generally as a desirable partner for future social interactions.

A different line of research suggests that third party punishment has evolved as a stable behavioral mechanism to signal dominance with the ultimate goal of deterrence. This theory posits

that because humans evolved in small-scale societies where the most common type of interactions were face-to-face, the human mind infers that mistreatment of a third party predicts later mistreatment of oneself (Roos, Gelfand, Nau, & Carr, 2014). Hence, third party punishment serves as a signal of dominance aimed at deterring personal mistreatment. Krasnow, Delton, Cosmides and Tooby (2016) found that when punishers don't have information about how they will be treated, they infer that mistreatment of a third party predicts mistreatment of themselves, and those inferences in turn, predict third party punishment. In a different study, Krasnow and Delton (2017) found that punishers in a traditional third party punishment game punished the most when they predicted that they themselves would be treated worst by dictators. In direct support of this dominance-signaling theory, it has been shown that engaging in third party punishment can effectively signal dominance without many of the negative social consequences associated with other forms of aggressive behavior (Gordon et al., 2014), and that moralistic punishment allows dominant individuals to maintain their high status position, while failure to punish leads to that position being at risk (Gordon & Lea, 2016). The indirect personal benefits from such signaling can also be attained if the punisher possesses some sort of welfare interdependence with the actual or potential victims of the aggressor (Pedersen, McAuliffe & McCullough, under review). In other words, third-party punishment can be used to deter aggressors from harming individuals with whom the punisher shares a fitness goal (J. W. Martin, Jordan, Rand, & Cushman, 2019; Pedersen, McAuliffe, Shah, et al., 2018). In sum, this line of research highlights the possibility that third-party punishment serves as an honest signal of the capability of retaliation preemptively directed at would be offenders to avoid potential mistreatment of the self.

The numerous studies supporting these two lines of research have helped to highlight third-party punishment as a signaling mechanism of either trustworthiness or dominance. However, even

though researchers on both sides recognize that third-party punishment could be used to signal more than one trait type, there are currently no studies that attempt to find out how and when third-party punishment signals one over the other. In the next section I describe how these two signal types fit within social perception models allowing for exploration of trustworthiness and dominance in tandem.

3.4. Third-party punishment within social perception models

These two distinct lines of research which point to third-party punishing as a mechanism to signal trustworthiness / general desirability on the one hand, and dominance / capability of retaliation on the other, align with commonly observed orthogonal dimensions of person perceptions. Specifically, they line up with the warmth and competence dimensions at the center of the Stereotype Content Model (Cuddy, Fiske, & Glick, 2008), and the trustworthiness and dominance dimensions of functional human face evaluation (Oosterhof & Todorov, 2008). Both of these frameworks map onto a primary assessment of affiliation underlying approach or avoidance mechanisms (warmth and trustworthiness), and a secondary (both in importance and temporal sequence) assessment of potential action and agency (competence and dominance). According to these orthogonal dimensions of social perception, actors could first obtain information regarding the intentions of third-party punishers towards them (the warmth dimension) and then figure out the corresponding capability of said third-party punishers to pursue those intentions (the competence dimension).

However, Goodwin (2015) argues that the warmth dimension actually conflates morality traits (such as trustworthiness, honesty and sincerity) with sociability traits (such as friendliness, extroversion and warmth itself). Research by Goodwin and colleagues suggests that moral traits

predominate social perception and evaluation (Goodwin, Piazza, & Rozin, 2014) and that sociability and competence traits are perceived as positive or negative contingent upon judgments of morality traits (Landy, Piazza, & Goodwin, 2016). Even though this three-dimensional model expands on (but does not contradict with) the warmth and competence model (Goodwin, 2015), I will only rely on the two-dimensional model to analyze third-party punishment because it is rarely deemed entirely immoral and more importantly because it potentially conveys very little sociability information.

With regards to morality, third-party punishment would arguably only be regarded as completely immoral in very specific instances that are not commonly observed in everyday life. Namely, it would take the intentionally severe punishment of an innocent person (where the punisher knows the alleged offender is completely innocent). In most real-world scenarios, third-party punishment is considered at least partially moral. Regarding sociability, it would most likely be very difficult to infer any of the sociability traits that this model specifically alludes to (i.e. sociable, warm, friendly, easy-going, extroverted and playful) (Landy et al., 2016) from a third-party punishment scenario alone. Previous research, though, has found robust evidence that third-party punishers can be independently perceived as trustworthy (e.g. Barclay, 2006; Jordan et al., 2016) and friendly (Nelissen, 2008) which are key measures of warmth (Cuddy et al., 2008) and as dominant and capable of retaliating (e.g. Delton & Krasnow, 2017; Gordon et al., 2014) which indicate competence. These two dimensions can be harnessed to analyze how third-party punishers are socially perceived in a systematic manner. By studying how third-party punishers are perceived along these two dimensions (while at the same time avoiding the questions that exclusively allude to sociability traits), it is possible to gain a better understanding of the signaling content of third-

party punishment within a well-established social perception framework and examine how such perceptions vary as a function of proportionality.[2]

Nevertheless, previous research has shown the possibility of act-person dissociations whereby observers judge a given act as positive but nonetheless make negative attributions about the moral character of the agents that carry out those actions (Uhlmann, Zhu, & Tannenbaum, 2013). A cause of these dissociations lies in the attributional ambiguity of the actions and the asymmetry of the assumed motivations driving said action. Namely, an act that is viewed as morally correct, can be driven by moral or immoral motives, which in turn leads to corresponding positive or negative character evaluations of the person who carried it out. Therefore, the attributed motivations for engaging in a given action provide cues that inform social perception. The extent to which punishers are perceived as warm or competent, however, provides little information as to the attributed motives for third party-punishment. In the next section I explain why this is important in understanding the signaling of third-party punishing.

3.5. The perceived motives of third-party punishers

According to the expressivist view, a third party that enacts a severe punishment on a transgressor is ostensibly motivated to do so, in large part, because he really cares about the underlying moral norm that was violated. In this sense, third-party punishment is useful for the punisher because it allows him to express his moral condemnation and makes a statement about which moral norms he cares about. However, it could also be leveraged by the third party to appear as if they care about the moral norm in question. Disingenuous punishment allows the third party

[2] Even though the SCM provides me with a social perception framework within which to situate the perceptions of punishers along the orthogonal dimensions of warmth and competence, I do not necessarily commit to the other theoretical components of the SCM, such as its relation with prejudiced emotions or the four patterns of behavioral responses that derive from it

to signal warmth and / or competence and gain the corresponding indirect social benefits associated with such signals. Previous research has shown that moral condemnation of a transgression is a better signal of moral behavior and overall character perception than direct statements of moral behavior (Jordan, Sommers, Bloom, & Rand, 2017). Because it is a comparatively costlier signal (it entails a significant cost for the punisher compared to moral condemnation), third-party punishment is arguably a more effective signal of character trait and moral behavior. This makes it particularly attractive for committed dishonest signallers who wish to obtain the corresponding indirect benefits.

Therefore, the attributed motives for punishment can provide key additional information that explains what is driving the warmth and competence perceptions of third-party punishers. More specifically, if an individual is judged as engaging in third-party punishment out of genuine moral concern, it is more likely that she be perceived as trustworthy than if she is judged as engaging in third-party punishment for reputational-seeking interests. Likewise, observers will arguably deem a disingenuous third-party punisher (i.e. punishing for reputational seeking interests) as less dominant and less likely to retaliate than an honest third-party punisher. Furthermore, if the punishment is motivated by real moral principle, observers will be more likely to infer that the punisher cares about the underlying moral norm and is less likely to violate it himself.

By taking into account the motives for punishment to examine the perceived warmth and competence of third-party punishers, the signaling function of moralistic punishment can be more comprehensively mapped out. How third-party punishment signals trustworthiness and warmth versus dominance and competence helps to clarify how third-party punishers are socially perceived, and to what extent they will be taken into consideration for future social interactions.

However, previous research has shown that person judgments can depend highly on the social role they are intended to play (i.e. leaders vs non-leaders) (Uhlmann et al., 2013). Similarly, the social roles that a third-party punisher could fill are arguably very different, and as I'll expand in the next section, potentially opposed with regards to warmth and competence perceptions.

3.6. Third-party punishers as potential friends or leaders

Given their social role and social objective, a friend and a leader are very likely to possess distinct character traits. Thus, when choosing a friend, individuals will look for character traits that are arguably unlike those they look for when choosing a leader. From an evolutionary perspective, people choose friends based on their potential for cooperation and mutual aid (Tooby & Cosmides, 1996) and their potential as allies that can provide support during interpersonal and intergroup conflict (DeScioli & Kurzban, 2009; DeScioli & Kurzban, 2012). In other words, most people want a friend who'll be there for them to provide some sort of material or psychological support in case they need it (regardless of whether or not the need is actualized).

Leaders, on the other hand, tend to be sought out as orchestrators of solutions to collective problems. From the evolutionary perspective, leaders are central for the successful navigation of problems related to group living and people often choose them for this purpose (Price & Van Vugt, 2015). Consequently, leaders tend to possess a set of character traits that enable them to enforce collective action in the face of social conflicts (Tooby & Cosmides, 2010; von Rueden, Gurven, Kaplan, & Stieglitz, 2014). In other words, most people want a leader who'll help them resolve social problems and mobilize their group towards a common goal. To ensure contributions to group objectives and other types of collective action, one of the most effective tools at the disposal of

leaders is punishment (Fehr & Gächter, 2000; Tooby, Cosmides, & Price, 2006; von Rueden et al., 2014). Therefore, it could be argued that people look for the capability to punish non-contributors (or cheaters) as a focal trait of would-be leaders.

Given that an essential trait of competent leaders is the enforcement of participation in collective action, the benefits of competence and dominance are not necessarily tied to the benefits commonly expected from friendship. The value of a friend is not derived from their capability to coerce or convince others to invest in collective goods and actions, but from their willingness to invest in us. Moreover, the character traits typically valued in leaders, such as impartiality and dominance (Boehm, 2000; von Rueden et al., 2014), can be counter-adaptive to value when choosing friends. Specifically, the likelihood that a dominant leader will engage in within-group exploitation (Boehm, 2000), makes it costlier to have a dominant friend than a dominant leader (J. K. Snyder et al., 2011).

Research conducted by Laustsen and Petersen (2015) provides empirical evidence for this divergence between dominant trait preference for leaders and friends in the realm of face perceptions. The authors found that people tend to choose dominant leaders over non-dominant ones (in terms of dominant versus non-dominant face perceptions according to Oosterhof and Todorovs' two-dimensional model (2008)), especially when confronted with a social problem or threat context (versus a natural disaster context). They also found an interesting difference in leader type dependent on political ideology. Namely, conservative participants were significantly more likely to choose dominant leaders than liberal participants irrespective of context. More importantly, the researchers found that irrespective of context and ideology, participants strongly prefer non-dominant friends. Interestingly, the dominant faces were also rated as significantly more unfriendly than the non-dominant faces (Laustsen & Petersen, 2015), which coupled with

research that shows that dominant features are inversely related to perceptions of trustworthiness (Buckingham et al., 2006; Jensen & Petersen, 2011) points to a possible dissociation between trustworthiness (a key component of warmth) and dominance (a key component of competence) perceptions at work when considering another individual as a potential leader or friend.

In conjunction, the perceptions of warmth and competence, in addition to the perceived motives for punishment as well as the potential social role best suited to the third-party punisher, represent a comprehensive framework to study the content, honesty and purpose of moralistic punishment signaling. Nonetheless, when considering the determinants of the character traits of a third-party punisher, there are only a handful of variables that can significantly alter the signaling content of the punishment itself. It appears that, everything else equal, the only variable at the punisher's disposal is the proportionality of the punishment. In the next section I describe how proportionality plays a central role in the signaling function of moralistic punishment and how it can help understand the different perceptions observers might have of third-party punishers.

3.7. Moralistic punishment as a signal determined by proportionality.

Proportionality is a basic component of punishment (Bedau & Kelly, 2017). It is a necessary concept of the act of punishing and is regularly decided by the punisher herself. Given its fundamental role within punishment, it can be understood as a crucial determinant of what is signaled by the act of punishing. Punishment is a response with a normative value. It can be correct or incorrect, and it is intuitively thought that a person is right to exact punishment when it is fair (or wrong to do it when it is unfair). Fairness, however, is a relational property of punishment and thus can be more precisely operationalized as proportionality

On the one hand, punishment can only be logically imposed on individuals who are believed to have acted wrongly (Bedau & Kelly, 2017). Being found guilty (regardless of its adequacy) is a necessary condition for justified punishment (Bedau & Kelly, 2017). Deservedness is a fundamental constituent of proportionality, and hence of punishment in that regardless of the justification adjudicated for the punishment (e.g. deterrence or retribution), most theorists agree that to deem someone guilty is to imply responsibility of the corresponding offense, thus making punishment permissible (Wasserstrom, 1964). Moreover, research by Barclay (2006) indicates that punishment imposed on non-offenders (i.e. not directed at free riders on public goods games) is not rewarded monetarily in posterior trust games. This suggests that deservedness can determine the signaling value of punishment.

On the other hand, severity is a measure of the fairness of the punishment (von Hirsch, 1992). People have a sense that punishments scaled to the gravity of offenses are fairer than punishments that are not (von Hirsch, 1992). Likewise, Darley, Carlsmith and Robinson (2000) have presented evidence which suggests that people are motivated to inflict punitive measures on the offender in proportion to the severity of the crime. Carlsmith, Darley and Robinson (2002) have also demonstrated that the seriousness of the crime predicts the severity of the punishment that people assign. Furthermore, evidence presented by Goodwin and Benforado (2015) suggests that most people interpret punishment as owed to an offender in proportion to the severity of their wrong, and Hofmann, Brandt, Wisneski, Rockenbach and Skitka (2018) found that people seek to punish in proportion to the perceived wrongness of the transgression.

Moreover, a robust line of research suggests that people look for behavioral patterns that may indicate the presence of underlying positive or negative traits in others (Fudenberg, Rand, Dreber, Ellingsen, & Nowak, 2009; Schweitzer, Hershey, & Bradlow, 2006). Given that

proportionality is determined by the punisher, and that it is a property of all moralistic punishments, people could use information about proportionality to infer the presence or lack of underlying character traits. Firstly, proportionality in third-party punishment can be used as a proxy to infer the punisher's underlying trustworthiness (and general warmth) and dominance (and general competence) traits. Secondly, the perceived motives of the punisher could be inferred from the punishment's proportionality. Finally, it can also convey something about the potential roles the punisher could or could not fulfil in the future. Intuitively, a disproportionately lenient punishment could be said to signal a lack of interest in the underlying moral norm that was violated, a corresponding attempt at disingenuous moral signaling, and / or relatively low competence. Such a disproportionately lenient punishment might deem the punisher a poor candidate for a leadership position, but it could also be interpreted as a forgiving characteristic desirable in a friend. A disproportionately severe punishment on the other hand, might make the punisher appear as caring too much about the underlying moral norm and be perceived as a corresponding genuine signal of competence (unless its interpreted as moral grandstanding to seek status (Tosi & Warmke, 2016)), but not the kind of person you might want as friend.

Furthermore, proportionality is arguably the only element that can be controlled by the punisher in a third-party punishment. An observer first has to decide whether or not to punish and then decide how severe that punishment will be. The former is a categorical decision and determines if the observer becomes a punisher or not. The latter designates the proportionality of the punishment. Granted, the particular manner and delivery of the punishment can vary tremendously, but both are directly connected to and interpreted as part of the punishment's proportionality. Even though the cost incurred by the punisher can be theoretically controlled in isolation by a third-party punisher (and in fact it has been manipulated in isolation within third-

party punishment economic experiments (Nelissen, 2008)), in real world scenarios, the proportionality of the punishment is inextricably tied to its cost. The harsher the punishment, the higher the cost incurred, and a higher cost typically results in a more severe punishment.

In conclusion, different research strands suggest that punishment carries key information about the punisher and the characteristics of the situation that make the punishment itself possible. Additionally, proportionality is a fundamental component of punishment and as such plays a potentially determinant role in the information conveyed by the punishment. All else being equal, the proportionality of a third-party punishment has the potential to determine the overall content of what is signaled by moralistic punishment, the perceived motives of the punisher and the potential social role that the punisher could play in future interactions. Recent research has not, however, integrated these aspects in the study of moralistic punishment.

4. Research Questions

In order to effectively express moral outrage and communicate the moral norms that matter to an agent, the punishment must function as a credible signal of moral traits. So far, the existing research has presented evidence that moralistic punishing can signal trustworthiness, and dominance. Nevertheless, the signaling function of moralistic punishment has been limited and greatly determined by the experimental paradigms used thus far. Namely, proportionality, arguably a crucial component of punishment, has not been taken into account as a determinant of the content, honesty and purpose of the signal. Therefore, the general objective of this research is to identify how the signaling function of third-party punishing is determined by the punishment's proportionality and to evaluate the extent to which third-party punishing effectively signals the

underlying motivations (genuine moral concern or self interest), the character traits of the punisher (warmth or competence), and determines which social role the punisher is considered for (friend or leader).

Accordingly, I address 5 specific questions:

(1) How do the reputational benefits of third-party punishing (TPP) in an economic game vary according to the punishment's proportionality? Specifically, experiments 1A and 1B attempt to find out how the trustworthiness signaled via TPP is affected by two levels of deservedness (undeserved vs deserved) and two levels of punishment severity (no punishment vs high punishment severity).

(2) How do the reputational benefits of TPP in an economic game vary according to severity when the punishment is deserved? Experiment 2 focuses only on deserved punishment to address how the trustworthiness signaled via TPP is affected by four levels of severity (low, medium-low, medium-high and high).

(3) What is the prevailing content of the signal emitted via TPP and does the content vary depending on proportionality? Concretely, experiment 3 assesses to what extent TPP signals warmth versus competence as a function of 3 levels of punishment severity (disproportionately lenient, proportional and disproportionately severe).

(4) Do the attributed motives for TPP vary depending on proportionality and do the attributed motives mediate the perceived warmth and competence traits? This question is also addressed in experiment 3 focusing on self-interest motives vs genuine moral principle for the same three levels of proportionality.

(5) Does the social role assigned to the moralistic punisher vary as a function of proportionality in TPP? This question, also addressed in experiment 3, aims to find

out how the proportionality of TPP affects how a punisher is considered for a leadership role vs a friendship role for the same three levels of proportionality?

5. Experiments

In addition to vignette evaluations, the following studies make use of economic game paradigms. In particular, experiments one (A and B) and two make use of third-party punishing games (TPP) and trust games (TG). The vast majority of moralistic signaling research has been conducted with these economic games (e.g. Barclay, 2006; Everett, Pizarro, & Crockett, 2016; Gordon & Lea, 2016; Jordan et al., 2016; Nelissen, 2008), so it makes sense to test the effect of proportionality on third-party punishment within such paradigms.

A conventional TPP involves a traditional Dictator Game but introduces a third player as an observer and potential punisher. Specifically, in a TPP a sender is endowed with a given amount of money and must decide whether to share it with a receiver. If the sender does not share, the observer must decide whether to pay to punish the sender. A punishment entails a cost to the observer (effectively lowering her/his endowment) but results in a loss to the Sender (lowering her / his endowment as well). This research will make use of TPP situation results as input for the subsequent trust game (TG) of interest. Even though TPP does not encapsulate all forms of punishing, it does allow modelling of the most common type of punishment which institutions and individuals engage in under law and especially in liberal constitutional democracies. As in TPP, punishments are most often determined by and administered by directly unaffected third parties and most societies incur considerable costs carrying them out (Bedau & Kelly, 2017). Moreover, decisions that involve costs for the decision maker are perceived as especially informative about character (Ohtsubo & Watanabe, 2009). Hence, costly TPP works both as a fairly good

representation of a large portion of punishments in everyday social interactions and has the potential to express underlying traits such as but not necessarily limited to trustworthiness.

The studies rely on variations of the trust game (TG) similar to that used by Jordan et al. (2016). In the TG senders are endowed with a given amount of money and must then decide how much of that money to send to a receiver. Whatever amount is sent is multiplied by 3 and the receiver must then decide how much of that larger amount to return to the sender. As such, the optimal decision entails cooperation, with the sender sending her entire endowment and the receiver returning 50% of the now larger amount (Camerer, 2003). Nevertheless, there are incentives for selfish behavior both on the part of the sender (especially if she doesn't trust the receiver) as well as on the part of the recipient (given that it's a one-shot game and there would be no retribution whatsoever from keeping all the money).

5.1. Experiment 1A

Given that most of the existing research has focused on trustworthiness signaled by punishment (Barclay, 2006; Jordan et al., 2016), analyzing which factors influence the extent and effectiveness of the signaling of trustworthiness was a reasonable starting point. Therefore, the purpose of the first experiment was to test how the deservedness and severity of a punishment within a third-party punishment (TPP) paradigm influenced the perceived trustworthiness of the punishers measured in a subsequent trust game (TG).

In experiment 1A, subjects played as senders in the TG with a recipient that "played" as an observer in the previous TPP game. In the TG, senders were informed of the "behavior" of the recipient in the previous TPP game. However, unlike the TG in Jordan et al. (2016) recipients and their behavior were manipulated. Specifically, participants first read about a given TPP including

the amount sent by the sender (either half of or none of the money; independent variable capturing the punishment's deservedness).

With the purpose of reducing ambiguity with regards to deservedness, I informed participants of the average amount sent by most senders in the TPP, thereby establishing a baseline for comparison against which participants can make an informed deservedness judgment. A meta-analysis of 616 treatments analyzing 261 publications involving dictator games (Engel, 2011), found that 64% of senders send at least some of their money to recipients, and on average they send (approximately) 30% of their money. Therefore, given the dichotomous nature of the present study's TPP game (where senders can choose to share or not 50% of their endowment) I informed participants that over 60% of senders share their money with recipients. Participants also read about the corresponding punishment exacted by the observer (capturing the severity of the punishment).

Next, participants played a TG as senders with the observers in the previous TPP as receivers. They were endowed with 20 pence and had to decide how much (if any) of their money to send to the receiver (dependent variable). The amount of money sent served as a proxy of the level of trust placed on the receiver as a function of the punishment's deservedness and severity in the TPP game. This allowed me to measure how effectively punishing signals trustworthiness in the face of these two factors.

Following Everett, Pizarro, and Crockett (2016), after the Trust Game (TG) I also provided participants with the punishment decisions of four other hypothetical TG partners corresponding to the four possible deservedness-severity conditions and then asked them who of the four possible partners they preferred to play with in another TG game. This served as an additional explicit measure of partner choice.

5.1.1 Hypotheses

The findings obtained by Jordan et al. (2016), suggest that justly punishing an offender will signal trustworthiness. Moreover, people tend to rely on traits such as trustworthiness and fairness to infer subsequent cooperation (Walker & Hennig, 2004), and a lack of trustworthiness suggests a person will defect in joint endeavors, while unfair treatment suggests they will not divide resources equitably (Uhlmann, Pizarro, & Diermeier, 2015). In addition, given that behaviors that are statistically rare or otherwise extreme are perceived as highly informative about character traits (Ditto & Jemmott, 1989; Fiske, 1980; Kelley, 1967) it is reasonable to assume that extreme punishments (with regard to severity) or uncommon punishments (with regards to deservedness) will lead to stronger trait inferences. In other words, in line with Jordan et al. (2016), it is safe to assume that fair punishers will be trusted more than unfair punishers and than those who turn a blind eye. People who don't punish an innocent person might also be deemed fair but may be judged as marginally less trustworthy compared with fair punishers (who incur a cost to execute the punishment, thus sending a comparatively stronger signal). To make it easier to present the hypotheses and subsequent results, I have provided each punisher/observer with a unique label according to the Deservedness - Severity condition each represents. The punisher in the Deserved-High Severity condition is labeled as the Fair Vigilante. The punisher in the Undeserved-High Severity condition is labeled as the Sadist. The observer in the Deserved-No Severity condition is labeled as the Forgiver. And the observer in the Undeserved-No Severity condition is labeled as the Fair Watcher.

Therefore, I hypothesize that:

H1: Fair Vigilantes will be perceived as more trustworthy than Sadists.

H2: Fair Watchers will be perceived as more trustworthy than Forgivers.

H3: Fair Vigilantes will be perceived as more trustworthy than Forgivers.

H4: Fair Watchers will be perceived as more trustworthy than Sadists.

H5: Participants will prefer to play another TG with Fair Vigilantes over all other alternatives.

To summarize, I expect a main effect of deservedness and severity, with more deserved and more severe punishers to be perceived as more trustworthy. I also expect an interaction, where the trustworthiness signaled via high severity punishing will depend on the deservedness of the punishment, and the trustworthiness signaled via deserved punishing will depend on the severity of the punishment.

5.1.2. Materials and procedure[3]

Participants first read the instructions for the Third-Party Punishment (TPP) game as well as the Trust Game (TG). Following the TPP instructions participants had to answer four comprehension questions designed to ensure they understood the logic and mechanics of the game. If a participant selected an incorrect answer, she/he was thusly notified, given a clue to help them arrive at the correct answer and asked the same question again. Only after correctly answering a given question could participants proceed to the next comprehension question. Likewise, the Trust Game instructions were followed by the same type and number of comprehension questions, but aimed at ensuring comprehension of the TG.

Both the TPP and the TG comprehension questions were exactly the same ones used by Jordan and colleagues (2016)[4], but the particular way in which they were deployed to ensure comprehension (i.e. only correct answer to proceed and clue provided after incorrect answer) was

[3] Study's procedure and analysis plan pre-registered at OSF: https://osf.io/82v5n/
[4] I'm very grateful to Jillian Jordan for kindly sharing with me the original materials used in her study

specific to the current study given the complex nature of the tasks. This deployment method was chosen because it provided pedagogical and practical advantages. Namely, participants who incorrectly answered a question received feedback on it (while with the method used by Jordan and colleagues (2016) a participant that answered incorrectly would have reasonably believed they had provided a correct response, thus compromising their overall comprehension) and were provided with a clue to help them better understand the mechanics of the games. Using this method also meant participants who answered all 8 questions had answered them correctly, thus eliminating the need to discard participant responses to the main task based on a potentially arbitrary number of incorrect comprehension questions.

The first part of this experiment had a 2X2 design with 2 levels for the deservedness factor (deserved and not deserved) and two levels for the severity factor (no punishment and high severity punishment). In the deserved condition, participants read about an observer who punished a sender that sent no money (0% of his endowment) to the recipient. In the undeserved conditions the participants read about an observer who punished a sender that sent 50% of his endowment to the recipient (i.e. 10 of his/her 20 pence).

For each deservedness condition there were also two levels of punishment. Participants read about an observer who either decided not to punish (i.e. spent no money to punish and hence didn't punish the sender) or exacted a harsh punishment (spent 5 of his 15 pence endowment to reduce 14 of the sender's 20 pence). Participants were randomly assigned to one of the four treatment groups and then played the TG with the punisher they had just read about (that punisher now playing as receiver and the participants as senders). Each condition included an animation which helped to illustrate the actions of all 3 players in the Third Party-Punishment game.

The second part of this experiment consisted of the explicit partner choice task. Following their decision in the Trust Game, participants were informed that in other first games, other sets of 3 different players had made different choices, corresponding to the four possible deservedness-severity combinations[5]. Participants were then asked to imagine they had to play a second Trust Game as they had done before. But, unlike the first Trust Game, participants were explicitly asked who of the 4 TPP observers they had been presented with they would rather have as receivers in this second Trust Game (participants chose just one of the 4 observers and didn't play the Trust Game).

5.1.3. Results

A total of 500 participants fluent in English were recruited via Prolific.co (sample size justification can be found in Appendix A). Participants took an average of 12.21 (+- 5.66) minutes to complete the whole task (including the explicit partner preference choice question). 8 participants who took less than 4 minutes, and 4 who took more than 45 minutes were excluded from the analysis, leaving a total sample size of 488 participants (55% female) with a mean age of 31.39 (+-10.26) years.

To measure participants' level of trust, the amount of money entrusted to receivers (previously observers or punishers in the TPP) was evaluated by condition. Table 1 reports the mean amounts entrusted by participants for each of the four deservedness-severity conditions.

[5] Just as with the first part of the experiment, each of the four combinations included an animation to illustrate the actions of all three players

Table 1

Amount of Money Entrusted by Condition

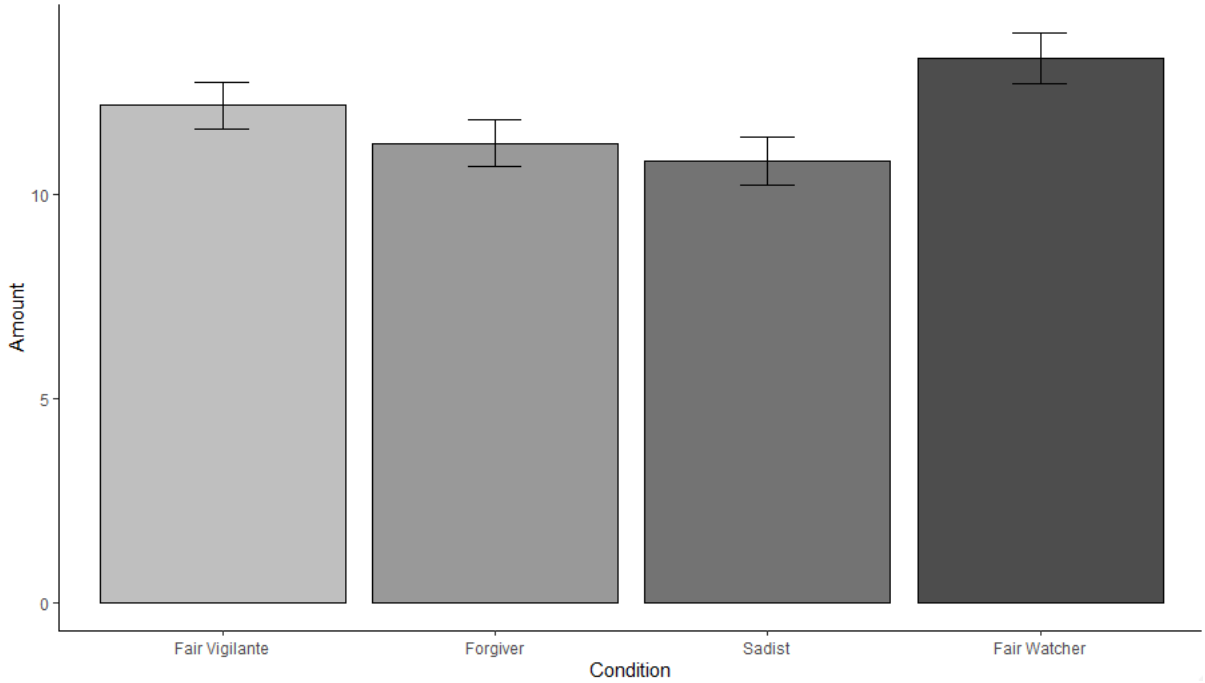
Condition	Mean	Median	S.D.	N
Fair Vigilante	12.16	10	6.28	125
Forgiver	11.24	10	6.46	124
Sadist	10.80	10	7.05	139
Fair Watcher	13.32	10	6.15	100

Note. Amounts in pence (hundreds of 1.00 £). S.D. = Standard Deviation

Levene's test confirmed homogeneity of variance, ($F(3,484) = 0.75$, $p = 0.52$) and the corresponding 2 (Deservedness: Deserved, and Undeserved) X 2 (Severity: High Severity and No Severity) ANOVA revealed a significant main effect of Condition on amount of money entrusted, $F(3,484) = 3.23$, $p = 0.02$, $\eta 2 = 0.02$. To test hypotheses H1, H2, H3 and H4 I conducted a post hoc contrast test which indicated that there was only a significant difference in the mean amount entrusted between the Sadist and the Fair Watcher $t(484) = 2.95$, $p = 0.02$, $d = 0.38$ (see Figure 1).

Figure 1

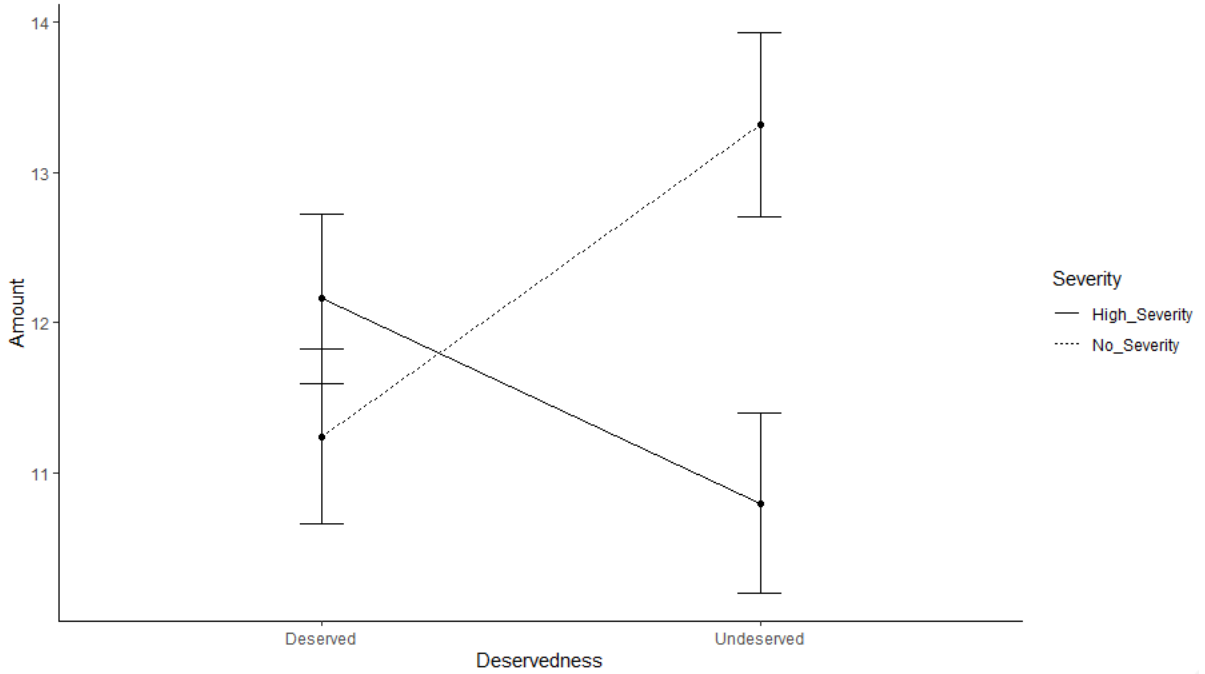
Mean Amounts Entrusted by Condition



Note. Error bars represent standard errors

Figure 2

Interaction Between Deservedness and Severity



Note. Error bars represent standard errors

Following the Trust Game, participants took part in the explicit partner preference choice task. Partner preference was evaluated observing the frequency with which a given partner was chosen from the available four alternatives (corresponding to the four possible Deservedness-Severity combinations). Table 2 summarizes the number of times each partner was chosen by condition.

Table 2

Number of times each player was chosen

Condition	Label	Number of times chosen	Percentage Point Difference
Deserved-High Severity	Fair Vigilante	125	0.61
Deserved-No Severity	Forgiver	99	-4.71
Undeserved-High Severity	Sadist	31	-18.65
Undeserved-No Severity	Fair Watcher	233	22.75

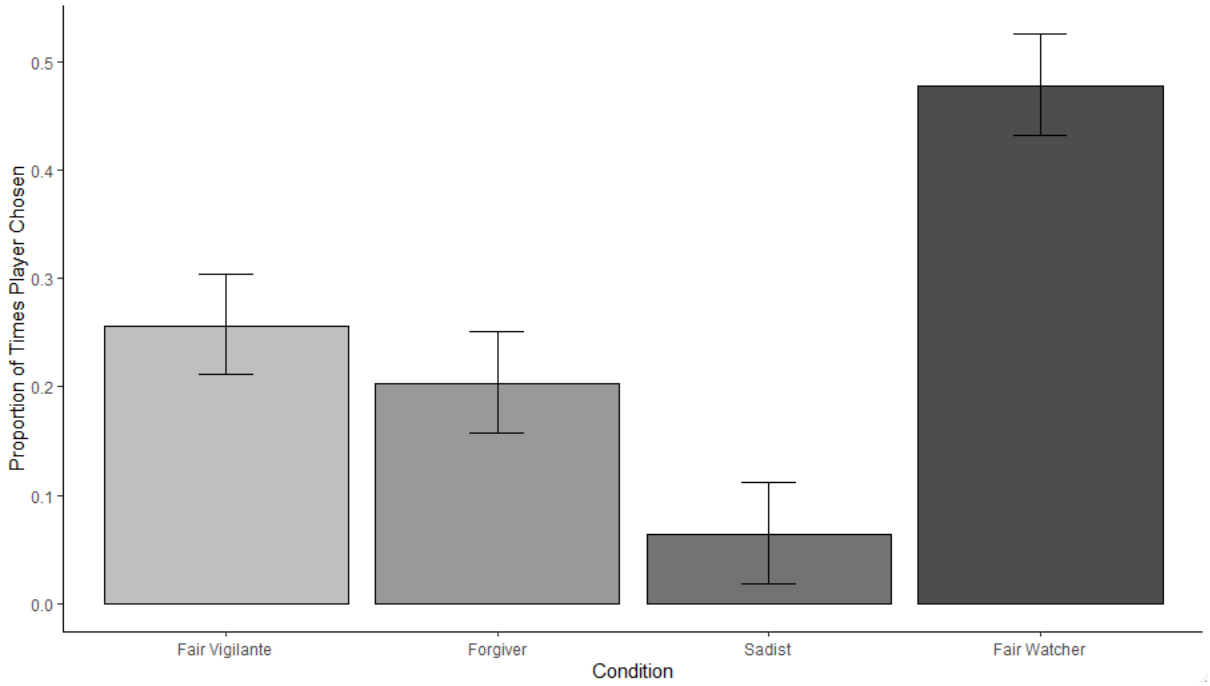
Note. Percentage point difference based on an expected equal split between the four alternatives

A chi-square test of independence was performed to examine the effect of Condition on the number of times a player in this second Third-Party Punishment game was preferred as a subsequent Trust Game partner, revealing a significant relation between these variables, X^2 (3, N = 488) = 173.28, p <0.0001. To test hypothesis H5, I conducted post hoc pairwise comparisons of explicit partner preference by Deserved-Severity conditions, which revealed significant differences between all conditions except between the Fair Vigilante and the Forgiver (that is, between both of the Deserved Conditions) p = 0.08 (see Table A1 in Appendix A). Moreover, an

odds ratio analysis revealed the Fair Watcher (Undeserved - No Severity) was much more likely to be chosen than all the other potential partners, *OR* = 9.44 (95% CI: 5.88 – 15.5), *p* <0.0001, while the Sadist (Undeserved – High Severity) was much less likely to be chosen, *OR* = 0.11 (95% CI: 0.06 – 0.17), *p* =<0.0001 (see Figure 3)(see Table A2 in Appendix A for full OR table). A multinomial model aimed at evaluating how assigned condition on the first task influenced likelihood to choose a given partner in the second task pointed to an overall increased likelihood to choose the Fair Watcher if a participant had been assigned to the Deserved No Severity or Undeserved High Severity conditions, and a decreased likelihood to choose the Sadist and the Forgiver if assigned to the Deserved High Severity condition (see Table A3 in Appendix A).

Figure 3

Proportion of Times a Player was Chosen



Note. Error bars represent confidence intervals

5.1.4. Discussion

The results of experiment 1A suggest that the deservedness and severity of a Third-Party punishment influence the degree of trust placed on the punisher (or observer in the case of no

severity punishments) measured in terms of (a) the money entrusted to them in a subsequent Trust Game paradigm, and (b) in the number of times they were chosen as preferred partners (i.e. as receivers) for a second Trust Game. Together, the results of both tasks of the experiment provide support for three out of the five hypotheses postulated earlier (i.e. H1, H2 and H4) and point to three clear tendencies. Firstly, a tendency to equally trust Deserved-High severity punishers and Deserved-No severity observers. Secondly, a tendency to distrust Undeserved-High severity punishers. And thirdly, a tendency to place an overall greater trust on Undeserved-No severity observers.

The lack of a difference between both of the deserved conditions (which is evidence against hypothesis H3) could be understood as a reflection, on the one hand, of the trust placed on those who fulfill their duties and do what they must to maintain the social order (i.e. the Deserved-High Severity condition), and on the other hand, of the trust placed on those who exhibit a forgiving quality (i.e. the Deserved-No Severity condition) in the face of an arguably not very serious offense (i.e. not sharing half of your endowed money with a complete stranger). Even though the characteristics signaled by punishers / observers of both deserved conditions are qualitatively very different (probably even diametrically opposed) they can both be construed as socially desirable, rendering their protagonists worthy of others' trust.

Secondly, the fact that people tend to distrust Undeserved-High severity punishers (i.e. Sadists) is not surprising, given that someone who severely punishes an innocent and arguably altruistic person (she/he did share half of her/his endowment with a complete stranger) essentially exhibits antisocial personality characteristics bordering on psychopathy. Previous research has found that such extreme behavior is particularly revealing of bad moral character and influences moral judgements (Tannenbaum, Uhlmann, & Diermeier, 2011). It is reasonable to expect that

such a person would behave in an antisocial manner again, hence leading people to distrust such a punisher. This evidence provides supports to hypotheses H1 and H4.

However, the reason people tend to place greater trust in the Undeserved-No severity observer is not as self-evident (evidence against hypothesis H5). The actions of this Fair Watcher do not constitute a strong signal, but it could be argued that his inaction sends a clearer signal that enables him to reap substantial reputational benefits. In particular, the mere fact that they do nothing puts them at an advantage over the other punishers/observers in the TPP. Compared with the Fair Vigilante, he is not as strict. Compared with the Forgiver, he does not turn a blind eye. And compared with the Sadist, he is more just. All of these potential person perceptions stem from the fact that the Fair Watcher's signaling is founded in inaction. Moreover, it could be argued that this inaction is the expected default in the case of undeserved punishment, as opposed to the Forgiver who signals a very specific quality via his inaction, but nonetheless a quality that might not be perceived positively by a segment of the population (i.e. forgive a guilty person). This introduces an omission bias that could be driving the overall greater trust placed in the Fair Watcher. To control for this potential inaction person perception bias, experiment 1B was conducted.

5.2. Experiment 1B

In experiment 1B I sought to replicate experiment 1A, while controlling for the potential inaction quality of the No Severity conditions (including the Fair Watcher and Forgiver conditions in the explicit partner preference task). To that end, experiment 1B followed exactly the same mechanics and content as experiment 1A, except for the fact that in the Third-Party Punishment game, if an observer chose not to punish they had to pay a tax of the exact same value as the cost

to punish. In particular, participants were informed that if an observer chose not to punish they would have to pay a tax of 5 pence, and in both No Severity conditions, participants read that the observer had been effectively charged the 5 pence tax. The same was true of the two No Severity conditions in the explicit partner preference task (i.e. both the Fair Watcher and the Forgiver paid the tax). By introducing an amount that had to be paid in the case of exerting no punishment, the inactions of the Forgiver and the Fair Watcher were eliminated because those two players had to effectively decide between paying to punish or paying to not punish.[6]

5.2.1. Results

A total of 600 participants fluent in English were recruited via Prolific.co (sample size justification can be found in Appendix A). Participants took an average of 13.16 (+- 6.19) minutes to complete the whole task (including the explicit partner preference choice question). 13 participants who took less than 4 minutes, and 2 who took more than 45 minutes were excluded from the analysis, leaving a total sample size of 585 participants (56% female) with a mean age of 29.95 (+-11.3) years.

As in experiment 1A, the amount of money entrusted to receivers (previously observers or punishers in the TPP) was evaluated by condition as a proxy of trust. Table 3 reports the mean amounts entrusted by participants for each of the four deservedness-severity conditions.

Table 3

Amount of Money Entrusted by Condition

Condition	Mean	Median	S.D.	N
Fair Vigilante	12.66	11.5	6.77	146

[6] Study's procedure and analysis plan pre-registered at OSF: https://osf.io/w6f32/

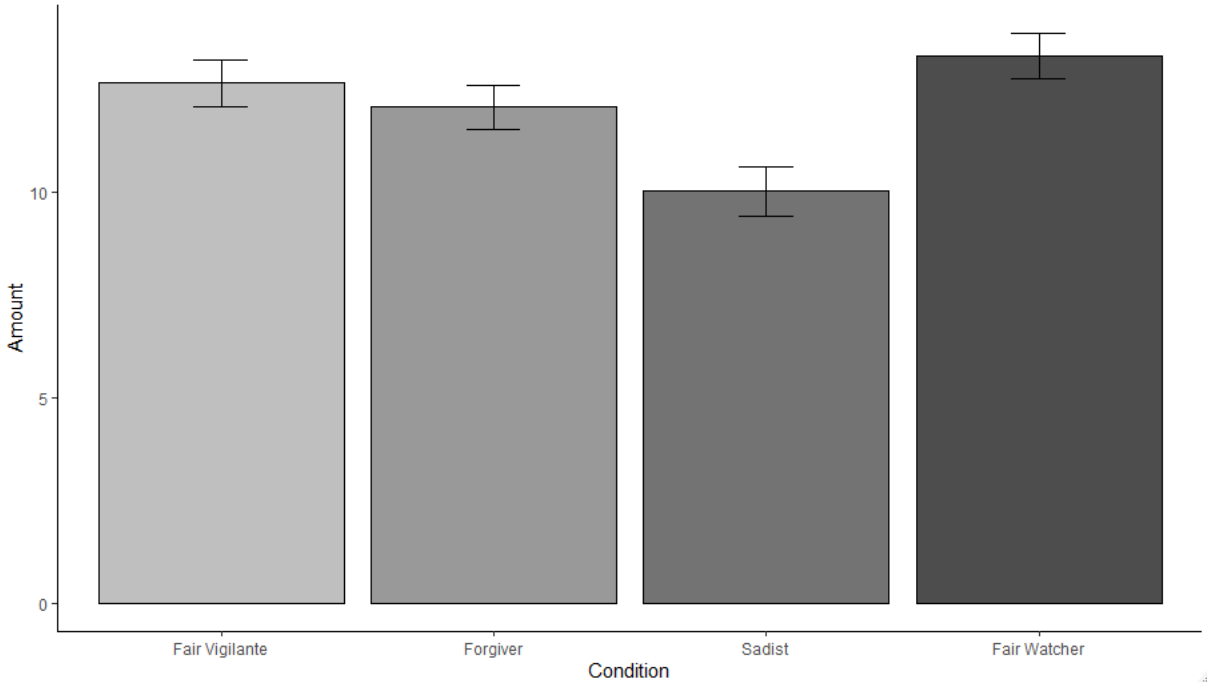
Forgiver	12.08	10	6.77	159
Sadist	10.03	10	6.86	133
Fair Watcher	13.31	10	6.78	147

Note. Amounts in pence (hundreds of 1.00 £). S.D. = Standard Deviation

Levene's test confirmed homogeneity of variance, ($F(3,581) = 0.87$, $p = 0.46$) and the corresponding 2 (Deservedness: Deserved, and Undeserved) X 2 (Severity: High Severity and No Severity) ANOVA revealed a significant main effect of Condition on amount of money entrusted, $F(3,585) = 3.23$, $p < 0.0001$, $\eta 2 = 0.03$. As with experiment 1A, a post hoc contrast test aimed at evaluating hypotheses H1, H2, H3 and H4 indicated significant differences in the money entrusted between the Sadist and Fair Watcher conditions $t(585) = 4.04$, $p < 0.0001$, d = 0.48, but unlike experiment 1A, it also found a significant difference between the Fair Vigilante and Sadist conditions $t(585) = -3.23$, $p = 0.007$, d = 0.39 (see Figure 4).

Figure 4

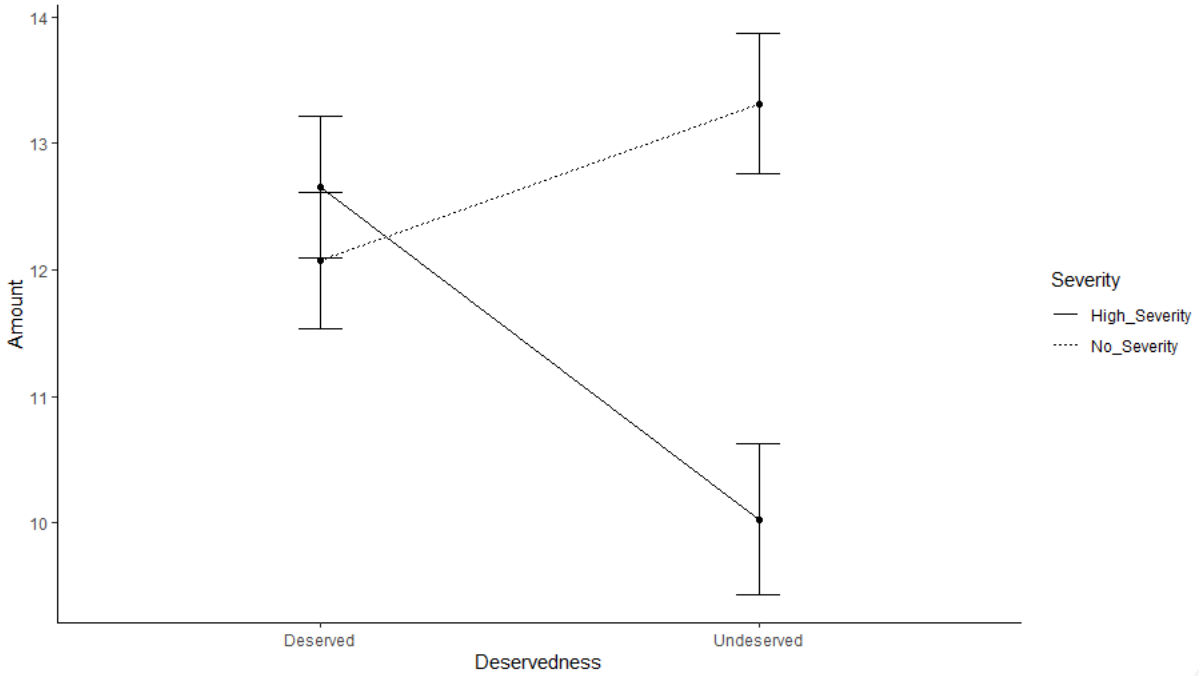
Mean Amounts by Condition



Note. Error bars represent standard errors

Figure 5

Interaction Between Deservedness and Severity



Note. Error bars represent standard errors

As with experiment 1A, participants took part in the explicit partner preference choice task after the Trust Game and partner preference was equally evaluated observing the frequency with which a given partner was chosen from the available four alternatives. Table 4 summarizes the number of times each partner was chosen by condition:

Table 4

Number of times each player was chosen

Condition	Label	Number of times chosen	Percentage Point Difference
Deserved-High Severity	Fair Vigilante	149	0.47
Deserved-No Severity	Forgiver	145	-0.21
Undeserved-High Severity	Sadist	31	-19.70
Undeserved-No Severity	Fair Watcher	260	19.44

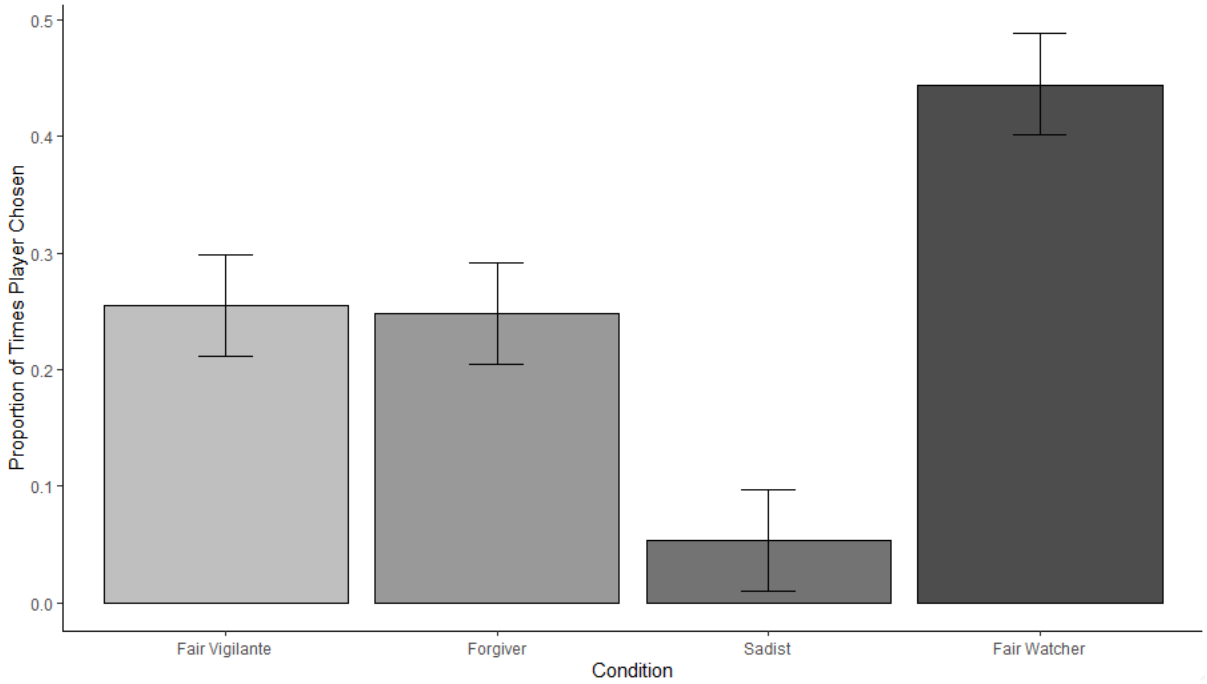
Note. Percentage point difference based on an expected equal split between the four alternatives

A chi-square test of independence was performed to examine the effect of Condition on the number of times a player in the Third-Party Punishment game was preferred as a subsequent Trust Game partner, revealing a significant relation between these variables, X^2 (3, N = 585) = 179.36, p <0.0001. Replicating the results of experiment 1A, post hoc pairwise comparisons of explicit partner preference by Deserved-Severity conditions, aimed at testing hypothesis H5, revealed significant differences between all conditions except between the Fair Vigilante and the Forgiver (that is, between both of the Deserved Conditions) p = 0.82 (see Table A4 in Appendix A). And in similar fashion to the results of experiment 1A, the Fair Watcher (Undeserved - No Severity) was much more likely to be chosen than all the other potential partners, *OR* = 8.58 (95% CI: 5.48

– 13.79), p =<0.0001, while the Sadist (Undeserved – High Severity) was much less likely to be chosen, *OR* = 0.12 (95% CI: 0.07 – 0.18), p =<0.0001 (see Figure 6)) (see Table A5 in Appendix A for full OR table). A multinomial model aimed at evaluating how assigned condition on the first task influenced likelihood to choose a given partner in the second task pointed to an overall increased likelihood to choose the Fair Watcher if a participant had been assigned to the Deserved High Severity condition, and a decreased likelihood to choose the Sadist if assigned to the Deserved High Severity or Deserved No Severity conditions (see Table A6 in Appendix A).

Figure 6

Proportion of Times a Player was Chosen



Note. Error bars represent confidence intervals

5.2.2. Discussion

The results of experiment 1B confirm the main results of experiment 1A and point to the same three main tendencies. Namely, a tendency to (a) equally trust Deserved-High severity punishers and Deserved-No severity observers, to (b) distrust Undeserved-High severity punishers,

and (c) to place an overall greater trust on Undeserved-No severity observers. However, it also found a tendency for participants to trust Deserved-High Severity punishers (Fair Vigilantes) significantly more than Undeserved-No Severity ones (Sadists). This last finding does not contradict the results of experiment 1A; it simply bears statistical significance on this observed tendency in experiment 1A (in the same direction), thus evidencing the fact that people tend to trust just punishers over unjust ones, which should not be surprising.

More importantly, the results of experiment 1B replicate the results of 1A while controlling for the possible inaction driving the trust placed on No-Severity observers. The fact that participants in experiment 1B tended to send larger average amounts of money to Undeserved-No severity observers in the Trust Game, and preferred Fair Watchers as potential partners above all other punishers/observers even when they actively decided (and paid a price) to not punish, indicates that people place greater trust in observers who make an active choice not to punish an innocent person. This suggests that people indeed trust just over unjust individuals, but given the choice will place greater trust on individuals who justly (and actively) avoid punishment over those who justly exert punishment.

This preference might be explained if we take into account the particular characteristics of the offense in the case of a deserved punishment and the limited window to signal personal traits allowed by the Third-Party Punishment game (TPP). A deserved punishment within the TPP game paradigm employed involves not sharing half of your money (money that you have been endowed with) with a complete stranger. No observer is given any additional information regarding the financial situation of the sender or receiver in the TPP, nor of their previous behavior in other games (or any other contexts for that matter). Given such an information vacuum, some participants might question whether the sender in the TPP can afford to send half of his money, or

whether the receiver actually deserves it. In such a case and without access to further information on which to base their person perceptions, participants might question the deservedness of the punishment and therefore not give the Fair Vigilante the benefit of the doubt, hence favoring the Fair Watcher. Likewise, the observed preference is related to omission bias, which refers to the preference for harm caused by omissions over equal or lesser harm caused by acts (Baron & Ritov, 2004). Extant research has found that people tend to rate harmful omission as less immoral than harmful acts (Haidt & Baron, 1996; Spranca, Minsk, & Baron, 1991), as well as a tendency to favor indirectly harmful options over directly harmful alternatives (Royzman & Baron, 2002).

While the binary nature of the severity employed in experiments 1A and 1B (punish with high severity or not punish at all) was very useful to tease out the broad main effect of this component of punishment on perceived trust, it pigeonholes punishers into harsh-punisher and not-punisher categories that might impede a more fine-grained analysis. To circumvent these obstacles and gain a better perspective on how severity affects perceived trust, experiment 2 was designed to center around varying degrees of severity for exclusively deserved punishment.

5.3. Experiment 2

In experiments 1A and 1B, I looked at the effect of deservedness and severity on the signaled trustworthiness as measured in the subsequent TG. However, a large majority of punishments in social interactions carry at least some level of deservedness. It is only in extreme cases that wholly underserved punishments (that is, where everyone except the punisher deems there is no transgression that justifies a punishment of any kind) are promoted and administered (Bedau & Kelly, 2017). Conversely, there appears to be a great deal of variability regarding the adequate severity of a punishment for a given transgression. It varies across individuals, cultures

and time periods (Bedau & Kelly, 2017). As a matter of fact, the issue of severity weighs heavily on the arguments for and against deterrence, retribution, restitution, rehabilitation and even the expressive quality of punishment (e.g. Carlsmith et al., 2002; Feinberg, 1965; Kahan, 1996). Therefore, the object of experiment 2 is to focus only on how the trustworthiness signal varies across different levels of severity.

5.3.1 Materials and procedure[7]

The general logic and mechanics of experiment 2 were the same as those employed in experiments 1A and 1B. Namely, participants read about a Third-Party Punishment game and then participated in a Trust Game. However, in experiment 2 all TPP results presented to participants were deserved (i.e. in all conditions the punisher had punished because Player 1 had behaved selfishly). Participants first read four scenarios of TPP where the sender decided to keep all the money (i.e. shared 0 pence with the recipient) and the corresponding punishment exacted by the observer (in four different levels of severity). Participants then played four different TGs as senders with the observers in the previous TPP as receivers. Therefore, experiment 2 gained statistical power by turning the procedure into a within-subjects design. Participants were endowed with 20 pence and had to decide how much (if any) of their money to send to the receiver (dependent variable). The amount of money sent served as a proxy of the level of trust placed on the receiver as a function of the punishment's severity in the TPP game.

Experiment 2 evaluated trustworthiness dependent on 4 different levels of punishment severity. Specifically, participants played four sequential trust games with fictional recipients that allegedly played the Third-Party Punishment game as punishers (i.e. recipients and their behavior

[7] Study's procedure and analysis plan pre-registered at OSF: https://osf.io/5h926

in the TPP were manipulated by the experimenter just like in experiments 1A and 1B). Each trust game was preceded by a TPP that participants read about, each with a different level of punishment severity exerted by the then recipient in the TG. The order in which each TPP and corresponding TG were presented was randomized across participants to avoid order effects.

In the low severity condition, participants read about a punisher who in the TPP game paid 3 pence to reduce the sender's amount by 5 pence. In the medium-low severity condition, participants read about a punisher who in the TPP game paid 7 pence to reduce the sender's amount by 10 pence. In the medium-high severity condition, participants read about a punisher who in the TPP game paid 11 pence to reduce the sender's amount by 15 pence. In the high severity condition, participants read about a punisher who in the TPP game paid all her / his 15 pence to completely reduce the sender's amount (i.e. by 20 pence).

Materials were modelled after those used by Jordan et al. (2016). All conditions and a sample of the illustration presented to participants (medium-high severity) can be found in Appendix C.

5.3.2. Hypotheses

Given that a higher punishment implies a higher incurred cost for the signaler, and that decisions that involve costs for the decision maker are perceived as especially informative about character (Ohtsubo & Watanabe, 2009), I expect higher severity to signal greater trustworthiness. Accordingly, I hypothesize that:

H1: Punishers in the high severity condition will be trusted more (i.e. sent more money in the TG) than punishers described in the medium-high, medium-low and low severity

conditions respectively. This relationship would be described by a positive linear function between severity and trustworthiness.

Nonetheless, given the novel nature of the study and the relative lack of research that integrates severity in the study of punishment I consider alternative hypotheses. Firstly, it might be the case that severe punishing signals moralization. Moralization refers to the perception that someone sees an issue as morally relevant (Kreps & Monin, 2014). The expressivist theory conceives of punishment as a way to communicate moral condemnation (Kahan, 1996). Speakers perceived to moralize an issue are treated differently from those whose positions seem merely pragmatic (Kreps & Monin, 2014). In the context of the current study, a pragmatic position could be to keep most of the money one has been endowed with and spend little of it punishing another individual. According to Kreps and Monin (2014), moralizers can appear inflexible or self-righteous, hence punishing (as opposed to not punishing) could arguably have a detrimental effect on perceived trustworthiness. This would mean that:

H2: Severity, as a proxy of the cost incurred to punish, would signal trustworthiness linearly only up to a certain point. Beyond that point, more severe punishments (and thus a higher incurred cost) would show an inverse relationship to trustworthiness as the punishment starts to signal moralization instead of trustworthiness; very severe punishers would be perceived as inflexible and self-righteous instead of trustworthy. Such a relationship would imply an inverse U-shaped function between severity and trustworthiness.

It could also be the case that the relationship between severity and trait inference only exists after a given threshold. It is possible that more severe punishments do not lead to stronger trait inferences at any point, but instead depend on a threshold detection of a moral norm violation that

deserves to be punished, which in turn triggers perceptions of certain moral traits on punishers. Encountering moral norm violations in person is actually not very common: between 5% and 10% (Wilhelm Hofmann et al., 2018). The fact that people apparently do not encounter moral norm violations on a regular basis in their daily life could be interpreted as indicating that most people don't deem the vast majority of norm violations as blameworthy. This suggests that whenever people identify an action as violating a moral norm, they immediately conceptualize it as deserving of punishment. In turn, any severity increases beyond that point would not lead to stronger trait inferences. This would mean that:

> **H3**: Punishing signals trustworthiness equally regardless of the severity of the punishment inflicted, as long as people deem the punishment as deserved. In other words, lenient punishers would be trusted just as much as heavy-handed ones when the punishment is perceived as deserved.

5.3.3. Results

A total of 401 participants fluent in English were recruited via Prolific.co (sample size justification can be found in Appendix B). Participants took an average of 11.43 (+- 5.94) minutes to complete the whole task. 4 participants who took less than 4 minutes, and 2 who took more than 45 minutes were excluded from the analysis, leaving a total sample size of 395 participants (50% female) with a mean age of 30.07 (+-10.54) years.

To measure participants' level of trust, the amount of money entrusted to receivers (previously observers or punishers in the TPP) was evaluated by condition within subjects. Table 5 reports the mean amounts entrusted by participants for each of the four deservedness-severity conditions.

Table 5

Amount of Money Entrusted by Condition

Condition	Mean	Median	S.D.	N
Low Severity	9.94	10	6.18	395
Medium Low Severity	10.97	10	5.93	395
Medium High Severity	11.95	11	6.22	395
High Severity	12.38	11	6.90	395

Note. Amounts in pence (hundreds of 1.00 £). S.D. = Standard Deviation

A linear mixed model with a random intercept for each participant was fitted to predict the amount entrusted as a function of the condition, using the lmer library (Bates, Mächler, Bolker, & Walker, 2015). A type II Wald F test of the model indicated a significant main effect of condition on entrusted amount, $F(3,395) = 28.42$, $p < 0.0001$, $\eta_p 2 = 0.067$ (see Table B1 in Appendix B). To test hypotheses H1, H2 and H3, I conducted post hoc pairwise comparisons, which found significant differences between all but the Medium High and High severity conditions (see Table 6).

Table 6

Pairwise Comparisons of Severity Conditions

Severity Condition	Mean Difference	p-value	CI lower bound	CI upper bound
High – Medium High	0.43	0.48	-0.31	1.18
High – Medium Low	1.41	<0.0001	0.67	2.15
High – Low	2.44	<0.0001	1.70	3.19
Medium High – Medium Low	0.98	0.004	0.24	1.72
Medium High – Low	2.01	<0.0001	1.27	2.75

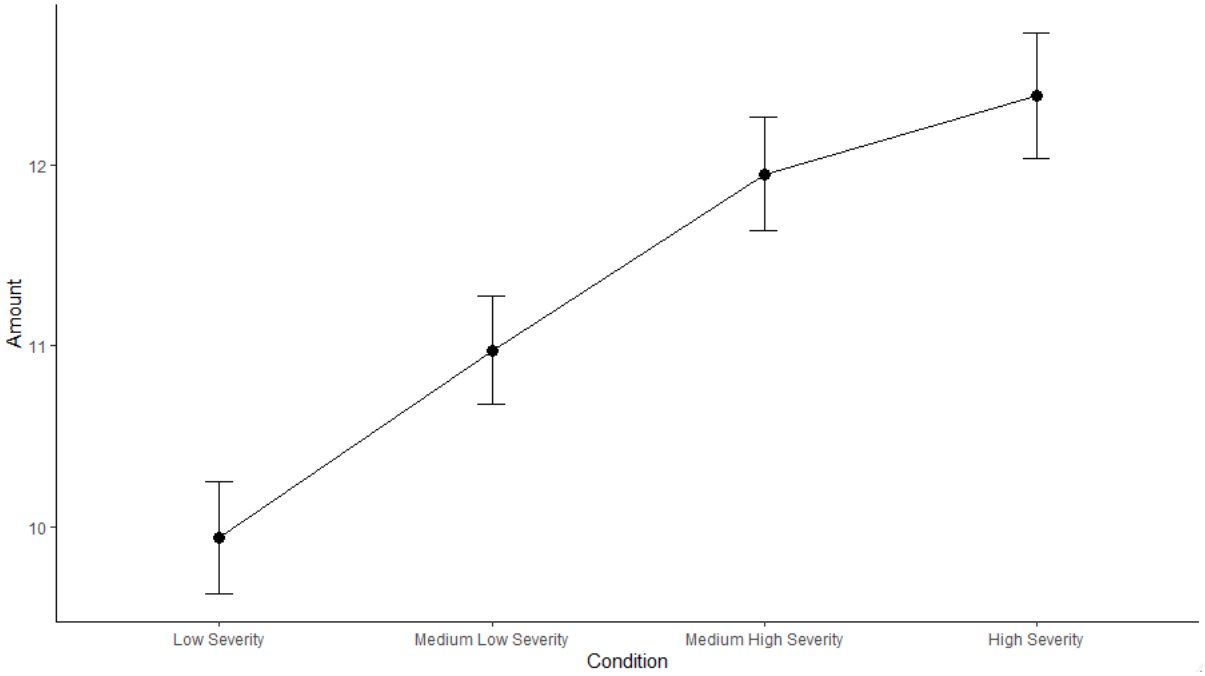
Medium Low – Low	1.03	0.002	0.29	1.78

Note. Mean Difference Amounts in pence (hundreds of 1.00 £). CI = Confidence interval

In general, as the severity of the punishment increased, the amount of money entrusted to the punisher also increased but with the aforementioned ceiling effect once Medium-High severity was reached (see Figure 7). Controlling for demographic variables did not result in any change in the pattern observed (see Table B2 in Appendix B).

Figure 7

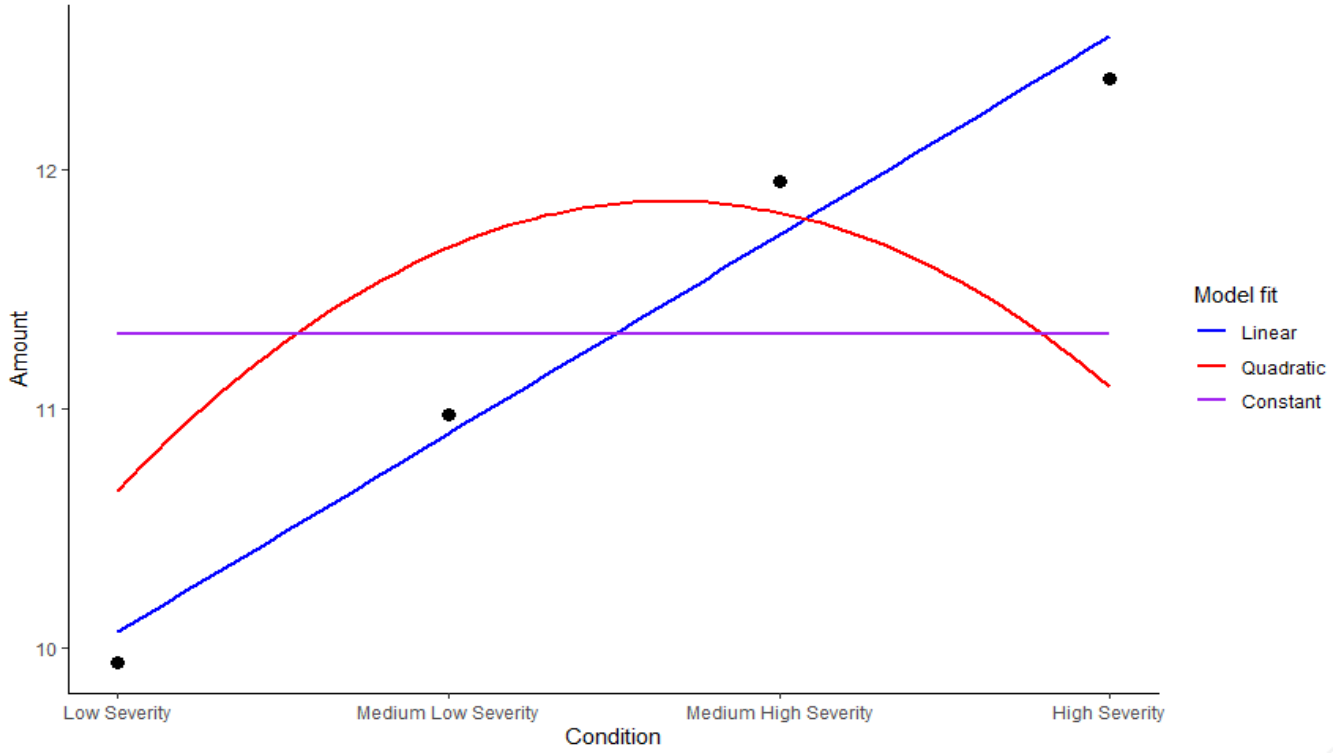
Mean Amounts Entrusted by Condition



Note. Error bars represent standard errors

Figure 8

Model Fit for Mean Entrusted Amounts by Condition



5.3.4. Discussion

The results of experiment 2 suggest that severity has a direct influence over the perceived trustworthiness of a given punisher (at least within the Third-Party Punishment and Trust Game paradigms employed). More specifically, the results point to a significant increase in the trust placed in a punisher as the severity increased from low to high. Of particular note was the substantial difference in entrusted money between the extremes (low and high severity) of 2.44 pence. This overall trend supports hypothesis H1, whereby greater severity is a proxy of greater incurred cost and therefore a more informative signal of character traits (trustworthiness in this case). In other words, lenient punishers (i.e. low and medium low severity conditions) either send

a weaker signal that they care about selfish behavior (compared with medium high and high severity conditions), or the lenient response could have actually been interpreted by participants as a clear signal that they partially condone selfish behavior (and are therefore not to be trusted).

Nevertheless, the lack of a statistically significant difference in mean entrusted money between the uppermost levels of severity (i.e. medium high and high) provides nuance to this interpretation as the relationship is not perfectly positive and linear. Even though the results found no evidence to support hypothesis H3, the theoretical principles behind this hypothesis could help us to understand the observed behavior. Namely, the fact that high severity punishers were not deemed more trustworthy than medium high punishers could be due to a threshold effect of appropriate severity reached at the medium high severity level. In other words, participants could have interpreted a medium high severity as the appropriate level of punishment given the offense described. In particular, in the medium high severity condition, the punisher spent 73% of her endowment to cause a loss of 75% of the sender's money (because the sender had refused to share 50% of his money with another person). And any additional cost incurred to punish more severely (i.e. to cause a 100% loss of money) did not add strength to the signaled trustworthiness.

The results of the experiment found no evidence for hypothesis H3, as there was not an observed inverse relationship between severity and perceived trustworthiness (measured in terms of the amounts of money entrusted). However, this does not mean that a shift from signaling trustworthiness to signaling moralization could not happen at some point once a high enough severity level is reached. The characteristics of the experimental paradigm used in experiments 1A, 1B and 2 make this difficult to ascertain given that the transgressions and punishments are all monetary and limited to the amounts endowed (with a maximum total loss of 20 pence for the receiver and 15 pence for the punisher). Nonetheless, one could easily imagine instances in

everyday life where the severity of the punishment far exceeds the perceived magnitude of the transgression. Such heavy-handed punishers would be perceived as moralizing, overly strict and maybe even self-righteous, thus leading to decreased trustworthiness as severity increases.

The aim of experiment 3 is to examine these sorts of situations and explore the effect of severity on perceived trustworthiness in addition to other character traits from a more naturalistic perspective. To that end, experiment 3 will make use of different experimental methods that transcend the economic games employed thus far.

5.4. Experiment 3

In the previous experiments the focus was on how moralistic punishment signals trustworthiness. As mentioned before however, the literature suggests that in addition to trustworthiness, third-party punishment can also be used to signal dominance. As one of the most important, if not the only variable a third-party punisher can manipulate, proportionality has the potential to directly influence the extent to which third-party punishment signals trustworthiness and / or dominance. Moreover, according to the definition I put forth, moralistic punishment signals the moral norms that the punisher cares about. Proportionality can also determine the extent to which observers perceive the punisher as motivated by genuine moral concern or personal interest (i.e. how much she really cares about the underlying moral norm that was violated). Likewise, proportionality can also affect whether a third-party punisher is considered for a leadership or a friendship role. In experiment 3, I use participants' judgments of vignettes to complement the economic game paradigms used in experiments 1A, 1B and 2. This method enables analysis of third-party punisher perceptions in more familiar everyday contexts and allows

for signaling of traits beyond trustworthiness. In addition, it provides convergent evidence of the effect of proportionality in third-party punisher signaling with a different methodology.

In experiment 3, I examine how proportionality affects three distinct (but possibly related) aspects of social perception. Firstly, experiment 3 aims to find out how proportionality affects perceived warmth versus perceived competence. Because the extant literature focuses on trustworthiness versus dominance signaling in third-party punishment, these two concepts take center stage but are also evaluated within the broader warmth and competence dimensions of the Stereotype Content Model given how closely they track its core underlying concepts (Cuddy et al., 2008). To that end, the individual measures of trustworthiness and dominance are analyzed both in isolation and as part of aggregate indexes of warmth and competence as a function of proportionality. Secondly, experiment 3 evaluates how proportionality affects the perceived motives people attribute to third-party punishers. More specifically, it assesses how proportionality affects attributed motives of genuine moral concern versus motives of self-interest. Finally, it examines how proportionality affects the social role the punisher is considered for. Concretely, it evaluates how proportionality affects whether the punisher is considered for a friendship or a leadership role.

5.4.1 Materials and procedure[8]

In experiment 3, each participant was randomly assigned to read four vignettes. Each vignette depicted a transgression perpetrated on a victim and a corresponding punishment enacted by a third-party not directly affected by the transgression. There were four different transgression scenarios, and each scenario had three possible alternatives corresponding to the three levels of

[8] Study's procedure and analysis plan pre-registered at OSF: https://osf.io/rvnxg/

punishment proportionality. That is, for each scenario there was a disproportionately lenient punishment version, a proportionate punishment version, and a disproportionately severe punishment version, for a total of 12 vignettes given the 4 (scenario) X 3 (level) possible combinations.

After reading each vignette, participants had to answer 22 questions aimed at measuring warmth, competence, attributed motives and potential social role of the punisher. In order to avoid possible demand effects, in experiment 3 each participant was randomly assigned to read just 4 of the 12 possible vignettes. Thus, a given participant could have read the same scenario twice, but never with the same level of punishment proportionality. Sample size was determined taking into account the fact that each participant was not going to be exposed to all 12 combinations. The following is an example of one of the vignettes, adapted from Martin, Jordan, Rand and Cushman's materials (J. W. Martin et al., 2019) for the stutter scenario, proportional punishment level (all the vignettes are in Appendix D):

> "One day before work, Fred stops by a coffee shop close to the office where he works and sees one of his colleagues just finishing posting copies of a cartoon around the shop that makes fun of another man at the company who stutters. Fred's colleague looks angry and like someone you wouldn't want to mess with. Fred tells his colleague, in a loud enough voice so that half the people in the coffee shop can hear, that making fun of someone's stutter and posting an offensive cartoon is not OK and tears the cartoon down."

The vignettes were tested to assess and calibrate the perceived severity of each level within each scenario. All scenario levels tested demonstrated statistically significant severity differences in the expected direction (i.e. disproportionately severe levels as more severe than proportionate

and lenient levels, and proportionate levels as more severe than lenient levels) (see Figure E1 in Appendix E) (severity ratings differences between levels by scenario for all vignettes can be found in Table E1 in appendix E).

As mentioned earlier, after reading each vignette, participants answered questions aimed at measuring warmth versus competence, attributed motives and the potential social role of the punisher. The order of the questions was counterbalanced across participants and randomized within each block. The warmth and competence traits followed some of the most frequently used questions of the Stereotype Content Model (Cuddy et al., 2008) along with some commonly used questions in previous third-party punisher perception research (Barclay, 2006; Gordon et al., 2014; Nelissen, 2008) that equally track the two dimensions. Taking into account the fact that attributed motives were separately analyzed, I purposely left out traits from the Stereotype Content Model that could be interpreted as related to motivations (e.g. “well intentioned”, “good natured” and “sincere”). Traits that might have an ambiguous interpretation in the context of a third-party punishment have also been left out (e.g. “intelligent”, “independent” and “skillful”).

Consequently, participants rated the punisher on a seven-point scale, on the following five traits that map on to the warmth dimension:

a) trustworthiness (1 = extremely untrustworthy, 7 = extremely trustworthy)
b) friendliness (1 = extremely unfriendly, 7 = extremely friendly)
c) warmth (1 = extremely cold, 7 = extremely warm)
d) kindness (1 = extremely unkind, 7 = extremely kind)
e) niceness (1 = extremely mean, 7 = extremely nice)

Likewise, participants rated the punisher on the following five traits that map on to the competence dimension:

a) dominance (1 = extremely nondominant, 7 = extremely dominant)

b) competence (1 = extremely incompetent, 7 = extremely competent)

c) confidence (1 = extremely unconfident, 7 = extremely confident)

d) capability (1 = extremely incapable, 7 = extremely capable)

e) efficacy (1 = extremely ineffective, 7 = extremely effective).

As mentioned earlier both the trust and dominance questions were analyzed in isolation, and as part of their corresponding warmth and competence dimensions. Following Uhlmann, Zhu and Tannenbaum (2013), participants also reported on the perceived underlying motives for the punishment. In order to assess the extent to which participants regarded the punisher as driven by real moral reasons, they reported, on a seven-point scale (1 = definitely not, 7 = definitely yes), whether the third-party punisher acted out of:

a) genuine moral concerns

b) moral principle

c) a genuine moral stand

Similarly, as a measure of the degree to which the punisher was deemed to be driven by self-interest reasons, participants reported, in the same seven-point scale, whether the third-party punisher acted based on:

a) personal self-interest

b) what was good for them personally

c) selfish reasons

Finally, participants reported the extent to which they considered the punisher a potential friend or leader. Concretely, participants reported on a seven-point scale (1 = strongly disagree, 7 = strongly agree), whether the punisher:

a) would make a good leader
b) has the potential to be a good leader
c) shows leadership
d) would make a good friend
e) has the potential to be a good friend
f) shows the qualities of a friend

5.4.2. Hypotheses

5.4.2.1. Warmth and Competence

The results of Experiment 2 provide an evidence-backed foundation from which to generate hypotheses based on signal strength. Namely, the overall positive linear relationship between severity and signaled trust found in experiment 2, suggests that a similar trend might be found between severity and trust in different contexts and could reasonably extend to other signaled traits such as warmth, dominance and competence. Therefore, taking into account the extant research on trustworthiness signaling via TPP (Barclay, 2006; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015) as well as the results of Experiment 2, it is reasonable to expect that proportionate third-party punishers will be trusted more than disproportionately lenient third-party punishers. Likewise, it could be argued that the greater severity of disproportionately severe punishment compared with proportionate punishment renders the trustworthiness signal weaker in the latter case compared to the former. This would lend additional credence to the prediction that disproportionately severe third-party punishers will be rated as more trustworthy than their proportionate and disproportionately lenient counterparts. Considering how critical trustworthiness is to the approach versus avoidance concept at the core of the warmth dimension,

this prediction in trustworthiness between proportionate and disproportionate punishers could also be reasonably extended to warmth in general. This same relative signal strength hypothesis would also lead one to expect proportionate punishers to be perceived as more dominant and competent than disproportionately lenient ones, and disproportionately severe punishers as more dominant and competent than proportionate ones. This hypothesis would be described as:

> **H1**: An overall positive relationship between proportionality and the two dimensions of warmth and competence.

However, research on the trustworthiness perceptions of consequentialist versus deontological decision-makers (Everett et al., 2016) could point in a different direction. According to this research, people generally place greater trust on deontological decision-makers than on consequentialist ones. If participants understand proportionate punishers as punishing a complete stranger for the sake of enforcing a social norm that is beneficial to the majority of people, they could be perceived as pursuing a more consequential goal than disproportionately lenient punishers. By the same token, disproportionately lenient punishers could be perceived as more empathic (forgiving of the transgressor) and aversive to harm, both of which are characteristically deontological qualities (Kahane, Everett, Earp, Farias, & Savulescu, 2015; Uhlmann et al., 2013). Consequently, disproportionately lenient punishers could be trusted more and be perceived as warmer than proportionate punishers. If this type of person perception carries on linearly, proportionate punishers would also be considered as more trustworthy and warmer than disproportionately severe punishers. Disproportionately severe punishment might also be interpreted as an attempt at moralization (Kreps & Monin, 2014), and therefore, disproportionately severe punishers perceived as inflexible or self-righteous, thus leading to comparatively lower

perceptions of warmth. This outlines an overall negative relationship between severity and warmth.

Nonetheless, perceptions of competence would still follow the same trend as indicated by the relative signal strength hypothesis. Namely, disproportionately severe punishers would still be perceived as more competent than proportionate punishers, and these in turn as more competent than disproportionately lenient punishers. These perceptions align with the conceptual notion of competence described in the Stereotype Content Model, whereby competence indicates the capability or capacity of an individual to carry out her intentions (Cuddy et al., 2008). As such, competence is more directly related to the magnitude of the action, and in this case the severity of the punishment is a proxy of the capacity of the individual to execute the corresponding intention. This contrasting hypothesis would be described as:

> **H2**: A negative relationship between severity and warmth and, at the same time, a positive relationship between severity and competence.

A third possibility could be explained by a combination of the signaling strength hypothesis with the consequentialist and moralizing perceptions of disproportionately severe punishment. More specifically, from this perspective one would predict proportionate punishers to be perceived as higher in warmth and competence than disproportionately lenient ones because they provide a stronger signal of both traits. At the same time proportionate punishers would be perceived as lower in competence than disproportionately severe punishers because of the same relative signal strength, but they would be perceived as higher in warmth than disproportionately severe punishers because they are not seen as overtly consequentialist or moralizing. This third possibility is perhaps the most intuitive hypothesis and is described by:

H3: An inverse U relationship between severity and warmth and an overall positive relationship with competence.

However, given the novel nature of the present study and the lack of research on punishment proportionality and person perception, I think it is possible to observe additional variations on the three possibilities described above. More specifically it is possible that:

The relationship between severity and warmth and competence is positive (or negative) only up to a point, after which it is observed to stay relatively the same. This alternative outcome would result in proportionate punishers being grouped with disproportionate ones (lenient or severe) at the same level with regards to warmth or competence. Taking into account the results of experiment 2, the most likely of these outcomes would involve, on the one hand, proportionate and disproportionately severe punishers being perceived as equally higher in trust, warmth, dominance and competence than disproportionately lenient punishers.

5.4.2.2 Self-Interest and Genuine Moral Motives

Given that proportionality is usually tied to signaling cost, proportionate punishers should be attributed more genuine moral motivations, while disproportionately lenient punishers should be attributed more self-interested motivations. In other words, because the lenient third-party is not willing and/or unable to incur a greater cost to punish, participants could interpret leniency as an attempt to signal disingenuously and to opportunistically gain from the situation without truly committing. Likewise, in the case of the disproportionately severe punisher, the greater incurred cost might also be interpreted as a cue of genuine moral motivations. This hypothesis would be described as:

H4: A positive relationship between severity and genuine moral motivations (i.e. as severity increases punishers will be deemed to have acted based on more genuine moral motivations).

However, it could also be the case that participants interpret the excessive cost incurred and the correspondingly disproportionate severity as an overt attempt at moral grandstanding (Tosi & Warmke, 2016). Namely, disproportionately severe third-party punishing might be perceived as motivated mostly by a desire to appear more moral than is really the case. Within costly signaling theory this is similar in nature to excessive displays that imply high costs intended to signal certain character traits for personal benefit (Grafen, 1990; Zahavi, 1975) such as conspicuous consumption and wealth displays (Griskevicius et al., 2007; Hardy & Van Vugt, 2006). From this perspective, disproportionately severe punishers could be perceived as more motivated by status seeking motives (i.e. self-interest) than genuine moral concerns, when compared with proportionate punishers. This hypothesis would imply:

H5: An inverse U-shaped relationship between severity and genuine moral motivations (i.e. as severity increases from disproportionately lenient to proportional, punishers will be deemed to have acted based on more genuine moral motivations, but as severity increases from proportional to disproportionately severe, punisher will be deemed to have acted based on less genuine moral concerns).

On the other hand, the difference, between disproportionately lenient and disproportionately severe punishers is harder to predict. If disproportionately lenient punishment is construed as a form of forgiveness and / or empathy, it could be the case that lenient punishers are perceived as more driven by true moral principles compared to disproportionately severe punishers. However, if disproportionately severe punishers are perceived as particularly

committed to and concerned with the moral norms in questions, it's reasonable to assume that disproportionately severe punishers would be attributed more moral motivations compared to disproportionately lenient punishers.

Nonetheless, in line with the more intuitive prediction of proportionality and the two dimensions of warmth and competence, the more intuitive relationship between proportionality and moral attributions would entail that proportionate punishment is consistently perceived as more driven by moral principle than either disproportionately lenient or severe punishment. More importantly, in line with the evidence that an asymmetry in self interest versus moral principle motivation leads to dissociations in character judgments (Uhlmann et al., 2013), I predict that the relationship between severity and the two dimensions of warmth and competence will be effectively mediated by the type of moral attributions bestowed upon third-party punishers. Namely, it is reasonable to assume that motive attributions of genuine moral concern will lead to greater perceptions of warmth, and motive attributions of self-interest will lead to greater perceptions of competence.

5.4.2.3 Friend and Leader

Finally, with regards to social role, it is reasonable to expect a similar relationship between proportionate and disproportionate punishers dependent on warmth and competence perceptions. Based on the evidence that people prefer non-dominantly looking friends and dominantly looking leaders (Laustsen & Petersen, 2015), one could reasonably expect warmer perceptions to be correlated with potential friend roles, and more competent perceptions with potential leader roles. Accordingly, if an overall positive relationship between severity and the two dimensions of warmth and competence is observed, one would expect:

H6: Disproportionately severe punishers to be considered both more as friends and leaders than proportionate punishers, and these in turn to be considered more as friends and leaders than disproportionately lenient punishers.

Conversely, a negative relationship between severity and warmth but positive between severity and competence, would lead one to expect the same type of social role preference as described above in terms of leaders but not of friends. In particular:

H7: Disproportionately lenient punishers would be considered more as potential friends than proportionate punishers, and these in turn would be considered more as friends than disproportionately severe punishers.

Another possibility entails an overall preference of proportionate punishers both as leaders and friends compared with disproportionate punishers (lenient or severe). This, however, is less likely, especially in light of previous findings on role preference (Laustsen & Petersen, 2015), because it would entail an interaction effect between warmth and competence resulting in proportionate punishers being regarded as better friend candidates that disproportionately lenient punishers, but also better leader candidates than disproportionately severe punishers.

To recapitulate the main hypothesis by dimensions, I expect to see an overall positive (H1), negative (H2) or inverse U-shaped (H3) relationship between severity level and Trust, Warmth, Dominance and Competence. Likewise, I expect to see an overall positive (H4) or inverse U-shaped (H5) relationship between severity level and attributed moral motives. Finally, I expect to see either an overall positive (H6) or negative relationship (H7) between severity level and the two potential social roles (leadership and friendship) of punishers.

5.4.3. Results

A total of 747 participants fluent in English were recruited via Prolific.co (sample size justification can be found in Appendix F). Participants took an average of 11.52 (+- 5.22) minutes to complete the whole task. 7 participants who took less than 4 minutes, and 17 who took more than 35 minutes were excluded from the analysis, leaving a total sample size of 723 participants (54% female) with a mean age of 29.33 (+-10.72) years.

5.4.3.1. Trust

As mentioned before, trustworthiness and dominance were first analyzed in isolation and compared by both proportionality level and scenario. Likewise, as previously indicated, participants provided their trust rating on a 7-point Likert scale, where the greater the rating provided by participants, the more trustworthy the punisher was deemed. Table 7 reports the mean ratings for each of the three proportionality levels within each of the four scenarios.

Table 7

Trust Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	4.57	4	0.98	239
Cut in Line	Proportional	5.37	5	1.17	244
	Severe	4.52	4	1.28	229
Eggs	Lenient	4.37	4	1.02	262
	Proportional	5.33	5	1.10	213
	Severe	4.02	4	1.42	264
Stutter	Lenient	4.58	5	1.28	243
	Proportional	5.38	5	1.29	239
	Severe	4.05	4	1.60	248

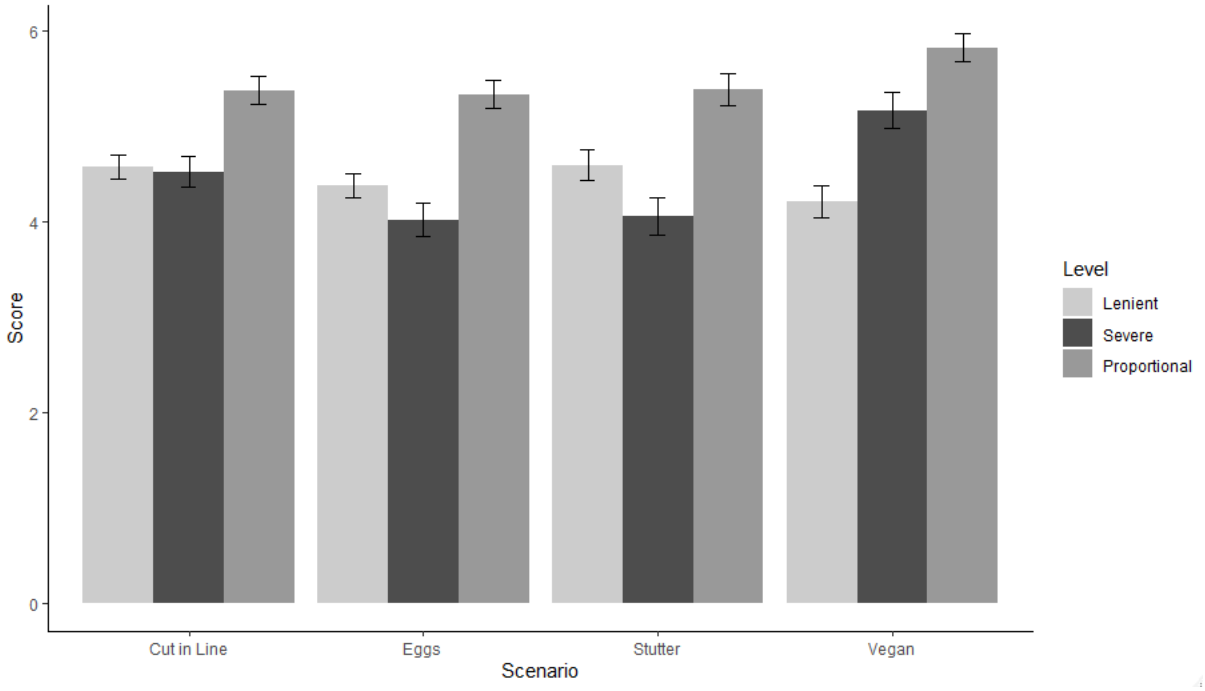
Vegan	Lenient	4.21	4	1.33	244
	Proportional	5.82	6	1.14	229
	Severe	5.16	5	1.47	238

Note. S.D. = Standard Deviation

Three mixed effects models (Bates et al., 2015) were fitted to analyze these data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs, Cutting in Line) as between participants factors, and Score (rating of trust in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 409.63$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 185.91$, $p < 0.0001$] and better fit (AIC = 10058.3, 9652.6 and 9484.7, respectively) (see details of the models fitted in Table F1 in Appendix F). Post hoc pairwise comparisons were conducted to test hypotheses H1, H2 and H3 on the trust dimension and significant trust rating differences between all proportionality levels within each of the four scenarios were identified, except for the lenient and severe levels within the Cut in Line scenario (see Figure 9).

Figure 9

Mean Trust Ratings by Scenario and Level



Note. Error bars represent standard errors

These results indicate that participants placed significantly greater trust on proportionate punishers compared with lenient and severe ones. There is an interaction of proportionality level and scenario, so that participants also placed greater trust on lenient punishers over severe ones in the Eggs and Stutter scenarios, but went in the opposite direction in the Vegan scenario (as mentioned, in the Cut in Line scenario there was no statistically significant difference between lenient and severe levels). Proportionality is clearly a contextual matter that allows for variations when it comes to its consequences. All pairwise comparison mean differences can be found in Table F2 in Appendix F.

5.4.3.2. Warmth

The five items of the Warmth index (including the trust item) were evaluated for internal consistency by scenario and level, and all 12 combinations had adequate internal consistency, with

Cronbach's Alphas of 0.84 or greater and McDonald's Omegas of 0.87 or greater (all Cronbach's Alphas and McDonald's Omegas by scenario and level can be found in Appendix G). They were therefore aggregated and used to measure Warmth ratings, where the greater the rating provided by participants, the warmer the punishers were deemed to be. Table 8 reports the mean warmth ratings for each of the three proportionality levels within each of the four scenarios.

Table 8

Warmth Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	4.45	4.4	0.85	239
Cut in Line	Proportional	4.99	4.8	0.98	244
	Severe	3.89	4	1.2	229
Eggs	Lenient	4.28	4.2	0.78	262
	Proportional	4.88	4.8	0.84	213
	Severe	3.21	3.2	1.26	264
Stutter	Lenient	4.53	4.4	1.04	243
	Proportional	4.95	5	1.15	239
	Severe	3.69	3.8	1.28	248
Vegan	Lenient	4.32	4.2	1.02	244
	Proportional	5.31	5.2	0.96	229
	Severe	4.79	4.8	1.23	238

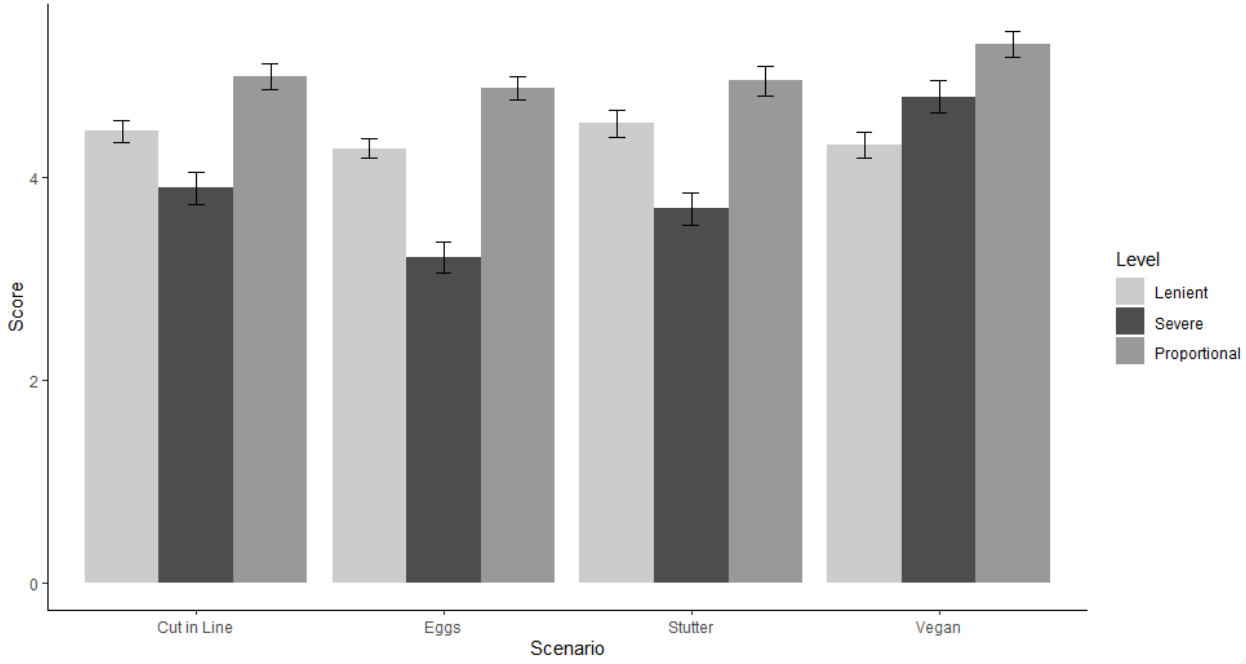
Note. S.D. = Standard Deviation

As with trust, three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs,

Cutting in Line) as between participants factors, and Score (rating of warmth in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 526.66$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 342.23$, $p < 0.0001$] and better fit (AIC = 9224.2, 8701.6 and 8461, respectively). In order to test hypotheses H1, H2 and H3 on the warmth dimension, I conducted post hoc pairwise comparisons and found significant warmth rating differences between all proportionality levels within each of the four scenarios (see Figure 10).

Figure 10

Mean Warmth Ratings by Scenario and Level



Note. Error bars represent standard errors

Overall, these results indicate that participants deemed proportionate punishers as warmer than lenient and severe ones. Participants also rated lenient punishers as warmer than severe ones, except for the case of the Vegan scenario, where disproportionately severe punishers were rated

as warmer than lenient punishers (but still less warm the proportionate punishers). Details of the fitted models can be found in Table F3 in Appendix F. These findings parallel those evidenced by the trust only ratings (except for the warmth difference between lenient and severe punishers within the Cut in Line scenario) and reflect the notion that trust is a fundamental component of the Warmth dimension (see Table F4 in Appendix F for all pairwise comparisons).

5.4.3.3. Dominance

As with trust, dominance was first analyzed in isolation by both proportionality level and scenario. Participants provided their dominance ratings on a 7-point Likert scale, where the greater the rating provided by participants, the more dominant they deemed the punisher. Table 9 reports the mean ratings for each of the three proportionality levels within each of the four scenarios.

Table 9

Dominance Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	3.54	4	1.30	239
Cut in Line	Proportional	5.22	5	1.12	244
	Severe	5.76	6	1.18	229
Eggs	Lenient	3.68	4	1.31	262
	Proportional	5.21	5	1.01	213
	Severe	5.73	6	1.23	264
Stutter	Lenient	3.59	4	1.40	243
	Proportional	5.67	6	1.17	239
	Severe	5.69	6	1.36	248
Vegan	Lenient	3.43	3	1.37	244

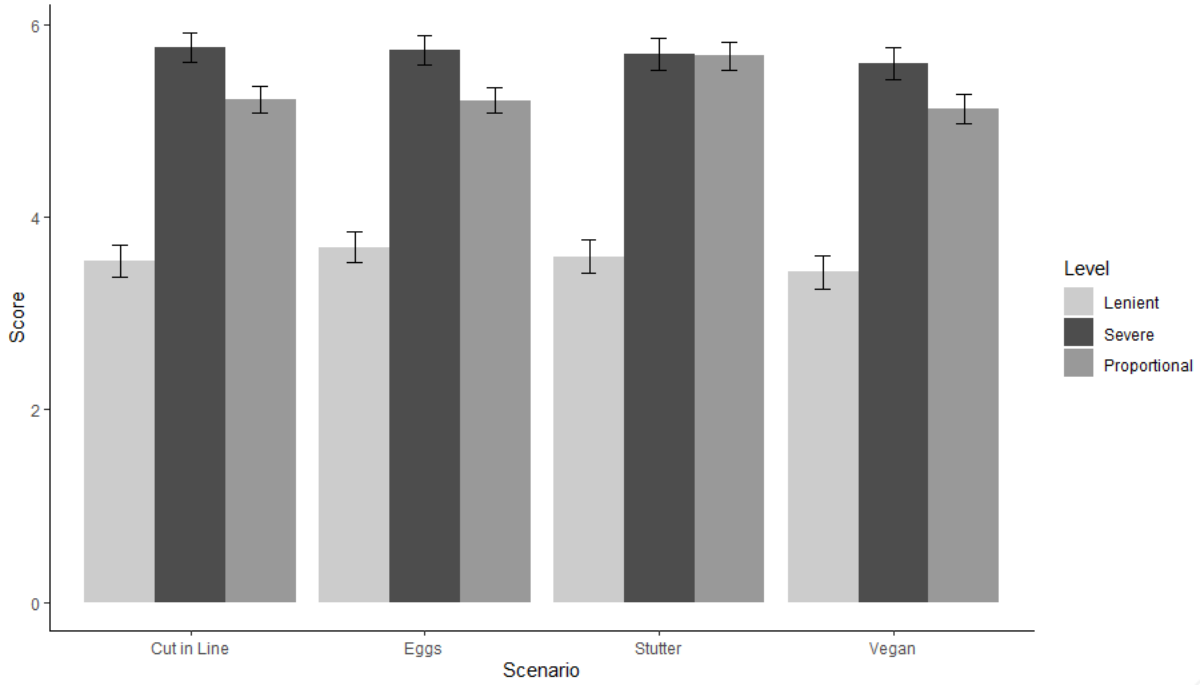
Proportional	5.12	5	1.13	229
Severe	5.59	6	1.30	238

Note. S.D. = Standard Deviation

Once again, three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs, Cutting in Line) as between participants factors, and Score (rating of dominance in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 1377.64$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 38.022$, $p < 0.0001$] and a slightly better fit compared with the model with just the fixed effect (AIC = 10787.5, 9413.9 and 9393.9, respectively). Hypotheses H1, H2 and H3 were put to the test on the dominance dimension with post hoc pairwise comparisons, which identified significant dominance rating differences between all proportionality levels within each of the four scenarios, except for the proportional and severe levels within the Stutter scenario (see Figure 11).

Figure 11

Mean Dominance Ratings by Scenario and Level



Note. Error bars represent standard errors

These results indicate that participants rated severely disproportionate punishers as more dominant than all other punishers by scenario (except for the aforementioned Stutter scenario where there was no statistically significant difference between proportional and severe levels). And point to a general positive relationship between severity and dominance, with all lenient punishers in all scenarios rated as the least dominant. Details of the fitted models and all pairwise comparisons can be found in Tables F5 and F6 respectively in Appendix F.

5.4.3.4. Competence

The five items which compose the Competence index (including the dominance item) were evaluated for internal consistency by scenario and level, and all 12 had adequate internal consistency, with Cronbach's Alphas of 0.75 or greater and McDonald's Omegas of 0.83 or greater (all Cronbach's Alphas and McDonald's Omegas by scenario and level can be found in Appendix G). They were therefore aggregated and used to measure Competence ratings, where the greater

the rating provided by participants, the more competent the punishers were deemed to be. Table 10 reports the mean competence ratings for each of the three proportionality levels within each of the four scenarios.

Table 10

Competence Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	3.95	4	1.09	239
Cut in Line	Proportional	5.62	5.8	0.88	244
	Severe	5.20	5.2	1.10	229
Eggs	Lenient	3.98	4	1.10	262
	Proportional	5.40	5.4	0.81	213
	Severe	4.81	5	1.01	264
Stutter	Lenient	3.98	4	1.10	243
	Proportional	5.57	5.6	0.97	239
	Severe	4.96	5	1.12	248
Vegan	Lenient	3.68	3.8	1.16	244
	Proportional	5.57	5.6	0.90	229
	Severe	5.53	5.6	1.07	238

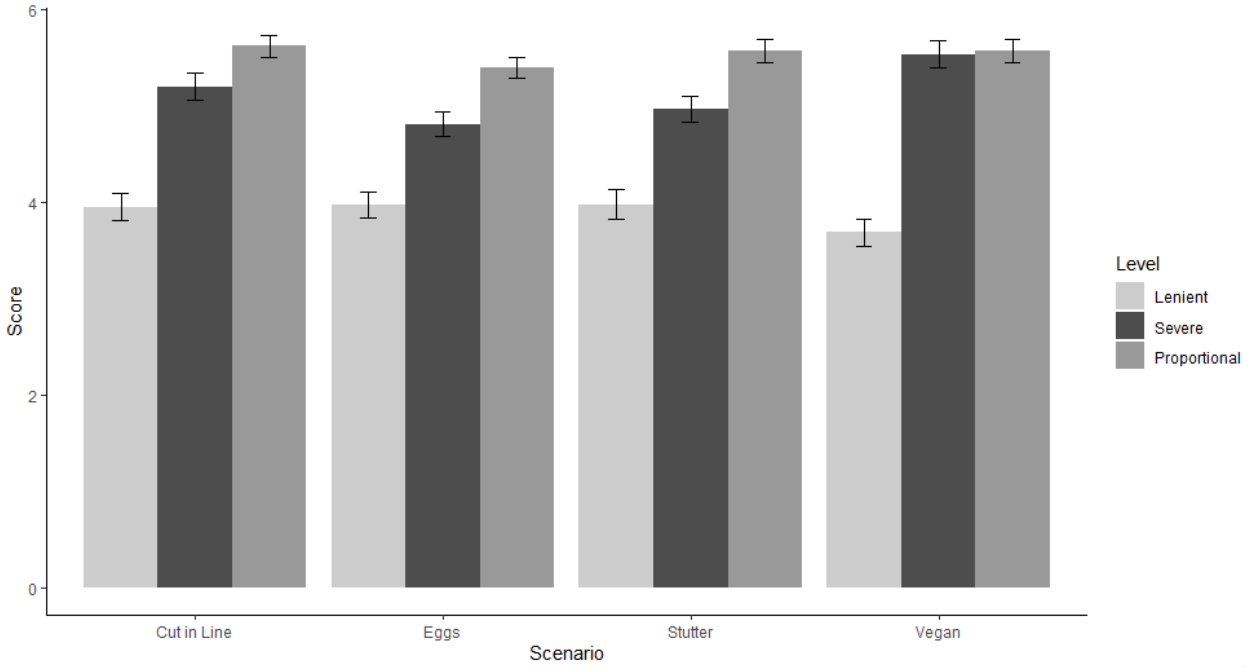
Note. S.D. = Standard Deviation

As with dominance, three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs, Cutting in Line) as between participants factors, and Score (rating of competence in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions

of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 1187.38$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 109.28$, $p < 0.0001$] and a better fit (AIC = 9519, 8335.6 and 8327.9, respectively). Details of the fitted model can be found in Table F7 in appendix F. Post hoc pairwise comparisons to test hypotheses H1, H2 and H3 in the competence dimension were conducted and identified significant competence rating differences between all proportionality levels within each of the four scenarios, except for the proportional and severe levels within the Vegan scenario (see Figure 12). All pairwise comparison mean differences can be found in Table F8 in appendix F.

Figure 12

Mean Competence Ratings by Scenario and Level



Note. Error bars represent standard errors

In most cases, participants deemed proportionate punishers as more competent than lenient and severe ones (expect for the lack of a statistically significant difference between proportional and severe punishers in the Vegan scenario). Likewise, participants always rated severe punishers

as more competent than lenient ones (all pairwise comparison mean differences can be found in Table F7 in Appendix F). Unlike the parallel between trust and warmth ratings, dominance and competence do not seem to track the same overarching person perception constructs, at least when measured in the context of third person punishers. The possible explanation for this observed phenomenon as well as its implications will be developed in the discussion section.

5.4.3.5. Moral Motives

In order to evaluate the internal consistency of the six items which compose the Moral Motives index, three of the questions were reverse coded (the third-party punisher acted based on personal self-interest, what was good for them personally, selfish reasons). All 12 level X scenario combinations were found to have adequate internal consistency, with Cronbach's Alphas of 0.76 or greater and McDonald's Omegas of 0.88 or greater (all Cronbach's Alphas and McDonald's Omegas by scenario and level can be found in Appendix G). They were therefore aggregated and used to measure Moral Motive ratings, where the greater the rating provided by participants, the more based on genuine moral motivation the punishment was deemed to be (and conversely, the lower the rating the more based on self-interested motivation the punishment was deemed to be). Table 11 reports the mean moral motivation ratings for each of the three proportionality levels within each of the four scenarios.

Table 11

Moral Motive Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
Cut in Line	Lenient	4.59	4.50	1.08	239
	Proportional	5.35	5.50	1.04	244
	Severe	4.71	4.67	1.16	229

Eggs	Lenient	4.64	4.67	1.05	262
	Proportional	5.50	5.67	1.04	213
	Severe	4.38	4.33	1.14	264
Stutter	Lenient	4.86	4.83	1.20	243
	Proportional	5.43	5.67	1.15	239
	Severe	4.45	4.50	1.01	248
Vegan	Lenient	4.49	4.50	1.33	244
	Proportional	5.69	5.83	1.01	229
	Severe	5.39	5.50	1.11	238

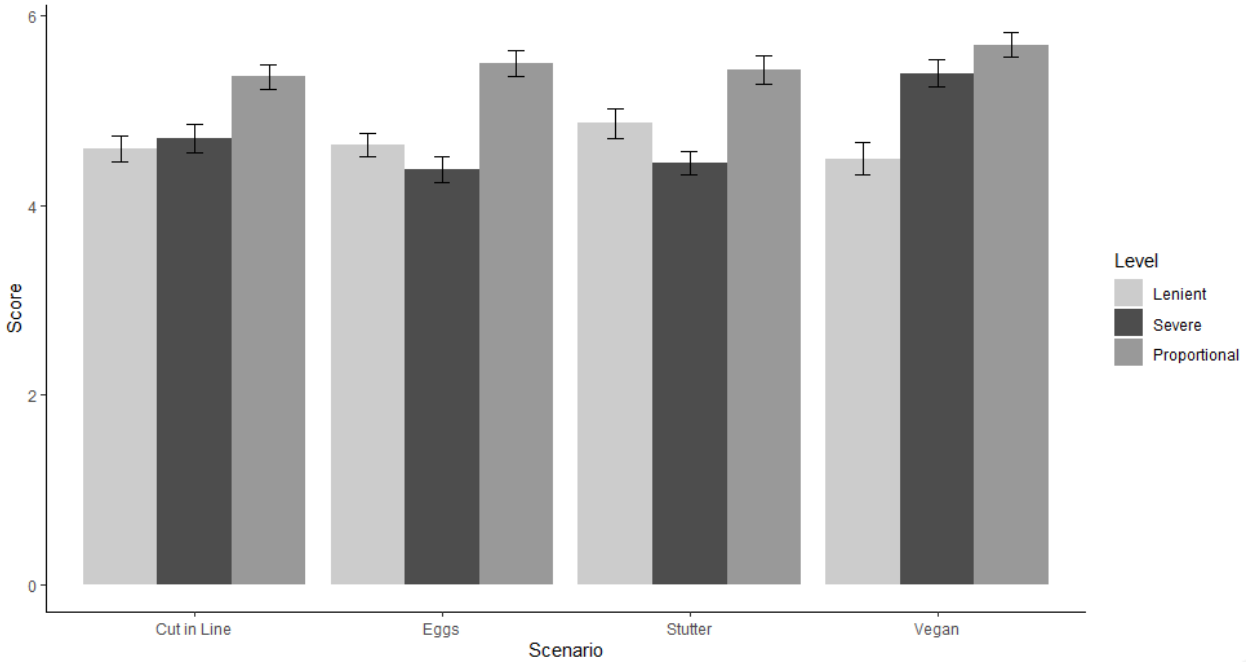
Note. S.D. = Standard Deviation

Again, three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs, Cutting in Line) as between participants factors, and Score (rating of moral motive in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 368.21$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 183.79$, $p < 0.0001$] and a slightly better fit (AIC = 9056, 8691.8 and 8526, respectively). Details of the fitted model can be found in Table F9 in Appendix F. In order to test hypotheses H4 and H5, I conducted post hoc pairwise comparisons, which identified significant moral motive rating differences between all proportionality levels within each of the four scenarios, except for the lenient and severe levels

within the Cut in Line scenario (see Figure 13). (See Table F10 in Appendix F for all pairwise comparisons).

Figure 13

Mean Moral Motive Ratings by Scenario and Level

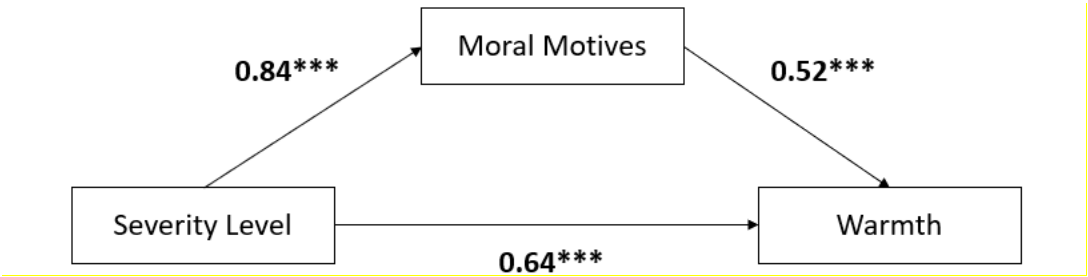


Note. Error bars represent standard errors

Overall, participants rated proportionate punishers as having punished based more on genuine moral motives (and less based on self-interest motives) compared with lenient and severe punishers. However, the relationship between lenient and severe punisher differed depending on the scenario. More specifically, participants deemed lenient punishers in the Eggs and Stutter scenarios as more driven by genuine moral motives than severe punishers, but went in the opposite direction in the Vegan scenario (and found no difference between lenient and severe punishers in the Cut in Line scenario). All pairwise comparison mean differences can be found in Table F10 in Appendix F.

In order to evaluate if the effects of severity level on perceived warmth and competence are mediated by attributed moral motives a mediation analyses were conducted. The results revealed that the effect of severity level on warmth is indeed partially mediated by moral motives. As Figure 14 illustrates, the regression coefficient between severity level and perceived warmth and the regression coefficient between attributed moral motives and perceived warmth was significant. The indirect effect was (0.84)*(0.52) = 0.44. I tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 1000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was 0.44, and the 95% confidence interval ranged from 0.38 to 0.50. Thus, the indirect effect was statistically significant ($p < 0.001$).

Figure 14



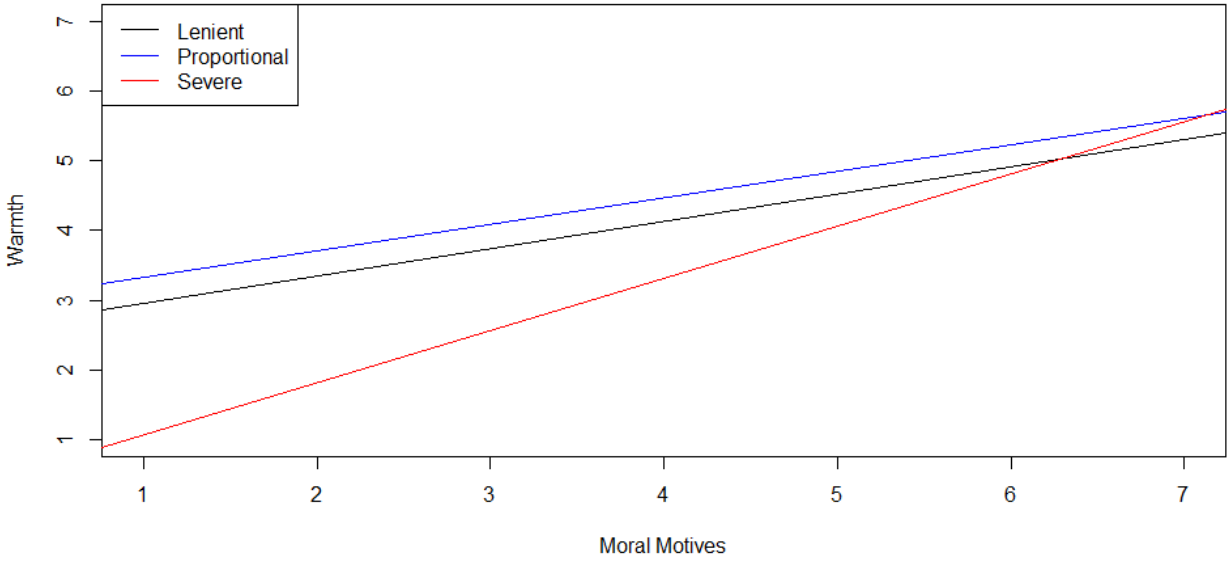
*Note. Standardized Regression Coefficients for the relationship between Severity Level and Warmth as Mediated by Attributed Moral Motives. ***p < 0.001*

Over two thirds (68%) of the effect of severity are accounted by the mediation by moral motives. That is, it is feasible to interpret the severity manipulation effect on warmth as determined by inferences people make on moral motives of the actors. When I analyzed the effect of attributed moral motives on perceived warmth for each of the three severity levels independently, I found steeper slopes for the lenient and especially the severe punishment levels compared with the

proportional level (see Figure 15). This indicates that the effect of proportional punishment relies on a weaker connection between moral motives and warmth than for severe punishments.

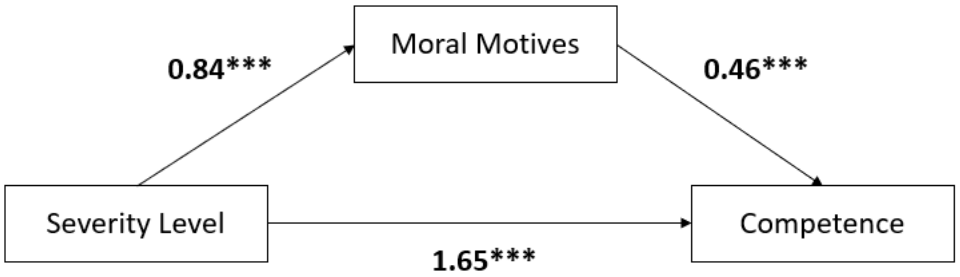
Figure 15

Perceived Warmth regressed on Moral Motives for each of the Severity Levels



When looking at the relationship between level and competence, the mediation analysis revealed that the effect of severity level on competence is also partially mediated by moral motives. As Figure 16 illustrates, the regression coefficient between severity level and perceived competence and the regression coefficient between attributed moral motives and perceived competence was significant. The indirect effect was (0.84)*(0.46) = 0.39. The bootstrapped unstandardized indirect effect was 0.39, and the 95% confidence interval ranged from 0.34 to 0.45.

Figure 16



*Note. Standardized Regression Coefficients for the relationship between Severity Level and Competence as Mediated by Attributed Moral Motives. ***p < 0.001*

The mediation effect of moral motives on the relationship between severity level and competence is considerably weaker than its effect on perceived warmth. Only 23% of the effect of severity level on perceived competence goes through moral motives. In this case it seems that the moral motives attributed to punishers did not determine their perceived competence. Contrary to perceived warmth, perceived competence seems to be more directly dependent on severity level or mediated by a third variable not part of the general theoretical framework of the present study.

5.4.3.6. Social Role

The social role dimension consisted of three items aimed at measuring the potential of the punisher of being deemed a leader, and three items aimed at measuring his potential as a friend. Given that these do not necessarily entail the extreme ends of one same continuum, the leader and friend items were separately evaluated for internal consistency. All 12 level X scenario combinations for leader had adequate internal consistency, with Cronbach's Alphas and McDonald's Omegas of 0.89 or greater, and all combinations for friend shown to have Cronbach's Alphas and McDonald's Omegas of 0.87 or more (all Cronbach's Alphas and McDonald's Omegas by scenario and level can be found in Appendix G). They were therefore aggregated and used to measure, on the one hand Leadership ratings (where the greater the rating provided by participants, the more they considered the punisher a leader), and on the other hand Friendship ratings (where

the greater the rating provided by participants, the more they considered the punisher a friend). Table 12 reports the mean leadership ratings for each of the three proportionality levels within each of the four scenarios.

Table 12

Leadership Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	3.85	4	1.44	239
Cut in Line	Proportional	5.69	6	1.07	244
	Severe	4.70	5	1.59	229
Eggs	Lenient	3.86	4	1.29	262
	Proportional	5.46	5.33	1.02	213
	Severe	3.65	3.67	1.62	264
Stutter	Lenient	3.88	4	1.48	243
	Proportional	5.41	5.67	1.34	239
	Severe	4.13	4.33	0.70	248
Vegan	Lenient	3.49	3.33	1.45	244
	Proportional	5.66	6	1.13	229
	Severe	5.35	5.67	1.36	238

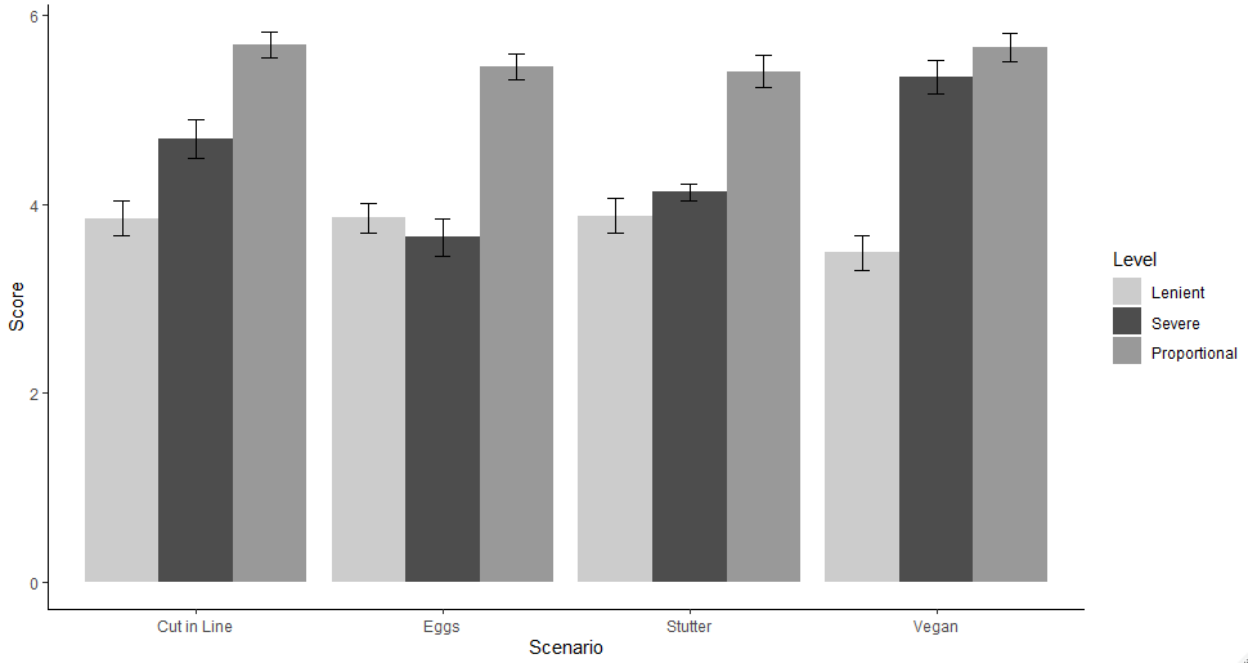
Note. S.D. = Standard Deviation

Three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan, Eggs, Cutting in Line) as between participants factors, and Score (rating of leadership in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed

effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 798.61$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 277.82$, $p < 0.0001$] and a better fit (AIC = 10734, 9939.4 and 9679.6, respectively)(details of the fitted models can be found in Table F11 in Appendix F). To evaluate hypotheses H6 and H7 with regards to leadership, I conducted post hoc pairwise comparisons, which identified significant leadership rating differences between all proportionality levels within each of the four scenarios, except for the lenient and severe levels in the Eggs scenario (see Figure 17). All pairwise comparison mean differences can be found in Table F12 in appendix F.

Figure 17

Mean Leadership Ratings by Scenario and Level



Note. Error bars represent standard errors

The proportionate punisher is consistently hailed as the best leader. Collapsing by scenario, I found that contrary to what might have been expected (based on previous research), proportionate punishers are considered better leaders overall than severe leaders.

Table 13 reports the mean friendship ratings for each of the three proportionality levels within each of the four scenarios.

Table 13

Friendship Ratings by Scenario and Level

Scenario	Level	Mean	Median	S.D.	N
	Lenient	4.49	4.67	1.08	239
Cut in Line	Proportional	5.36	5.33	1.08	244
	Severe	4.17	4.33	1.33	229
Eggs	Lenient	4.37	4.33	1.08	262
	Proportional	5.35	5.33	1.01	213
	Severe	3.40	3.33	1.50	264
Stutter	Lenient	4.67	4.67	1.38	243
	Proportional	5.58	5.67	1.30	239
	Severe	4.11	4	1.58	248
Vegan	Lenient	4.24	4.33	1.36	244
	Proportional	5.79	6	1.05	229
	Severe	5.19	5.33	1.44	238

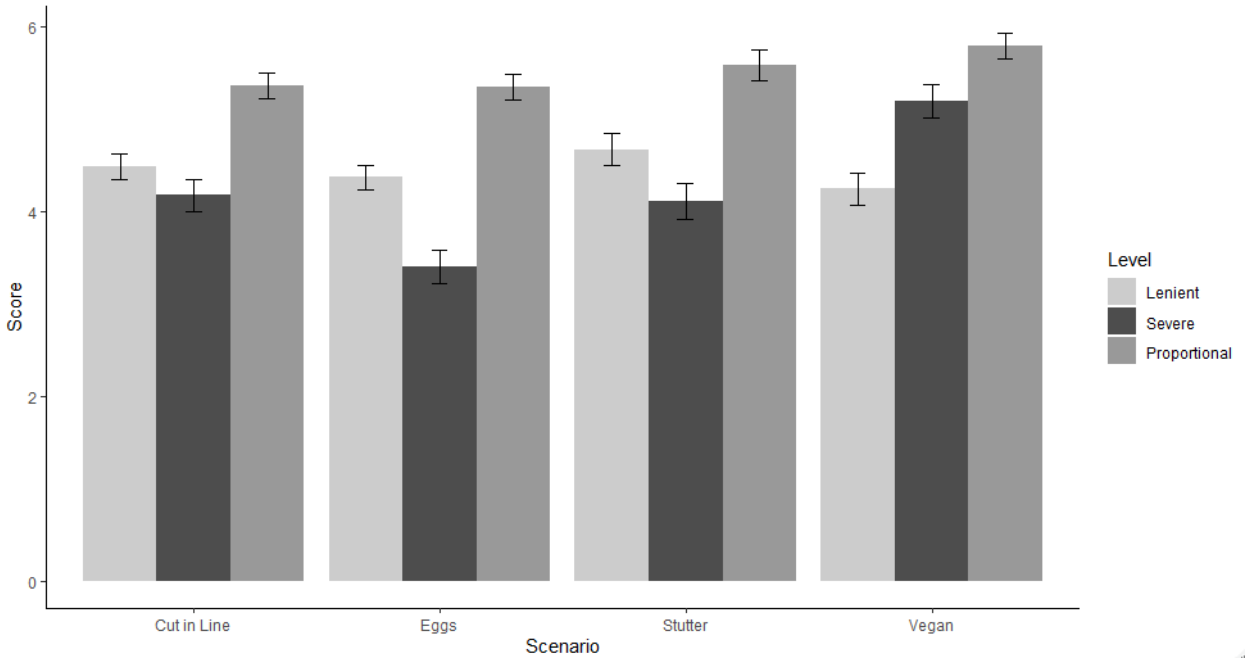
Note. S.D. = Standard Deviation

As with Leadership, three mixed effects models were fitted to analyze the data with a random intercept for participant, level (lenient, proportional, severe) and scenario (Stutter, Vegan,

Eggs, Cutting in Line) as between participants factors, and Score (rating of friendship in the 7 point Likert scale) as a dependent variable. The difference between the models were the inclusions of the level fixed effect and an interaction term between level and scenario respectively. The inclusion of the level fixed effect resulted in a significant change [$\chi^2(2) = 520.64$, $p < 0.0001$]. Moreover, the inclusion of the interaction term resulted in a significant change [$\chi^2(9) = 296.85$, $p < 0.0001$] and a better fit (AIC = 10332.8, 9816.2 and 9537.3, respectively). Details of the model fit can be found in Tabble F13 in Appendix F. Post hoc pairwise comparisons to test hypotheses H6 and H7 regarding potential friendship roles were conducted and identified significant friendship rating differences between all proportionality levels within each of the four scenarios (see Figure 18). (See Table F14 in Appendix F for all pairwise comparisons).

Figure 18

Mean Friendship Ratings by Scenario and Level

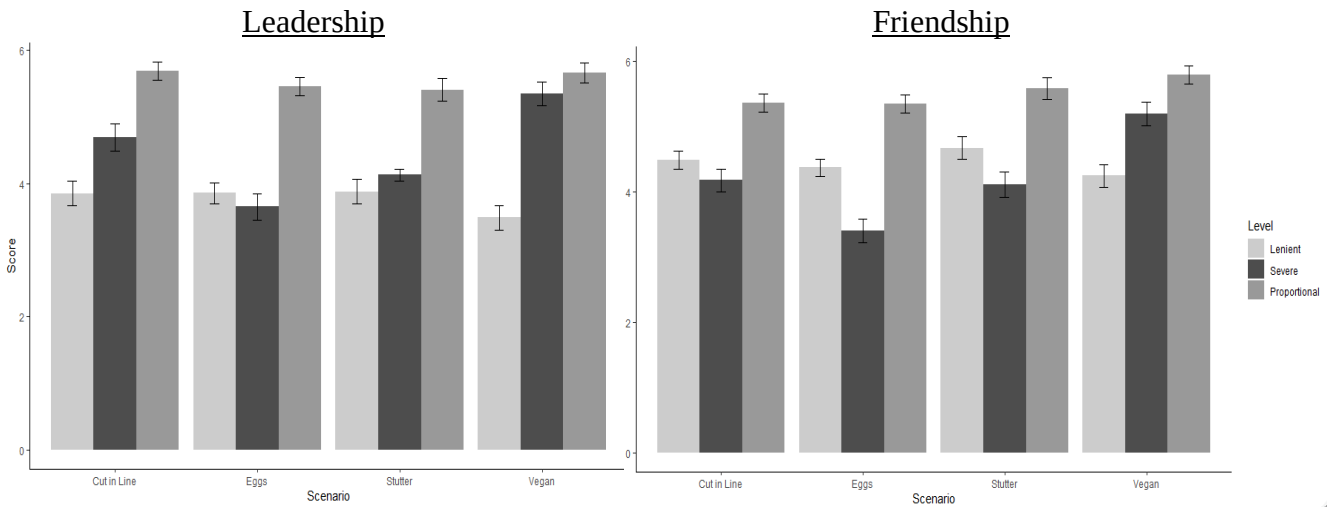


Note. Error bars represent standard errors

The results of the Leadership and Friendship ratings indicate that participants find proportionate punishers to be better suited for potential leadership and friendship roles compared with lenient and severe punishers, and surprisingly, they also rate them more as friends versus lenient and severe punishers. Such counterintuitive results might be better understood if one considers the multifaceted qualities a good leader is often supposed to have, which might include good, proportionate judgment. This will be further developed in the discussion section. Furthermore, the results indicate that in most cases, participants rate severe punishers more as leaders than lenient punishers (except for the Eggs scenario), while in most cases they rate lenient punishers more as friends than severe ones (except for the Vegan scenario) (see Figure 19).

Figure 19

Mean Friendship and Leadership Ratings by Scenario and Level



Note. Error bars represent standard errors

5.4.4. Discussion

5.4.4.1. Trust and Warmth

In support of hypothesis H3, the results of experiment 3 suggest that proportionate punishers are deemed to be as both more trustworthy and warmer than disproportionately lenient

and / or disproportionately severe punishers. This inverse-U shaped function between punishment severity and trust (or warmth) might be explained by a combination of relative signal strength as severity increases from lenient to proportionate, and excessive severity interpreted as moralization or stubborn consequentialism. Firstly, it is possible that proportionate punishers were perceived as more trustworthy and warmer than disproportionately lenient ones because the comparatively higher severity implied a greater incurred cost and an overall stronger signal of trust and warmth. Secondly, it is feasible that disproportionately severe punishers efforts were perceived as the by-products of moralization (Kreps & Monin, 2014), hence rendering them as self-righteous in the eyes of observers, and / or as overtly consequentialist (Everett et al., 2016), making them appear as too cold and calculating. The mediation analysis provides support for this notion (and hypothesis H9), as perceived warmth depends to a large extent on the moral motives attributed to punishers, and attributions of genuine moral concern lead to greater perceptions of warmth.

The fact that the three levels of proportionality produced essentially the same results in trust as well as in warmth is a testament to how central the trust component is to the broader concept of warmth. The only difference between the two results was the greater perceived warmth of disproportionately lenient punishers compared with disproportionately severe ones in the Cut in Line scenario specifically (a difference that simply had not reached statistical significance when measuring trust).

Nevertheless, there was a surprising and interesting result observed when measuring both trust and warmth. Namely, the fact that, opposite to what was observed in all other scenarios, in the Vegan scenario, disproportionately severe punishers were rated as more trustworthy and warmer than disproportionately lenient ones. This apparently counterintuitive result might be the product of the characteristics of the Vegan scenario in general and of its disproportionately severe

punishment version in particular. Since similar, apparently counterintuitive results were evidenced when evaluating moral motives for this scenario and version (i.e. severe punishers were attributed more genuine moral motives than lenient ones), as well as when evaluating friendship potential (i.e. severe punishers were rated more as potential friends than lenient ones), this specific vignette grants a more detailed analysis:

> "Harry is at a barbecue. A man purposefully cooks a beef hamburger and feeds it to a vegetarian woman, telling her it is a vegan "incredible burger" (a vegan product that is marketed as tasting just like beef). The vegetarian woman eats the burger, after which the man privately reveals what he has done to Harry, who does not know the vegetarian woman personally. The man looks angry and like someone you wouldn't want to mess with. Harry tells the man that deceiving people about what they are eating is not OK, then tells the host of the barbecue what the man did and asks the man to leave the barbecue. Afterwards, he hires a lawyer and convinces the woman to file a lawsuit against the man."

Unlike the other disproportionately severe versions of the other three scenarios (all vignette version can be found in Appendix D), this punishment entails an action that could be easily interpreted as a socially desirable characteristic, even if its deemed excessively severe (all vignette scenarios were pre-tested to ensure adequate severity perceptions and differences in severity levels in the expected direction). More specifically, the protagonist, Harry, goes out of his way to hire a lawyer (potentially incurring a high monetary cost), which might be interpreted as very sensible course of action within the possible avenues for punishment available to someone desiring to inflict a severe punishment (Harry could have easily assaulted, or otherwise verbally or psychologically abused the offender in an attempt to severely punish him). Moreover, filing a lawsuit could be

interpreted as a just way to stand up for something one believes in while maintaining a degree of impartiality (because in the end it will be an impartial judge or a jury that makes the final decision in the matter). And even if Harry loses the lawsuit, he will be sending a very strong signal of blame, and hence of his strong commitment to a specific set of moral norms (Shoemaker & Vargas, 2019) in a non-violent manner.

Therefore, especially when compared with the disproportionately lenient version of the punishment in this scenario (i.e. Harry makes eye contact with the man and shakes his head in disapproval), the disproportionately severe punishment sends a much stronger signal while maintaining a positive person perception quality. Nonetheless, the main takeaway from all scenarios (including the Vegan scenario) is that participants prefer (i.e. rate as more trustworthy and warmer) proportionate punishers over all other punishers. This means that proportionality is fundamental in the trust and warmth perceptions of third-party punishers.

5.4.4.2. Dominance and Competence

With regards to dominance, the results of Experiment 3 also lend support to hypothesis H3, whereby the relationship between severity and dominance are largely described by a positive linear function. This might be explained by the fact that punishment severity is a very direct and clear proxy of dominance. This reflects the notion of third-party punishment espoused by Krasnow, Delton, Cosmides and Tooby (2016) whereby individuals interpret mistreatment of a third party as potential mistreatment of themselves, leading third-party punishment to be utilized as a signal of dominance aimed at deterring personal mistreatment. Under this view, the harsher the punishment, the stronger the signaling of dominance. The only exception was the observed lack of a difference between the proportional and disproportionately severe punishers within the Stutter scenario, which might be explained by a ceiling effect of severity on dominance reached with the

proportional punisher. It might be reasonable to argue that, even though there was a difference in severity between these two levels (confirmed in the pretests of the vignettes), the resulting perceived dominance did not significantly increase via the additional actions that the punisher took in the disproportionately severe case.

The results of the experiment, when looking at the dimension of competence as a whole did not parallel those of trust and warmth. This might indicate that dominance is not as central a concept of competence as trust is of warmth, at least when evaluating the construct in the context of third-party punishment employed in this study (or that the actions presented in the vignettes do not allow for an adequate observation of dominance as part of the broader competence dimension).[9] While proportionate punishers were rated as more dominant and more competent than disproportionately lenient punishers, disproportionately severe punishers were rated as more dominant but as less competent than proportionate ones. This means that the relationship between severity and competence is better described by an inverse-U function (while the relationship between severity and dominance is better described by a positive linear one). Unlike the relationship between severity and warmth, the relationship between severity and competence does not appear to be mediated by moral motive attributions. This implies that the perceived competence of the punishers did not seem to depend on whether the they were deemed to have been acting based on genuine moral concerns or not. Perceived competence then, might be more closely linked to a general ability to engage than to the reasons for engaging.

The only exception to the inverse-U shaped function between severity and competence, was the observed lack of a statistically significant difference between proportionate and

[9] A one factor analysis conducted on the five items of Competence was not significant [$\chi^2(5) = 943.69$, $p < 0.0001$] with particularly high Uniqueness (0.63) and low factor loadings (0.61) for dominance compared with the other four items.

disproportionately severe punishers in the Vegan scenario, likely due to a competence ceiling effect similar to the one observed in the Stutter scenario when evaluating dominance. These differences between dominance and competence are probably rooted in the qualitative differences between the two concepts. Namely, a competent person does not necessarily have to be dominant. In fact, the four items besides dominance that made up the competence index (i.e. competence, confidence, capability and efficacy) could arguably be negatively impacted by exhibitions of disproportionate severity (e.g. an excessively severe punisher might not be considered very efficacious), leading to the observed discrepancies.

Despite the mentioned differences between dominance and competence, the fact that disproportionately severe punishers were always rated as more competent than disproportionately lenient ones, does speak to a broader overarching tendency where harsher punishments generally lead to perceptions of greater dominance and competence. Again, this is likely driven by an interpretation of potential personal mistreatment, and hence of severity in third-party punishment as a deterrent instrument, construed as dominance and competence.

5.4.4.3. Moral Motives

In support of hypothesis H6, the results of experiment 3 point to an inverse-U shaped relationship between severity and moral motives (where the greater the score on the moral motives axis, the more the punisher was deemed to have acted based on genuine moral motives). This implies that proportionate punishers were rated as more motivated by genuine moral motives than either disproportionately lenient or disproportionately severe punishers. It would be safe to assume that while proportionate punishers are seen as incurring a greater cost and hence perceived as more morally motivated than disproportionately lenient ones, disproportionately severe punishers are

seen as engaging in moral grandstanding (Tosi & Warmke, 2016) and as more motivated by status seeking motives (i.e. self-interest) than proportionate punishers.

Furthermore, disproportionately lenient punishers are attributed more genuine moral motivations than disproportionately severe ones in the Eggs and Stutter scenarios, lending support to the notion that excessive punishment is viewed as more motivated by selfish concerns. In the Cut in Line scenario there was no observed difference in the motivations driving the actions of disproportionately lenient and severe punisher, which could be due to a slightly higher perceived severity of the lenient version in this scenario compared with other scenarios (all severity ratings and pairwise comparison can be found in Appendix E). As with competence, in the Vegan scenario, disproportionately severe punishers were also attributed more moral motivations compared with disproportionately lenient ones. Again, this is likely due to the particular characteristics of the severe punishment version of the Vegan scenario discussed earlier, which might have led participants to interpret the hiring of a lawyer as a particularly morally motivated action.

5.4.4.4. Leadership and Friendship Roles

Finally, the results of the experiment provide support for hypothesis H13, for the potential leadership as well as friendship roles of the punishers. This hypothesis posits that proportionate punishers are considered more as leaders and also more as friends compared with disproportionate punishers (lenient or severe). This is a counterintuitive result, especially in light of previous research on role preference (Laustsen & Petersen, 2015), which found that people tend to prefer non-dominantly looking friends and dominantly looking leaders. The results of experiment 3 not only indicate that disproportionately severe punishers are generally perceived as more dominant than proportionate punishers, but at the same time that they are not perceived as better leaders than proportionate punishers. This might be rooted in the same underlying mechanism which led to

lower ratings of competence for disproportionately severe punishers. In other words, what characterizes a good leader is bound to be very similar to what characterizes a competent individual, and it does not necessarily include dominance. Hence, while a good leader might have a decent amount of dominance (i.e. the proportionate punisher), an excessively dominant one (via severe punishment) could hinder some of the very qualities that make him a good leader (i.e. competence, confidence, capability and efficacy). Dominance might not be part of this same set because it denotes superior power over another which, in the case of excessive dominance, implies resorting to subjugation and control by force precisely because of a lack competence and leading ability.

Nonetheless, lending some support to hypothesis H12 (i.e. disproportionately lenient punishers considered more as potential friends than proportionate punishers, and these in turn considered more as friends than disproportionately severe punishers), the disproportionately severe punishers were generally considered more as leaders than disproportionately lenient ones (the only exception being the Eggs scenario, where participants went in the opposite direction, likely because in this scenario the disproportionately severe punisher acted in a manner that could have been interpreted as vengeful and obsessive; not characteristic of a leader).

The results on potential social role were especially surprising because the same inverse U-shaped function found between severity and leadership role, was also found to describe the relationship between severity and friend role. That is, proportionate punishers were considered more as friends than disproportionate punishers (lenient or severe), which suggests that participants highly valued proportionality as a defining quality of both friends and leaders. In the case of friendship, the counterintuitive higher rating of proportionate over disproportionately lenient punisher might boil down to a combination of better signal strength and the possibility that

lenient punisher were not perceived as forgiving (as in the case of the forgivers in experiments 1A and 1B) since they were exerting some sort of punishment little though it might have been.[10]

As with leadership, however, there was some support for hypothesis H12 with regards to friendship. In the vast majority of cases, disproportionately lenient punishers were considered more as friends than disproportionately severe ones. This result finds backing in the notion that punishment severity can be used as an instrument of deterrence whereby people interpret harsh punishments of third parties as potential future mistreatment of themselves. In such a case, individuals prefer the lesser of two evils as a friend (i.e. an unjust person over one that could exert excessively severe punishment on oneself). The only exception, yet again, were the disproportionately severe punishers in the Vegan scenario being considered more as friends than the disproportionately lenient ones. As mentioned earlier, by hiring a lawyer these punishers were deemed to have acted more based on genuine moral motives and as possessing greater competence than their lenient counterparts, which could be argued are also socially desirable qualities of potential friends.

6. General Discussion

The overarching objective of the conducted studies was to identify how the signaling function of moralistic punishing (operationalized as third-party punishing) is determined by the punishment's proportionality and to evaluate the extent to which moralistic punishing effectively signals the underlying motivations (genuine moral concern or self interest), the character traits of

[10] An exploratory analysis found that while 55% of the relationship between severity level and friendship was accounted for by the mediation of moral motives, only 30% of the relationship between severity level and leadership was accounted for by attributed moral motives. This suggests a common aspect of perceived warmth and friendship on the one hand, and perceived competence and leadership on the other.

the punisher (warmth or competence), and determines which social role the punisher is considered for (friend or leader). The purpose and design of experiments 1A, 1B and 2 was to evaluate how deservedness and severity, two fundamental components of proportionality, affect the signaled trustworthiness within well-established economic game paradigms. Therefore, I will address the implications of these three experiments first, and then move on to the objective of experiment 3, which had a very different experimental approach aimed at capturing a broader set of character traits and person perceptions signaled via moralistic punishment.

6.1. Deservedness

A principal objective of experiments 1A and 1B was to evaluate the effect of a punishment's deservedness on the punisher's signaled trustworthiness. This constitutes a novel theoretical and methodological approach. Punishment deservedness (also referred to in the literature as deservingness) has been the subject of studies on its potential effects on the perpetrator and the victim (e.g. Evans, Galyer, & Smith, 2001), but not on the punisher. Likewise, deservedness has been analyzed as a mediator of the magnitude of pleasure at a third-person's misfortune (i.e. schadenfreude) (Feather & Sherman, 2003; van Dijk, Ouwerkerk, Goslinga, & Nieweg, 2005), which although not strictly a punishment exerted by a victim or third-person, could be construed as a punishment nonetheless. Finally, the effect of a punishment's deservedness has been evaluated as a determinant of contribution levels in public goods games (but deservedness is operationalized in terms of whether the punishment is directed at a specific individual versus the entire group). In none of these studies, however, was punishment deservedness examined as a determinant of the punisher's signaled character traits.

Moreover, the third-party punishment games used in the literature thus far have almost exclusively focused on deserved punishment, therefore ignoring a less frequent but no less

important type of punishment: the undeserved kind. Undeserved third-party punishment is arguably a very significant version of moralistic punishment as it embodies a moral norm violation (i.e. unjustified harm via punishment) and therefore a trigger for moral outrage (Crockett, 2017), which could in theory engender second-order moralistic punishment directed at the unjust punisher (Hofmann, Brandt, Wisneski, & Rockenbach, 2018; Jordan et al., 2016). Likewise, undeserved punishment (often referred to as antisocial punishment), has been studied almost exclusively in public goods paradigms (e.g. Herrmann, Thöni, & Gächter, 2008; Rand & Nowak, 2011; Szolnoki & Perc, 2017), and has been theorized as a problem for the evolution and stability of cooperation (Rand, Armao IV, Nakamaru, & Ohtsuki, 2010; Rand & Nowak, 2013). The only study, to my knowledge that has analyzed undeserved punishment in a third-party punishment paradigm (Rabellino, Morese, Ciaramidaro, Bara, & Bosco, 2016) did not investigate the reactions towards said punishers (e.g. via a posterior trust game) or their signaled character traits.

I found that when the punishment was undeserved, people tended to place greater trust on those who avoided punishment and distrust those who did punish. This difference in perceived trustworthiness is likely caused by two distinct underlying mechanisms. The first mechanism focuses on the greater trust placed on Fair Watchers. The comparatively greater trust placed on observers who do not punish, even when the avoidance of punishment entails an active choice (as well as an incurred cost) indicates that people tend to make character judgments by first focusing on a moralistic punisher's fairness and only then by considering the potential harms caused by said punisher. Hence, people first decide who to trust based on how fair a potential partner is (eliminating unfair individuals who might behave in a seemingly unfair fashion in future interactions with them), and only then choose a potential partner based on the actual harm caused by fair individuals. This seems to parallel the person perception assessment sequence of the

Stereotype Content Model, whereby a primary assessment of affiliation underlying approach or avoidance mechanisms is followed by a secondary assessment of potential action and agency (Cuddy et al., 2008).

The fact that no harm is preferred to harm (when both are fair) points to a general inclination for harms caused by omission over those caused by commission and reflects the presence of a possible omission bias, which has been previously evidenced in other contexts (Baron & Ritov, 1993; Haidt & Baron, 1996) as well as a preference for indirect harm over direct harm (Royzman & Baron, 2002). It also speaks to the potentially high attributional ambiguity (M. L. Snyder, Kleck, Strenta, & Mentzer, 1979) driving the act of punishment within the constraints of the third-person punishment game (TTP) experimental paradigm. In other words, the fact that it is difficult for a person to ascertain the motives or causes of the punishment in the TTP might render this act as of low informational value regarding the punisher's character traits (Uhlmann et al., 2015). Because people are provided with no other information about the punisher in the TPP, other than the fact that she exerted punishment on a complete stranger for not having shared half of the amount of money he was endowed with, opens up the possibility of alternative motives for punishment beyond the condemnation of selfishness (and hence the implicit belief in the associated moral norm ; Shoemaker & Vargas, 2019) that a person judging the punisher might consider.

The symbolic and expressive aspects of the punishment in this case might very well be interpreted as indicative of negative character traits. The sender in the TPP is essentially a person who has just experienced a windfall monetary gain and must immediately decide whether to share 50% of it with a stranger. Based on this limited information, a third person who punishes that sender could arguably be perceived as excessively punitive, strict and/or driven by an inherent pleasure to harm (via the punishment) or to appear as self-righteous.

The second mechanism driving the observed significant difference in trustworthiness between those who justly avoid punishment (i.e. Fair Watchers) and those who unjustly punish (i.e. Sadists) focuses on the distrust placed on the latter. The act of punishing a complete stranger who obviously does not deserve it is likely a statistically rare occurrence in daily life and many people would arguably deem it an extreme behavior, both of which are perceived as highly informative of character traits (Ditto & Jemmott, 1989; Fiske, 1980; Kelley, 1967; McKenzie & Mikkelsen, 2007). The fact that this act also involved a cost to the decision maker renders it especially informative of the character of the punisher (Ohtsubo & Watanabe, 2009). In particular, the rarity and extremeness of such an unjust punishment was likely very revealing of the punisher's poor moral character, and probably led to the observed moral judgments of distrust people made when considering whether to trust them in the trust game (Tannenbaum et al., 2011).

When the punishment was deserved, the observed lack of a difference between the conditions is likewise due to two mechanisms driving perceived trustworthiness in the same direction. The first mechanism drives perceived trust as a function of fairness and social status. On the one hand, trust placed on just punishers (i.e. Fair Vigilantes) reflects an overall preference for fair over unfair individuals (just as with the undeserved conditions, and again paralleling the sequential assessments of the Stereotype Content Model). On the other hand, deserved punishment has been linked to the maintenance of high-status positions. Moralistic punishment allows individuals to maintain their high-status position, while failure to punish leads to that position being at risk (Gordon & Lea, 2016). And more importantly, high-status individuals in third-party punishment games are expected to punish, whereas low-status individuals are not (Gordon & Lea, 2016), which would allow for perceived trust to be tied to punishment via perceived status. The link between status and perceived trust has been extensively established in previous research in

terms of socioeconomic position (Brandt, Wetherell, & Henry, 2015; Keijzer & Corten, 2016; Qi, Li, & Du, 2018), as well as in terms of job hierarchy, respect, prestige, organizational affiliation (Lount & Pettit, 2012), and even in terms of the relative status generated by distinct religions (Gupta, Mahmud, Maitra, Mitra, & Neelim, 2018). Here I propose that third-party punishment bestows an implicit high-status position compared with no punishment, thus generating an increase in perceived trust.

However, this is not a universal conditional assessment, as a considerable number of individuals placed greater trust on those who did not punish when it was deserved (i.e. Forgivers). Forgiveness can arguably be perceived as a desirable character trait. Firstly, it has been theorized to be the result of cognitive mechanisms designed to reduce the cost of deterrence while preserving valuable relationships (McCullough, Kurzban, & Tabak, 2013). More specifically, forgiveness has been conceptualized as a system that has evolved to counteract or limit the potential long-term detrimental effects of deterrence (i.e. retaliation), and the operation of this system depends on estimating the risk of future exploitation by the harm doer as well as the expected future value of the relationship with the harm doer (McCullough et al., 2013). Secondly, forgiveness has been posited as a powerful mechanism to restore mutual cooperation and as particularly apt strategy to avoid mutual recriminations (Axelrod, 2006). Thus, based on my findings I theorize that the trust placed on a forgiver in the TPP reflects the perceived likelihood that the forgiver has a vested interest in preserving relationships for future interactions, that he/she will cooperate in future interactions, and that he/she will behave in a non-recriminatory manner (in case that the receiver in the trust game thinks the amount entrusted by the sender is not sufficiently generous).

The observed split between participants who entrusted Fair Vigilantes and Forgivers reflects the lack of a preference for one over the other, at least within the confines of the economic

game paradigms used. Placed trust and preference of Fair Vigilantes over Forgivers is probably very highly dependent on the context and the types of moral norm infringements and punishments exerted. This represents a potentially prolific line of research for future studies that make use of methods that transcend traditional experimental economic games.

In addition, forgiving within the experimental designs of experiments 1A and 1B was constrained to an absolute lack of severity (i.e. no punishment), but this does not reflect the spectrum of forgiving available within punishment contexts in the real world. Forgiveness could conceivably be granted in any number of ways, including but not limited to postponing punishment, changing the type and quality of the punishment and by shifting part of the blame to environmental factors. None of these are contemplated in the studies comprising the present research and constitute one of its limitations, as well as one of the potential avenues for future research efforts. Nonetheless, another way to administer different degrees of forgiveness is attained by calibrating the severity of the punishment. I attempted to capture this, at least partially, in experiment 2, where the focus was put solely on severity as a determinant of trustworthiness signaled via moralistic punishment.

6.3. Severity

The vast majority of studies on moralistic punishment have considered only deserved punishment (e.g. Balafoutas & Nikiforakis, 2012; Balafoutas, Nikiforakis, & Rockenbach, 2014; Barclay, 2006; Fehr & Gächter, 2000; Hoffman, Hilbe, & Nowak, 2018; Jordan et al., 2016, 2017; Kahneman, Knetsch, & Thaler, 1986; Kurzban et al., 2007; J. Martin, Jordan, Rand, & Cushman, 2018; Nelissen, 2008; Pedersen, McAuliffe, & McCullough, 2018) which probably reflects the

much higher incidence of deserved moralistic punishment in daily life compared with undeserved punishment. Despite this, to date no studies on third-party punishment, either in field or lab settings (including third-party punishment game, public goods, prisoner dilemma or vignette based experimental approaches) have focused on how the punishment's severity affects the reaction of the individuals involved. The only study that has, to my knowledge, varied punishment severity in some way was conducted by Balafoutas and colleagues (2014), where in addition to direct third-party punishment the researchers included indirect third-party punishment via withholding help (which could be understood as a different, probably lower level of severity). In this case the researchers found that indirect punishment was more prevalent than direct punishment.[11] Regardless, the effect of punishment severity on the perceived character traits of the punisher has not been previously examined.

Taking into account the likely higher occurrence of deserved punishment in real life situations, experiments 2 and 3 focused exclusively on deserved moralistic punishment in order to examine how differing levels of severity influence perceived character traits. It should be noted, though, that the purpose of experiments 1A and 1B was to tease out the existence of an effect of deservedness and severity on perceived trust. To that end, both experiments used binary notions of deservedness and severity. Even though it could be argued that deservedness as a concept is not strictly binary, within most punishment scenarios the attributions of deservedness that perpetrator, victim, and punisher make are indeed often binary (irrespective of their actual and potentially contradicting directions).

Severity is different in this respect. Any given transgression or rule violation can theoretically be punished with very different levels of severity, and if it is the same type of

[11] Nevertheless, examining how punishment severity affected the outcome of third-party punishment was not the purpose of that study.

punishment, then those severity levels usually lie within a continuum. In this sense, the two levels of severity used in experiments 1A and 1B constitute a limitation, since the No Severity condition could ostensibly not be considered as a severity level. Nonetheless, the aim of this No Severity condition was, as mentioned earlier, to reveal whether there was an effect at all by allowing for clear comparison with the High Severity level. Furthermore, no-punishment conditions have been used previously in the literature, most notably in the study by Jordan and colleagues (2016)[12], where the researchers employed a similar experimental approach to the one used in experiments 1A and 1B, albeit with different conclusions. Namely, they found that participants entrusted larger average amounts to punishers over non-punishers (both deserved) in a trust game followed by a third-party punishment game. On the contrary, I found no statistically significant difference[13] between these two conditions (both in the trust game as well as in the explicit partner preference task). However, the differences in the methodology employed make the direct comparison of these conflicting results very difficult.[14] At the very least, I think the conflicting results grant an additional examination that incorporates key experimental design features of both studies.

Nevertheless, one of the main objectives of experiment 2 was, precisely, to tackle the potential problems a No Severity condition poses in terms of its direct comparability with High Severity conditions. Therefore, experiment 2 compared four levels of severity (of the same type of punishment), in order to assess their effect on perceived trustworthiness. This constitutes a novel methodological approach since third-party punishment games used thus far in the literature have been strictly binary in nature (punish or no punishment). This allows for a more nuanced

[12] Barclay (2006) also used no-punishment conditions, albeit in a public goods paradigm as opposed to a third-party punishment game.
[13] Although the direction was the same.
[14] Which include the use of the "strategy method", the difference in how the comprehension questions were operationalized, and more importantly, the fact that they did not manipulate the behavior of the observer / punisher in the TPP, while I did.

understanding of the signaling function of moralistic punishment and captures a critical component of punishment with regards to proportionality. It enables punishers to calibrate the exerted severity so that they can impose what they consider a punishment that adequately fits the crime[15], and more importantly, it allows for the finetuning of the signaling function of moralistic punishment (and more broadly of the expressivist function of punishment).

The results of experiment 2 describe an overall positive linear relationship between punishment severity and perceived trustworthiness. The differences in mean entrusted amounts between the lowest and highest punishment severity conditions were particularly stark, outlining a linear trend that could be explained in terms of signal strength. In other words, the higher the severity of the punishment, the stronger the signal emitted by the punisher, and hence the stronger the perceived trustworthiness. This theory aligns with the tenets of costly signaling theory, whereby the more excessive the signaling display and thus the implied cost incurred, the more difficult it is to fake the signal which makes it stronger by virtue of its inherent honesty (Bird, Smith, & Bird, 2001; McAndrew, 2018). A caveat to this interpretation is that the characteristics of the TPP in experiment 2 made the cost incurred to punish explicit. This is not necessarily a limitation since it does not hinder the honesty of the signal as a function of its cost, but it does represent a qualitative difference with most moralistic punishments that takes place in real life settings where the cost incurred must be inferred by observers dependent only on the severity exerted.

Within the constraining features of the third-party punishment and trust games used, the positive linear relationship observed does present a ceiling effect, whereby the difference in signaled trustworthiness between the medium-high and high severity levels is negligible. This lack

[15] In this experiment severity was manipulated but it does permit different degrees of punishment severity within TPP and opens the door for other researchers to use TPP beyond binary decision-making constraints.

of a difference can be explained via two different mechanisms that could work in tandem and reinforce each other. Firstly, a mechanism that sets an honesty threshold rendering additional severity and cost inconsequential in terms of the signal's strength. In other words, the increased severity of the high level compared to the medium-high level does not add to the signal's honesty (i.e. observers do not consider a high severity punisher to display a greater handicap than the medium-high severity punisher). Likewise, the severity and cost incurred by the medium-high severity punisher could be enough for the signal to be considered by observers as un-fakeable.

Secondly, a mechanism whereby additional severity partially signals moralization could work in detriment of the signal's honesty. According to Kreps and Monin (2014), moralizing arises when individuals are perceived as purposefully treating an issue as moral way beyond its pragmatic features. As I mentioned before, high severity punishers might be seen as taking an especially non-pragmatic course of action. By spending the entire amount of their endowment to exert punishment (taking away the sender's entire endowment), high severity punishers end up with no money in an attempt to signal how much they care about the injustice committed by the sender. A more pragmatic approach would arguably entail incurring a considerable, but crucially lower cost to teach the selfish sender a lesson while at the same time sending a strong enough signal about the moral norms the punisher cares about. Therefore, I posit that moralizing high severity punishers can be perceived as overtly inflexible and self-righteous (Kreps & Monin, 2014).

The forgiving hypothesis formulated as an explanation for the behavior observed in experiments 1A and 1B, however, does not seem to play a role in this case. Forgiving, then, would only have a positive relationship with signaled trustworthiness when it is total. That is, when forgiving entails no punishment at all. On the contrary, as the results of experiment 2 indicate, if forgiving is construed as reduced severity, it does not bestow on the punisher the augmented

perceived trustworthiness (or at the very least the reduced signal strength resulting from the diminished severity eclipses any trust gains obtained via forgiveness).[16]

The obvious limitation of the moralizing interpretation lies in the inability of experiment 2 to demonstrate larger reductions in perceived trust as severity increases. Given that the experiment only had four severity levels, and that the highest severity level was in essence determined by the maximum amount of money a punisher could pay to exert punishment, the potential degree of moralization that could be displayed was also limited. If, for example, a punisher could somehow use some of his own (not endowed) money to punish the sender with increasing severity (e.g. by reducing some of the sender's own money, or by imposing non-monetary penalties), then one would be able to assess if and by how much perceived trustworthiness decreased, lending additional credence to the moralization hypothesis.

One of the objectives of experiment 3 was to circumvent this limitation. By going beyond the monetary punishments available within typical third-party punishment games, experiment 3 allows for severity increases that could be considered more disproportionate (compared with the high severity condition of experiment 2) and that would correspondingly be interpreted as more moralizing. With regards to signaled trust, the results of experiment 3 do in fact indicate a reduction as severity increases beyond what would be considered proportional. This points to a possible moralization of disproportionately severe punishers as they engage in excessive displays of punishment (with their correspondingly excessive incurred costs). This same trend was observed for almost all character trait signals measured in experiment 3. More specifically, signal strength increased as a function of punishment severity up to the proportional punishment level (i.e. from disproportionately lenient to proportional), and then decreased as severity increased (i.e. from

[16] This also represents a promising line for future research, carefully calibrating degrees of forgiveness, both quantitatively and qualitatively, to examine its effects on signaled character traits.

proportional to disproportionately severe). This general trend points to the importance of proportionality, at least when it is calibrated via the punishment's severity, in determining the signaling of character traits, which will be discussed in greater detail later.

Nevertheless, the relationship between severity and signaled trustworthiness is no exception to this trend and appears to depend highly on the perceived proportionality of the punishment. This implies that the high severity condition in experiment 2 might not carry the same connotations of disproportionate severity that the highest severity conditions carried in experiment 3. This, in turn, implies that the general logic and mechanics of traditional third-party punishment games have some important limitations. Namely, that they might not allow for disproportionately severe displays of punishments and its corresponding signaling, despite the fact that these do occur in everyday life and have therefore neglected its effect in human cooperation. This issue is compounded by the fact that previous studies have never used different levels of severity in traditional TTP games, focusing exclusively on binary decision-making (punish or no punish).[17]

The limitations of third-party punishment and trust games in the study of character trait signaling extend beyond the aforementioned lack of severity levels. Firstly, the underlying obligation to share placed on the sender in the TPP game is not necessarily in line with normative codes of conduct or is necessarily considered a moral imperative. Failing to share money with a complete stranger one has never met, and knows nothing about, does not necessarily constitute a moral norm for everyone. More so when the amount in question is very small and has been previously gifted to the sender. Hence, it is possible that, at least for some observers, punishing for not sharing in this context is undeserved (or by the same token that failing to punish is appropriate),

[17] To my knowledge, disproportionately severe punishment has also been ignored in moralistic punishment studies employing public goods paradigms.

which could have significant impacts on the perceived trustworthiness of the punisher / non-punisher.

Secondly, the trust game paradigm allows for nothing except signaling and interpretations about trustworthiness. This might sound like an obvious observation given the name of the game, but it is still significant, especially when it is followed by the TPP. The only message a punisher can ostensibly convey in the TPP is whether or not she condemns not sharing (with the caveats mentioned above). Within the confines of the trust game, that not-sharing assessment must necessarily be translated into perceptions of trustworthiness. Assuming an observer truly believes not-sharing in the TPP is indicative of a negative character trait (e.g. selfishness), and further that punishing such an act is an honest signal of the moral norms the punisher believes in (i.e. that selfishness is wrong), he still has to translate that belief into an assessment of the punisher's trustworthiness. Moreover, this definition of trustworthiness is solely limited to trusting that the punisher will not behave in a selfish manner himself (selfish exclusively in terms of whether or not he will share in the trust game). In this sense, the trust game is especially well designed to capture signals of trustworthiness (in terms of sharing), but it does so at the expense of capturing other potential character traits. If, for example, punishing in the TPP signals the punisher's pragmatism or empathy, the trust game will not capture that signal. It will only translate low amounts of money (or no money at all) sent to the receiver as indicative of low trust.

On of the main objectives of experiment 3 was to broaden the range of character traits a moralistic punisher could signal. Therefore, the third-party punishment game paradigm used in the previous experiments had to be transcended. Likewise, the methodology employed to measure any potential signals had to be more flexible than the trust game, so that different types of signals could be captured. To that end, experiment 3 made use of vignettes that depicted scenarios closer to

everyday life occurrences and measured moralistic punishment signal content with questions about the punisher. Still, the scope of the signalling context (including setting, norm infringement and protagonists) as well as the set of signals was potentially infinite. Hence, it had to be constrained, focusing on the main signals that have been featured in extant moralistic punishment research (i.e. trustworthiness and dominance), with the support of a social perception framework that at the same time allowed for the structured exploration of additional signals (i.e. warmth, competence, potential social role and moral motives).

6.3. Moralistic Punishment Signals Beyond Trust

In experiment 3 I first looked at trust and warmth signaled via moralistic punishment. I found that proportionality had a very similar effect on both signaled trustworthiness and warmth. This reflects the centrality of trust to the warmth dimension, as predicted by the Stereotype Content Model. Furthermore, the overall trend points to a careful calibration between the severity of the exerted punishment and the strength of the signaled trust and warmth. More specifically, the relationship seems to be marked by a sort of goldilocks effect, where too much or too little severity has detrimental effects on the signal. As mentioned in the discussion of experiment 3, the underlying mechanisms for the improvement in signal from lenient to proportional, and the deterioration of the signal from proportional to severe, are likely due to signal strength and moralization respectively. In other words, disproportionately lenient moralistic punishments affect mostly the signal's honesty, where the low cost implied is a cue of potential false signaling. Disproportionately severe punishment, on the other hand, affects mostly the signal's content, where high incurred costs are indicative of other character traits besides trust or warmth (e.g. inflexibility, self-righteousness, strictness, hotheadedness). To be fair, lenient punishments can

arguably affect signal content (e.g. greediness, laziness), just as severe ones can affect the signal's honesty (e.g. an attempt at moral grandstanding), but I propose that these are secondary assessments that arise as a result of the deteriorated signal's honesty and content transformation respectively.

I found the same overall goldilocks trend for virtually all of the other signals. The results of competence and social role (both friendship and leadership) also speak to the careful calibration of proportionality for the sake of signal strength. As with trust and warmth, the mechanisms driving the inverse U-shaped function between severity and these character traits are tied to signal strength and moralization. The case of dominance is different because it increases in a linear fashion as a function of severity. Such distinctiveness reflects that dominance, as measured in experiment 3, represents a slightly different psychological construct from the one used in the Stereotype Content Model (SCM) and does not map onto the broader competence dimension as well as it does in the SCM. This is likely due to the particular operationalization of the construct in my experiment and the focus on moralistic punishment.

More importantly, the fact that as severity increased dominance also increased is of particular significance when analyzed together with the signaling of trust. Extant research has found evidence for the signaling of these two seemingly contradicting and opposed signals as a product of third-party punishing. One line of research claims that third-party punishers obtain reputational benefits in terms of increased trustworthiness in future interactions, thus compensating for the cost incurred to punish (Barclay, 2006; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015). The other line of research claims that third-party punishers compensate for the cost incurred to punish by signaling dominance to would be offenders with the ultimate goal of deterrence (Delton & Krasnow, 2017; Krasnow et al., 2016; J. W. Martin et al., 2019;

Pedersen, McAuliffe, Shah, et al., 2018). My findings indicate that these two signals are not mutually exclusive within third-party punishment and establishes a theoretical bridge between the two lines of research, as third-party punishment can signal trustworthiness and dominance depending on the proportionality of the punishment. The key then, lies in the proportionality of the punishment (at least in terms of its severity) to modify the signal's content. In order to effectively signal dominance (as a clear signal of the capability of retaliation preemptively directed at would be offenders to avoid potential mistreatment of the self) the punishment has to be perceived as disproportionately severe. Otherwise it will signal trustworthiness (when it is proportional) or be conceived as a dishonest signal (when it is disproportionately lenient).

I expected to find the same relationship between severity and dominance when examining signaled leadership. The research establishing a link between perceived dominance and leadership (and perceived trust and friendship) has been exclusively conducted using face evaluations (Laustsen & Petersen, 2015; Oosterhof & Todorov, 2008). To my knowledge, my research is the first to evaluate the link between perceived dominance (and trust) and perceived leadership (and friendship) using situational vignettes and moralistic punishment. Based on those previous studies, I hypothesized that if dominance displayed a positive linear relationship with severity (which it did), I would see a similar pattern between severity and leadership (which I did not). The inverse U-shaped relationship between severity and leadership mirrored the pattern observed with most other signals in experiment 3. Before I comment on the possible explanation for these results, I would like to point out that the relationship between disproportionately severe and disproportionately lenient punishment actually did follow the expected positive linear trend with regards to leadership (i.e. severe punishers generally perceived as better leaders than lenient ones),

as well as with regards to friendship (i.e. lenient punishers perceived more as potential friends than severe ones).

The observed disparity between my findings and previous research, besides the obvious methodological differences, likely stems from the conceptualizations of leadership used and the introduction of proportionality. The concept of leadership used by Laustsen and Petersen (2015) was heavily context dependent, characterizing a leader within a very particular hierarchical structure in the midst of a potentially life-threatening event. Specifically, participants in that study were presented with a scenario where they had to imagine they were on board a ship in the 18th century and were faced with either an attack by pirates or a storm that would put the voyage at risk. Then, from a set of dominant and non-dominant looking faces (extracted from the validated face trustworthiness vs face dominant images developed by Oosterhof and Todorov, 2008) participants had to choose a captain (reflecting choice of leader) and a cabin-mate (reflecting choice of friend).

Beyond the fact that this was a paired choice task, the qualities needed in a ship captain that can lead people aboard a ship faced with the mentioned threats are very particular. They are arguably similar leadership qualities to those a lot of people seek out in war-time presidents, military leaders, and probably some sports coaches, emergency care physicians, and managers in industries / cultures characterized by vertical organizational structures. These all likely fall within the directive, autocratic and transactional styles of leadership, which focus on behaviors related to giving detailed directions, expecting subordinates to follow those instructions, and making decisions with limited subordinate input (Lorinkova, Pearsall, & Sims, 2013). These leadership styles can be attractive in certain contexts (especially when resource availability is critically limited and/or role ambiguity is potentially dangerous) because it makes task accomplishment

easier for subordinates by providing them with specific, role-relevant directions and it helps them focus their efforts toward their individual tasks (Kahai, Sosik, & Avolio, 2004). Such leadership styles typically employ management-by-exception practices, whereby leaders take steps to follow closely any divergence from planned results and take any necessary actions to correct the situation (Bass, 1998). Severe third-party punishment has the ability to correct unwanted behavior, help reaffirm subordinate and leader roles (via dominance), and signal deterrence so that it does not happen again. Thus, one could argue that disproportionately severe moralistic punishment is a tool that characterizes a very specific concept of leadership, but it does not necessarily encapsulate all types of leadership.

Empowering, democratic and transformational leadership styles focus on stimulating change in subordinates' attitudes and values through strategies of empowerment, augmenting subordinate's self-efficacy beliefs and fostering the internalization of the leader's vision (Conger & Kanungo, 1998). These leadership styles, characterized by greater freedom to innovate and role flexibility, are more attractive when intrinsic motivation is important for goal achievement and typically lead to better long-term performance via improved team learning and coordination (Lorinkova et al., 2013). From the subordinate's point of view, this type of leadership would allow for greater freedom, increased flexibility, and an inherent motivation to perform and collaborate, all of which make it a more desirable leadership style as long as there in no imminent threat or contextual emergency on the horizon.

Given the information provided by the vignettes in experiment 3, an observer making a judgment about the leadership qualities of a moralistic punisher likely attributes greater leadership skills to a proportionate punisher over a disproportionately lenient punisher based on their ability to take action (in line with my observations of perceived competence which mirror the predictions

of the SCM). More importantly, since none of the vignettes describe an imminent environmental threat or contextual urgency, an observer would arguably not see the need for a directive / autocratic type of leader, opting instead for a fair and just proportionate punisher as a potential leader instead of an excessively punitive bully.

The underlying moral motives of disproportionate vs proportionate punishers provide support for this explanation, as proportionate punishers were deemed to be driven more by genuine moral concerns compared with disproportionately severe ones. The findings with regards to moral motives, though, offer preliminary evidence of a potentially broader signaling aspect of proportionality.

6.5. Moral Motives and Proportionality

Just as with warmth, competence, leadership and friendship, punishment proportionality (in terms of severity), had a significant and considerable effect on the moral motives attributed to the punisher. Proportionate punishers were hailed as being more driven by true moral principles and less by self-interest compared with disproportionate punishers at both ends of the severity spectrum. This again, provides evidence for the content of the signal molded in response to the punisher's behavior, with lenient and severe punishments serving as cues of underlying self-interests driving behavior. The fact that proportionate punishers are seen as clearly acting on moral principles (while disproportionate punishers are not) is particularly interesting as it implies that proportionality provides powerful cues about morality. Whether this relationship between proportionality and morality drives other character trait perceptions such as warmth is certainly interesting, but what really seems to stand out is the apparent tight link between proportionality and attributions of morality. Perhaps more interesting is when certain positive or desirable

character traits, such as competence, are also linked to proportionality without necessarily having to go through attributions of morality. In those cases, it seems even more evident that being a proportionate punisher is associated with being good in a broader sense of the word.

Taken together, the findings on experiment 3 suggest that proportionality carries key information about the punisher. More importantly, it seems like proportionality is interpreted as an overall predictor of a punisher's goodness. Goodness in this case refers to a broad set of character traits that exemplify virtuous action, ability and character. Proportionate punishers were perceived as both warmer and more competent than disproportionate punishers, which are orthogonal person perception dimensions (according to the Stereotype Content Model) but both are arguably desirable character traits. In fact, the SCM predicts that the elicited emotional reaction for individuals high in both warmth and competence is admiration (Cuddy et al., 2008). Proportionate punishers were also rated as both potential good leaders and potential good friends. In many contexts and situations, such as when impartiality and dominance become key features of great leadership, a good leader will show qualities that don't necessarily make for a good friend and vice versa (Boehm, 2000; von Rueden et al., 2014). However, proportionate punishers were hailed as fit candidates for both social roles, revealing the overall positive connotation of proportionality.

Nevertheless, the extent to which a proportionate moralistic punisher signals positive and desirable character traits is highly dependent on the context. Though proportionality can be assessed beyond the characteristics of the scenarios used in experiment 3, its consequences cannot. As I showed in the analysis of experiment 3, the signal content varied importantly as a function of the scenario (see page 86 on notes on Vegan scenario). Punishment in different social spheres relies on different sets of assumptions and values shared by actors and observers, that have subtle

but significant impacts on the inferences we make. What is deemed proportional in one context might be deemed highly disproportionate in another, and the corresponding positive content signaled in the former can become negative in the latter.

In this sense, proportionality seems to be closely aligned with the Aristotelian notion of intermediate ethical virtues. Aristotle argues that all virtues can be conceptualized as being condition intermediate between two other states, one involving excess, and the other deficiency, but that the so-called golden mean is always determined by taking into account the particular circumstances of the individual (Kraut, 2018). According to this notion, whenever a person chooses to perform a virtuous act, he can be described as aiming at an act that is in some way or other intermediate between the aforementioned alternatives that he rejects (Kraut, 2018). I argue that observers consider the intermediate condition of a moralistic punishment (i.e. its proportionality) to assess how virtuous of an act it is, and by extension, to evaluate the virtuosity of the punisher. As I said above, this depends on the context and is associated set of assumptions held by moralistic punishers as well as observers. Nonetheless, when a punishment is deemed proportional, it signals the virtuosity of the punisher, which manifest more specifically under several socially desirable person perceptions or character traits (e.g. trustworthiness, warmth, competence, leadership, friendship).

This conceptualization of proportionality fits with the notion that people are in essence naïve virtue theorists and the person-centered approach to moral judgments (Uhlmann et al., 2015), whereby morality is best understood at the level of persons rather than acts. More importantly however, the intermediate condition of an act would be a fundamental aspect we all evaluate in assessing the character traits of the agent involved. The results of my studies seem to suggest this is indeed the case with moralistic punishment. Nonetheless, it is entirely possible that the perceived

intermediate condition of any act can critically determine its perceived morality and influence the perceived character traits of the people involved. This is particularly significant because the study of moral judgment in psychology has traditionally focused on moral value trade-offs which typically leave no room for the study of moral reasoning of intermediate courses of action among excessive and deficient alternatives. If behaviors that are statistically rare or otherwise extreme are perceived as highly informative about character traits (e.g. Ditto & Jemmott, 1989; McKenzie & Mikkelsen, 2007), then behaviors that are perceived as condition intermediate (or proportional) might also be highly informative about character traits (in particular of the individual's goodness) and should be taken into consideration in the moral reasoning research in general and in a person-centered approach to moral judgment in particular.

7. References

Axelrod, R. (2006). The evolution of cooperation. 1984 New York. *NY: Basic Books*.

Balafoutas, L., & Nikiforakis, N. (2012). *Norm enforcement in the city: A natural field experiment*. Retrieved from Faculty of Economics and Statistics, University of Innsbruck website: https://econpapers.repec.org/RePEc:inn:wpaper:2012-12

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, Vol. 137, pp. 594–615. https://doi.org/10.1037/a0023489

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. *Journal of Risk and Uncertainty*, *7*(1), 17–33. https://doi.org/10.1007/BF01065312

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85. https://doi.org/10.1016/j.obhdp.2004.03.003

Bass, B. M. (1998). Transformational leadership: industrial, military, and educational impact, 1998. *Mahwah, NJ: Laurence Erlbaum*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedau, H. A., & Kelly, E. (2017). Punishment. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Bentham, J. (1780). *An Introduction to the Principles of Morals and Legislation* (Vol. 45). Dover Publications.

Bird, R. B., Smith, E. A., & Bird, D. W. (2001). The hunting handicap: Costly signaling in human foraging strategies. *Behavioral Ecology and Sociobiology*, *50*(1), 9–19. https://doi.org/10.1007/s002650100338

Boehm, C. (2000). Conflict and the evolution of social control. *Journal of Consciousness Studies*, *7*(1–2), 79–101.

Bone, J. E., McAuliffe, K., & Raihani, N. J. (2016). Exploring the motivations for punishment: Framing and country-level effects. *PLoS ONE*, *11*(8), 1–14. https://doi.org/10.1371/journal.pone.0159769

Boyd, R., & Mathew, S. (2015). Third-party monitoring and sanctions aid the evolution of language. *Evolution and Human Behavior*, *36*(6), 475–479. https://doi.org/10.1016/j.evolhumbehav.2015.06.002

Brandt, M. J., Wetherell, G., & Henry, P. J. (2015). Changes in Income Predict Change in Social Trust: A Longitudinal Analysis. *Political Psychology*, *36*(6), 761–768. https://doi.org/10.1111/pops.12228

Buckingham, G., DeBruine, L. M., Little, A. C., Welling, L. L. M., Conway, C. A., Tiddeman, B. P., & Jones, B. C. (2006). Visual adaptation to masculine and feminine faces influences generalized preferences and perceptions of trustworthiness. *Evolution and Human Behavior*, *27*(5), 381–389.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in Cognitive Sciences, 7*(5), 225–231. https://doi.org/10.1016/S1364-6613(03)00094-9

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. https://doi.org/10.1016/j.jesp.2005.06.007

Carlsmith, K. M., & Darley, J. M. (2008). Psychological Aspects of Retributive Justice. *Advances in Experimental Social Psychology*, *40*(07), 193–236. https://doi.org/10.1016/S0065-2601(07)00004-4

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). *Why Do We Punish ? Deterrence and Just Deserts as Motives for Punishment*. *83*(2), 284–299. https://doi.org/10.1037//0022-3514.83.2.284

Chance, P. (2013). *Learning and Behavior*. Retrieved from https://books.google.com.co/books?id=dYzsXLN4je0C

Conger, J. A., & Kanungo, R. N. (1998). *Charismatic leadership in organizations*. Sage Publications.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1–3. https://doi.org/10.1038/s41562-017-0213-3

Crombag, H., Rassin, E., & Horselenberg, R. (2003). On vengeance. *Psychology, Crime & Law*, *9*(4), 333–344. https://doi.org/10.1080/1068316031000068647

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. *Advances in Experimental Social Psychology*, *40*(07), 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). *Incapacitation and just deserts as motives for punishment*. *24*(6), 659–683.

Darley, J. M., & Pittman, T. S. (2003). The Psychology of Compensatory and Retributive Justice. *Personality and Social Psychology Review*, *7*(4), 324–336. https://doi.org/10.1207/S15327957PSPR0704_05

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734–743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

DeScioli, P., & Kurzban, R. (2009). The alliance hypothesis for human friendship. *PloS One, 4*(6), e5802.

DeScioli, P., & KURzBAN, R. (2012). The company you keep: Friendship decisions from a functional perspective. *Social Judgment and Decision Making*, 209–225.

Ditto, P. H., & Jemmott, J. B. (1989). From rarity to evaluative extremity: Effects of prevalence information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology*, Vol. 57, pp. 16–26. https://doi.org/10.1037/0022-3514.57.1.16

Dreber, A., & Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *Behavioral and Brain Sciences, 35*(1), 24. https://doi.org/DOI: 10.1017/S0140525X11001221

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature, 452*(7185), 348–351. https://doi.org/10.1038/nature06723

Duff, A., & Hoskins, Z. (2018). Legal Punishment. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University.

Durkheim, E., & Simpson, G. (1933). *Emile Durkheim on The division of labor in society;* New

York: Macmillan.
Ellsworth, P. C., & Ross, L. (1983). Public Opinion and Capital Punishment: A Close Examination of the Views of Abolitionists and Retentionists. *Crime & Delinquency*, *29*(1), 116–169. https://doi.org/10.1177/001112878302900105
Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7
Evans, I. M., Galyer, K. T., & Smith, K. J. H. (2001). Children's perceptions of unfair reward and punishment. *Journal of Genetic Psychology*, *162*(2), 212–227. https://doi.org/10.1080/00221320109597962
Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772–787. https://doi.org/10.1037/xge0000165
Feather, N. T., & Sherman, R. (2003). Envy, Resentment, Schadenfreude, and Sympathy: Reactions to Deserved and Undeserved Achievement and Subsequent Failure. *Personality and Social Psychology Bulletin*, *28*(7), 953–961. https://doi.org/10.1177/01467202028007008
Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4
Fehr, E., & Ga, S. (2002). *Altruistic punishment in humans.pdf*. 137–140.
Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *The American Economic Review*, *90*(4), 980–994. https://doi.org/10.1016/j.jmoneco.2005.10.016
Feinberg, J. (1965). The Expressive Function of Punishment. *The Monist*, *49*(3).
Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889
French, P. A. (2001). *The Virtues of Vengeance*.
Frey, K. S., Pearson, C. R., & Cohen, D. (2015). Revenge is seductive, if not sweet: Why friends matter for prevention efforts. *Journal of Applied Developmental Psychology*, *37*, 25–35. https://doi.org/10.1016/j.appdev.2014.08.002
Frijda, N. H. (2004). Emotions and action. *Feelings and Emotions: The Amsterdam Symposium*, 158–173.
Fudenberg, D., Rand, D. G., Dreber, A., Ellingsen, T., & Nowak, M. A. (2009). Positive Interactions Promote Public Cooperation. *Science*, (325), 1272–1275.
Geeraets, V. (2018). Two Mistakes about the Concept of Punishment. *Criminal Justice Ethics*, *37*(1), 21–35. https://doi.org/10.1080/0731129X.2018.1441227
Goodwin, G. P. (2015). Moral Character in Person Perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709
Goodwin, G. P., & Benforado, A. (2015). Judging the goring ox: Retribution directed toward animals. *Cognitive Science*, *39*(3), 619–646. https://doi.org/10.1111/cogs.12175
Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168. https://doi.org/10.1037/a0034726
Gordon, D. S., & Lea, S. E. G. (2016). Who punishes? The status of the punishers affects the perceived success of, and indirect benefits from, "'moralistic'" punishment. *Evolutionary Psychology*, *14*(3), 1–14. https://doi.org/10.1177/1474704916658042

Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS ONE*, *9*(10), 1–10. https://doi.org/10.1371/journal.pone.0110045

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546.

Griskevicius, V., Tybur, J. M., Sundie, J. M., Cialdini, R. B., Miller, G. F., & Kenrick, D. T. (2007). *Blatant Benevolence and Conspicuous Consumption : When Romantic Motives Elicit Strategic Costly Signals*. *93*(1), 85–102. https://doi.org/10.1037/0022-3514.93.1.85

Gupta, G., Mahmud, M., Maitra, P., Mitra, S., & Neelim, A. (2018). Religion, minority status, and trust: Evidence from a field experiment. *Journal of Economic Behavior & Organization*, *146*, 180–205.

Haidt, J., & Baron, J. (1996). Social Roles and the Moral Judgment of Acts and Omissions. *European Journal of Soical Psychology*, *26*(2), 201–218.

Hampton, J. (1984). The moral education theory of punishment. *Philosophy & Public Affairs*, *13*(3), 208–238. https://doi.org/10.2307/2265412

Hardy, C., & Van Vugt, M. (2006). Giving for glory in social dilemmas: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*, 1402–1413.

Hart, H. L. A. (1968). *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hoffman, M., Hilbe, C., & Nowak, M. A. (2018). The signal-burying game can explain why we obscure positive traits and good deeds. *Nature Human Behaviour*, *2*(6), 397–404. https://doi.org/10.1038/s41562-018-0354-z

Hofmann, W, Brandt, M., Wisneski, D., & Rockenbach, B. (2018). Moral Punihsment in Everyday Life. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167204271575

Hofmann, Wilhelm, Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral Punishment in Everyday Life. *Personality and Social Psychology Bulletin*, 0146167218775075. https://doi.org/10.1177/0146167218775075

Jensen, N. H., & Petersen, M. B. (2011). To defer or to stand up? How offender formidability affects third party moral outrage. *Evolutionary Psychology*, *9*(1), 147470491100900130.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1111/j.1558-5646.2011.01232.x

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

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.

Kahai, S. S., Sosik, J. J., & Avolio, B. J. (2004). Effects of participative and directive leadership in electronic groups. *Group & Organization Management*, *29*(1), 67–105.

Kahan, D. M. (1996). What do alternative sanctions mean? *The University of Chicago Law Review*, *63*(2), 591–653. https://doi.org/10.2307/1600237

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian"

judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. https://doi.org/10.1016/j.cognition.2014.10.005

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the Assumptions of Economics. *The Journal of Business*, Vol. 59, p. S285. https://doi.org/10.1086/296367

Kant, I., & Hastie, W. (2002). *The Philosophy of Law: An Exposition of the Fundamental Principles of Jurisprudence as the Science of Right*. Retrieved from https://books.google.com.co/books?id=_tjsapYzAgAC

Keijzer, M. A., & Corten, R. (2016). *In status we trust: A vignette experiment on socioeconomic status and reputation explaining interpersonal trust in peer-to-peer markets*. 1–45.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192–238.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. *Psychological Science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Kraut, R. (2018). Aristotle's Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Kreps, T. A., & Monin, B. (2014). Core Values Versus Common Sense: Consequentialist Views Appear Less Rooted in Morality. *Personality and Social Psychology Bulletin*, *40*(11), 1529–1542. https://doi.org/10.1177/0146167214551154

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When It's Bad to Be Friendly and Smart: The Desirability of Sociability and Competence Depends on Morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272–1290. https://doi.org/10.1177/0146167216655984

Laustsen, L., & Petersen, M. B. (2015). Does a competent leader make a good friend? Conflict, ideology and the psychologies of friendship and followership. *Evolution and Human Behavior*, *36*(4), 286–293. https://doi.org/10.1016/j.evolhumbehav.2015.01.001

Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin*, *24*(6), 563–574. https://doi.org/10.1177/0146167298246001

Levitt, S. D. (2004). Understanding Why Crime Fell in the 1990s: Four Factors that Explain the Decline and Six that Do Not . *Journal of Economic Perspectives*, *18*(1), 163–190. Retrieved from http://www.aeaweb.org/articles?id=10.1257/089533004773563485

Lorinkova, N. M., Pearsall, M. J., & Sims, H. P. (2013). Examining the differential longitudinal performance of directive versus empowering leadership in teams. *Academy of Management Journal*, *56*(2), 573–596. https://doi.org/10.5465/amj.2011.0132

Lount, R. B., & Pettit, N. C. (2012). The social context of trust: The role of status. *Organizational Behavior and Human Decision Processes*, *117*(1), 15–23. https://doi.org/10.1016/j.obhdp.2011.07.005

Martin, J., Jordan, J., Rand, D. G., & Cushman, F. (2018). When do we punish people who don't? *SSRN Electronic Journal*, *193*(July), 104040. https://doi.org/10.2139/ssrn.3080990

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition*, *193*, 104040.

McAndrew, F. T. (2018). Costly signaling theory. *Encyclopedia of Evolutionary Psychological*

Science, 1–8.
McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160
McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61. https://doi.org/10.1016/j.cogpsych.2006.04.004
Moore, M. S. (2010). *Placing Blame: A Theory of the Criminal Law*. https://doi.org/10.1093/acprof:oso/9780199599493.001.0001
Morris, H. (1981). A Paternalistic Theory of Punishment. *American Philosophical Quarterly*, *18*(4).
Nelissen, R. M. A. (2008). The price you pay: cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001
Nozick, R. (1981). *Philosophical Explanations* (Vol. 92). Harvard University Press.
Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, *30*(2), 114–123. https://doi.org/10.1016/j.evolhumbehav.2008.09.004
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.
Osgood, J. M. (2017). Is revenge about retributive justice, deterring harm, or both? *Social and Personality Psychology Compass*, *11*(1), 1–15. https://doi.org/10.1111/spc3.12296
Osgood, J. M., & Muraven, M. (2015). Does counting to ten increase or decrease aggression? The role of state self-control (ego-depletion) and consequences. *Journal of Applied Social Psychology*, *46*(2), 105–113. https://doi.org/10.1111/jasp.12334
Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants With and Without a Sword: Self-Governance is Possible. *The American Political Science Review*, *86*(2), 404–417. https://doi.org/10.2307/1964229
Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? a closer look. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758). https://doi.org/10.1098/rspb.2012.2723
Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology General*. https://doi.org/10.1037/xge0000410
Pedersen, E. J., McAuliffe, W. H. B., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2018). *When and why do third parties punish outside of the lab? A cross-cultural recall study*.
Price, M. E., & Van Vugt, M. (2015). The service-for-prestige theory of leader–follower relations: A review of the evolutionary psychology and anthropology literatures. *Biological Foundations of Organizational Behavior*, 397–477.
Qi, Y., Li, Q., & Du, F. (2018). Are rich people perceived as more trustworthy? Perceived socioeconomic status modulates judgments of trustworthiness and trust behavior based on facial appearance. *Frontiers in Psychology*, *9*(APR), 1–9. https://doi.org/10.3389/fpsyg.2018.00512
Rabellino, D., Morese, R., Ciaramidaro, A., Bara, B. G., & Bosco, F. M. (2016). Third-party punishment: altruistic and anti-social behaviours in in-group and out-group settings. *Journal of Cognitive Psychology*, *28*(4), 486–495. https://doi.org/10.1080/20445911.2016.1138961

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., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, *265*(4), 624–632. https://doi.org/10.1016/j.jtbi.2010.06.010

Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications*, *2*(1). https://doi.org/10.1038/ncomms1442

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

Roos, P., Gelfand, M., Nau, D., & Carr, R. (2014). High strength-of-ties and low mobility enable the evolution of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776), 20132661.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184. https://doi.org/10.1023/A:1019923923537

Santos, M. dos, Rankin, D. J., & Wedekind, C. (2010). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371–377.

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005

Sefton, M., Shupp, R., & Walker, J. (2007). THE EFFECT OF REWARDS AND SANCTIONS IN PROVISION OF PUBLIC GOODS. *Economic Inquiry*, *45*(4), 671–690. Retrieved from https://econpapers.repec.org/RePEc:bla:ecinqu:v:45:y:2007:i:4:p:671-690

Shoemaker, D., & Vargas, M. (2019). Moral torch fishing: A signaling theory of blame. *Nous*, (May), 1–22. https://doi.org/10.1111/nous.12316

Singer, R. G. (1979). *Just deserts: sentencing based on equality & desert*. Retrieved from https://books.google.com.co/books?id=MFIqAQAAMAAJ

Sloan, J. J., & Miller, J. L. (1990). Just Deserts, The Severity Of Punishment And Judicial Sentencing Decisions. *Criminal Justice Policy Review*, *4*(1), 19–38. https://doi.org/10.1177/088740349000400102

Snyder, J. K., Fessler, D. M. T., Tiokhin, L., Frederick, D. A., Lee, S. W., & Navarrete, C. D. (2011). Trade-offs in a dangerous world: Women's fear of crime predicts preferences for aggressive and formidable mates. *Evolution and Human Behavior*, *32*(2), 127–137.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*(12), 2297–2306. https://doi.org/10.1037/0022-3514.37.12.2297

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

Szolnoki, A., & Perc, M. (2017). Second-order free-riding on antisocial punishment restores the effectiveness of prosocial punishment. *Physical Review X*, *7*(4), 1–11. https://doi.org/10.1103/PhysRevX.7.041027

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*(6), 1249–1254. https://doi.org/10.1016/j.jesp.2011.05.010

Tooby, J., & Cosmides, L. (1996). Friendship and the banker's paradox: Other pathways to the evolution of adaptations for altruism. *Proceedings of the British Academy*, *88*, 119–143. https://doi.org/10.1002/(SICI)1520-6300(1998)10:5<681::AID-AJHB16>3.3.CO;2-I

Tooby, J., & Cosmides, L. (2010). Groups in mind: The coalitional roots of war and morality. *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, 91–234.

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive adaptations for n-person exchange: the evolutionary roots of organizational behavior. *Managerial and Decision Economics*, *27*(2-3), 103–129.

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy and Public Affairs*, Vol. 44. https://doi.org/10.1111/papa.12075

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Van den Haag, E. (1975). *Punishing criminals : concerning a very old and painful question*. Retrieved from http://books.google.com/books?id=cMAfAAAAIAAJ

van Dijk, W. W., Ouwerkerk, J. W., Goslinga, S., & Nieweg, M. (2005). Deservingness and Schadenfreude. *Cognition and Emotion*, *19*(6), 933–939. https://doi.org/10.1080/02699930541000066

Vidmar, N. (2001). Retribution and revenge. In *Handbook of justice research in law.* (pp. 31–63). Dordrecht, Netherlands: Kluwer Academic Publishers.

von Hirsch, A. (1992). Proportionality in the Philosophy of Punishment. *Crime and Justice*, *16*, 55–98. Retrieved from http://www.jstor.org/stable/1147561

von Rueden, C., Gurven, M., Kaplan, H., & Stieglitz, J. (2014). Leadership in an egalitarian society. *Human Nature*, *25*(4), 538–566.

Walker, L. J., & Hennig, K. H. (2004). Differing Conceptions of Moral Exemplarity: Just, Brave, and Caring. *Journal of Personality and Social Psychology*, Vol. 86, pp. 629–647. https://doi.org/10.1037/0022-3514.86.4.629

Wilson, J. Q. (1983). *Thinking about crime*. Retrieved from https://books.google.com.co/books?id=jPa-r5lA7egC

Yamagishi, T. (1988). The Provision of a Sanctioning System in the United States and Japan. *Social Psychology Quarterly*, *51*(3), 265–271. https://doi.org/10.2307/2786924

Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214.

Zaibert, L. (2006). *Punishment and Retribution*. Retrieved from https://books.google.com.co/books?id=5go7I84EGI0C

Appendix A

Statistical Analyses – Experiments 1A and 1B

Sample size justification for experiments 1A and 1B

Given that these studies aimed to test a rather novel hypothesis, and since I was mainly interested in the interaction effect between deservedness and severity, I conducted an a priori power analysis with a minimum detectable effect of 0.3, an alpha level of 0.05 and a 0.9 power aiming to reasonably reduce Type 2 errors. This rendered a total sample size of N = 480. Hence total number of participants to collect data from was rounded to 500 (experiments 1A) and 600 (experiment 1B) to account for potential attrition, task time limits (participants taking too long or too little on the task) and responses with mistakes (e.g. wrong Prolific ID).

Table A1

Pairwise Comparisons of Partner Preference Choice Frequency – Experiment 1A

Proposed Label based on Condition	Sadist	Forgiver	Fair Vigilante
Forgiver	$p < 0.0001$	-	-
Fair Vigilante	$p < 0.0001$	$p = 0.082$	-
Fair Watcher	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$

Table A2

Odds Ratios for Partner Preference Choice Frequency – Experiment 1A

Table A3

Multinomial Model of Partner Preference Dependent of Previous Task Condition – Experiment 1A

	Deserved High Severity		**Deserved No Severity**		**Undeserved High Severity**		**Undeserved No Severity**	
Player	OR	SE	OR	SE	OR	SE	OR	SE
Player 15	-76.19***	0.35	45.38	0.52	12.00	0.53	-37.78	0.64
Player 6	-45.23*	0.26	110.70	0.37	70.43	0.37	21.74	0.40
Player 9	19.05	0.21	90.61*	0.31	104.40*	0.30	58.67	0.32

Note. OR = odds ratio. SE = standard error.

* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Proposed Label based on Condition	**OR**	**95% CI**	***p*-value**
Fair Watcher	9.44	5.88-15.50	< 0.0001
Fair Vigilante	1.71	1.06-2.81	0.02
Forgiver	0.58	0.36-0.94	0.02
Sadist	0.11	0.06-0.17	< 0.0001

Note. OR = Odds Ratio

CI = Confidence Interval

Table A4

Pairwise Comparisons of Partner Preference Choice Frequency – Experiment 1B

Proposed Label based on Condition	Sadist	Forgiver	Fair Vigilante
Forgiver	$p < 0.0001$	-	-
Fair Vigilante	$p < 0.0001$	$p = 0.082$	-
Fair Watcher	$p < 0.0001$	$p < 0.0001$	$p < 0.0001$

Table A5

Odds Ratios for Partner Preference Choice Frequency – Experiment 1B

Proposed Label based on Condition	**OR**	**95% CI**	***p*-value**
Fair Watcher	8.58	5.48-13.79	< 0.0001
Fair Vigilante	1.71	1.06-2.81	0.02
Forgiver	0.37	0.23-0.59	< 0.0001
Sadist	0.12	0.07-0.18	< 0.0001

Note. OR = Odds Ratio

CI = Confidence Interval

Table A6

Multinomial Model of Partner Preference Dependent of Previous Task Condition - Experiment 1B

	Deserved High Severity		**Deserved No Severity**		**Undeserved High Severity**		**Undeserved No Severity**	
Player	OR	SE	OR	SE	OR	SE	OR	SE
Player 9	117.86***	0.23	-34.88	0.30	-20.28	0.31	-8.20	0.31
Player 12	64.29*	0.24	-27.81	0.32	-61.56**	0.35	-52.66*	0.35
Player 15	-60.71***	0.36	-76.32*	0.63	-66.51	0.59	-22.22	0.50

Note. OR = odds ratio. SE = standard error.

* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Appendix B

Statistical Analyses - Experiment 2

Sample size justification for experiment 2

Given the within-subjects design of this study, I conducted an a priori power analysis for between effects with a minimum detectable effect of 0.2, an alpha level of 0.05 and a 0.9 power aiming to reasonably reduce Type 2 errors. This rendered a total sample size of N = 358. Hence total number of participants to collect data from was rounded to 401 to account for potential attrition, task time limits (participants taking too long or too little on the task) and responses with mistakes (e.g. wrong Prolific ID).

Table B1

Mixed Effects Model for Entrusted Amounts – Experiment 2

	Amount		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	12.38	11.76 – 13.01	**<0.001**
Low_Severity	-2.44	-3.01 – -1.88	**<0.001**
Medium_High_Severity	-0.43	-1.00 – 0.13	0.134
Medium_Low_Severity	-1.41	-1.98 – -0.84	**<0.001**
Random Effects			
σ^2	16.45		
$\tau_{00\ Prolific\_ID}$	23.47		
ICC	0.59		
$N_{Prolific\_ID}$	395		
Note. Observations	1580		
Marginal R^2 / Conditional R^2	0.022 / 0.597		

Table B2

Mixed Effects Model for Entrusted Amounts with Demographics – Experiment 2

	Amount		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	18.59	5.24 – 31.95	**0.006**
Low_Severity	-2.44	-3.01 – -1.88	**<0.001**
Medium_High_Severity	-0.43	-1.00 – 0.13	0.134
Medium_Low_Severity	-1.41	-1.98 – -0.84	**<0.001**
18	-3.79	-14.56 – 6.97	0.490
19	-3.92	-14.49 – 6.64	0.467
1§	-6.47	-21.02 – 8.09	0.384
20	-3.77	-14.38 – 6.83	0.486
21	-2.83	-13.48 – 7.81	0.602
22	-5.09	-15.64 – 5.46	0.345
23	-6.39	-16.95 – 4.17	0.236
24	-3.56	-14.06 – 6.94	0.507
25	-4.17	-14.98 – 6.65	0.450
26	-3.78	-14.41 – 6.84	0.485
27	-1.21	-11.73 – 9.30	0.821
28	-3.07	-13.62 – 7.49	0.569
29	-4.14	-14.74 – 6.46	0.444
30	-4.35	-15.00 – 6.29	0.423
31	-3.82	-14.93 – 7.29	0.501
32	-2.88	-13.60 – 7.84	0.598
33	-1.83	-12.50 – 8.84	0.737
34	-2.59	-13.96 – 8.79	0.656
35	-3.94	-14.94 – 7.06	0.483

36	-3.68	-14.73 – 7.36	0.513
37	-2.74	-13.80 – 8.31	0.626
38	-2.00	-13.58 – 9.57	0.734
39	-3.60	-15.50 – 8.30	0.553
40	-8.15	-19.35 – 3.05	0.154
41	-3.84	-15.10 – 7.41	0.503
42	2.99	-8.97 – 14.96	0.624
43	-8.42	-19.67 – 2.84	0.143
44	-4.08	-15.00 – 6.84	0.464
45	4.58	-7.39 – 16.55	0.453
46	-1.31	-12.84 – 10.21	0.823
47	-0.56	-13.23 – 12.12	0.931
48	-6.27	-19.12 – 6.58	0.339
49	-2.22	-13.77 – 9.33	0.706
50	9.93	-5.55 – 25.41	0.209
52	2.15	-10.54 – 14.84	0.740
53	-8.72	-21.89 – 4.45	0.194
54	-4.89	-17.69 – 7.90	0.453
55	2.91	-9.78 – 15.61	0.653
56	-4.04	-15.32 – 7.25	0.483
57	-3.88	-16.45 – 8.68	0.544
59	-13.24	-27.94 – 1.45	0.077
60	-6.99	-21.69 – 7.71	0.351
62	-0.05	-12.75 – 12.66	0.994
63	-5.77	-17.75 – 6.22	0.346

64	-4.17	-18.84 – 10.51	0.578
68	-4.42	-18.85 – 10.01	0.548
70	-3.89	-18.46 – 10.68	0.601
Male	1.29	0.20 – 2.39	**0.021**
Other	-1.62	-6.16 – 2.92	0.484
No Qualifications	2.14	-1.13 – 5.41	0.199
O-Level / GCSE or similar	-1.43	-3.84 – 0.97	0.243
Postgraduate degree or similar	0.02	-1.69 – 1.73	0.981
Undergraduate degree or similar	-0.33	-1.73 – 1.07	0.645
£10,001 - £15,000	-2.84	-11.61 – 5.92	0.525
£15,001 - £25,000	-2.85	-11.55 – 5.86	0.522
£25,001 - £35,000	-2.77	-11.43 – 5.89	0.530
£35,001 - £50,000	-4.32	-13.15 – 4.51	0.338
£5,000 - £10,000	-2.04	-10.76 – 6.68	0.647
£50,001 - £65,000	-1.16	-10.15 – 7.84	0.801
£65,001 - £80,000	-4.73	-14.34 – 4.88	0.335
above £80,001	-9.97	-20.28 – 0.33	0.058
under £5,000	-3.31	-12.00 – 5.39	0.456

Random Effects

σ^2	16.45
$\tau_{00\ Prolific\_ID}$	22.73
ICC	0.58
$N_{\ Prolific\_ID}$	395
Observations	1580
Marginal R^2 / Conditional R^2	0.130 / 0.635

Appendix C

Experiment 2 Materials

Low Severity Condition

• Player 1 shared NONE of his 20 pence with Player 2.

• Player 3 then decided to spend 3 pence to punish Player 1 (reducing Player 1's amount by 5 pence)

Medium-Low Severity Condition

• Player 1 shared NONE of his 20 pence with Player 2.

• Player 3 then decided to spend 7 pence to punish Player 1 (reducing Player 1's amount by 10 pence)

Medium-High Severity Condition

• Player 1 shared NONE of his 20 pence with Player 2.

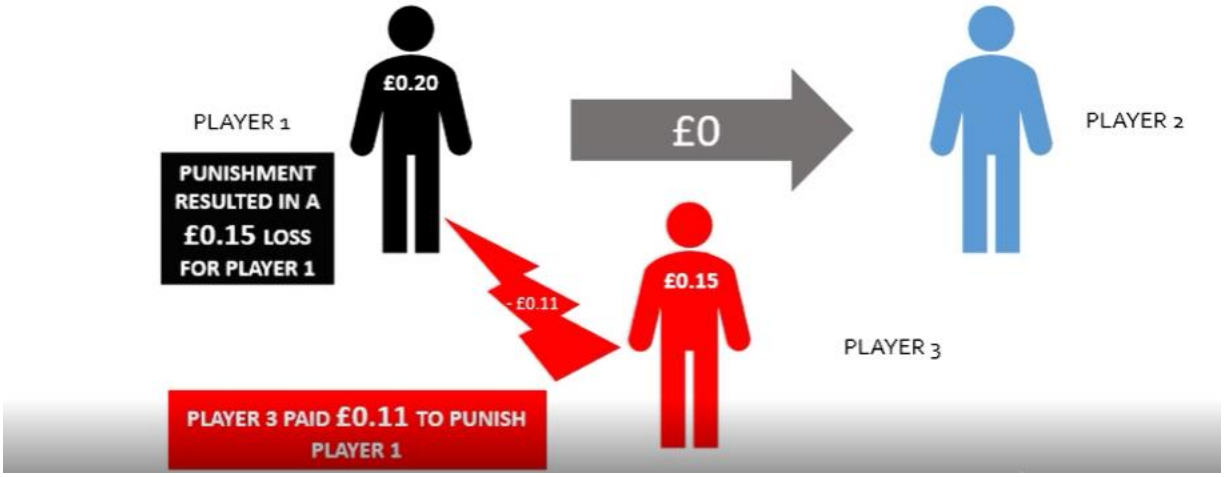
• Player 3 then decided to spend 11 pence to punish Player 1 (reducing Player 1's amount by 15 pence)

High Severity Condition

• Player 1 shared NONE of his 20 pence with Player 2.

• Player 3 then decided to spend 15 pence to punish Player 1 (reducing Player 1's amount by 20 pence)

Sample Illustration
Medium-High Severity Condition



Note. All illustrations were animated in MS Powerpoint and presented as short videos.

Appendix D

Vignettes – Experiment 3

STUTTER

Lenient

One day before work, Fred stops by a coffee shop close to the office where he works and sees one of his colleagues just finishing posting copies of a cartoon around the shop that makes fun of another man at the company who stutters. Fred's colleague looks angry and like someone you wouldn't want to mess with. Fred makes eye contact with his colleague and shakes his head in disapproval.

Proportional

One day before work, Fred stops by a coffee shop close to the office where he works and sees one of his colleagues just finishing posting copies of a cartoon around the shop that makes fun of another man at the company who stutters. Fred's colleague looks angry and like someone you wouldn't want to mess with. Fred tells his colleague, in a loud enough voice so that half the people in the coffee shop can hear, that making fun of someone's stutter and posting an offensive cartoon is not OK and tears the cartoon down.

Severe

One day before work, Fred stops by a coffee shop close to the office where he works and sees one of his colleagues just finishing posting copies of a cartoon around the shop that makes fun of another man at the company who stutters. Fred's colleague looks angry and like someone you wouldn't want to mess with. Fred tells his colleague, in a very loud voice so that everyone at the coffee shop can hear, that making fun of someone's stutter and posting an offensive cartoon is not OK and tears the cartoon down, then sends an email to everyone at work telling them about the

incident and posts a picture of his colleague on Twitter accusing him of being a horrible human being.

VEGAN

Lenient

Harry is at a barbecue. A man purposefully cooks a beef hamburger and feeds it to a vegetarian woman, telling her it is a vegan “incredible burger” (a vegan product that is marketed as tasting just like beef). The vegetarian woman eats the burger, after which the man privately reveals what he has done to Harry, who does not know the vegetarian woman personally. The man looks angry and like someone you wouldn't want to mess with. Harry makes eye contact with the man and shakes his head in disapproval.

Proportional

Harry is at a barbecue. A man purposefully cooks a beef hamburger and feeds it to a vegetarian woman, telling her it is a vegan “incredible burger” (a vegan product that is marketed as tasting just like beef). The vegetarian woman eats the burger, after which the man privately reveals what he has done to Harry, who does not know the vegetarian woman personally. The man looks angry and like someone you wouldn't want to mess with. Harry tells the man that deceiving people about what they are eating is not OK, then tells the host of the barbecue what the man did and asks the man to leave the barbecue.

Severe

Harry is at a barbecue. A man purposefully cooks a beef hamburger and feeds it to a vegetarian woman, telling her it is a vegan “incredible burger” (a vegan product that is marketed as tasting just like beef). The vegetarian woman eats the burger, after which the man privately reveals what he has done to Harry, who does not know the vegetarian woman personally. The man looks angry

and like someone you wouldn't want to mess with. Harry tells the man that deceiving people about what they are eating is not OK, then tells the host of the barbecue what the man did and asks the man to leave the barbecue. Afterwards, he hires a lawyer and convinces the woman to file a lawsuit against the man.

EGGS

Lenient

Martha is jogging around her neighborhood one evening. She is on her second loop of the neighborhood when she spots a teenage girl who she does not know throwing the last egg from a carton at a house. The girl looks angry and like someone you wouldn't want to mess with. Martha makes eye contact with the girl and shakes her head in disapproval.

Proportional

Martha is jogging around her neighborhood one evening. She is on her second loop of the neighborhood when she spots a teenage girl who she does not know throwing the last egg from a carton at a house. The girl looks angry and like someone you wouldn't want to mess with. Martha confronts the girl and tells her that throwing eggs at houses is not OK.

Severe

Martha is jogging around her neighborhood one evening. She is on her second loop of the neighborhood when she spots a teenage girl who she does not know throwing the last egg from a carton at a house. The girl looks angry and like someone you wouldn't want to mess with. Martha confronts the girl and tells her that throwing eggs at houses is not OK. After the incident and every day for the following month, Martha looks for the girl while she is waiting for her school bus and tells her that she should be ashamed of herself.

CUTTING IN LINE

Lenient

Jennifer is taking care of her niece who is playing with other children in the park. Most of the children are between 4 and 7 and impatiently wait in line for their turn at the slide. Jennifer sees a boy cutting in line in front of all the other children. The boy's father looks angry and like someone you wouldn't want to mess with. Jennifer makes eye contact with the boy and shakes her head in disapproval.

Proportional

Jennifer is taking care of her niece who is playing with other children in the park. Most of the children are between 4 and 7 and impatiently wait in line for their turn at the slide. Jennifer sees a boy cutting in line in front of all the other children. The boy's father looks angry and like someone you wouldn't want to mess with. Jennifer tells the boy that cutting in line is not OK and makes him go back to the end of the line so that he waits for his turn like everybody else**.**

Severe

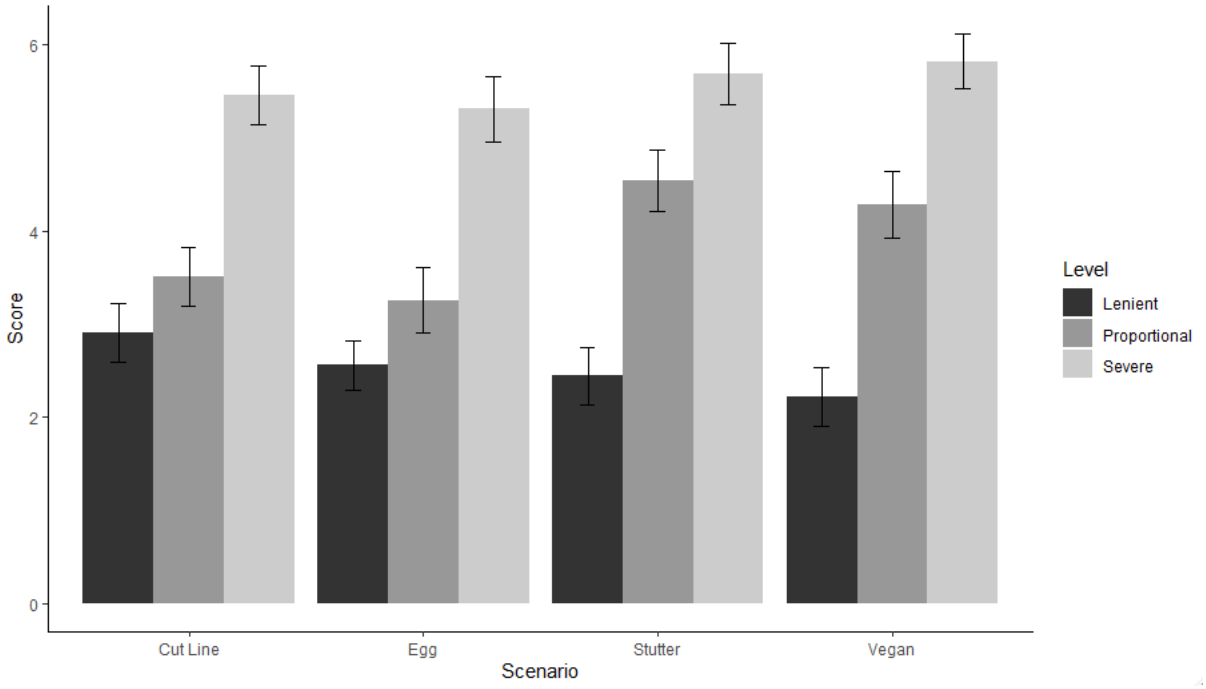
Jennifer is taking care of her niece who is playing with other children in the park. Most of the children are between 4 and 7 and impatiently wait in line for their turn at the slide. Jennifer sees a boy cutting in line in front of all the other children. The boy's father looks angry and like someone you wouldn't want to mess with. Jennifer tells the boy that cutting in line is not OK, then takes him out of the line and makes him stand in a corner for 15 minutes while he watches the other children playing before he can go again.

Appendix E

Vignettes Stimulus Pretest for Severity – Experiment 3

Figure E1

Mean Severity Ratings by Level and Scenario



Note. Error bars represent standard errors

Table E1

Severity Ratings by Level and Scenario

		Score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.90	2.59 – 3.22	**<0.001**
Level Proportional	0.60	0.16 – 1.04	**0.008**
Level Severe	2.55	2.10 – 3.00	**<0.001**
Scenario Egg	-0.35	-0.79 – 0.09	0.118
Scenario Stutter	-0.47	-0.92 – -0.01	**0.043**
Scenario Vegan	-0.69	-1.14 – -0.24	**0.003**
LevelProportional:ScenarioEgg	0.10	-0.52 – 0.72	0.758

LevelSevere:ScenarioEgg	0.20	-0.44 – 0.84	0.540
LevelProportional:ScenarioStutter	1.50	0.87 – 2.13	**<0.001**
LevelSevere:ScenarioStutter	0.70	0.06 – 1.34	**0.033**
LevelProportional:ScenarioVegan	1.47	0.83 – 2.10	**<0.001**
LevelSevere:ScenarioVegan	1.05	0.42 – 1.69	**0.001**
Observations	1016		
R^2 / R^2 adjusted	0.434 / 0.427		

Appendix F

Statistical Analyses – Experiment 3

Sample size justification for experiment 3

Given that this study aimed to test a series of hypothesis where participants were randomly assigned to the 12 possible level X scenario combinations, I conducted an a priori power analysis for general linear models with a an effect size (f2) of 0.03, an alpha level of 0.05 and a 0.9 power. This rendered a total sample size of N = 725. Hence total number of participants to collect data from was rounded to 747 to account for potential attrition, task time limits (participants taking too long or too little on the task) and responses with mistakes (e.g. wrong Prolific ID).

Table F1

Mixed Effects Model for Trust

		Score	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.54	4.39 – 4.70	**<0.001**
Proportional	0.79	0.58 – 1.00	**<0.001**
Severe	-0.00	-0.22 – 0.21	0.983
Eggs	-0.14	-0.35 – 0.07	0.198
Stutter	-0.00	-0.22 – 0.21	0.974
Vegan	-0.34	-0.56 – -0.13	**0.002**
LevelProportional:ScenarioEggs	0.13	-0.17 – 0.44	0.391
LevelSevere:ScenarioEggs	-0.38	-0.68 – -0.09	**0.011**
LevelProportional:ScenarioStutter	0.03	-0.27 – 0.34	0.824
LevelSevere:ScenarioStutter	-0.46	-0.76 – -0.15	**0.003**

LevelProportional:ScenarioVegan	0.85	0.55 – 1.16	**<0.001**
LevelSevere:ScenarioVegan	0.98	0.68 – 1.29	**<0.001**

Random Effects

σ^2	1.30
$\tau_{00\ Prolific\_ID}$	0.32
ICC	0.20
$N_{\ Prolific\_ID}$	723
Observations	2892
Marginal R^2 / Conditional R^2	0.169 / 0.332

Table F2

Pairwise Comparisons of Trust by Level

Scenario	**Level**	**EMM**	**CI**	***p* value**
	Lenient	4.54	4.39-4.70	< 0.0001
Cut in Line	Proportional	5.33	5.18-5.49	0.99
	Severe	4.54	4.38-4.70	< 0.0001
	Lenient	4.41	4.26-4.56	< 0.0001
Eggs	Proportional	5.33	5.16-5.50	0.0006
	Severe	4.02	3.87-4.17	< 0.0001
	Lenient	4.54	4.38-4.70	< 0.0001
Stutter	Proportional	5.36	5.21-5.52	0.0001
	Severe	4.08	3.93-4.23	< 0.0001
	Lenient	4.20	4.04-4.35	< 0.0001
Vegan	Proportional	5.84	5.68-6.00	< 0.0001
	Severe	5.18	5.02-5.34	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Table F3
Mixed Effects Model for Warmth

		WARMTH_	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.43	4.30 – 4.56	**<0.001**
Proportional	0.55	0.37 – 0.72	**<0.001**
Severe	-0.53	-0.71 – -0.35	**<0.001**
Eggs	-0.10	-0.28 – 0.07	0.233
Stutter	0.06	-0.11 – 0.24	0.493
Vegan	-0.13	-0.30 – 0.05	0.157
LevelProportional:ScenarioEggs	0.02	-0.23 – 0.27	0.865
LevelSevere:ScenarioEggs	-0.58	-0.82 – -0.33	**<0.001**
LevelProportional:ScenarioStutter	-0.10	-0.34 – 0.15	0.450
LevelSevere:ScenarioStutter	-0.24	-0.49 – 0.01	0.059
LevelProportional:ScenarioVegan	0.47	0.22 – 0.72	**<0.001**
LevelSevere:ScenarioVegan	1.01	0.76 – 1.26	**<0.001**
Random Effects			
σ^2	0.86		
$\tau_{00\ Prolific\_ID}$	0.27		
ICC	0.24		
$N_{Prolific\_ID}$	723		
Observations	2892		
Marginal R^2 / Conditional R^2	0.231 / 0.414		

Table F4
Pairwise Comparisons of Warmth by Level

Scenario	Level	EMM	CI	*p* value
	Lenient	4.43	4.30-4.56	< 0.0001
Cut in Line	Proportional	4.98	4.85-5.10	< 0.0001
	Severe	3.90	3.77-4.03	< 0.0001
	Lenient	4.33	4.20-4.45	< 0.0001
Eggs	Proportional	4.89	4.76-5.03	< 0.0001
	Severe	3.22	3.09-3.34	< 0.0001
	Lenient	4.49	4.36-4.62	< 0.0001
Stutter	Proportional	4.94	4.81-5.07	< 0.0001
	Severe	3.72	3.59-3.85	< 0.0001
	Lenient	4.30	4.18-4.43	< 0.0001
Vegan	Proportional	5.32	5.19-5.45	< 0.0001
	Severe	4.78	4.65-4.91	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Table F5
Mixed Effects Model for Dominance

	Score		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.56	3.40 – 3.71	**<0.001**
Proportional	1.66	1.45 – 1.88	**<0.001**
Severe	2.20	1.98 – 2.41	**<0.001**
Eggs	0.14	-0.07 – 0.34	0.194
Stutter	-0.04	-0.25 – 0.17	0.688
Vegan	-0.16	-0.37 – 0.05	0.139

LevelProportional:ScenarioEggs	-0.14	-0.44 – 0.16	0.359
LevelSevere:ScenarioEggs	-0.16	-0.46 – 0.13	0.279
LevelProportional:ScenarioStutter	0.49	0.19 – 0.78	**0.001**
LevelSevere:ScenarioStutter	-0.02	-0.32 – 0.28	0.906
LevelProportional:ScenarioVegan	0.07	-0.23 – 0.37	0.633
LevelSevere:ScenarioVegan	0.04	-0.26 – 0.34	0.804

Random Effects

σ^2	1.28
$\tau_{00\ Prolific\_ID}$	0.27
ICC	0.18
$N_{Prolific\_ID}$	723
Observations	2892
Marginal R^2 / Conditional R^2	0.371 / 0.482

Table F6

Pairwise Comparisons of Dominance by Level

Scenario	**Level**	**EMM**	**CI**	***p* value**
	Lenient	3.56	3.40-3.71	< 0.0001
Cut in Line	Proportional	5.22	5.07-5.38	< 0.0001
	Severe	5.76	5.60-5.91	< 0.0001
	Lenient	3.70	3.55-3.84	< 0.0001
Eggs	Proportional	5.22	5.06-5.38	< 0.0001
	Severe	5.73	5.58-5.88	< 0.0001
	Lenient	3.52	3.36-3.67	< 0.0001
Stutter	Proportional	5.66	5.51-5.82	< 0.0001
	Severe	5.69	5.54-5.85	0.96

	Lenient	3.40	3.25-3.55	< 0.0001
Vegan	Proportional	5.14	4.98-5.30	< 0.0001
	Severe	5.63	5.48-5.79	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Table F7
Mixed Effects Model for Competence

	COMPETENCE_		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.95	3.83 – 4.08	**<0.001**
Proportional	1.64	1.47 – 1.81	**<0.001**
Severe	1.26	1.08 – 1.43	**<0.001**
Eggs	0.06	-0.10 – 0.23	0.467
Stutter	-0.03	-0.20 – 0.14	0.705
Vegan	-0.30	-0.47 – -0.13	**<0.001**
LevelProportional:ScenarioEggs	-0.25	-0.49 – -0.00	**0.047**
LevelSevere:ScenarioEggs	-0.46	-0.69 – -0.22	**<0.001**
LevelProportional:ScenarioStutter	0.01	-0.23 – 0.25	0.955
LevelSevere:ScenarioStutter	-0.19	-0.44 – 0.05	0.117
LevelProportional:ScenarioVegan	0.29	0.05 – 0.53	**0.019**
LevelSevere:ScenarioVegan	0.65	0.41 – 0.89	**<0.001**
Random Effects			
σ^2	0.81		
$\tau_{00\ Prolific\_ID}$	0.27		
ICC	0.25		
$N_{Prolific\_ID}$	723		
Observations	2892		
Marginal R^2 / Conditional R^2	0.329 / 0.499		

Table F8
Pairwise Comparisons of Competence by Level

Scenario	Level	EMM	CI	*p* value
	Lenient	3.95	3.83-4.08	< 0.0001
Cut in Line	Proportional	5.59	5.47-5.72	< 0.0001
	Severe	5.21	5.08-5.34	< 0.0001
	Lenient	4.01	3.89-4.13	< 0.0001
Eggs	Proportional	5.41	5.28-5.54	< 0.0001
	Severe	4.82	4.70-4.94	< 0.0001
	Lenient	3.92	3.79-4.04	< 0.0001
Stutter	Proportional	5.57	5.44-5.69	< 0.0001
	Severe	4.98	4.86-5.11	< 0.0001
	Lenient	3.65	3.52-3.77	< 0.0001
Vegan	Proportional	5.58	5.45-5.71	< 0.0001
	Severe	5.56	5.43-5.69	0.97

Note. EMM = estimated marginal means. CI = confidence interval.

Table F9
Mixed Effects Model for Moral Motives

		MORAL_	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.57	4.43 – 4.70	**<0.001**
Proportional	0.77	0.59 – 0.94	**<0.001**
Severe	0.13	-0.05 – 0.31	0.143
Eggs	0.13	-0.04 – 0.30	0.144
Stutter	0.26	0.08 – 0.43	**0.004**

Vegan	-0.10	-0.28 – 0.07	0.243
LevelProportional:ScenarioEggs	0.04	-0.21 – 0.29	0.755
LevelSevere:ScenarioEggs	-0.42	-0.66 – -0.17	**0.001**
LevelProportional:ScenarioStutter	-0.17	-0.42 – 0.08	0.182
LevelSevere:ScenarioStutter	-0.48	-0.73 – -0.23	**<0.001**
LevelProportional:ScenarioVegan	0.49	0.24 – 0.74	**<0.001**
LevelSevere:ScenarioVegan	0.79	0.54 – 1.04	**<0.001**

Random Effects

σ^2	0.85
$\tau_{00\ Prolific\_ID}$	0.39
ICC	0.31
$N_{Prolific\_ID}$	723
Observations	2892
Marginal R^2 / Conditional R^2	0.142 / 0.410

Table F10

Pairwise Comparisons of Moral Motives by Level

Scenario	**Level**	**EMM**	**CI**	***p* value**
	Lenient	4.57	4.43-4.70	< 0.0001
Cut in Line	Proportional	5.33	5.20-5.46	0.31
	Severe	4.70	4.57-4.82	< 0.0001
	Lenient	4.69	4.62-4.81	< 0.0001
Eggs	Proportional	5.50	5.36-5.64	< 0.0001
	Severe	4.41	4.28-4.54	< 0.0001
	Lenient	4.83	4.69-4.96	< 0.0001
Stutter	Proportional	5.42	5.29-5.55	< 0.0001
	Severe	4.48	4.35-4.61	< 0.0001

	Lenient	4.46	4.33-4.59	< 0.0001
Vegan	Proportional	5.72	5.58-5.85	< 0.0001
	Severe	5.39	5.25-5.52	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Table F11
Mixed Effects Model for Leadership

	LEADERSHIP_		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.82	3.66 – 3.98	**<0.001**
Proportional	1.86	1.64 – 2.08	**<0.001**
Severe	0.90	0.67 – 1.12	**<0.001**
Eggs	0.05	-0.17 – 0.26	0.661
Stutter	-0.01	-0.24 – 0.21	0.895
Vegan	-0.37	-0.59 – -0.15	**0.001**
LevelProportional:ScenarioEggs	-0.28	-0.59 – 0.04	0.084
LevelSevere:ScenarioEggs	-1.10	-1.41 – -0.79	**<0.001**
LevelProportional:ScenarioStutter	-0.28	-0.59 – 0.03	0.082
LevelSevere:ScenarioStutter	-0.52	-0.83 – -0.20	**0.001**
LevelProportional:ScenarioVegan	0.38	0.06 – 0.69	**0.018**
LevelSevere:ScenarioVegan	1.03	0.72 – 1.35	**<0.001**
Random Effects			
σ^2	1.38		
$\tau_{00\ Prolific\_ID}$	0.35		
ICC	0.20		
$N_{Prolific\_ID}$	723		
Observations	2892		
Marginal R^2 / Conditional R^2	0.290 / 0.433		

Table F12
Pairwise Comparisons of Leadership by Level

Scenario	Level	EMM	CI	*p* value
	Lenient	3.82	3.66-3.98	< 0.0001
Cut in Line	Proportional	5.68	5.52-5.85	< 0.0001
	Severe	4.72	4.55-4.88	< 0.0001
	Lenient	3.87	3.71-4.02	< 0.0001
Eggs	Proportional	5.46	5.28-5.63	0.14
	Severe	3.66	3.51-3.82	< 0.0001
	Lenient	3.81	3.65-3.97	< 0.0001
Stutter	Proportional	5.39	5.23-5.56	< 0.0001
	Severe	4.18	4.02-4.34	< 0.0001
	Lenient	3.45	3.29-3.61	< 0.0001
Vegan	Proportional	5.69	5.53-5.86	< 0.0001
	Severe	5.38	5.22-5.54	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Table F13
Mixed Effects Model for Friendship

	FRIENDSHIP_		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.45	4.30 – 4.61	**<0.001**
Proportional	0.90	0.68 – 1.11	**<0.001**
Severe	-0.27	-0.48 – -0.05	**0.018**
Eggs	-0.04	-0.25 – 0.17	0.701
Stutter	0.16	-0.06 – 0.37	0.158

Vegan	-0.22	-0.44 – -0.01	**0.041**
LevelProportional:ScenarioEggs	0.04	-0.27 – 0.34	0.812
LevelSevere:ScenarioEggs	-0.74	-1.04 – -0.44	**<0.001**
LevelProportional:ScenarioStutter	0.07	-0.23 – 0.38	0.646
LevelSevere:ScenarioStutter	-0.21	-0.51 – 0.10	0.186
LevelProportional:ScenarioVegan	0.69	0.38 – 0.99	**<0.001**
LevelSevere:ScenarioVegan	1.24	0.93 – 1.54	**<0.001**

Random Effects

σ^2	1.33
$\tau_{00\ Prolific\_ID}$	0.32
ICC	0.19
$N_{Prolific\_ID}$	723
Observations	2892
Marginal R^2 / Conditional R^2	0.226 / 0.376

Table F14

Pairwise Comparisons of Friendship by Level

Scenario	**Level**	**EMM**	**CI**	***p* value**
	Lenient	4.45	4.29-4.61	< 0.0001
Cut in Line	Proportional	5.35	5.19-5.51	0.04
	Severe	4.19	4.03-4.35	< 0.0001
	Lenient	4.41	4.26-4.56	< 0.0001
Eggs	Proportional	5.35	5.18-5.51	< 0.0001
	Severe	3.41	3.26-3.56	< 0.0001
	Lenient	4.61	4.45-4.77	< 0.0001
Stutter	Proportional	5.58	5.42-5.74	< 0.0001
	Severe	4.14	3.98-4.29	< 0.0001

	Lenient	4.23	4.07-4.39	< 0.0001
Vegan	Proportional	5.81	5.65-5.98	< 0.0001
	Severe	5.20	5.04-5.36	< 0.0001

Note. EMM = estimated marginal means. CI = confidence interval.

Appendix G

Reliability Analyses for Dimensions – Experiment 3

Warmth

Dimension	Scenario & Level	Cronbach's alpha	McDonalds' omega
Warmth Overall		0.93	0.93
	Stutter – Lenient	0.92	0.93
	Stutter – Proportional	0.91	0.93
	Stutter – Severe	0.92	0.93
	Vegan – Lenient	0.92	0.93
	Vegan – Proportional	0.89	0.90
	Vegan – Severe	0.92	0.93
	Eggs – Lenient	0.86	0.87
	Eggs – Proportional	0.86	0.89
	Eggs – Severe	0.91	0.93
	Cut – Lenient	0.89	0.91
	Cut – Proportional	0.89	0.92
	Cut – Severe	0.84	0.94

Competence

Dimension	Scenario & Level	Cronbach's alpha	McDonalds' omega
Competence Overall		0.89	0.93
	Stutter – Lenient	0.92	0.94
	Stutter – Proportional	0.85	0.94
	Stutter – Severe	0.80	0.87
	Vegan – Lenient	0.91	0.94
	Vegan – Proportional	0.85	0.89
	Vegan – Severe	0.86	0.90
	Eggs – Lenient	0.88	0.91
	Eggs – Proportional	0.83	0.86
	Eggs – Severe	0.75	0.83
	Cut – Lenient	0.88	0.92
	Cut – Proportional	0.86	0.89
	Cut – Severe	0.84	0.89

Moral Motives

Dimension	Scenario & Level	Cronbach's alpha	McDonalds' omega
Moral Motives Overall		0.82	0.90
	Stutter – Lenient	0.83	0.94
	Stutter – Proportional	0.81	0.89
	Stutter – Severe	0.81	0.91
	Vegan – Lenient	0.88	0.93
	Vegan – Proportional	0.81	0.90
	Vegan – Severe	0.82	0.93
	Eggs – Lenient	0.77	0.88

	Eggs – Proportional	0.80	0.94
	Eggs – Severe	0.79	0.90
	Cut – Lenient	0.79	0.90
	Cut – Proportional	0.76	0.89
	Cut – Severe	0.83	0.92

Leadership

Dimension	**Scenario & Level**	**Cronbach's alpha**	**McDonalds' omega**
Leadership Overall		0.89	0.90
	Stutter – Lenient	0.94	0.94
	Stutter – Proportional	0.94	0.94
	Stutter – Severe	0.94	0.94
	Vegan – Lenient	0.93	0.93
	Vegan – Proportional	0.92	0.92
	Vegan – Severe	0.93	0.93
	Eggs – Lenient	0.90	0.91
	Eggs – Proportional	0.89	0.89
	Eggs – Severe	0.92	0.92
	Cut – Lenient	0.94	0.94
	Cut – Proportional	0.93	0.93
	Cut – Severe	0.93	0.94

Friendship

Dimension	**Scenario & Level**	**Cronbach's alpha**	**McDonalds' omega**
Friendship Overall		0.94	0.94
	Stutter – Lenient	0.93	0.94
	Stutter – Proportional	0.93	0.94
	Stutter – Severe	0.93	0.93
	Vegan – Lenient	0.93	0.93
	Vegan – Proportional	0.89	0.90
	Vegan – Severe	0.94	0.91
	Eggs – Lenient	0.88	0.88
	Eggs – Proportional	0.87	0.87
	Eggs – Severe	0.94	0.94
	Cut – Lenient	0.89	0.89
	Cut – Proportional	0.91	0.91
	Cut – Severe	0.93	0.93