Technische Universität München
TUM School of Social Sciences and Technology

An examination of third-party punishment and its boundaries:

Between the lab and real life

Daniel Toribio-Flórez

Vollständiger Abdruck der von der TUM School of Social Sciences and Technology der Technischen Universität München zur Erlangung eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitzende: Prof. Dr. Doris Holzberger

Prüfende der Dissertation:

1. Prof. Dr. Anna-Julietta Baumert
2. Prof. Dr. Azzurra Ruggeri
3. Prof. Dr. Sebastian Berger

Die Dissertation wurde am 19.11.2020 bei der Technischen Universität München eingereicht und durch die TUM School of Social Sciences and Technology am 21.04.2022 angenommen.

Abstract

Third-party punishment (3PP) refers to costly reactions against perpetrators of norm violations by uninvolved third parties. This dissertation is inspired by the discrepancies in the levels of 3PP between field and lab settings, which question the role of 3PP in the maintenance of social norms. Thus, I investigated an explanatory boundary condition of 3PP –namely, the situational ambiguity– and how it affected the external validity of the main paradigm used to study 3PP in the lab. Moreover, I examined the relationship of 3PP with people's norm perceptions in a natural context.

Abstrakt

Die Bestrafung durch Dritte (Third-party Punishment, 3PP) bezieht sich auf kostspielige Reaktionen von unbeteiligten Dritten gegenüber Tätern, die eine Normverletzung begehen. Die vorliegende Dissertation ist inspiriert durch die Diskrepanzen in den Niveaus von 3PP zwischen Feld- und Laborsituationen, die die Rolle von 3PP bei der Aufrechterhaltung von sozialen Normen in Frage stellen. Daher habe ich eine erklärende Randbedingung von 3PP untersucht – die situative Mehrdeutigkeit - und wie sich diese auf die externe Validität des Hauptparadigmas auswirkt, welches zur Untersuchung des 3PP im Labor verwendet wurde. Darüber hinaus habe ich die Beziehung zwischen 3PP und der Normwahrnehmung von Menschen in einem natürlichen Kontext untersucht.

Table of Contents

General Introduction

From forming queues at the ticket office of the movie theater, to splitting the costs of a dinner equally among a group of close friends, to the acquired custom of wearing masks in public spaces during times of global pandemics: social norms guide our social behavior. Social norms are rules or standards that configure our social context by providing collective consciousness about the appropriateness (and inappropriateness) of certain types of behavior (Chung & Rimal, 2016). In this sense, social norms have an important informative value, as they establish shared expectations of conformity (Bicchieri, 2005). Specifically, they inform us about how we are expected to behave and about how others will likely behave, and consequently reduce the uncertainty of our social environment (Cialdini & Goldstein, 2004).

Beyond their informative value, social norms facilitate cooperation among people by establishing external contingencies to incentivize individuals to behave cooperatively and against their self-interests (Legros & Cislaghi, 2020; Szekely et al., 2021; Young, 2015). These external contingencies can take the form of social punishment towards those who transgress social norms. Consider the norm of wearing masks during a global pandemic. To control the spread of the pandemic, it is essential that individuals coordinate in the implementation of protection measures, among them the proper use of masks in public spaces. If a part of the population does not comply with this norm, the likelihood that the pandemic continues to spread and to endanger the entire population will grow. To ensure cooperation in this situation, institutions may have already shared healthcare recommendations and even formally enforced norms (e.g., legal norms). Yet, individuals who wear masks – and thus, comply with the norm – may additionally be participants of the enforcement of the social norm by reprimanding or confronting those who do not wear masks – and thus, transgress the norm. This type of reaction would exemplify a mechanism of informal, decentralized social punishment – as opposed to the mechanisms of formal, centralized punishment (e.g., law enforcement). Informal social punishment can take different forms, some of them more direct (e.g., public reprimanding, confronting or withholding future help) and some more indirect (e.g., ridiculing, socially excluding or gossiping; Balafoutas, Nikiforakis, et al., 2014; Molho et al., 2020). Critically, this type of informal social punishment has been theorized to be a fundamental element for the maintenance of social norms (Bicchieri, 2005; Cialdini & Trost, 1998; Fehr & Fischbacher, 2004).

In the last two decades, the study of punishment behavior as a reaction against the violation of social norms has received extensive attention (Boyd et al., 2003; Carpenter, 2007; Egas & Riedl, 2008; Eriksson et al., 2021; FeldmanHall et al., 2014; Molho et al., 2020) and, in particular, the case of *third-party punishment* (Fehr & Fischbacher, 2004; Janssen & Bushman, 2008; Krasnow et al., 2016; Lewisch et al., 2015; Riedl et al., 2012; Tan & Xiao, 2018). Third-party punishment refers to the punitive reactions against the violation of social norms by third parties, who are unaffected by the respective norm violation. This definition may remind the reader to situations of bystander interventions, where uninvolved third parties address the perpetrator of a racist aggression, a sexual assault, a situation of bullying or a case of organizational corruption, among other

examples. In these situations, third parties may incur personal costs by investing time and resources for their intervention, but additionally by risking their physical integrity, their social reputation, or their job. Indeed, a relevant, yet puzzling feature of third-party punishment as a behavioral phenomenon is that it does not provide immediate benefits to third parties, but instead, it usually entails immediate personal costs. Thus, one may find third-party punishment often termed altruistic or *costly third-party punishment* (henceforth, 3PP) in the literature (e.g., Marlowe et al., 2008; Ohtsubo et al., 2010). More importantly, the costly character of 3PP bring into question why people engage in this type of behavior. As an answer to this matter, some researchers proposed that 3PP could function as a fundamental mechanism for the reinforcement of social norms (Fehr & Fischbacher, 2004) and hence, that 3PP arguably have an important social value despite not offering immediate benefits for the individual.

Previous research has provided extensive evidence of the prevalence of 3PP through lab experiments (e.g., Balafoutas, Grechenig, et al., 2014; Fehr & Fischbacher, 2004; FeldmanHall et al., 2014) and across different cultures (Henrich et al., 2006; Marlowe et al., 2008). These experiments generally assess 3PP as reactions against the violation of fairness in the distribution of monetary resources in economic games, as I will explain later in more detail. Yet, other lab studies showed that 3PP also emerges as a reaction against the violation of other social norms, such as being dishonest (Ohtsubo et al., 2010) or socially rejecting others without justification (Dimitroff et al., 2020). A limited but steadily growing number of studies offer records of 3PP as a response to daily norm violations in the field, such as littering in public spaces (Balafoutas et al., 2016; Balafoutas, Nikiforakis, et al., 2014; Balafoutas & Nikiforakis, 2012; Winter & Zhang, 2018). Furthermore, 3PP is not only present among adults, but also among young children, suggesting that the sensitivity to react punitively against the violation of social norms emerges in early stages in life (Jordan et al., 2014; McAuliffe et al., 2015).

Despite the extensive evidence, some critical work (Baumard, 2010; Guala, 2012; Pedersen et al., 2013) has pointed out that considerable discrepancies regarding the prevalence of 3PP exist between lab experiments (e.g., Fehr & Fischbacher, 2004) and field studies (e.g., Balafoutas & Nikiforakis, 2012). As I will detail below, 3PP is more rarely observed in the field in comparison to in the lab. This observation raises several questions about research on 3PP with theoretical and methodological implications. First, have lab experiments overlooked relevant situational boundary conditions that hinder 3PP in the field? If so, do the experimental methods generally used in the lab for studying 3PP have limited external validity (i.e., do not offer generalizable findings)? And lastly, if 3PP is not frequently observed in daily life, can it still be considered a critical mechanism for the reinforcement of social norms?

In the present dissertation, I aim to provide answers to these three questions. Regarding the first question, I will examine the effect of one of the potential boundary conditions of 3PP, namely the level of *situational ambiguity* often characterizing situations of norm violations in the field. Specifically, I will look at

how the ambiguity affecting the interpretation of the norm violation (Chapters 1) and the uncertainty about the costs associated to 3PP (Chapter 2) affect the latter. To address the second question, I will further examine how these two situational boundaries may affect the generalizability of the 3PP observed in lab studies (Chapter 3). And for trying to answer the third question, I will investigate whether 3PP relates to people's perceptions of social norms and whether this relationship change as a function of an institutional signal promoting normative change (Chapter 4).

How 3PP is studied in the lab: The third-party punishment game (3PPG)

To delve into the proposed questions, it is firstly necessary to understand how researchers have studied 3PP in the lab. For this purpose, let us use the example about the social norm of wearing masks in public spaces in times of a pandemic. Imagine the following situation: You are commuting by bus and you observe an old woman sitting a few seats from you. At the next stop, a young man enters the bus, chatting on his phone. He stands right beside where the old woman is seated. He wears a mask but the mask is hanging from his chin instead of covering his nose and mouth. What would you do?

In this situation, we shall highlight important situational elements and roles. First, we have the young man, who is the *perpetrator* of a norm violation. In this case, the *norm violation* is not wearing a mask properly in a public space. Second, we have the old woman, for whom the norm violation can have negative consequences as a member of a vulnerable group. Therefore, the old woman can be considered the *victim* of the norm violation. Finally, you are an uninvolved *third party*, who witnesses the norm violation but who is not necessarily affected by it, since you are at safe distance from the young man. These different elements configure the basic structure of the so-called *third-party punishment game* (3PPG), the experimental setup that researchers have commonly used to assess 3PP in the lab.

The 3PPG is an economic game, first introduced in the seminal paper by Fehr and Fischbacher (2004). As in other economic games (for a review, see Thielmann et al., 2021), in the 3PPG people make decisions with real financial consequences. As the situation described above, the 3PPG involves three people, who play in the roles of Person A (the potential *perpetrator*), Person B (the potential *victim*) and Person C (the *third party*). Person A receives a monetary endowment (e.g., 10 euros) and has to decide how they wish to distribute it with Person B. Person A can therefore decide to split this endowment with Person B or keep all the money to themselves. Previous research suggests that, in this situation, people apply the norm of fairness and thus perceive that the more uneven for the benefit of Person A, the more socially inappropriate the distributions are (Krupka & Weber, 2013). Therefore, the *norm violation* in the 3PPG is to distribute the endowment unfairly. In a second stage, Person C learns about Person A's decision and has the option to punish Person A by imposing a cost on the latter (e.g., subtracting money from Person A's final endowment). However, as I previously highlighted, 3PP is usually costly. This is also the case in the 3PPG, where the

punishment decision of Person C entails a cost for themselves. Specifically, Person C has to invest some money from their own endowment to punish Person A.

In the 3PPG, researchers generally assess 3PP by examining its prevalence (i.e., percentage of third parties who exert 3PP) and its intensity (i.e., how much money third parties subtract from the perpetrator). In their original study, Fehr and Fischbacher (2004) reported that roughly 60% of people engaged into 3PP and, in line to the perceived norm of fairness, third parties applied more intense punishment the more uneven the distributions were. Other lab studies have replicated similar results (e.g., Henrich et al., 2006).

Situational ambiguity as a boundary condition of 3PP

From the relatively high level of 3PP shown in lab experiments, it is surprising that relatively low levels of 3PP that researchers have recorded in the field. For instance, the prevalence of 3PP against littering in public spaces does not exceed 10-15% (Balafoutas et al., 2016; Balafoutas & Nikiforakis, 2012; Winter & Zhang, 2018). Other studies using ambulatory assessment (Molho et al., 2020) or recall methodologies (Pedersen et al., 2020), with the intention to capture a wider variety of daily norm violations, have shown similarly low rates of 3PP. In the lab, there are several parameters that have been found to decrease 3PP, for instance, the severity of the norm violation (i.e., with lower 3PP the less severe the norm violation; e.g., Fehr & Fischbacher, 2004; Henrich et al., 2006) or the costs of punishment (i.e., with lower punishment the higher the cost; Egas & Riedl, 2008). Still, lab studies may have neglected other critical boundary conditions that explain 3PP in the field but that are not present in experimental paradigms like the 3PPG. For example, the anticipation of a counter reaction by the punished perpetrator could discourage 3PP beyond its regularly associated costs (e.g., time, resources). When this possibility was introduced in the 3PPG, researchers observed 3PP to decrease drastically (Balafoutas, Grechenig, et al., 2014).

A relevant and unexplored difference between lab experiments and daily situations of norm violations in the field refers to the quality of the information that third parties receive. Whereas in the 3PPG researchers provide third parties with clear information about the norm violation and the actual costs that 3PP entails, daily situations of norm violations often entail *situational ambiguity* (for similar arguments, see Wu et al., 2021). With situational ambiguity, I refer to noisy, incoherent, or incomplete information about critical situational elements that may hinder the decision making of third parties. As an example, consider once again the situation about the young man not using a mask in a public space. Under some circumstances, it may be difficult to assess whether this man is transgressing the norm. For instance, if he was with his back toward you, his position may make it difficult to assure whether he is properly wearing his mask or not. Alternatively, you could be certain about him not wearing his mask properly, but this could be due to a medical condition that justifies why he does not wear the mask continuously. In these alternative scenarios, where there is ambiguity about whether his behavior actually entails a norm violation or whether there is an acceptable

justification, you may hesitate to address him. Moreover, the costs of addressing him might not be easy to anticipate. For instance, it is uncertain whether this person is infected (and thus, whether approaching him would put you at risk) or whether he can react aggressively against you. All these considerations may ultimately affect how you, as the third party, decide to react or not.

The relevance of examining the effect of situational ambiguity resides in the potential mismatch between "the intended outcomes and the actual outcomes" of the decision of the third parties that the ambiguity may bring about (Wu et al., 2021). Under ambiguity, third parties may commit mistakes, such as punishing others who have not violated a norm or leaving unpunished those who violated a norm. Similarly, they may punish when their personal costs are unexpectedly high or refrain from punishing when the costs were actually minimal. Some of these possible outcomes may raise individual concerns that affect the decision making of third parties and ultimately lead them to opt for remaining passive when witnessing the (potential) norm violation.

In this dissertation, I will examine how the introduction of situational ambiguity in the 3PPG affects 3PP. More concretely, I will assess the effect of the ambiguity affecting the identification of the norm violation on 3PP (Chapter 1). Specifically, I will argue that the ambiguity of the norm violation induce concerns about punishing unfairly that may discourage third parties from punishing. Moreover, I will investigate whether the uncertainty about the costs associated to 3PP hinder this behavior (Chapter 2).

On the issue of external validity

After arguing that situational ambiguity may be a neglected boundary condition of 3PP in the field and testing whether this is the case in the 3PPG, a follow-up question is whether the consideration of ambiguity in the 3PPG diminishes the discrepancies between field and lab findings. In other words, does the 3PP we observe in the 3PPG under ambiguity resemble more closely the 3PP that people show in the field? This question directly relates to the *external validity* of findings from lab studies using the 3PPG, which refers to the generalizability of findings across different contexts, populations, and methodological operationalizations (Campbell & Stanley, 1963).

In the case of 3PP, the close examination of the external validity of findings from the 3PPG can be incredibly informative for at least two different reasons. First, the exploration of 3PP across different real-life situations and field settings may facilitate the identification of situational boundary conditions potentially overlooked by experimental paradigms like the 3PPG. In the end, the observation that people exert 3PP as a response to norm violations should generalize from the 3PPG to the field as long as these critical boundary conditions (and their associated psychological factors) converge across both contexts (Levitt & List, 2007). Thus, if the 3PPG incorporated the critical situational factors that arguably moderate 3PP in the field – as I propose regarding situational ambiguity –, one would expect people's 3PP in the 3PPG to converge with their

3PP in the field and, consequently, the external validity of the 3PPG to increase. Second, the examination of the external validity of the 3PPG may help to refine previous theoretical claims about the implications of 3PP for the reinforcement of social norms (Fehr & Fischbacher, 2004). If the 3PP observed in the 3PPG does not accurately reflect how people react to violations of social norms in the field, we should cautiously revise whether 3PP can indeed be considered a key mechanism to explain the maintenance of the system of social norms.

In this dissertation, I will not limit the examination of the external validity to the comparison between the prevalence of 3PP in the lab and in the field. Instead, I will examine how people's individual behavior in the lab correlates with their own behavior in the field, similarly to previous work on the external validity of other economic games (e.g., Benz & Meier, 2008; Galizzi & Navarro-Martinez, 2018). Specifically, I will test whether the 3PP that people show in the 3PPG predicts how they behave (or intend to behave) in a field-like situation of a norm violation and whether the presence or absence of ambiguity moderates this relationship (Chapter 3).

Is 3PP a key element for the system of social norms?

As I previously mentioned, some researchers have argued that 3PP is an important element for the reinforcement of social norms (Fehr & Fischbacher, 2004). To test this idea, many researchers focused on examining the role of 3PP in the maintenance of cooperation (e.g., Egas & Riedl, 2008; Fowler, 2005; Wu et al., 2009). Yet, another approach to shed light on this issue is to assess 3PP in a context of normative change. Social norms change over time, and, to reinforce those changes, the external contingencies that social norms establish to incentivize compliance (Legros & Cislaghi, 2020; Young, 2015) should also change. Thus, a context of normative change should not only reflect a change in people's perceptions of social norm, but additionally, a change in people's willingness to punish the types of behavior that the changing norms reframe as socially inappropriate. Put it more simply, 3PP should accompany any changing trend in the perception of social norms.

A convenient opportunity to test this idea would be to assess the effect of a given *institutional signal*. Institutional signals are decisions or rules made by public institutions that have been theorized to function as a source of normative information (Tankard & Paluck, 2016). When people consider the signaling institution as legitimate, institutional signals are likely perceived as indicators of social norms (Licht, 2008; McAdams, 2000) and, in turn, they should affect people's perceptions of social norms. For instance, previous work has shown how a decision by a major legal court about gay marriage change the perceived social norms about this societal issue (Tankard & Paluck, 2017) or how political campaigns may change the acceptability of prejudices towards immigrants (Crandall et al., 2018).

In line with the rationale proposed above, I argue that 3PP should be similarly affected by an institutional signal that promotes a normative change if 3PP was to be considered a relevant mechanism for the reinforcement of social norms. Thus, if the institutional signal indicates that a given behavior is no longer considered appropriate, people's willingness to intervene against that behavior should congruently increase to reinforce the changing social norm. In this dissertation, I will use the context of the Covid-19 pandemic to test this notion. More concretely, I will examine whether the governmental rules of physical distancing introduced to control the spread of the virus had a parallel effect on the perception of social norms and people's willingness to punish transgressions of physical distancing (Chapter 4).

The present dissertation

The overall goal of the present dissertation is to shed light on the behavioral phenomenon of 3PP. Based on the discussed discrepancies between 3PP in the lab and in the field, I investigate conditions of situational ambiguity under which 3PP may be hindered. In particular, I will study the effect of ambiguity of the norm violation and cost uncertainty in the 3PPG, the experimental setting commonly used to assess 3PP in the lab. Then, I will use these first steps to examine the issue of the external validity of the 3PPG. The inclusion of these potential boundary conditions in the 3PPG may be informative with regard to the generalizability of 3PP to other situations of norm violations outside the lab. I will check if this is the case by comparing people's behavior in the 3PPG with their behavior in a field-like situation involving a norm violation. Finally, this dissertation will examine the notion that 3PP is an important reinforcing element of social norms. Specifically, I will use a case study within the context of the Covid-19 pandemic to analyze the effect of an institutional signal indicating a normative change on people's perceptions of social norms, as well as on their willingness to punish the violation of the new norms.

Overview of Chapters

The present dissertation includes four empirical chapters and a general discussion. Chapter 1, 2, and 3 will present data from a common set of experimental studies. These studies addressed several research questions, and therefore, I will present their results separately in different chapters.

Chapter 1 presents the results of six lab studies, examining the effect of ambiguity affecting the identification of the norm violation on 3PP. In this chapter, I will discuss how, under ambiguity of the norm violation, third parties may experience concerns about punishing unfairly. I argue that these concerns may discourage third parties from engaging into 3PP. To delve into this idea, I will evaluate the role of interindividual differences in justice concerns, as well as the inclination of third parties to resolve the ambiguity when they are given the opportunity to do so.

Chapter 2 reports the data from three of the studies to examine whether uncertainty about the costs associated with 3PP further hinders this behavior. Here, I will also assess the role of interindividual

differences in justice concerns, as these could shed light on different sensitivities to the effect of cost uncertainty.

Chapter 3 focuses on the issue of the external validity of the 3PPG. Specifically, I will test whether the 3PP that people show in the 3PPG predicts how they behave (or intend to behave) in a field-like situation of a norm violation. Critically, I further check whether the inclusion of ambiguity of the norm violation and cost uncertainty in the 3PPG moderate this relationship. This chapter will report data from two of the studies. The first study examines the relationship between people's 3PP in the 3PPG and their intervention behavior against a staged embezzlement of lab funds. The second study presents the embezzlement in a vignette format and I will analyze the relationship between 3PP and people's intervention intentions.

Chapter 4 presents a case study in Germany during the first stage of the Covid-19 pandemic. Following a pre-post quasi-experimental design, the goal of this case study was to examine the effect of the introduced governmental rules of physical distancing (i.e., institutional signal) on people's perceptions of social norms and their willingness to punish (or intervene) against violations of the new norms. Importantly, I will also analyze whether the expected relationship between perceptions of social norms and willingness to punish changes as a function of the introduction of the governmental rules.

In the **general discussion** of this dissertation, I will summarize and elaborate on the empirical findings from each of these chapters and their respective contributions to the investigation of 3PP, as well as on those questions that may remain unanswered. I will make further emphasis on the theoretical and methodological implications that the present research agenda could have for future research.

Chapter 1

Ambiguity of the norm violation as boundary condition of third-party punishment

Costly third-party punishment (3PP) against norm violations has barely been investigated under situational ambiguity. In six studies, we introduce *ambiguity of the norm violation* in a third-party punishment game and consistently observe it to decrease *costly third-party punishment* (3PP). Our research suggests two plausible mechanisms of this effect. First, under ambiguity, people might refrain from punishing due to the risk of punishing unfairly. Thus, people with higher (vs. lower) other-oriented justice sensitivity (Observer JS) reduced 3PP more pronouncedly (Studies 1.1-1.3). Moreover, those who decided to resolve the ambiguity (hence, removing the risk of punishing unfairly) punished to greater extents (Studies 1.4-1.5). Second, people aiming at avoiding the costs of 3PP might use ambiguity to justify their intention of remaining passive. Supporting this notion, 3PP also decreased among those with low Observer JS, and those who decided to keep the ambiguity showed the lowest levels of 3PP. Therefore, ambiguity entails a relevant situational boundary that sheds light on the prevalence of 3PP and its preceding decision-making.

Chapter 1 is based on: Toribio-Flórez, D., Saße, J., & Baumert, A. (2022). "Proof *under* reasonable doubt": Ambiguity of the Norm Violation as Boundary Condition of Third-Party Punishment. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672211067675

Third-party punishment can manifest itself in a wide range of phenomena, from confronting discrimination to speaking up against (cyber)bullying. At its core, third-party punishment refers to sanctioning reactions against someone who violates a norm (i.e., a perpetrator) by an uninvolved witness. These reactions are considered highly desirable for the reinforcement and maintenance of social norms (Yamagishi, 1986); however, they usually entail costs for the third party, either physical (e.g., violence), social (e.g., ostracism), or economical (e.g., dismissal). Thus, the investigation of *costly third-party punishment* (henceforth, *3PP*) has raised special interest among biologists, anthropologists, economists, and psychologists (e.g., Janssen & Bushman, 2008; Krasnow et al., 2016; Lewisch et al., 2015; Riedl et al., 2012).

Researchers commonly investigate 3PP as the financially costly sanctioning of others who distribute monetary resources unequally between themselves and second parties. Empirical evidence from lab studies has shown that 50-60% of people engage in 3PP, with higher sanctions the more unequal the distributions are (e.g., Fehr & Fischbacher, 2004; Henrich et al., 2006). Critically, most of these studies provided decision-making settings with perfect situational information, allowing third parties to identify swiftly whether particular distributions constituted violations of fairness or equity norms. In real-life situations outside the lab, this is unlikely to occur. Individuals often receive noisy, incoherent, or incomplete information, which can create *ambiguity* about whether the perpetrator's behavior actually adheres to or violates a norm. Resonating with this discrepancy, researchers have failed to observe comparable levels of 3PP in the field (e.g., Balafoutas et al., 2014; Brauer & Chekroun, 2005).

In line with theoretical models on bystander intervention (Baumert et al., 2013; Osswald et al., 2010), we assume that the interpretation of a norm violation as such is a necessary requirement for 3PP to occur. Therefore, we argue that ambiguity of a norm violation could constitute a pivotal boundary condition that hinders the decision of third parties to act against the norm violation. Previous research on 3PP has generally neglected the role of ambiguity of the norm violation. To our knowledge, only a recent set of studies (Jordan & Kteily, 2020, Punishment Experiments 1-3) has provided empirical evidence in this regard. In these studies, third parties showed less willingness to engage in indirect punishment (i.e., donation to an organization protesting against the perpetrator) against an ambiguous (vs. unambiguous) case of sexual harassment.

In the present research, we investigate how ambiguity of the norm violation influences direct 3PP and aim to shed light on distinct motivations underlying the effect of ambiguity.

Ambiguity of norm violations and 3PP

As an early critical step for 3PP, the interpretation of the situation has downstream effects on any further decision-making (Baumert et al., 2013; Osswald et al., 2010). If third parties access clear situational information, they can readily interpret the perpetrator's behavior as a norm violation and then turn to ponder whether and how to react against it. Conversely, if the situational information is ambiguous, the interpretation

of the norm violation should be hampered. At least, two psychological explanations make it plausible that ambiguity of the norm violation reduces 3PP.

First, third parties might refrain from punishing due to concerns of unfairly sanctioning someone who actually did not violate any norm (Grechenig et al., 2010). They might be aware that handling ambiguous information entails the risk of wrongly assuming that a norm violation has occurred when actually it did not (i.e., *type I error*). The motivation to avoid committing type I errors could be fueled by anticipated feelings of guilt and reputational or moral concerns, as undeserved punishment might be negatively judged by others and by oneself. Findings from Jordan and Kteily (2020, Punishment Experiment 2) suggest that such concerns would be warranted. They showed that punishing an ambiguous (vs. unambiguous) norm violation entailed lower reputational benefits. We argue that the punishment of an ambiguous norm violation that turns out not to have occurred could bring with it the prospect of reputational losses and negative self-evaluations that third parties might try to avoid.

Second, avoiding the costs of punishing could be enticing to third parties. Individuals whose primary motivation is to avoid costs might use the ambiguity of the norm violation as a justification to remain passive. Supporting this reasoning, previous research has shown that people act less prosocially if the situation provides a justification for it (Dana et al., 2007; Haisley & Weber, 2010). For instance, in a "dictator game", researchers concealed how much money the recipient would actually receive from participants playing as dictators. In this setting, where uncertainty masked the dictator's decision, participants were more likely to choose the option that was more beneficial to themselves. Furthermore, when researchers gave them the opportunity to reveal the concealed information, some participants deliberately avoided doing so (Dana et al., 2007). In light of these results, researchers suggested that some people might exploit situational ambiguity as "moral wiggle room", which allows them to hide or justify selfish motives (Dana et al., 2007; Haisley & Weber, 2010). We propose that this explanation could apply to 3PP as well. If a norm violation is ambiguous, some third parties who would punish under no ambiguity may use the situational ambiguity to justify their passiveness, thereby avoiding own costs (Kriss et al., 2016).

Taken together, we expect that ambiguity of the norm violation would reduce 3PP because ambiguity introduces a risk of punishing unfairly and a potential justification to avoid incurring costs. To investigate these two underlying mechanisms, we employed two approaches: the examination of a) inter-individual differences in concerns about injustice as moderator of the effect of ambiguity and b) the inclination of third parties to resolve the ambiguity. Both approximations aimed to distinguish between those who, under ambiguity, remain passive due to the risk of punishing unfairly and those who do so to avoid own costs, as we describe below.

Justice sensitivity as moderator of reactions to ambiguity

People may differ in the extent to which they are susceptible to the effects of ambiguity. Specifically, dispositional concerns about injustice should relate to individual motivations to avoid committing injustice and accept own costs in order to restore justice. Thus, we investigated the moderating role of inter-individual differences in *Justice Sensitivity* (JS).

JS is a multidimensional personality construct that captures the strength of cognitive, emotional, and behavioral reactions to perceived injustice (Baumert & Schmitt, 2016). Researchers have distinguished four facets of JS, according to the perspectives from which people can experience injustice and react to it (Schmitt et al., 2010). What is important for the present research is both the perspective of a *perpetrator* who inflicts injustice on others (i.e., Perpetrator JS) and the perspective of an uninvolved *observer* (i.e., Observer JS). Both perspectives refer to sensitivity to injustice done to others, and their combination has been observed to predict third-party reactions to norm violations (Lotz, Baumert, et al., 2011; Niesta Kayser et al., 2010). Critically, they relate conceptually to the mechanisms discussed to underlie the predicted effect of ambiguity on 3PP.

Perpetrator JS captures concerns about committing any injustice oneself (Baumert & Schmitt, 2016). Hence, people with high Perpetrator JS should be particularly concerned about punishing unfairly, since unjustified punishment would constitute an act of injustice in and of itself. Thus, when ambiguity increases the danger of misinterpreting the norm violation (i.e., committing a type I error), individuals with high (vs. low) Perpetrator JS should be more hesitant to punish.

For its part, Observer JS predisposes individuals to perceive injustice and be motivated to act against it, out of a genuine other-oriented concern for justice (Baumert et al., 2011; Schmitt et al., 2010). Observer JS has been found to relate negatively to selfish behavior (Edele et al., 2013; Fetchenhauer & Huang, 2004; Lotz, Baumert, et al., 2011), even when the situation excused acting selfishly (Lotz et al., 2013). Therefore, people with low (and not high) Observer JS should readily exploit the ambiguity of the norm violation as justification to remain passive and avoid own costs.

Resolving the ambiguity and its underlying motivations

If third parties faced an ambiguous norm violation, gaining information that resolves the ambiguity would allow for a more informed decision about whether or not to punish. Especially, third parties who wanted to avoid the risk of punishing unfairly should resolve the ambiguity. This would alleviate their type I error concerns, and therefore, it should facilitate exerting 3PP if a norm violation had actually occurred.

Conversely, those third parties whose main goal is to avoid incurring costs do not gain from resolving the ambiguity. On the contrary, we argue that some might find keeping the situation ambiguous beneficial to uphold a situational justification for remaining passive (Dana et al., 2007; Stüber, 2020).

Consequently, we examined whether third parties who decide to resolve the ambiguity do so to punish subsequently the potential norm violation and whether those who do not resolve the ambiguity actually remain passive.

Research overview

In six studies, we investigated the effect of ambiguity of the norm violation and its underlying mechanisms on 3PP. In Studies 1.1 and 1.2, we tested the main effect of ambiguity on 3PP and the moderating role of JS. Studies 1.4 and 1.5 aimed to examine whether and why third parties would resolve the ambiguity before deciding to punish. Studies 1.3 and 1.6 served to rule out potential confounds of the experimental manipulation used in the other studies.

In every study, we used the third-party punishment game (3PPG) as experimental paradigm (Fehr & Fischbacher, 2004). The 3PPG follows the structure of a dictator game to the extent that the dictator (Person A) can share an endowment with a passive recipient (Person B). A third party (Person C), unaffected by the dictator's decision, is informed about the dictator's distribution and can influence it by deducting coins from the dictator's final payoff (henceforth, *punishment* of Person A). In our studies, Person C simultaneously could decide to add coins to the recipient's final payoff (henceforth, *compensation* of Person B; for results about compensation, see Supplementary Material). The addition of compensation to the 3PPG aimed to counter experimental demand effects potentially occurring in settings where third parties can solely punish (Lotz, Okimoto, et al., 2011; Pedersen et al., 2018; van Doorn et al., 2018). Critically, both the punishment of Person A and the compensation of Person B, were associated with known costs for Person C.

We experimentally manipulated ambiguity of the norm violation by providing Person C with perfect or imperfect information about the endowment of Person A. In the *no ambiguity* conditions, participants learnt the exact endowment that Person A had received. In the *ambiguity* conditions, the endowment of Person A was randomly determined. While Person A was informed about their exact endowment before making their respective decision, Person C only learned about the range of possible endowments. Therefore, for Person C, it was ambiguous whether Person A's distribution was unequal or not. Note that, in this setting, people generally regard unequal distributions by Person A as norm violations (Erkut et al., 2015; Krupka & Weber, 2013).

Studies 1.1-1.2

First, we tested our main hypothesis, which held that ambiguity of the norm violation would reduce levels of 3PP (H1). Furthermore, we examined whether JS moderated this effect. Specifically, we expected that levels of 3PP under *ambiguity* (vs. *no ambiguity*) would decrease more pronouncedly among third parties with high (vs. low) Perpetrator JS (H2a). We also predicted that 3PP under *ambiguity* (vs. *no ambiguity*) would decrease more pronouncedly among those with low (vs. high) Observer JS (H2b). We preregistered these hypotheses for Study 1.1 (https://osf.io/ubnzm) and Study 1.2 (https://osf.io/etgq9). Informed by the results of Study 1.1, the preregistration of Study 1.2 additionally included the competing hypothesis that the expected decrease of 3PP under ambiguity would be more pronounced among third parties with high (vs. low) Observer JS (H2b').

Method

Design

In Studies 1.1 and 1.2, each participant played four rounds of the 3PPG. In two rounds (*no ambiguity*), participants, playing as Person C, learnt that Person A received a fixed endowment of 10 experimental currency units (ECUs). In the other two rounds (*ambiguity*), they learnt that Person A would receive a randomly determined endowment that ranged from 2 to 10 ECUs.

The reason behind having four rounds was that these studies included, besides ambiguity, a second within-subject factor to address a further, yet unrelated research question; namely how uncertain (vs. certain) costs for Person C affected 3PP, independently of the ambiguity of the norm violation. Two of the four rounds established a certain cost of ½ ECU (per 1 ECU that Person C wished to punish or compensate), whereas in the other two rounds, this cost was uncertain, as it randomly varied between 0.01 and 1 ECUs. This research question will be addressed in detail in Chapter 2.

Participants

Studies 1.1 and 1.2 were part of a larger research project and aspects unrelated to the research questions at hand determined the sample size. Yet, we conducted sensitivity analyses to determine the minimum effect size that our recruited samples allowed us to detect with 90% statistical power and $\alpha = .05$ (see below).

Study 1.1. We recruited 175 participants and excluded data of 11 who failed or did not respond to preregistered data-quality items (e.g., "Honestly speaking, should we use your responses for our research?"). The final sample consisted of 164 undergraduate students from diverse disciplines (77% women; age range from 18 to 33, $M = 22.79$, $SD = 2.92$). The smallest effect size we could detect with this sample size was a

standardized regression weight of β = .10. Participants could earn up to €10 in the 3PPG.[1] They also received a fixed monetary reward of €30, due to their participation in a subsequent lab session aimed to address the research question discussed in Chapter 3.

Study 1.2. We recruited 228 participants and excluded data of 2 based on the same preregistered data-quality items used in Study 1.1. The final sample consisted of 226, mainly undergraduate students from diverse disciplines (73% women, age range from 18 to 68, M = 23.26, SD = 5.44). The smallest effect size we could detect with this sample size was a standardized regression weight of β = .074. Participants could earn up to €10 in the 3PPG. As in Study 1.1, they further received a fixed monetary reward, in this case of €20, due to their participation in a subsequent lab session aimed to address the research question discussed in Chapter 3.

Procedure

Participation was either in the lab or online. After providing informed consent, participants completed the JS inventory, presented among different personality questionnaires (see preregistration for all materials).

Then, participants played the 3PPG. They learned that their decisions would have real financial consequences for themselves and others. They played the different rounds of the 3PPG in a fixed sequential order – Round 1 and Round 2 (*no ambiguity*), Round 3 and Round 4 (*ambiguity*). Several comprehension questions ensured that participants understood the general rules of the 3PPG and the manipulated elements of each round (e.g., "How many [ECUs] does Person A receive?" to tap into the ambiguity manipulation). When participants answered them incorrectly, we repeated the instructions.

In the 3PPG, we implemented a strategy vector method (e.g., Oxoby & McLeish, 2004). Participants sequentially decided in the role of Person A, B, and C. In the critical role of Person C, participants received an endowment of 10 ECUs. They made decisions that were conditional on seven different possible distributions from Person A (i.e., "Given that Person A transfers [0 to 6] ECUs to Person B, how many ECUs do you wish to *deduct* from Person A's / *add* to Person B's endowment?"). Participants could punish Person A between 0 and the maximum remaining amount of money that Person A would have after each distribution, given an initial endowment of 10 ECUs. In the *ambiguity* condition, Person C did not know what the initial endowment of Person A was. Consequently, the participants' punishment decision could in principle exceed the actual remaining endowment of Person A. However, participants were informed that their punishment would only become effective until Person A's endowment was reduced to 0 ECUs, while they would still incur costs for the punishment they decided to apply. At the end of the experiment, we

[1] In Studies 1.1-1.3, 1 ECU = €1.00.

randomly grouped participants into triads and assigned them to one of the three roles. We then calculated their payoff based on the decisions they made in one round selected at random and the corresponding decisions of the other two members of the triad. For the complete instructions, see "Methodology File.docx" at https://osf.io/2q9vm/.

Measures

3PP. We computed a continuous measure of 3PP in each condition by summing up the amount deducted by Person C in those decisions in the role of Person C that implied a reaction to an unequal split of a 10-ECU endowment (i.e., Person A sent [0, 1, 2, 3, or 4] ECUs to Person B). We excluded the decisions entailing a fair split (i.e., Person A sent [5 or 6] ECUs) under the assumption that these would not generally be perceived as norm violations (Erkut et al., 2015; Krupka & Weber, 2013). If participants did not report more than one decision in a round, their data were not included in the analyses for that round.

Justice sensitivity. We assessed JS with the German *Justice Sensitivity Short Scales* (Baumert, Beierlein, et al., 2014), which include two items each for measuring Perpetrator JS (e.g., "I feel guilty when I enrich myself at the cost of others") and Observer JS (e.g., "I am upset when someone is undeservingly worse off than others"). For descriptive results, see Table 1.1.

Table 1.1

Descriptive statistics and Cronbach's a for Observer and Perpetrator Justice Sensitivity (JS) across studies and their correlation within each study.

	M	***SD***	**Cronbach's α**	***r* [95% CI]**
Study 1.1				
Observer JS	3.14	1.17	.75	.32*** [.17, .45]
Perpetrator JS	3.69	1.14	.75	
Study 1.2				
Observer JS	3.06	1.16	.75	.42*** [.30, .52]
Perpetrator JS	3.64	1.15	.71	
Study 1.3				
Observer JS	3.13	0.84	.87	.55*** [.47, .63]
Perpetrator JS	3.62	0.96	.91	
Study 1.4				
Observer JS	2.87	0.94	.90	.62*** [.57, .67]
Perpetrator JS	3.52	0.97	.91	
Study 1.5				
Observer JS	2.98	0.98	.93	.55*** [.50, .60]
Perpetrator JS	3.53	1.04	.94	
Study 1.6				
Observer JS	3.22	0.83	.89	.52*** [.46, .57]
Perpetrator JS	3.76	0.86	.91	

Note. Response options ranged from 0 (*Not at all)* to 5 (*Absolutely*).
*** $p < .001$.

Statistical analyses

We followed a multilevel modelling approach, clustering punishment decisions in each round of the 3PPG (Level 1) within participants (Level 2). The model included the participants' ID as a random factor and ambiguity of the norm violation as a Level-1 fixed factor, Observer JS and Perpetrator JS as Level-2 fixed factors (grand-mean centered), and the respective two cross-level interactions.

Results and discussion

Table 1.2 displays the descriptive statistics of 3PP in each experimental condition.

The multilevel models from the two studies (see Table 3) consistently showed that the ambiguity of the norm violation significantly reduced 3PP (supporting H1). Moreover, in both studies, Observer JS and not Perpetrator JS significantly moderated the effect of ambiguity. Simple slope analyses indicated that, under ambiguity, participants with high Observer JS (i.e., 1*SD* above the mean) reduced their level of 3PP more pronouncedly than those with low Observer JS (i.e., 1*SD* below the mean; see Figure 1.1). We had initially predicted this specific pattern for Perpetrator JS (H2a). However, the findings indicated that, first and foremost, Observer JS captured relevant inter-individual differences in the reaction to ambiguity of the norm violation (supporting competing H2b').

Notwithstanding the consistency of the results across Studies 1.1 and 1.2, we considered necessary to replicate them while excluding that the effect of ambiguity derived from the fixed order of the rounds in the 3PPG. Given that we introduced ambiguity in the two last rounds, our findings could have resulted from a decay in prosocial behavior across game rounds – as previously observed in repeated-game designs (e.g., Fehr

Table 1.2

Descriptive statistics of punishment per condition in Studies 1.1-1.3.

	Punishment		
	M	*SD*	*Perc.* (%)
Study 1.1			
No ambiguity	10.83	7.71	84.62
Ambiguity	6.11	6.04	75.23
Study 1.2			
No ambiguity	9.69	7.74	77.43
Ambiguity	4.99	6.18	65.93
Study 1.3			
No ambiguity	6.43	7.32	59.46
Ambiguity	3.35	5.02	52.62

Note. Punishment = Amount of ECUs (1 ECU = 1 Euro) subtracted from Person A. *M* and *SD* = mean and standard deviation of the sum of Euros punished across decisions to unequal splits from Person A (i.e., €[0 to 4] coins) to Person B. *Perc.* (%) = Percentage of participants who punished at least 1 ECU.
For descriptive statistics split by decision of the strategy method, including those to fair distributions by Person A, see Supplementary Material.

& Gächter, 2002) – or, alternatively, from end-of-game effects (Andreoni, 1988). Study 1.3 addressed these concerns.

Study 1.3

Study 1.3 followed the same within-subject design as Studies 1.1 and 1.2, but we introduced a "pseudo-randomization" of the order of the rounds in the 3PPG. We counterbalanced across participants which of the four rounds they played first. After this first round, we randomly presented the three other rounds. A practical matter was behind this pseudo-randomization. In case we had identified any order effects (e.g., decay of 3PP over time or end-of-game effect), the pseudo-randomization would have allowed us to test our hypotheses using the first round to analyze between-subject differences in 3PP across experimental conditions. We explain any other methodological deviation from Studies 1.1 and 1.2 below.

Method

Participants

We conducted a priori power simulations to plan our data collection. We used the standardized estimates observed in Study 1.1 to generate random multilevel data (1000 iterations). Specifically, these estimates corresponded to a simplified multilevel model including the effects of ambiguity of the norm violation, Observer JS and their two-way interaction. We assumed Observer JS to be normally distributed and a critical $\alpha = .05$. The results of the simulations indicated that the targeted sample size of 300 participants would suffice to detect a significant Ambiguity x Observer JS with 95.1% statistical power. The script for reproducing this simulation is available at https://osf.io/2q9vm/ (i.e., "Study 1/Data/Rscript/4. Power Simulation for Study 3.R")

We recruited 311 participants and excluded data of 22 based on one attention check and on the data-quality items used in Studies 1.1 and 1.2. The final sample consisted of 284 participants, mainly undergraduate students from diverse disciplines (68% women, age range from 18 to 73, M = 23.37, SD = 6.19). Participants received a fixed monetary reward of €2.50 and they could additionally earn up to €10 in the 3PPG.

Procedure

We conducted the study online. After providing informed consent, participants completed the JS scales (embedded between two filler questionnaires) and the 3PPG. With the exception of the pseudorandomized rounds of the 3PPG, any other procedural detail was identical to Studies 1.1 and 1.2.

Measures

Justice sensitivity. In this study, we assessed JS using the 40-item version of the JS Inventory (Schmitt et al., 2010). Ten items served for measuring each JS perspective, including the same two items of the short version.

Statistical analyses

We tested the same model as in Studies 1.1 and 1.2. Next, we created two factors for examining order effects and introduced them as covariates into our model. At Level 1, the factor *Position* captured the position at which a particular round of the game was presented to a participant (i.e., 0 – Round 1, 1 – Round 2, 2 – Round 3, 3 – Round 4). This factor accounted for potential linear effects on 3PP over time. Moreover, we wanted to pay special attention to the effects of the order of presentation of the *ambiguity* conditions. Thus, at Level 2, the factor *Ambiguity Order* captured the six possible randomized orders in which the *ambiguity* (yes) and *no ambiguity* (no) conditions were presented to participants (i.e., 0 – no, no, yes, yes; 1 – yes, yes, no, no; 2 – no, yes, no yes; 3 – yes, no, yes, no; 4 – yes, no, no, yes; 5 – no, yes, yes, no). We tested the main effects of these two factors on 3PP and, more importantly, whether Ambiguity Order moderated the effect of ambiguity or the cross-level interactions with Observer and Perpetrator JS.

Results and discussion

As in Studies 1.1 and 1.2, ambiguity significantly reduced 3PP, and Observer JS significantly moderated this effect (see Table 1.3). Simple slope analysis indicated that the effect of ambiguity was more pronounced among those with high (vs. low) Observer JS (see Figure 1.1). In contrast to Studies 1.1 and 1.2, Perpetrator JS significantly moderated the effect of ambiguity. Ambiguity decreased 3PP more pronouncedly among those with high Perpetrator JS, $\beta = -.52$, $t(828.90) = -11.291$, $p < .001$, 95% CI [-.61, -.43], than among those with low Perpetrator JS, $\beta = -.33$, $t(825.69) = -7.360$, $p < .001$, 95% CI [-.42, -.20].

Entering the factors Position and Ambiguity Order as covariates did not significantly increase the model fit, $\Delta AIC = 20.1$, $\chi^2(31) = 41.846$, $p = .092$. For simplifying the report of the results (as the number of model parameters increased to 36), we present the omnibus tests of the main and interaction effects in an ANOVA table (Table 1.4). The main effects of Position and Ambiguity Order were not significant. More importantly, Ambiguity Order did not significantly moderate any main effect or interaction. Note, however, that the Ambiguity x Perpetrator JS interaction was no longer significant in this model.

Study 1.3 closely replicated the findings from Studies 1.1 and 1.2, while the experimental setup served to unconfound ambiguity effects from potential order effects. Thus, Studies 1.1-1.3 provided consistent evidence for a decrease of 3PP under ambiguity (H1), especially among participants with high Observer JS (H2b'). Previous research suggests that individuals with high other-oriented JS (e.g., Observer and Perpetrator JS) are less driven by selfish temptations (Baumert, Schlösser, et al., 2014; Fetchenhauer & Huang, 2004; Lotz et al., 2013). Hence, it seemed implausible that individuals high in Observer JS reduced 3PP because ambiguity offered them moral wiggle room to avoid incurring costs. Instead, it is more conceivable that these people hesitated to punish because ambiguity introduced a risk of becoming unfair themselves (i.e., type I error concerns).

Table 1.3

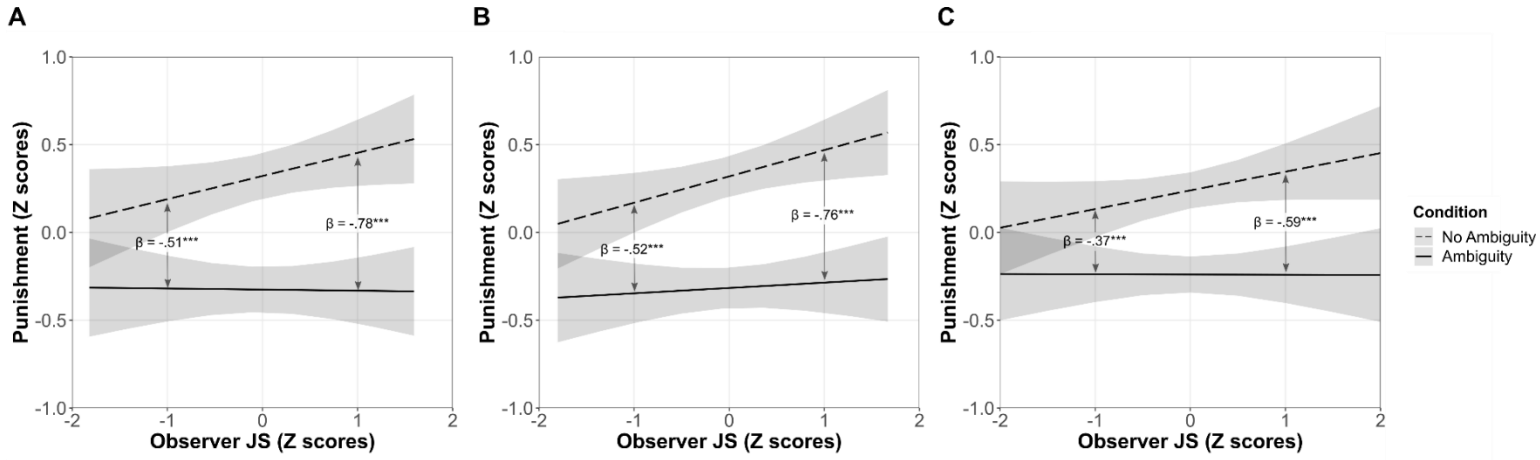
Tested multilevel model on punishment in Studies 1.1-1.3.

	Study 1.1			Study 1.2			Study 1.3		
Parameters	β [95% CI]	***t***	***p***	β [95% CI]	***t***	***p***	β [95% CI]	***t***	***p***
Fixed effects									
Ambiguity of norm violation	-.64 [-.74, -.55]	-13.06	**<.001*** **	-.63 [-.70, -.56]	-17.38	**<.001*** **	-.48 [-.54, -.41]	-14.47	**<.001*** **
Perpetrator JS	.06 [-.07, .20]	0.92	.356	-.03 [-.15, .10]	-0.44	.661	.19 [.06, .31]	2.96	**.003** **
Observer JS	.13 [.00, .27]	1.89	.060	.14 [.01, .26]	2.17	**.031***	.11 [-.02, .23]	1.70	.089
Ambiguity x Perpetrator JS	.01 [-.09, .12]	0.26	.795	.05 [-.03, .13]	1.25	.212	-.11 [-.18, -.03]	-2.67	**.008** **
Ambiguity x Observer JS	-.13 [-.23, -.03]	-2.57	**.011***	-.12 [-.20, -.04]	-2.95	**.003** **	-.11 [-.19, -.03]	-2.73	**.006** **
Random effects									
σ^2	20.99			16.19			12.55		
$\tau_{00\ ID}$	26.85			32.99			25.39		
ICC_{ID}	0.56			0.67			0.67		
N_{ID}	163			224			281		
Observations	648			896			1108		
Marginal / Conditional R^2	0.118 / 0.613			0.108 / 0.706			0.094 / 0.700		

Note. JS = Justice Sensitivity, σ^2 = Residual variance; $\tau_{00\ ID}$ = Variance of the intercept; ICC_{ID} = Intraclass correlation coefficient; N_{ID} = Total number of individuals.
*** $p < .001$, ** $p < .01$, * $p < .05$.

Figure 1.1

Two-way interaction between ambiguity and Observer JS in Studies 1.1 (A), 1.2 (B), and 1.3 (C).



Note. Standardized regression coefficients represent the effect of ambiguity at +1SD and -1SD Observer JS based on simple slope analyses.
*** $p < .001$. Band widths 95% CIs.

Table 1.4

ANOVA table of multilevel model accounting for order effects in Study 1.3.

Parameters	Punishment			
	F	***df***	***p***	η_p^2
Ambiguity of norm violation	205.02	1, 812	**< .001*** **	.197
Perpetrator JS	3.55	1, 264	.061	.004
Observer JS	0.57	1, 264	.452	.001
Position	0.93	1, 811	.336	.001
Ambiguity order	1.76	5, 264	.122	.010
Ambiguity x Perpetrator JS	3.08	1, 811	.080	.004
Ambiguity x Observer JS	7.39	1, 812	**.007****	.009
Ambiguity x Ambiguity order	2.03	5, 812	.072	.012
Perpetrator JS x Ambiguity order	0.78	5, 264	.567	.005
Observer JS x Ambiguity order	0.7	5, 264	.623	.004
Ambiguity x Perpetrator JS x Ambiguity order	1.68	5, 811	.136	.010
Ambiguity x Observer JS x Ambiguity order	0.32	5, 812	.904	.002

Note. JS = Justice Sensitivity, *df* = Numerator and denominator degrees of freedom calculated with Satterthwaite's method.
*** $p < .001$, ** $p < .01$.

In sum, the examination of the moderating role of JS was fruitful, highlighting the importance of justice concerns for understanding the decrease of 3PP under ambiguity. However, it only offered indirect evidence regarding the motivational underpinnings of the effect of ambiguity. Therefore, in Studies 1.4 and 1.5, we used an alternative experimental approach to fill this gap by investigating whether and why third parties would resolve the ambiguity.

Study 1.4

In Study 1.4, we introduced a third condition to our experimental design, where third parties had the opportunity to resolve the ambiguity by revealing perfect information about the norm violation before their punishment decision. Giving third parties the opportunity to resolve the ambiguity should help to distinguish more clearly the aforementioned motivational mechanisms underlying the effect of ambiguity. Specifically, we argued that those who were motivated to react against unfairness in principle but hesitated to punish under ambiguity due to the risk of punishing unfairly, would take any chance to resolve the ambiguity to alleviate their type I error concerns. Conversely, those who aimed to remain passive to avoid incurring costs could keep and capitalize on the ambiguity as a situational justification to passiveness. Hence, the new condition

allowed us to examine a) whether those third parties who opted for resolving the ambiguity – arguably alleviating any type I error concern – punished the disambiguated norm violation, and b) whether those third parties who opted for keeping the situation ambiguous – and thus, a moral wiggle room to avoid incurring costs – remained passive.

We preregistered (https://osf.io/ym4r3) that, in the new condition, Observer and Perpetrator JS would positively predict whether third parties resolved the ambiguity (H3a-b), since we argued that both dispositional measures might capture type I error concerns. Moreover, we predicted that those third parties who chose to resolve the ambiguity would show higher levels of 3PP than those who did not resolve the ambiguity (H4). Under the assumption that those who resolved the ambiguity were genuinely motivated to disambiguate the situation to address any potential norm violation, we further hypothesized that, once they had resolved the ambiguity, they would show even higher levels of 3PP than those in the *no ambiguity* condition (H5). In contrast, we assumed that those who did not resolve the ambiguity aimed to avoid costs and exploit the ambiguity as moral wiggle room; therefore, we predicted that they would show the lowest levels of 3PP, even when compared with those in the *ambiguity* condition (H6).

Moreover, we administered a post-experimental questionnaire to gauge different considerations that third parties might have had when deciding a) whether to resolve the ambiguity and b) whether to punish, including the discussed type I error concerns and cost avoidance.

Method

Design

Study 1.4 followed a between-subject design, with three experimental conditions: *no ambiguity, ambiguity,* and the additional condition where participants faced the same ambiguous norm violation, but, prior to their punishment decision, they could *resolve/not resolve* the ambiguity.

Different from the strategy method used in Studies 1.1-1.3, here, participants only decided in the role of Person C and reacted to a specific distribution by Person A (i.e., a distribution of 1 ECU, given an endowment of 10 ECUs; see below for details about how we elicited this distribution from Person A). The cost of punishment and compensation was fixed (i.e., for every 1 ECU punished or compensated, a cost of ½ ECU).

Participants

We conducted an a priori sample size estimation based on the ambiguity effect observed in Study 1.3. Since in Study 1.4 we did not use the strategy method, we took as a reference the between-subject difference between the *no ambiguity* and *ambiguity* conditions (i.e., using only the first round of the 3PPG) for the decision of the strategy method in which the dictator distributed 1 ECU (i.e., $d = -66$). A sample size of $N = 102$

already guaranteed to detect this effect size with 95% statistical power in a one-tailed t-test, which we originally preregistered. However, before beginning the data collection, we considered a second power estimation to guarantee the potential replication of the Ambiguity x JS interaction. We used as reference the significant Ambiguity x Perpetrator JS interaction that we observed for the single decision of distributing 1 ECU in Study 1.3 ($\beta = .297$). A sample size of $N = 300$ would suffice to detect this interaction effect with 80% statistical power. Since we additionally planned to compare the *no ambiguity* and *ambiguity* conditions with the subset of participants who decided to *resolve/not resolve* the ambiguity in our third experimental condition, we planned to collect double the sample size ($N = 600$) to balance the number of participants in each compared group.

We collected data from 711 participants. We excluded data of 8 who did not finish the study and 70 who failed preregistered comprehension checks about the 3PPG. Therefore, the resulting sample was 633 participants, mainly undergraduate students from diverse disciplines (59% women; age range from 18 to 66, $M = 25.47$, $SD = 5.51$). They received a fixed monetary reward of €2.00 and they could additionally earn up to €5.00 in the 3PPG.[2]

Procedure

Study 1.4 followed a two-session procedure. In the first session, we elicited the targeted norm transgression (i.e., distributions of 1 ECU from an endowment of 10 ECUs). The second session was the actual experiment, where we collected the reactions from participants playing as Person C (for a similar procedure, see Kurzban et al., 2007).

For the first session, we recruited a small group of participants who made decisions in the roles of Person A and Person B (n = 20). To elicit the targeted norm transgression, we simplified Person A's decision to a dichotomous choice between a fair distribution (i.e., 5 ECUs – 5 ECUs) and an unfair distribution (i.e., 1 ECU – 9 ECUs). Note that in the conditions involving ambiguity of the norm violation (i.e., *ambiguity* and *resolve/not resolve* conditions), Person A's endowment was randomly determined between 2 and 10 ECUs. If the endowment was different from 10 ECUs, we could not compare these conditions with the *no ambiguity* condition. Therefore, in order to determine random endowments while assuring that most of them were 10 ECUs, we tweaked the underlying probability distribution, so that only in 10% of the cases Person A would receive an amount different from 10 ECUs. We would later inform participants playing as Person C that Person A's endowment was determined according to a probability distribution that was unknown to them. Once we collected the decisions, we formed dyads and randomly assigned participants to the role of Person A and Person B. Those dyads in which Person A distributed fairly or Person A received an endowment different from 10 ECUs were paid according to Person A's decision. In contrast, for those dyads in which

[2] In Studies 1.4 to 1.6, 1 ECU = €0.50.

Person A decided on the unfair distribution, we applied the average punishment and compensation in each experimental condition as participants in the role of Person C would later decide in the second session.

The second session was, indeed, our actual experiment. After participants provided informed consent and demographic information, they answered the JS scales, embedded in between two filler questionnaires. Then, they were randomly assigned to one of the three conditions and played the 3PPG as Person C.

The *no ambiguity* and *ambiguity* conditions were identical to those in Studies 1.1-1.3. In the *resolve/not resolve* condition, participants received the same information as those in the *ambiguity* condition, with the difference that, before the punishment and compensation decisions, we asked them whether they would like to know how many ECUs Person A had originally received. If they answered "yes", we disclosed that Person A had received an endowment of 10 ECUs; if they answered "no", this information remained unknown. Participants then made their punishment and compensation decisions. Finally, they completed the post-experimental questionnaire.

Measures

Decision to resolve ambiguity. We coded participants' choice to reveal the information regarding Person A's endowment as No – 0 / Yes – 1.

3PP. The measure of 3PP was the total amount of ECUs that participants wished to subtract from Person A (from 0 to 10 ECUs).

Justice sensitivity. We used the 40-item JS Inventory (Schmitt et al., 2010).

Post-experimental questionnaire

The post-experimental questionnaire included ad-hoc-generated items to capture potential concerns regarding the decisions of whether or not a) resolving the ambiguity, and b) punishing Person A. Table 1.5 summarizes these items and the respective dimensions on which they loaded in a Principal Component Analysis (PCA). The complete questionnaire, its detailed psychometric information, and secondary results are included in the Supplementary Material.

Concerns related to resolving the ambiguity. Participants in the *resolve/not resolve* condition reported their level of agreement with several items capturing considerations underlying their decision to resolve the ambiguity, namely, wanting to avoid making an unfair decision (i.e., *type I error*), a sense of *curiosity*, experiencing *situational accountability,* and *cost avoidance.*

Concerns related to punishment. Participants in all conditions reported their level of agreement with items assessing considerations underlying their punishment decision. We tailored three of these items to capture concerns about the *unfairness of their punishment decision* and its potential consequences for *social-image* (items 1, 3 and 4), whereas two other items aimed at capturing *self-image* concerns (items 5 and 6). We

employed additional items to explore concerns about *lacking information* to make the decision, *inequality aversion,* and *cost avoidance*. Note that we originally included further items for type I error concerns (item 2) or lack of information (item 8). However, the PCA showed unexpected loadings for these items and we therefore dropped them.

Table 1.5

Selected items from post-experimental questionnaire after Principal Component Analysis with oblimin rotation.

Concerns related to resolving ambiguity	**Items**
Unfair decision	1. I've been thinking about the risk of being unfair to Person A.
Curiosity	2. I made my decision out of curiosity.
Situational accountability	3. I felt responsible for what happened between Person A and Person B.
Cost avoidance	4. My priority was to avoid costs.
Concerns related to punishment	**Items**
Type I error	
Unfair Decision/ Social-Image	1. I was concerned that my decision about Person A could be unfair.
	3. I was concerned that Person A might see me as an unfair Person based on my decision. 4. I was concerned with whether Person A could perceive me as mean.
Self-image	5. I was concerned about feeling like a malefactor. 6. I would feel bad myself if my decision had been unfair.
Lack of information	7. I felt that my decision was not informed.
Inequality aversion	11. I intended to split the points between Person A and Person B more fairly. 12. I ideally wanted Person A and Person B to receive the same number of points.
Cost avoidance	17. My priority was to avoid costs. 18. I have barely taken into account the cost of my decision. **(R)**

Note. Every item used response options from 0 (*Not at all)* to 5 (*Absolutely*).
From concerns related to punishment, we excluded two items due to unexpected loadings in PCA:
2. "I was worried about whether Person A deserved to have points deducted"
8. "I had all the information I needed to make the decision"

Statistical analyses

First, we used the data from the *no ambiguity* and *ambiguity* conditions in a multiple regression model to test whether our previous findings replicated. The model regressed 3PP on ambiguity, Observer JS, Perpetrator JS, and the Ambiguity x Observer JS and Ambiguity x Perpetrator JS interactions.

Second, with the data from the *resolve/not resolve* condition, we used logistic regression to test the effect of Observer and Perpetrator JS on the decision to resolve the ambiguity.

Finally, to test for the expected differences in 3PP across conditions (*ambiguity* and *no ambiguity*) and/or self-selected groups (*resolved* and *not resolved*) we fitted three independent regression models with different data subsets. Each model included one dummy-coded variable as predictor, which captured the subset of groups that the model aimed to compare: Dummy 1 (0 – Not resolved, 1 – Resolved), Dummy 2 (0 – No Ambiguity, 1 – Resolved), and Dummy 3 (0 – Ambiguity, 1 – Not resolved), respectively.

Results

Main results

Comparing the *no ambiguity* and the *ambiguity* conditions, we observed that 3PP was significantly lower in the latter than in the former (see Table 1.6, Model 1). Furthermore, we observed a significant positive effect of Observer JS, but not Perpetrator JS. The two-way interactions were not significant. However, in a model without Perpetrator JS, the Ambiguity x Observer JS interaction was significant (see Model 2) and showed a similar pattern as in previous studies (see Figure 1.2).

Table 1.6

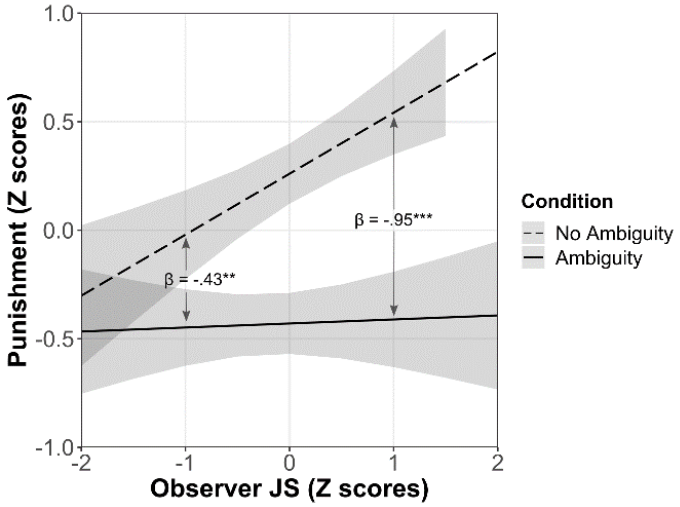
Tested multiple regression models on punishment in Study 1.4.

	Model 1 (preregistered)			**Model 2**		
Parameters	β [95% CI]	***t***	***p***	β [95% CI]	***t***	***p***
Ambiguity of norm violation	-.68 [-.88, -.48]	-6.65	**< .001*** **	-.68 [-.88, -.48]	-6.68	**< .001*** **
Perpetrator JS	.13 [-.05, .32]	1.39	.165	-	-	-
Observer JS	.20 [.02, .38]	2.14	**.033***	.28 [.14, .42]	3.87	**< .001*** **
Ambiguity x Perpetrator JS	-.08 [-.33, .17]	-0.61	.545	-	-	-
Ambiguity x Observer JS	-.21 [-.45, .04]	-1.67	.097	-.26 [-.46, -.06]	-2.57	**.011***
Observations	328			328		
R^2 / Adj. R^2	.176 / .163			.170 / .162		

Note. *** $p < .001$, ** $p < .01$, * $p < .05$.

Figure 1.2

Two-way interaction between ambiguity and Observer JS in Study 1.4.



Note. Standardized regression coefficients represent the effect of ambiguity at +1SD and -1SD Observer JS based on simple slope analysis.
*** $p < .001$. Band widths 95% CIs.

Next, we examined the subset of participants in the *resolve/not resolve* condition. Most chose to resolve the ambiguity (87.9%), whereas a minority did not (12.1%). The logistic regression model showed that this decision was not predicted by Observer JS, Wald (1) = -0.19, $p = .849$, OR = 0.96, 95% CI [0.60, 1.54], nor Perpetrator JS, Wald (1) = -1.49, $p = .137$, OR = 0.72, 95% CI [0.46, 1.11]; thus, H3 was not supported.

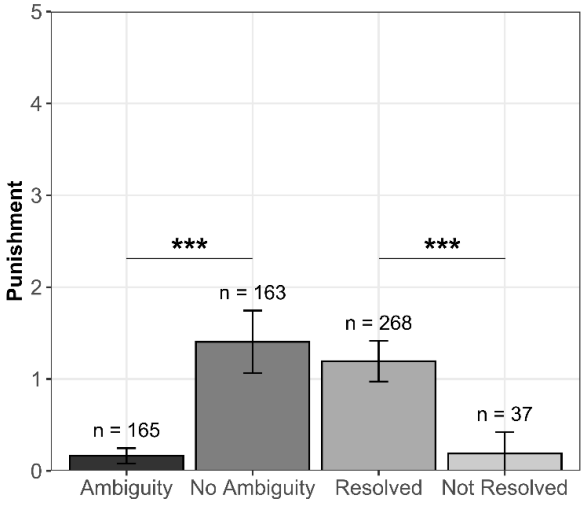
Furthermore, we compared the levels of 3PP across the different subsets of participants (see Figure 1.3). The first regression model showed those who resolved the ambiguity punished significantly more than those who did not – supporting H4; Dummy 1, β = -.57, $t(303) = -3.29$, $p = .001$, 95% CI [-.91, -.23]. The second model showed that those who resolved the ambiguity did not punish more than those in the *no ambiguity* condition – not supporting H5; Dummy 2, β = -.11, $t(429) = -1.07$, $p = .285$, 95% CI [-.30, .09]. The third model showed that those who did not resolve the ambiguity did not punish less than those in the *ambiguity* condition – not supporting H6; Dummy 3, β = .04, $t(200) = 0.24$, $p = .807$, 95% CI [-.32, .40].

Exploratory results

In the post-experimental questionnaire, we observed that participants who resolved the ambiguity reported significantly higher concerns about making an unfair decision and being accountable for the situation, significantly higher curiosity and significantly lower concerns about avoiding costs than those who did not resolve the ambiguity (see Figure 1.4A). Despite these group differences, a visual comparison indicated that the primary considerations of those who resolved the ambiguity were curiosity and cost avoidance.

Figure 1.3

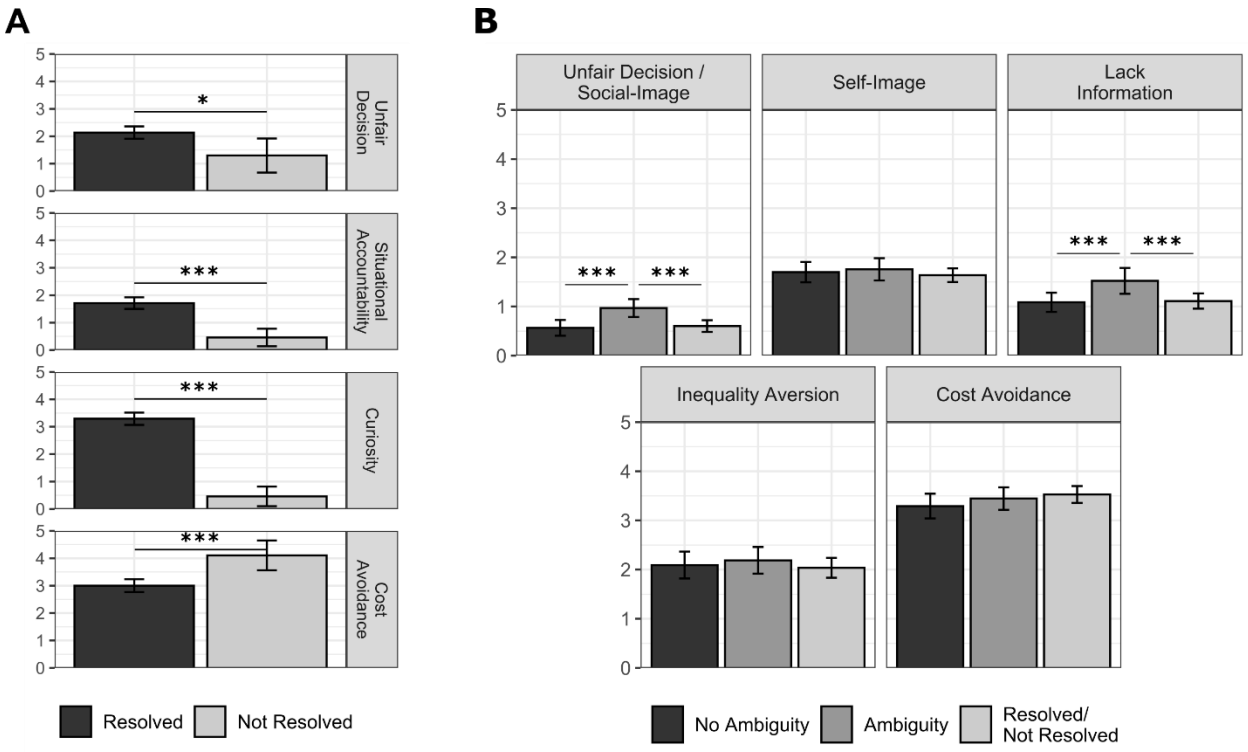
Levels of punishment in experimental conditions and self-selected groups in Study 1.4.



Note. *** $p < .001$. Error bars 95% CIs.

Figure 1.4

Differences in types of concerns associated with the decision to resolve the ambiguity (A) and the decision to punish (B) in Study 1.4.



Note. *p*-values correspond to Welch independent sample t-tests (A) and linear regressions including Condition as dummy-coded predictor (B).
*** $p < .001$, ** $p < .01$, * $p < .05$. 95% CIs error bars.

Regarding the considerations about the punishment decision (see Figure 1.4B), participants in the *ambiguity* condition reported a significantly higher perceived lack of information and higher concerns about making an unfair decision and social image than in the *no ambiguity* and *resolve/not resolve* conditions. The *no ambiguity* and the *resolve/not resolve* conditions did not differ significantly in any scale. This is not surprising, since most participants in the *resolve/not resolve* condition decided to resolve the ambiguity. Therefore, the psychological situation for them when making the punishment decision was effectively the same as for those in the *no ambiguity* condition.

Discussion

In Study 1.4, we replicated the negative effect of ambiguity on 3PP and the moderating role of Observer JS (in the model without Perpetrator JS).

Additionally, Study 1.4 shed new light on whether ambiguity induced type I error concerns that potentially hindered 3PP. Those who had the opportunity, and decided to, resolve the ambiguity subsequently punished to similar extents than those in the *no ambiguity* condition. This indicated that resolving the ambiguity alleviated any concerns that otherwise made participants hesitate to punish. Results from the post-experimental questionnaire supported our argument that such concerns under ambiguity related to the possibility of making an unfair decision and being perceived as unfair. In a similar vein, those participants who decided to resolve the ambiguity also reported that this decision was (at least in part) motivated by the possibility of making an unfair decision and perceived situational accountability.

The post-experimental questionnaire also indicated that resolving the ambiguity had to do with a sense of curiosity. Thus, among those participants who resolved the ambiguity, it was difficult to differentiate between those who mainly intended to inform their decision to react against the (potential) norm violation and those who were merely curious. This issue could have weakened, for example, the expected – yet unobserved – relationship between JS and the decision to resolve the ambiguity. Similarly, it could explain why those who resolved the ambiguity did not punish more than those in the *no ambiguity* condition did, as we had expected.

Finally, cost avoidance seemed to be a default priority across conditions, but those who did not resolve the ambiguity reported to have weighted in costs even more. This observation aligned with the argument that some third parties may exploit ambiguity as moral wiggle room to avoid incurring costs.

Study 1.5

In our next study, we used the setup of Study 1.4, with the crucial difference that the decision to resolve the ambiguity was costly. With this modification, we aimed to distinguish more clearly those third parties who would resolve the ambiguity merely out of curiosity from those who would do so to inform their punishment decision to avoid punishing unfairly. We reasoned that, for the former, the associated cost would

generally discourage them from resolving the ambiguity, whereas for the latter, resolving ambiguity presumably had further value as means to obtain critical information that alleviates their concerns about punishing unfairly, and therefore, it remained a worthwhile choice.

Method

Participants

Taking as reference the same a priori power estimation we conducted for Study 1.4, we collected data from 857 participants from a German online panel. We excluded data from 52 who did not finish the study and 119 who failed preregistered comprehension checks about the 3PPG and the manipulation of ambiguity (e.g., "How many [ECUs] does Person A receive?"). The resulting sample was 686 participants, overall with high education (62%) and a more representative distribution of gender and age than our previous studies (50% women; age range from 18 to 82, $M = 46.53$, $SD = 14.41$). Participants received a fixed monetary reward of €2.00, and they could earn up to €5.00 in the 3PPG.[3]

Procedure

To introduce a cost in the *resolve/not resolve* condition without altering the payoff structure of the 3PPG, participants could win a €5.00 voucher in a raffle with 1/20 probability at the end of the study. We provided this information in every experimental condition. In the *resolve/not resolve* condition, the probability of winning the raffle was conditional on the decision to resolve the ambiguity. Specifically, if participants decided to resolve the ambiguity, the probability of earning the voucher would decrease to 1/30.

Any other detail about the procedure was identical to Study 1.4, including the elicitation of the norm violation in a different session with a small sample making decisions as Person A and B (n = 11).

Design, measures, and statistical analyses

The design, measures, and performed statistical analyses were identical to Study 1.4.

Post-experimental questionnaire

Some items in the post-experimental questionnaire slightly differed from Study 1.4. Here, we summarize the relevant changes for the interpretation of the results presented below. The complete questionnaire and information on its psychometrics are included in the Supplementary Material.

First, the two items intended to capture *self-image* concerns, did not load onto the same factor in the PCA. To facilitate a clearer comparison between studies, we decided to drop this dimension from our exploratory analyses.

[3] In Studies 4, 5 and 6, 1 ECU = €0.50.

Second, we used two new items to measure perceived *lack of information*, which the PCA showed to load onto the same factor: "I felt that I didn't have enough information to make my decision" and "I had all the information I needed to make the decision" (R).

Results

Main results

We found that 3PP was significantly lower in the *ambiguity* condition than in the *no ambiguity* condition, supporting H1 (see Table 1.7). We did not observe significant effects of Observer JS or Perpetrator JS. Different from previous studies, the Ambiguity x Observer JS interaction was not significant, not supporting H2b'. Instead, we observed a significant Ambiguity x Perpetrator JS interaction. Simple slopes indicated that the effect of ambiguity was only significant among those with high Perpetrator JS (see Figure 1.5).

In the *resolve/not resolve* condition, a third of the participants resolved the ambiguity despite the incurred cost (36.1%), whereas most participants decided not to resolve the ambiguity (63.9%). The logistic regression model showed that this decision was not predicted by Observer JS, Wald (1) = -0.61, p = .539, OR = 0.91, 95%CI [0.69, 1.21], nor Perpetrator JS, Wald (1) = -0.77, p = .439, OR = 0.90, 95%CI [0.68, 1.18], not supporting H3a-b.

Next, we compared the levels of 3PP across the different subsets of participants (see Figure 1.6). The first regression model indicated that those who resolved the ambiguity punished significantly more than those who did not – supporting H4; Dummy 1 β = -1.11, t(339) = -11.58, p < .001, 95% CI [-1.29, -0.92]. The second model showed that those who resolved the ambiguity also punished significantly more than those in the *no ambiguity* condition – supporting H5; Dummy 2, β = .50, t(294) = 4.34, p < .001, 95% CI [.27, .72]. The third model showed that those who did not resolve the ambiguity punished significantly less than those in the *ambiguity* condition – supporting H6; Dummy 3, β = -.20, t(388) = -1.95, p = .052, 95% CI [-.40, .00].

Table 1.7

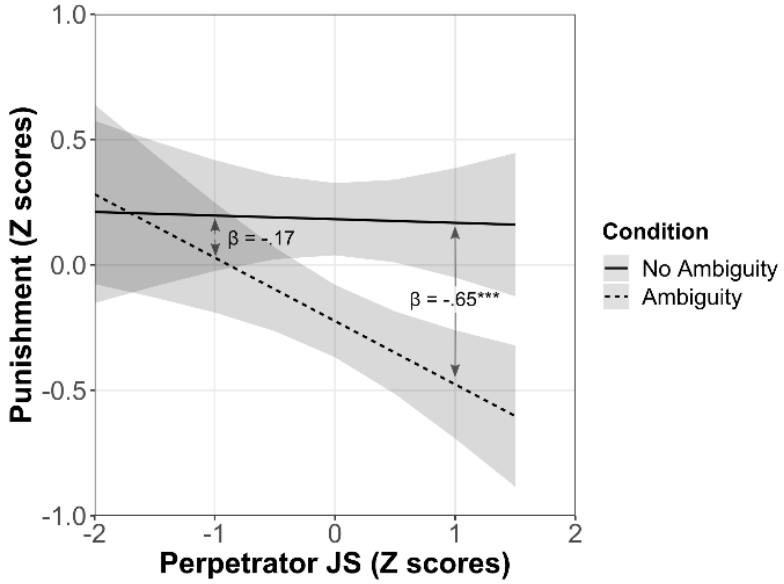
Tested multiple regression model on Punishment in Study 1.5.

Parameters	β [95% CI]	***t***	***p***
Ambiguity of norm violation	-.41 [-.62, -.21]	-3.94	**<.001*** **
Perpetrator JS	-.01 [-.19, .16]	-0.17	.865
Observer JS	.03 [-.14, .20]	0.34	.737
Ambiguity x Perpetrator JS	-.24 [-.48, -.01]	-2.01	**.045***
Ambiguity x Observer JS	.03 [-.21, .27]	0.27	.788
Observations		345	
R^2 / Adj. R^2		.070 / .056	

Note. *** p < .001, ** p < .01, * p < .05.

Figure 1.5

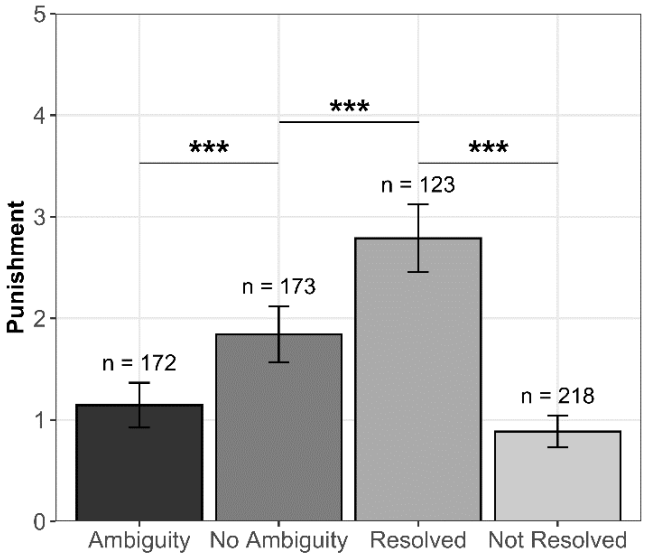
Two-way interaction between ambiguity and Perpetrator JS in Study 1.5.



Note. Standardized regression coefficients represent the effect of ambiguity at +1SD and -1SD Perpetrator JS based on simple slope analyses.
*** $p < .001$. Band widths 95% CIs.

Figure 1.6

Levels of punishment in experimental conditions and self-selected groups in Study 1.5.



Note. *** $p < .001$. Error bars 95% CIs.

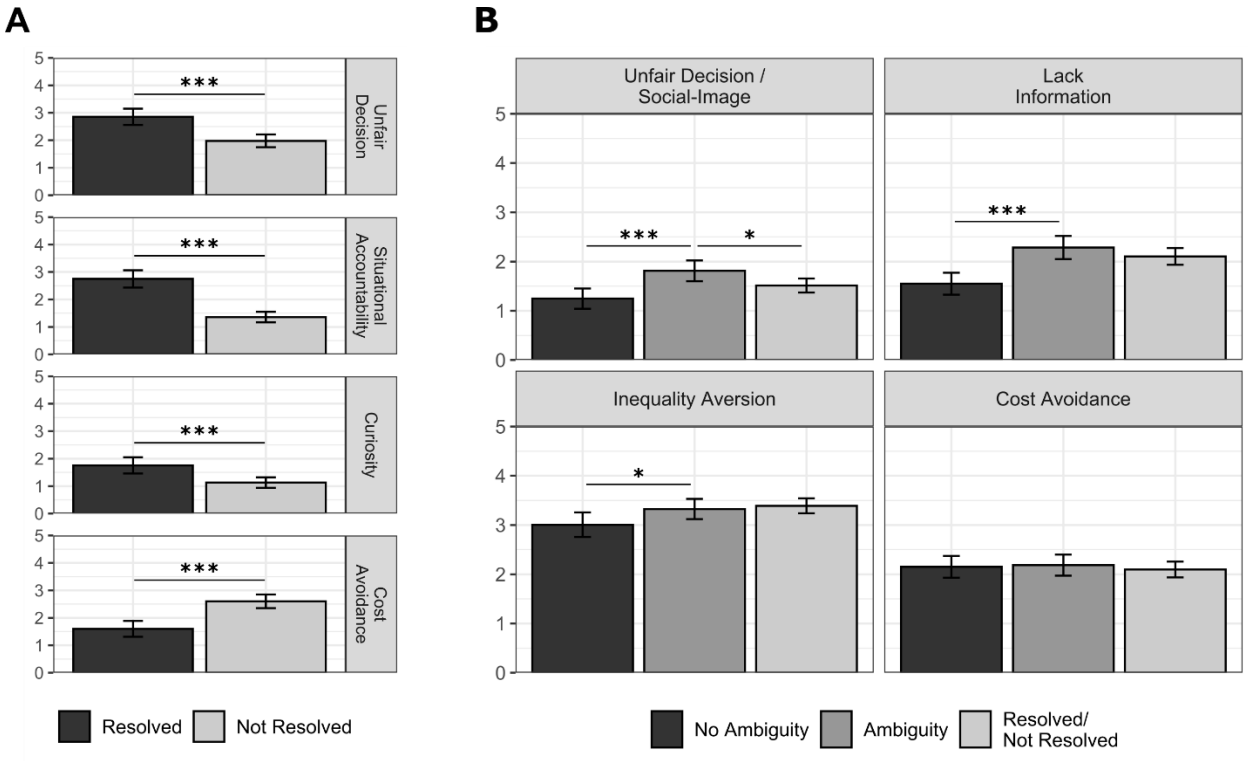
Exploratory results

In the *resolve/not resolve* condition, similarly to Study 1.4, we observed that participants who resolved the ambiguity reported significantly higher concerns about making an unfair decision and being accountable for the situation, significantly higher curiosity and significantly lower concerns about avoiding costs than those who did not resolve the ambiguity (see Figure 1.7A). In contrast to Study 1.4, the primary considerations of those who resolved the ambiguity were making an unfair decision and perceived situational accountability, which indirectly suggested that we succeeded in discouraging participants who would have resolved the ambiguity merely out of curiosity from doing so.

Regarding the considerations about the punishment decision (see Figure 1.7B), participants in the *ambiguity* condition reported a significantly higher perceived lack of information than those in the *no ambiguity* condition, but similar to those in the *resolve/not resolve* condition. We further observed that in the *ambiguity* condition, participants reported significantly higher concerns about making an unfair decision than in the *no ambiguity* and *resolve/not resolve* conditions.

Figure 1.7

Differences in types of concerns associated with the decision to resolve the ambiguity (A) and the decision to punish (B) in Study 1.5.



Note. *p*-values correspond to Welch independent sample t-tests (A) and linear regressions including Condition as predictor (B).
*** $p < .001$, ** $p < .01$, * $p < .05$. Error bars 95% CIs.

Discussion

We again replicated the effect of ambiguity and, in contrast to previous studies, we found Perpetrator and not Observer JS to moderate this effect. However, Study 1.5 extended our results by teasing apart people who were susceptible to ambiguity out of different motivations.

First, those participants who resolved the ambiguity were motivated to punish the disambiguated norm violation, as their higher levels of 3PP indicate. Yet, we infer from the lower 3PP in the *ambiguity* (vs. *no ambiguity*) condition that these participants would have refrained from punishing if they had not had the option to resolve the ambiguity (and, thus, to know whether their punishment was unfair). In fact, they willingly incurred additional costs for revealing that information, which highlights that they cared about addressing correctly a norm violation while avoiding being unfair themselves. The post-experimental questionnaire further supported this notion. This group of participants reported higher concerns about making an unfair decision and higher perceived situational accountability than those who did not resolve the ambiguity. Additionally, these type I error concerns resonated with how participants in the *ambiguity* condition perceived the situation. Specifically, this latter group perceived a higher lack of information and reported higher concerns for making an unfair decision or being perceived as unfair.

Second, those people who opted for keeping the situation ambiguous showed distinct motivations related to cost avoidance. In the current study, we disincentivized the option of resolving the ambiguity, although the costs we introduced were negligible (i.e., between €5.00 with probability 1/20 and 1/30, the expected value difference was €0.08). However, these minor costs were sufficient to lead a substantial percentage of third parties to avoid resolving the ambiguity. More crucially, this subgroup exerted the lowest levels of 3PP, even when compared to those in the *ambiguity* condition without the option to resolve the ambiguity. In line with the literature on moral wiggle room, we argue that these observations could indicate an intentional use of the situational ambiguity for the avoidance of individual costs.

Unexpectedly, as in Study 1.4, we did not find Observer and Perpetrator JS to predict the decision to resolve the ambiguity. Does this suggest that the JS scales did not capture type I error concerns? In the post-experimental questionnaire, both JS scales correlated significantly positively with concerns about making an unfair decision when it came to resolving the ambiguity or deciding to punish (see Supplementary Material). At this point, this set of results is puzzling. It could well be that people with high JS did not experience particularly high type I error concerns that motivated them to resolve the ambiguity, but it could also suggest that, for these people, other considerations might have played an opposing role to the type I error concerns when considering to resolve the ambiguity.

Study 1.6

We designed Study 1.6 to exclude a potential threat to the internal validity of our ambiguity manipulation. As a reminder, in our *no ambiguity* condition, we fixed Person A's endowment to 10 ECUs, whereas in the *ambiguity* condition Person A's endowment randomly varied between 2 and 10 ECUs. In the latter condition, third parties could make assumptions about the endowment's probability distribution (e.g., uniform distribution[4]) and consequently infer a lower expected value of the endowment in the *ambiguity* condition than in the *no ambiguity* condition. If this was the case, the lower expected value in the *ambiguity* condition could explain the reduction of 3PP.

For ruling out this explanation, in Study 1.6, we separately manipulated *ambiguity* (no ambiguity vs. ambiguity) and Person A's endowment *expected value* (low vs. high). Specifically, in the *no ambiguity/low expected value* condition, Person A received a fixed endowment of 6 ECUs, which corresponded with the expected value of our original *ambiguity* condition, now relabeled *ambiguity/low expected value* condition (i.e., endowment from 2 to 10 ECUs). In the *ambiguity/high expected value* condition, Person A received a random endowment from 2 to 18 ECUs, the expected value of which corresponded to the fixed endowment of our original *no ambiguity* condition, now relabeled *no ambiguity/high expected value* condition (i.e., endowment of 10 ECUs). Hence, our first preregistered hypothesis (https://osf.io/s7rza) was that the ambiguity effect would remain significant across expected values (H7), even if the interaction between ambiguity and expected value turned out significant.

In Study 1.6, we continued exploring the underlying mechanisms of the effect of ambiguity. For this purpose, we measured the third parties' type I error concerns through a post-experimental questionnaire similar to the one in Studies 1.4 and 1.5. We predicted that the introduction of ambiguity would relate to these type I error concerns, and that the latter would predict 3PP (H8). Moreover, we included Social Value Orientation (SVO; van Lange et al., 1997), as an additional validated measure of fairness concerns, to explore whether social preferences and, specifically, inequality aversion played a similar role under ambiguity to the one JS arguably played.

[4] In the post-experimental questionnaire of Studies 4 and 5, we asked participants in the ambiguity conditions to report their assumptions about Person A's endowments probability distribution (see Supplementary Material). Roughly, 38-54% reported to assume a uniform distribution (i.e., "any amount from 2 to 10 was equally likely"), whereas 28-31% reported not to have made any assumption.

Method

Design

We used a 2x2 between-subject design, manipulating *ambiguity* (no ambiguity vs. ambiguity) and the endowment's *expected value* (low vs high; see Table 1.8). As in Studies 1.4 and 1.5, participants made a single decision in response to a specific unfair distribution by Person A (i.e., a distribution of 1 ECU). The cost of punishment and compensation was ½ ECU per every ECU that they punished or compensated.

Table 1.8

Between-subject design of Study 1.6.

	No ambiguity	Ambiguity
Low expected value	6 ECUs	[2 ECUs-10 ECUs]
High expected value	10 ECUs	[2 ECUs-18 ECUs]

Note. The shaded cells correspond to the original conditions compared in Studies 1.1-1.5.

Participants

Under the possibility of observing a smaller ambiguity effect in the *low expected value* conditions, we conducted a safeguard power analysis. We considered the smallest ambiguity effect from our previous studies (i.e., d = -0.38 in Study 1.5) and took as a reference the lower bound of its 90% confidence interval (i.e., d = -.21). To detect this effect size in a one-tailed t-test (α = .05), we would need 780 observations to guarantee 90% statistical power.

The present study consisted of two stages (see justification below). To counter the dropout rates between stages and the exclusion based on preregistered criteria, we recruited 1116 participants for stage 1 from a German online panel. Of these, 829 completed stage 2 (i.e., 26% dropout rate). We excluded data of 51, who failed preregistered comprehension checks about the 3PPG. Therefore, the resulting sample consisted of 778 participants, mostly undergraduate students from diverse disciplines (77% women, age range from 18 to 65, M = 24.04, SD = 4.75). They received a fixed monetary reward of €0.50 per session. In addition, they could earn up to €1.00 in the SVO task and up to €5.00 in the 3PPG, but only if they completed both sessions.

Procedure

As in Studies 1.4 and 1.5, we independently recruited a small sample (n = 29) who made decisions as Person A and B in the different conditions to elicit the norm trasngressions.

Our actual experiment was divided in two stages in order to avoid any carry-over effects between the payoff of the SVO task and the payoff of the 3PPG. In stage 1, participants completed our measures of JS

and SVO and provided demographic information. We contacted participants 12 hours later to take part in stage 2; however, they generally took more than a day to do so (in hours, $M = 29.53$, $SD = 25.53$). During stage 2, we randomly assigned participants to play the 3PPG in the role of Person C in one of the experimental conditions. Subsequently, they completed the post-experimental questionnaire.

Measures

3PP. As in Studies 1.4. and 1.5, we assessed 3PP as the total amount of ECUs that participants wished to subtract from Person A (from 0 to 10 ECUs).

Justice sensitivity. We used the 40-item JS Inventory (Schmitt et al., 2010).

SVO. We used the SVO slider measure (Murphy et al., 2011), which consists of 15 items. In this study, the distributed points had a real monetary value of €0.01. At the end of the study, we grouped participants in dyads, we randomly assigned each dyad's member to the role of giver or receiver, and we paid them according to one of the 15 decisions selected at random. Following Murphy et al.'s (2011) indications, we used the nine secondary items to calculate the *index of inequality aversion*, which captured the average normalized difference between participants' choices and the option that maximized equality. Thus, low values of this index indicated high inequality aversion. For further information about the distributions of these measures, see Supplementary Material.

Post-experimental questionnaire

Type I error concerns. We measured type I error concerns with the same three items used in Studies 1.4 and 1.5 ($\alpha = .86$). We deviated from our preregistration and dropped a fourth item (i.e., "I was concerned about feeling like a malefactor") to keep our measure consistent across studies.

The post-experimental questionnaire assessed other considerations that participants could have regarding their decision to punish Person A. These included *lack of information*, *cost avoidance,* and *inequality aversion*, which we assessed with the same items used in Studies 1.4 and 1.5.

Additionally, the questionnaire measured participants' perceptions about the situation that we used as manipulation checks.

Perceived unfairness of Person A's decision. Since the degree of perceived unfairness of the distribution of Person A could differ across conditions with different expected values (e.g., 1 ECU out of 6 vs. 10 ECUs), we used two unipolar scales to measure how *fair* and how *unfair* participants found Person A's distribution (0 – *Not at all*, 5 – *Extremely*). We computed a difference score, with higher scores representing higher unfairness.

Perceived ambiguity of the norm violation. The no ambiguity and ambiguity conditions could differ in the degree of perceived ambiguity about the norm violation. Moreover, the ambiguity conditions

could also vary in this respect, since the ambiguity/high expected value condition used a wider range of endowment values (i.e., €[2-10] vs. €[2-18]). Thus, we used two items to measure the level of perceived ambiguity when judging Person A's decision: "How uncertain do you feel regarding your evaluation of Person A's decision?", "How difficult do you find to judge Person A's decision?" (Spearman-Brown's estimate, ρ = .83).

Statistical analyses

We first examined the manipulations checks through two independent regression models. We respectively regressed perceived unfairness and perceived ambiguity of the norm violation on ambiguity (0 – No ambiguity, 1 – Ambiguity), expected value (0 – Low expected value, 1 – High expected value) and their interaction.

Next, we tested H7 by regressing 3PP on ambiguity, expected value and their interaction.

With regard to H8, we examined the bivariate correlations between ambiguity, type I error concerns, and punishment. We did not test the preregistered mediation model for two reasons: first, because we did not find all expected bivariate correlations, and second, because we measured type I error concerns after punishment, which violated the presumption of temporal ordering of mediation analysis.

Results

Manipulation checks

For participants' perceived unfairness (see Figure 1.8A), we found a significant Ambiguity x Expected Value interaction term, β = -0.39, $t(768) = -2.88$, $p = .004$, 95% CI [-0.66, -0.13]. Participants perceived significantly less unfairness under *ambiguity* than under *no ambiguity* in the *low expected value* conditions – as the significant conditional main effect of Ambiguity indicated; β = -0.32, $t(768) = -3.31$, $p < .001$, 95% CI [-0.51, -0.13] –, and this difference was significantly more pronounced in the *high expected value* conditions. At the same time, participants perceived significantly more unfairness when Person A had 10 ECUs (*high expected value*) than when Person A had 6 ECUs (*low expected value*) in the *no ambiguity* conditions – as the significant conditional main effect of Expected Value indicated; β = .55, $t(768) = 5.76$, $p < .001$, 95% CI [0.37, 0.74]. Yet, this difference was significantly less pronounced in the *ambiguity* conditions.

Regarding perceived ambiguity (see Figure 1.8B), we did not find a significant Ambiguity x Expected Value interaction term. Participants perceived significantly higher ambiguity in the *ambiguity* than in the *no ambiguity* conditions – as the Ambiguity term indicated; β = .87, $t(769) = 9.66$, $p < .001$, 95% CI [0.70, 1.05] –, and significantly lower ambiguity in the *high* (vs. *low*) *expected value* condition – as the Expected Value term indicated; β = -0.22, $t(769) = -2.43$, $p = .015$, 95% CI [-0.40, -0.04].

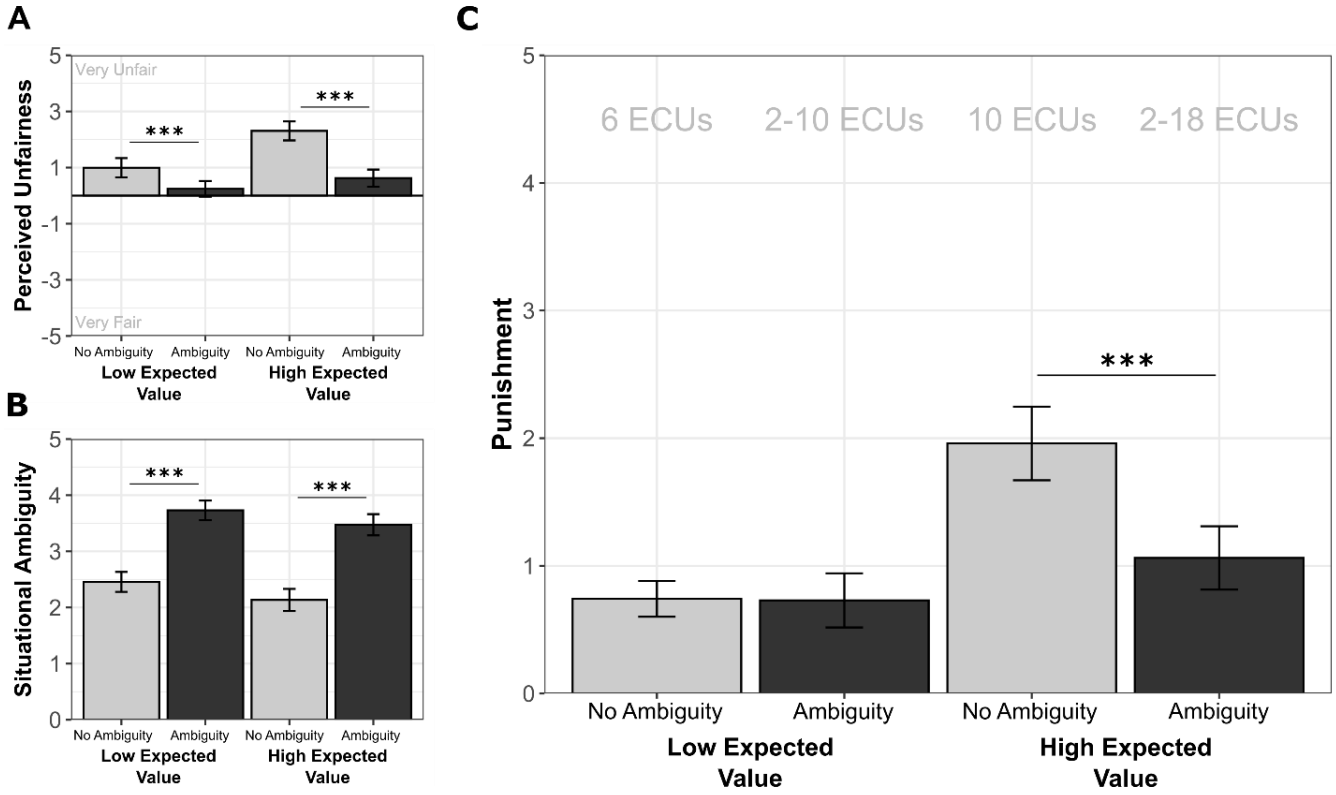
Main results

The model for 3PP revealed a significant Ambiguity x Expected Value interaction term, β = -0.52, $t(774) = -3.81$, $p < .001$, 95% CI [-0.79, -0.25] (see Figure 1.8C). Participants punished significantly less under *ambiguity* than under *no ambiguity* in the *high expected value* conditions. This was not the case in the *low expected value* conditions – as the non-significant Ambiguity term indicated; β = -0.01, $t(774) = -0.069$, $p = .945$, 95% CI [-0.20, 0.18]. Furthermore, participants punished significantly more in the *high* (vs. *low*) *expected value* condition under *no ambiguity* – as the significant Expected Value term indicated; β = 0.72, $t(774) = 7.40$, $p < .001$, 95% CI [0.53, 0.91] –, but this difference was less pronounced under *ambiguity*.

Since we observed the ambiguity effect in the *high expected value* conditions, we only used this data subset to check the bivariate correlations between the ambiguity manipulation, type I error concerns, and punishment. We found that ambiguity was significantly associated with type I error concerns, $r(386) = .15$, $p = .004$; however, type I error concerns were not associated with 3PP, $r(386) = -.02$, $p = .684$.

Figure 1.8

Levels of perceived unfairness (A), perceived situational ambiguity (B), and punishment (C) across experimental conditions in Study 1.6.



Note. *** $p < .001$, ** $p < .01$, * $p < .05$. Error bars 95% CIs.

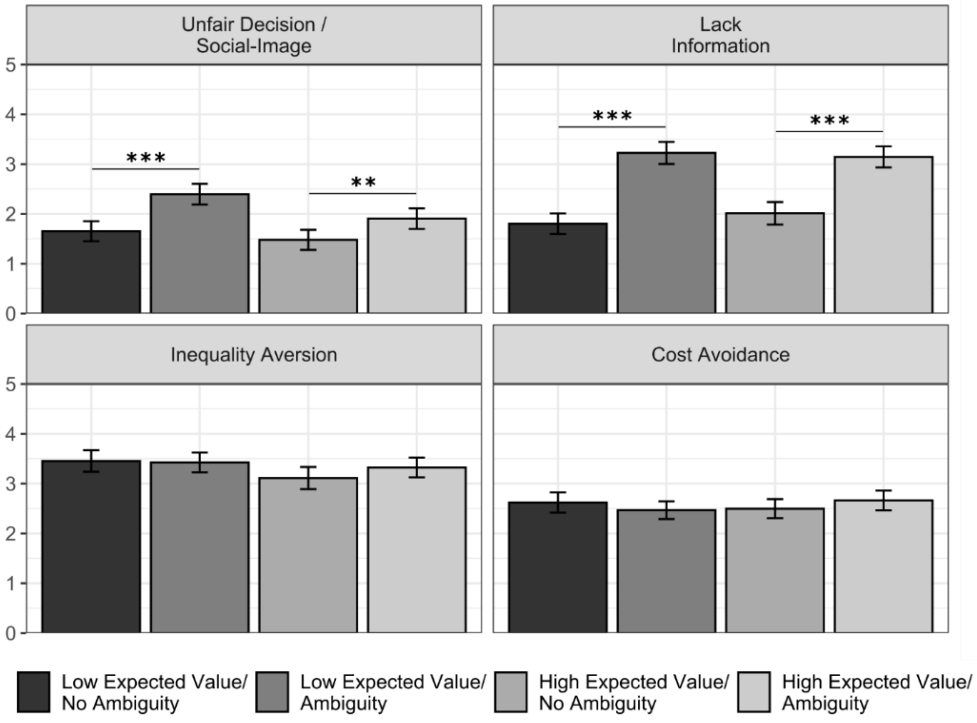
Exploratory results

We summarized the secondary findings from the post-experimental questionnaire in Figure 1.9. Since they offered a similar picture to Studies 1.4 and 1.5, we will not discuss them further.

Lastly, our exploratory results about the role of JS and the SVO index of inequality aversion showed that neither Observer JS nor SVO moderated the effect of ambiguity in the *high expected value* conditions.

Figure 1.9

Differences in types of concerns associated with the decision to punish in Study 1.6.



Note. *p*-values correspond to linear regressions including Ambiguity and Expected Value as predictors. ** $p < .01$, * $p < .05$. Error bars 95% CIs.

Discussion

Study 1.6 established that ambiguity of the norm violation reduced 3PP when holding constant the expected value of Person A's endowment, excluding the discussed threat to the internal validity of our ambiguity manipulation. Moreover, we found again that ambiguity heightened type I error concerns. Further aspects of our results need special attention.

The ambiguity effect emerged in the *high expected value* conditions, but not in the *low expected value* conditions. The pattern of results indicates that this was the case because 3PP was unexpectedly low in the *no ambiguity/low expected value* condition. In light of the low levels of perceived unfairness in the *low expected value* conditions, it seems plausible that most participants did not perceive the behavior of Person A as a norm

violation, even without ambiguity. By contrast, in the *high expected value* condition, people perceived higher unfairness under *no ambiguity* than under *ambiguity*, which could explain why the ambiguity effect held. These results suggest that ambiguity might decrease 3PP partly because people might be less likely to perceive the norm violation.

However, the ambiguity effect also related to how difficult people found to judge whether Person A behaved unfairly or not, as the levels of perceived ambiguity indicated. This was apparent in both the *high expected value* and *low expected value* conditions, with participants reporting to struggle more to judge Person A's behavior under *ambiguity* (vs. *no ambiguity*). If the ambiguity effect on 3PP only emerged in the *high expected value* conditions, it could be because, under *no ambiguity*, participants reported not to struggle when judging Person A's behavior as clearly unfair, and accordingly, showed higher 3PP. In contrast, in the *no ambiguity/low expected value* condition, participants reported not to struggle but perceived less unfairness. Taken together, our findings suggest that, under ambiguity, third parties are less likely to perceive norm violations because ambiguity hinders this perception, and therefore, levels of 3PP are expectedly lower.

Lastly, we observed higher type I error concerns under ambiguity. Yet, we did not find these concerns to correlate with 3PP. A possible explanation could be the post-experimental assessment of type I error concerns. By measuring these concerns after the punishment decision, some participants might have reported concerns about punishing unfairly that arose in reaction to their already exerted punishment – presumably correlating positively, instead of negatively, with 3PP.

General discussion

Despite the crucial role of 3PP for the maintenance of social norms (Fehr & Fischbacher, 2004), this behavior might be less prevalent than previous lab studies suggest, as some have discussed (Pedersen et al., 2018). In the setup of prior studies, a norm violation was generally easily identifiable. By contrast, in real-life settings, third parties often receive ambiguous information, rendering the interpretation of a situation as a norm violation complicated. Our research highlights that the ambiguity of the norm violation can indeed constitute a critical boundary of 3PP.

In six studies, we consistently observed that, under ambiguity (i.e., if third parties received imperfect information affecting the identification of the norm violation), levels of 3PP decreased, compared to when perfect information was provided. This effect of ambiguity has theoretical implications for understanding the decision-making preceding 3PP. Explanatory frameworks of bystander intervention have proposed that a critical step for third parties to react against norm violations is to interpret them as such (e.g., Baumert et al., 2013). If the situation entails ambiguity concerning the norm violation, we argued that its interpretation is hampered, which exerts downstream effects on how people ponder over engaging in 3PP. Specifically, ambiguity may elicit concerns about the risk of punishing unfairly, given that in one possible state of the

world a norm violation has not actually occurred. Simultaneously, ambiguity may provide a situational justification or "moral wiggle room" for those who intend to avoid costs (Dana et al., 2007; Stüber, 2020). Our research suggested that these two mechanisms could plausibly explain why 3PP decreases under ambiguity of the norm violation.

Across Studies 1.1-1.4, we observed that the effect of ambiguity was moderated by Observer JS, a construct capturing dispositional other-oriented concerns for justice (Baumert & Schmitt, 2016). Under ambiguity, third parties with high Observer JS reduced their level of 3PP more pronouncedly than those with low Observer JS. Why would stronger justice concerns lead third parties to refrain from punishing an ambiguous norm violation? A theoretically plausible explanation relates to the aforementioned risk of punishing unfairly.

Some people may refrain from punishing an ambiguous norm violation given the possibility that the norm violation has not actually occurred and that punishment is consequently unjustified. This notion received support from Studies 1.4 and 1.5, where third parties had the opportunity to resolve the ambiguity of the norm violation (and hence, to exclude that punishment was unfair). Those who resolved the ambiguity (even by incurring additional costs) reported being relatively more concerned about the possibility of punishing unfairly; therefore, resolving the ambiguity could have offered them the opportunity to alleviate these concerns. In fact, once they had resolved the ambiguity, their levels of 3PP matched (Study 1.4), or even exceeded (Study 1.5), those from third parties who did not face ambiguity of the norm violation. Taken together, our findings suggest that, for some third parties, ambiguity induces concerns about punishing unfairly that make them refrain from exerting 3PP.

Our research highlighted that a second important consideration that determined the behavior of third parties under (no) ambiguity was cost avoidance. If the third parties' main priority was to avoid costs by remaining passive, one could think that the ambiguity of the norm violation would be irrelevant for them. However, our findings suggest that, on the contrary, ambiguity also influenced the behavior of those third parties who were arguably more inclined to avoid costs. For instance, people with low Observer JS – thus, more prone to weight in the costs of 3PP (Lotz, Baumert, et al., 2011) – also reduced their 3PP under ambiguity, albeit to a lesser extent than those with high Observer JS (Studies 1.1-1.3). In other words, for some third parties, ambiguity facilitated to some extent the option of remaining passive. As discussed above, some people may deliberately use ambiguity as moral wiggle room to excuse their decision to remain passive (Dana et al., 2007; Stüber, 2020). In our research, this was apparent among those third parties who decided to keep the ambiguity (Studies 1.4 and 1.5). These people reported a heightened concern about avoiding costs and showed the lowest levels of 3PP.

More generally, the present research identified a situational boundary of 3PP in the lab – namely, the ambiguity of the norm violation – that could clarify when 3PP occurs in real-life settings. At present, we cannot ascertain that our findings would generalize to the field; however, we suspect that the role of ambiguity might be crucial in accounting for the already discussed discrepancies between lab and field studies (Guala, 2012; Pedersen et al., 2018). Ambiguity should be a relevant explanatory factor of 3PP in contexts where situational and normative information is scarce and/or contradictory. For example, previous research in the field has showed that conflicting (hence, arguably "ambiguous") injunctive and descriptive anti-littering norms discouraged third parties from punishing litterers (Berger & Hevenstone, 2016). Further research should offer systematic comparisons between lab and field experiments to corroborate that ambiguity certainly explains 3PP in more externally valid settings.

Limitations and future research

Despite the identified merits of the present research, our findings showed some inconsistencies and limitations that raise interesting questions for future research.

First, the moderating role of Observer JS did not replicate consistently across studies. Interestingly, it mainly replicated across those studies including the strategy method in the 3PPG (Studies 1.1-1.3). Some have discussed the strategy method as a methodological factor prompting punishment responses and eliciting a "cold" decision-making mindset due to the consideration of multiple hypothetical decisions (Brandts & Charness, 2011; Pedersen et al., 2013). Could this "colder" strategic mindset trigger concerns related to ambiguity (e.g., unfair punishment) that people high in JS would not normally consider in "hotter" decision-making setups? Future research should clarify if third parties who face an ambiguous norm violation and can afford a paused, reflective strategic (vs. immediate, impulsive) reaction, experience further concerns that make them hesitate and remain passive, especially people with dispositional justice concerns.

Second, the findings of Study 1.6 suggested an alternative explanation to why ambiguity decreased 3PP, namely that, under ambiguity, people may simply be less likely to perceive a norm violation, especially when the latter is potentially less severe (in our case, in the *low expected value* conditions). Specifically, third parties could just be inattentive and miss the situation under ambiguity or, instead, struggle to disentangle whether a norm violation has occurred or not due to the ambiguous information. In the second case, one could still expect third parties to experience concerns about punishing unfairly to determine the decision to punish. Future research should address this nuanced distinction.

Finally, research on 3PP has been criticized for the presence of strong demand effects (Pedersen et al., 2013, 2018). While the use of the strategy method could enhance these demand effects, our key results replicated in setups other than the strategy method (Studies 1.4-1.6). Moreover, demand effects could result from the limited behavioral options of lab experiments (Pedersen et al., 2018). We provided participants with

the opportunity to compensate the victim as an alternative behavioral reaction to counter this potential limitation. Yet, we acknowledge that our design does not fully resemble the behavioral repertoire that third parties might have in real-life situations. Therefore, we reemphasize the need to replicate the present findings in the field to rule out the potential influence of demand effects inherent to settings with higher experimental control.

Conclusion

The present research has taken an important step by establishing ambiguity of the norm violation as a critical situational boundary of 3PP. We demonstrated that, when a norm violation became ambiguous, third parties punished less. We argue that, when facing an ambiguous norm violation, an important consideration that ultimately prevented some third parties from punishing was to avoid engaging in unfair punishment. When possible, third parties could overcome this concern of punishing unfairly by means of resolving the ambiguity. However, an ambiguous norm violation also introduces moral wiggle room that some might exploit by keeping the situation ambiguous to remain passive and avoid incurring costs. Lastly, these findings emphasize how we can improve our estimation of the prevalence of types of behavior, such as 3PP, by considering situational factors that likely characterize many everyday situations in the field.

Supplementary material and open practices

The supplementary material of this chapter is available in the following OSF repository: https://osf.io/2q9vm/. This repository further includes the data, analysis code, codebook and research materials to reproduce the reported results. Note that in the repository, Studies 1.4, 1.5 and 1.6 are labelled differently (i.e., Study 4b, Study 4 and Study 5, respectively).

Chapter 2

Examining third-party punishment under cost uncertainty

Costly third-party punishment (3PP) against norm violations generally entails costs and little or no personal benefit for third parties. Previous work shows that the higher costs of 3PP are, the less willing are third parties to engage into 3PP. However, in real-life settings, third parties are often not able to anticipate the exact costs that 3PP will entail. In three studies, we investigate how this *cost uncertainty* affects 3PP in a third-party punishment game. We argue that, under cost uncertainty, third parties might tend to overestimate the occurrence of high costs and, hence, to be discouraged from punishing. Especially, we expect that those who are dispositionally indifferent to be the passive beneficiary of a situation of injustice should be more affected by cost uncertainty. Across studies, we do not observe cost uncertainty to affect 3PP, independently of third parties' Beneficiary Justice Sensitivity. We provide alternative explanations for this null finding, we examine potential limitations of our manipulation of cost uncertainty and we offer suggestions to improve the latter.

Chapter 2 is based on data from the three first studies presented in Chapter 1 (i.e., Studies 1.1-1.3). As explained in the general introduction, these studies addressed several independent research questions.

To engage in 3PP, third parties generally accept personal costs (in any of their forms, i.e., economic, social, physical) with little or no immediate personal benefits. These costs are a fundamental element to understand third-party punishment as a behavioral phenomenon as well as its underlying psychological processes. Specifically, previous work suggests that the psychological process of weighting the costs and benefits of 3PP is indeed an important determinant of 3PP (Baumert et al., 2013) and that when the costs of punishment behavior equal or exceed its benefits, punishment is less likely to occur (Anderson & Putterman, 2006; Carpenter, 2007; Egas & Riedl, 2008).

However, can third parties always accurately anticipate the costs that they would incur if they exerted punishment? Imagine the following scenario: You walk on the street and witness a person throwing a glass bottle onto the road. You may want to intervene and publicly reprimand this person; however, it is uncertain how this person will react. It is possible that they indeed feel reprimanded, show signs of shame and regret, clean up the pieces of glass and leave without causing further trouble. Yet, it is also possible that they react violently against you, insulting you or even physically attacking you. This uncertainty regarding the costs of your action could be frequently present when third parties face norm violation in real-life settings (Van Lange et al., 2012). The question that arises is how this *cost uncertainty* affects 3PP. To our knowledge, research on 3PP has neglected this situational aspect, since experimenters generally establish common knowledge about some fixed cost associated with 3PP (e.g., Fehr & Fischbacher, 2004; Leibbrandt & López-Pérez, 2012). We therefore aim to investigate whether the cost uncertainty is a situational boundary of 3PP.

Cost uncertainty and 3PP

Similarly to how people overestimate the probability of extreme life events (e.g., homicides; Lichtenstein et al., 1978), we propose that a situation in which costs are uncertain could lead third parties to overestimate, and therefore attribute more subjective weight to, the likelihood of potential high costs, even when these are rare (Barberis, 2013; Rozin & Royzman, 2001; Tversky & Kahneman, 1992).

The overestimation of extreme events results from heuristics of cognitive accessibility. Specifically, when an event is more accessible to memory, people tend to attribute a higher subjective probability to it (i.e., availability bias; Tversky & Kahneman, 1973). Despite its seeming irrationality, overestimating extreme events can actually turn to be a useful decision-making strategy, especially under conditions of uncertainty and time or cognitive constraints. In such circumstances, individuals might only be able to consider a sample of potential outcomes of their actions for their decision. The overestimation of the likelihood of extreme events serves as a sampling prioritization of the most potentially relevant outcomes, which previous work has shown to improve individuals' decision-making (Lieder et al., 2018).

In the case of 3PP, when the costs of this type of behavior are uncertain, third parties might similarly overestimate and rely on their subjectively estimated probability of highly costly events (e.g., being physically

assaulted, getting fired, suffering ostracism). The overestimation of the probabilities of high costs should bring about an increase of the overall cost-benefit ratio of 3PP, which, in turn, could plausibly hinder any intention from third parties of engaging in punishment behavior (Egas & Riedl, 2008). Thus, we hypothesize that facing uncertain (vs. certain) costs will discourage third parties from engaging in 3PP.

Cost uncertainty and Beneficiary Justice Sensitivity

Cost uncertainty may exert an overall effect on 3PP, but it could also influence differently the behavior of third parties with different personal dispositions. In particular, we further considered that the third parties' Justice Sensitivity (JS; Baumert & Schmitt, 2016) may explain inter-individual differences with regard to how cost uncertainty would affect 3PP. People who hold stronger justice concerns regarding others' injustices are more prone to engage in costly behavior that aim to restore justice (e.g., 3PP, Lotz et al., 2011; or protest behavior, Rothmund et al., 2014). As previous work suggests, people with high JS do so due to a stronger response of moral outrage against unfairness (Lotz et al., 2011), an emotional response that may incline third parties to address the unfairness despite the potential costs (Ginther et al., 2021; Gummerum et al., 2016).

Here, we specifically examined the moderating role of third parties' Beneficiary JS with regard to the effect of cost uncertainty on 3PP. Beneficiary JS captures individuals' predisposition to perceive and react against injustice from the perspective of a passive *beneficiary* of the unjust situation (Schmitt et al., 2010). In the setting of 3PP, avoiding own costs in the light of another person being unfairly disadvantaged could be construed as a kind of indirect benefit from the situation. People high in Beneficiary JS should be motivated to avoid this kind of benefit and thus punish the unfairness despite the costs and despite whether these are uncertain. Conversely, people low in Beneficiary JS should lack this motivation and be reluctant to incur uncertain and thus potentially high costs. Following this rationale, we hypothesize that those high in Beneficiary JS should show similar levels of costly third-party punishment regardless of the cost uncertainty, whereas those low in Beneficiary JS should punish less when costs are uncertain.

Research overview

The three first studies reported in Chapter 1 (i.e., Studies 1.1-1.3) addressed how the cost uncertainty affected 3PP and whether inter-individual differences in Beneficiary JS moderated this effect. Here, we exclusively present the results related to these two research questions (labelling the studies as Studies 2.1-2.3, respectively).

As a reminder, Studies 2.1-2.3 used the third-party punishment game (3PPG) as experimental paradigm, where third parties (Person C) had the possibility of *punishing* the dictator (Person A) and *compensating* the receiver (Person B). The three studies followed a within-subject design with four different game rounds, with the only difference that Studies 2.1 and 2.2 presented these rounds in a fixed order,

whereas Study 2.3 introduced a pseudo-randomization in the order of rounds. Furthermore, they all used the strategy vector method to assess 3PP (e.g., Oxoby & McLeish, 2004). For a matter of conciseness, the details regarding the experimental design, the participants' samples and the procedure and measures that were described in Chapter 1 will not be repeated here.

The three studies included an experimental manipulation of cost uncertainty. In the *certain costs* conditions, any third-party intervention (punishment or compensation) was associated to a fixed cost. In the *uncertain costs* conditions, the third-party interventions entailed a randomly determined cost that third parties did not know before making their decision.

Studies 2.1-2.3

In the three studies, we tested the hypothesis that uncertain (vs. certain) costs would reduce levels of 3PP (H1). We further hypothesized that the effect of cost uncertainty would be more pronounced among third parties with low (vs. high) Beneficiary JS (H2). We preregistered these hypotheses for Study 2.1 (https://osf.io/ubnzm) and Study 2.2 (https://osf.io/etgq9) and we tested the same preregistered hypotheses in Study 2.3.

Method

Design

As indicated in Chapter 1, the four rounds of the 3PPG that participants played corresponded to a 2x2 design, with cost uncertainty and ambiguity of the norm violation (see Chapter 1 for further details) as within-subject factors.

In the *certain costs* conditions, modifying the outcomes of Person A or Person B by 1 experimental currency unit (ECU) entailed a fixed cost of ½ ECUs for Person C. In the *uncertain costs* condition, the cost for Person C for modifying Person A's or B's outcome by 1 ECU was randomly determined within a range from 0.01 to 1 ECUs, based on a uniform probability distribution. Participants did not receive information about the underlying probability distribution. Thus, under cost uncertainty, the possible costs of punishment virtually varied from a scenario with almost no costs for the third party to a scenario where the effect of 3PP equaled its costs.

Measures

3PP. We computed continuous measures of 3PP, one measure per round of the 3PPG. Specifically, each measure represented the sum of ECUs deducted by Person C in those decisions from the strategy method that implied a reaction from Person C to an unequal split of a 10-ECU endowment (i.e., Person A sent [0, 1, 2, 3, or 4] ECUs to Person B). We excluded the decisions in reaction to a potential fair split (i.e., Person A sent [5 or 6] ECUs) because we assumed that these would not be perceived generally as norm

violations (Erkut et al., 2015; Krupka & Weber, 2013). If participants did not report more than one decision in a round, we did not include their data in the analyses for that round.

Beneficiary JS. In Studies 2.1 and 2.2, we assessed JS with the German *Justice Sensitivity Short Scales* (Baumert, Beierlein, et al., 2014), which include two items for measuring Beneficiary JS (e.g., "I feel guilty when I am better off than others for no reason."; Study 2.1, $M = 2.50$, $SD = 1.28$, $\alpha = .75$; Study 2.2, $M = 2.29$, $SD = 1.36$, $\alpha = .82$).

In Study 2.3, we measured JS with the 40-item version of the JS Inventory (Schmitt et al., 2010). Ten items served for measuring Beneficiary JS, including the same two items of the short version ($M = 2.80$, $SD = 0.96$, $\alpha = .88$).

In every study, the response options ranged from 0 (*Not at all)* to 5 (*Absolutely*).

Statistical analyses

We used a multilevel modelling approach, clustering punishment decisions in each round of the 3PPG (Level 1) within participants (Level 2). Every fitted model included participants' ID as random factor to account for the within-subject nature of our data.

To test whether cost uncertainty affected 3PP independently of ambiguity of the norm violation, we first fitted a model including cost uncertainty and ambiguity of the norm violation as Level-1 fixed factors (for further details about ambiguity of the norm violation, see Chapter 1). In a second model, we included cost uncertainty as Level-1 fixed factor, Beneficiary JS as Level-2 fixed factor (grand-mean centered) and the respective cross-level interaction to test the moderating role of Beneficiary JS.

In Study 2.3, an additional model included the factor *Position* (i.e., position at which a particular round of the game was presented to a participant; 0 – Round 1, 1 – Round 2, 2 – Round 3, 3 – Round 4) to account for potential linear effects on 3PP over time.

Results

Table 2.1 present descriptive statistics of 3PP clustered across conditions with certain and uncertain costs.

The first multilevel model did not show a significant effect of cost uncertainty, irrespective of the effect of ambiguity of the norm violation – in Study 2.1, $\beta = .04$, $t(487) = 0.73$, $p = .464$, 95%CI [-.06, .13]; in Study 2.2, $\beta = -.05$, $t(676) = -1.32$, $p = .186$, 95%CI [-.12, .02]; in Study 2.3, $\beta = -.03$, $t(828) = -0.89$, $p = .376$, 95%CI [-.10, .04]. As an additional model including the Uncertainty x Ambiguity interaction term indicated, cost uncertainty did not moderate the main effect of ambiguity of the norm violation – in Study 2.1, $\beta = .00$, $t(486) = -0.08$, $p = .935$, 95%CI [-.20, .19]; in Study 2.2, $\beta = .02$, $t(675) = 0.27$, $p = .786$, 95%CI [-.12, .16]; in Study 2.3, $\beta = -.04$, $t(827) = -0.54$, $p = .586$, 95%CI [-.17, .10].

The second model including Beneficiary JS did not show a significant effect of cost uncertainty, nor a significant Uncertainty x Beneficiary JS interaction in any of the three studies (see Table 2.2).

In Study 2.3, introducing the factor *Position* into the model did not increase the model fit $\Delta AIC = 0.20$, $\chi^2(1) = 1.806$, $p = .179$, nor it changed the main results (see Table 2.3).

General discussion

Despite the relevant theoretical implications that cost uncertainty could have for 3PP (Van Lange et al., 2012), we did not observe cost uncertainty to exert any effect on 3PP in our experimental setup. Yet, these null results should not lead us to reject the idea that cost uncertainty is an important situational boundary of 3PP. Instead, we identified different methodological explanations related to our experimental manipulation that could shed light onto the null findings.

Table 2.1

Descriptive statistics of punishment clustered by cost uncertainty conditions and split by ambiguity of norm violation conditions in Studies 2.1-2.3.

	M_{Sum}	SD_{Sum}	*Perc.* (%)
Study 2.1			
Certain costs	**8.34**	**7.23**	**79.14**
Certain costs + No amb. norm. viol.	10.67	7.49	85.28
Certain costs + Amb. norm. viol.	6.01	6.15	73.01
Uncertain costs	**8.60**	**7.40**	**80.67**
Uncertain costs + No amb. norm. viol.	11.00	7.94	83.95
Uncertain costs + Amb. norm. viol.	6.22	5.95	77.43
Study 2.2			
Certain costs	**7.52**	**7.31**	**73.23**
Certain costs + No amb. norm. viol.	9.90	7.47	80.09
Certain costs + Amb. norm. viol.	5.13	6.32	66.37
Uncertain costs	**7.16**	**7.45**	**70.13**
Uncertain costs + No amb. norm. viol.	9.47	8.01	74.78
Uncertain costs + Amb. norm. viol.	4.85	6.04	65.49
Study 2.3			
Certain costs	**4.99**	**6.36**	**54.69**
Certain costs + No amb. norm. viol.	6.58	7.16	61.15
Certain costs + Amb. norm. viol.	3.40	4.95	53.62
Uncertain costs	**4.79**	**6.56**	**57.40**
Uncertain costs + No amb. norm. viol.	6.27	7.48	57.76
Uncertain costs + Amb. norm. viol.	3.31	5.09	51.62

Note. Punishment = Amount of ECUs (1 ECU = 1 Euro) subtracted from Person A. *M* and *SD* = mean and standard deviation of the sum of Euros punished across decisions to unequal splits from Person A (i.e., €[0 to 4] coins) to Person B. *Perc.* (%) = Percentage of participants who punished at least 1 ECU.

Table 2.2

Tested multilevel model on punishment in Studies 2.1-2.3.

	Study 2.1			Study 2.2			Study 2.3		
Parameters	β [95% CI]	***t***	***p***	β [95% CI]	***t***	***p***	β [95% CI]	***t***	***p***
Fixed effects									
Cost uncertainty	.03 [-.08, .14]	0.54	.592	-.04 [-.13, .05]	-0.89	.375	-.03 [-.10, .04]	-0.849	.396
Beneficiary JS	.10 [-.04, .23]	1.44	.151	.05 [-.07, .17]	0.8´2	.412	.20 [.10, .31]	3.887	**<.001*** **
Cost uncertainty x Beneficiary JS	-.04 [-.16, .07]	-0.75	.454	.01 [-.08, .10]	0.22	.827	-.03 [-.11, .04]	-0.915	.361
Random effects									
σ^2	28.53			23.82			16.17		
$\tau_{00\ ID}$	25.74			31.53			24.12		
ICC_{ID}	0.47			0.57			0.60		
N_{ID}	160			218			281		
Observations	636			872			1108		
Marginal / Conditional R^2	0.007 / 0.478			0.003 / 0.571			0.036 / 0.613		

Note. JS = Justice Sensitivity, σ^2 = Residual variance; $\tau_{00\ ID}$ = Variance of the intercept; ICC_{ID} = Intraclass correlation coefficient; N_{ID} = Total number of individuals.
*** $p < .001$, ** $p < .01$, * $p < .05$.

Table 2.3

ANOVA table of multilevel model accounting for order effects in Study 2.3.

	Punishment			
Parameters	***F***	***df***	***p***	η_p^2
Cost uncertainty	0.732	827	.392	.001
Beneficiary JS	14.580	282	**<.001*** **	.049
Position	1.801	828	.180	.002
Cost uncertainty x Beneficiary JS	0.766	829	.382	.001

Note. JS = Justice Sensitivity, *df* = Numerator and denominator degrees of freedom calculated with Satterthwaite's method.
*** $p < .001$, ** $p < .01$.

One possibility relates to potentially large inter-individual differences in risk preferences. If participants drastically differed in their risk preferences, with some participants being very risk-seeking and others very risk-averse, opposing reactions to our uncertainty manipulation could have ultimately resulted in an overall null effect. Lacking of any measure of dispositional risk preferences, we could not test if this was the case. However, a visual inspection of the distribution of 3PP across experimental conditions did not seem to support such a bimodal trend.

Standard expected-utility theory (von Neumann & Morgenstern, 2007) could also explain the observed null effect under specific assumptions. More concretely, if our participants had neutral risk preferences and they had inferred a uniform probability distribution of the costs of punishment, then, the costs expected value would have been equivalent for certain and uncertain costs (i.e., approx. 0.5 ECUs). This would render 3PP equally costly across conditions, and therefore, it would explain why we did not observe an effect of uncertainty. The previous assumptions are not farfetched when closely considering our manipulation of cost uncertainty. With regard to risk-preferences, people generally show neutral risk preferences under uncertainty at small stakes (e.g., Bombardini & Trebbi, 2012). In the case of our manipulation, the range in which costs varied (i.e., from 0.01 to 1 ECUs) did not put third parties at very high stakes, which could have led participants to be indifferent about the degree of uncertainty.

Furthermore, our manipulation did not resemble the potential, yet unlikely, extreme costs, that could lead third parties to overestimate their probability of occurrence in real-life settings. The highest cost third parties could incur in our experiment (i.e., a maximum of 1 ECU per 1 ECU compensated or punished) might not have had enough subjective value to skew the costs probability distribution, and therefore, to discourage third parties from punishing. However, this highest cost was sufficiently high to establish a 1:1 cost-benefit ratio, which previous work has shown to have considerable effects on punishment behavior (Egas & Riedl, 2008).

Though we did not succeed in identifying an effect of cost uncertainty in our experiment, we join other researchers (e.g., Van Lange et al., 2012) in highlighting the need to continue examining the impact of the cost uncertainty for a deeper comprehension of 3PP and its situational boundaries. Researchers should consider alternative approaches to study the effects of cost uncertainty. For example, future research could benefit from simplifying the information about the potential costs (e.g., "you will incur a cost of either X or Y"), from assessing the inferences that third parties make regarding the underlying probability distribution of costs (e.g., "what costs do you find more likely to incur?"), and especially, from increasing the value of the highest potential cost. Some work by Balafoutas et al. (2014) suggests that this last option may be particularly fruitful. In their study, they manipulated whether third parties could be counterpunished by the dictator. That is, after being punished by the third party, the dictator could in turn lower the third party's income by withdrawing the ticket of the latter to participate in a lottery for a €200-voucher. Since getting

counterpunished is a probabilistic costly event – in this case, dependent on the decision of the dictator –, the manipulation of Balafoutas et al. (2014) could be considered a manipulation of cost uncertainty that substantially decreased 3PP.

Supplementary material and open practices

The supplementary material of this chapter is available in the following OSF repository: https://osf.io/2q9vm/. This repository further includes the data, analysis code, codebook and research materials to reproduce the reported results. Note that in the repository, Studies 2.1 and 2.2 are labelled differently (i.e., Study 1 and Study 2, respectively).

Chapter 3

On the external validity of third-party punishment in the lab

The high rates of *costly third-party punishment* (3PP) observed in lab experiments have led to theorize about 3PP as a critical sanctioning mechanism underlying the maintenance of social norms. However, critical views have questioned the external validity of these findings when considering how rare 3PP seems to be in the field. The present chapter aims to provide a test of the external validity of the *third-party punishment game* (3PPG). In two studies, we examine the relationship between people's 3PP in the 3PPG and their intervention behavior (Study 3.1) and intervention intentions (Study 3.2) in a field-like situation of a norm violation: an embezzlement of lab funds. We consider that the introduction of ambiguity of the norm violation and cost uncertainty into the 3PPG (i.e., factors that arguably affect 3PP in the field) would enhance this relationship and therefore increase the external validity of the 3PPG. We do not find 3PP to be associated with either people's intervention behavior or their intervention intentions. We offer a discussion about the implications of these results regarding the lack of external validity of research on 3PP and the situational specificity of this behavior.

Chapter 3 is based on data from the two first studies presented in Chapters 1 and 2 (i.e., Studies 1.1-1.2). As explained in in the general introduction, these studies addressed several independent research questions.

The first lab studies providing evidence of the prevalence of costly third-party punishment (3PP) warranted the raised interest in this behavioral phenomenon among researchers from different fields (e.g., biology, anthropology, economy, psychology; Janssen & Bushman, 2008; Krasnow et al., 2016; Lewisch et al., 2015; Riedl et al., 2012). In their seminal paper, Fehr and Fischbacher (2004) reported that around 60% of third parties engaged in 3PP of a dictator who made unfair monetary distributions in a dictator game. Shortly after, Henrich et al. (2006) showed similar average rates of 3PP across samples from different cultures.

Yet, the external validity of these results has been called into question. External validity refers to the degree of generalizability of a specific finding across populations, settings, and operationalizations of the critical variables under study (Campbell & Stanley, 1963). In the case of 3PP, several studies assessing reactions to different norm violations in the field (e.g., reactions to littering in public spaces) showed drastically lower rates of 3PP (i.e., 10-15%; Balafoutas et al., 2016; Balafoutas & Nikiforakis, 2012; Winter & Zhang, 2018), in comparison to previous lab studies. Similarly, studies that (retrospectively) assessed people's reactions against more diverse sets of norm violations by means of ambulatory assessment (Molho et al., 2020) or recall methodologies (Pedersen et al., 2020) also failed to show the high rates of 3PP observed in lab experiments. This has led some researchers to claim that the prevalence of 3PP in daily life and, consequently, the widely discussed role of 3PP in the maintenance of social norms and human cooperation have been overestimated (Guala, 2012; Pedersen et al., 2018).

In line with the rationale presented in Chapters 1 and 2, we argue that one potential explanation for an overestimation of 3PP in lab experiments may relate to the neglected role of situational ambiguity. In the field, information is usually noisy, incoherent, or incomplete, and may affect the decision-making of third parties, hindering, for example, the interpretation of the situation as a norm violation or the anticipation of potential costs. In contrast, third parties generally receive perfect situational information in the economic games used in lab experiments. In the case of the third-party punishment game (3PPG; Fehr & Fischbacher, 2004), this information allows third parties clearly to identify the norm violation (i.e., an unfair distribution of resources by the dictator) and the costs associated with 3PP (i.e., known a priori, fixed costs). These informationally "ideal" circumstances may facilitate people's engagement in 3PP.

In the present chapter, we build on the previous rationale to assess the external validity of lab research on 3PP. We propose that the introduction of situational ambiguity in the 3PPG improves the generalizability of lab findings regarding the occurrence of 3PP. Thus, we investigate whether the 3PP in a 3PPG under situational ambiguity better predicts people's punishment behavior against a norm violation in a field-like situation (i.e., an embezzlement of lab funds).

The importance of assessing the external validity of 3PP in the lab

Addressing the issue of the external validity of lab research on 3PP is relevant for two major reasons of phenomenological and theoretical nature, respectively.

At the phenomenological level, the study of the external validity of lab research on 3PP can help us to delve into the situational conditions under which 3PP actually occurs. In this respect, the aforementioned studies offering retrospective assessments of people's reactions to norm violations in the field were fruitful. They showed that direct confrontations of perpetrators (i.e., the closest type of reaction to the traditional operationalization of 3PP in the 3PPG) are more likely when third parties have more power, when they value the perpetrator (Molho et al., 2020), and when they perceive a high interdependence with the victim (Pedersen et al., 2020). These results exemplify how the close inspection of real-life settings is, therefore, of extreme importance to identify relevant situational factors moderating the occurrence of 3PP that have generally been overlooked in previous lab experiments.

From a theoretical standpoint, the assessment of the external validity of the 3PPG informs the question whether 3PP functions as an informal sanctioning mechanism for the reinforcement of social norms, as some initially argued (Fehr & Fischbacher, 2004; Yamagishi, 1986). The first studies demonstrating the widespread presence of 3PP as a reaction to unfairness in one-shot, anonymous lab experiments (Fehr & Fischbacher, 2004; Henrich et al., 2006) raised questions about the underlying role of this type of behavior. In the end, as it occurs in the lab, 3PP is in and of itself a suboptimal type of behavior from the individual's perspective, insofar as it leads third parties to incur costs on behalf of strangers without immediate personal benefits. On the other hand, if 3PP deterred others from free-riding and consequently reinforced cooperation among individuals, this indirect benefit could warrant the occurrence of 3PP. Following this rationale, many researchers focused on assessing the role that costly punishment played in the maintenance of human cooperation (e.g., Egas & Riedl, 2008; Fowler, 2005; Wu et al., 2009). Yet, skeptical voices have questioned whether costly punishment can indeed be considered a key reinforcing element of social norms and human cooperation, given that this type of behavior is rarely observed outside the lab (Baumard, 2010; Guala, 2012). If the frequency in which we observe 3PP in lab experiments does not correspond with what occurs in the field across an arguably wide variety of social norm violations (beyond violations of fairness), any interpretation about its role in the maintenance of social norms and human cooperation might be overstated. Thus, every contribution regarding the generalizability to the field of lab findings regarding the occurrence of 3PP will likely benefit this currently ongoing debate.

Assessing the external validity of the 3PPG

In line with Levitt and List's (2007) work on the generalizability of behavioral research, we think that the discrepancies between 3PP in the lab and in the field will remain, as the relevant factors affecting this

behavior are not considered across both contexts. In this vein, introducing factors that we suspect can plausibly affect 3PP outside the lab into lab experiments can be a means to increase the external validity of lab findings.

Here, we argue that situations of norm violations in the field generally entail a considerable level of ambiguity affecting the interpretation of the perpetrator's behavior as a norm violation and the anticipation of costs associated to 3PP. We propose that these two situational factors could hinder the occurrence of 3PP in the lab and in the field. If situational ambiguity was a critical boundary condition of 3PP in the field, the introduction of ambiguity in the 3PPG would arguably improve the external validity of the 3PP observed in this lab setting. In other words, the levels of 3PP observed in the 3PPG under ambiguity should be similar to those observed in the field. Thus, the first goal of the present chapter is to test whether 3PP in the 3PPG better predicts 3PP in the field when the 3PPG introduces ambiguity of the norm violation and cost uncertainty.

Furthermore, external validity entails that the interindividual differences observed in lab studies highly converge with those in the field, assuming similar situational and psychological factors across these two contexts (Levitt & List, 2007). Put differently, the behavior that individuals show in the lab should resemble how they behave in similar situations outside the lab. So far, we have been referring to discrepancies in the rates of 3PP between lab and field studies that used different ad-hoc samples (e.g., bystanders in a train station vs. undergraduate students). Although these discrepancies could suggest that different situational factors influence 3PP in the lab and in the field, the different rates of 3PP could also be attributed to differences between samples. To rule out this explanation, it is therefore critical to assess the external validity of the 3PPG at the individual level; that is, to observe whether the 3PP in the 3PPG of a given individual converges with their behavior in the field under similar situational factors (Levitt & List, 2007). Thus, as a second aim of the present chapter, we assess the relationship between people's 3PP in the 3PPG and in the field within the same group of individuals.

Research overview

To assess the external validity of the 3PPG, we used the data from two studies, already described in Chapters 1 and 2 (i.e., Studies 1.1-1.2 and 2.1-2.2, respectively). For purposes of clarity, in the present chapter we relabeled these studies as Studies 3.1 and 3.2.

Studies 3.1 and 3.2 each consisted of two lab sessions, which took place at least one week apart. In the first session, participants completed the 3PPG, where we experimentally manipulated within-subjects the ambiguity of the norm violation and the cost uncertainty (for further details and results, see Chapters 1 and 2). Therefore, we assessed the participants' 3PP under these manipulated situational factors. In the second lab sessions, we presented the same participants with a field-like norm violation, either in the form of a staged

situation in the lab (Study 3.1) or a vignette (Study 3.2). In both studies, the context of the norm violation consisted of the embezzlement of money from the lab funds, perpetrated by the experimenters conducting the study. The choice of an embezzlement as a norm violation suited our purposes in two ways. First, the embezzlement was a conceptually reasonable validation criterion regarding the generalization of the 3PPG to other situations of norm violations. According to a pretest (see below), people generally perceived it as a severe norm violation and an intervention against it, as a personally costly behavior (similarly to how 3PP is operationalized in the 3PPG). Second, the possibility of staging an embezzlement in the lab allowed us to standardize and exert some experimental control over how the norm violation occurred and was witnessed by participants.

To address the external validity of the 3PPG, we examined the relationship between people's 3PP in the 3PPG and their intervention against the embezzlement. Furthermore, we checked whether this relationship differed as a function of the ambiguity of the norm violation and the cost uncertainty introduced in the 3PPG.

In Study 3.1, we videotaped people's interventions against the embezzlement, which two independent raters later assessed. In Study 3.2, we used a proxy of intervention behavior commonly found in the literature, namely, people's intervention intentions in hypothetical vignettes (e.g., Niesta Kayser et al., 2010). Previous research suggests that people's intervention intentions might overestimate the extent to which people actually intervene against norm violations in real-life settings (Baumert et al., 2013; Karmali et al., 2017; Kawakami et al., 2009). As we discussed above, some researchers have raised similar arguments with regard to 3PP in lab settings and proposed that the levels of 3PP observed in economic games (e.g., 3PPG) likely overestimate the prevalence of this kind of behavior outside the lab (e.g., Guala, 2012; Pedersen et al., 2018). This suggests that both vignette and economic-game settings might share methodological features (e.g., availability of unambiguous situational information, "cold" or hypothetical decision-making) that make people report more frequent and stronger reactions to norm violations, compared to how they actually react in the field. We acknowledge that examining the relationship between 3PP and intervention intentions would not be, strictly speaking, an appropriate test of external validity. Yet, we think it would still be informative if people's 3PP in the 3PPG predicted their intervention intentions in the vignette settings better than their actual intervention behavior in real settings. In that case, the methodological commonalities between the 3PPG and vignette settings could inform us about the kind of context to which the 3PPG generalizes (e.g., contexts with unambiguous information or without constrains for deliberating about hypothetical outcomes). Hence, the examination of the relationship between 3PP in the 3PPG and intervention intentions in vignette settings could shed further light on conceptually relevant nuances of the external validity of the 3PPG.

Study 3.1

Study 3.1 examines whether 3PP predicts intervention behavior against the staged embezzlement. We argue that situations of norm violations in real-life settings can be characterized by ambiguity of the norm violation and cost uncertainty. Thus, we hypothesize that people's 3PP in the 3PPG will better predict their intervention behavior against the embezzlement, when the 3PPG introduces ambiguity of the norm violation and cost uncertainty.

We preregistered our hypothesis and analysis plan, procedure, video-based ratings, and a log file including every measure in the study in the OSF (https://osf.io/ubnzm). This study was part of a bigger research project with different research goals (see Sasse et al., 2020); thus, the preregistration includes further measures and procedural details that will not be discussed in this chapter. For a matter of conciseness, some details regarding the experimental design, the procedure and measures that were already described in Chapters 1 and 2 will not be repeated here.

Method

Design

The study followed the same 2x2 within-subject design used in the 3PPG (i.e., ambiguity of the norm violation, *no ambiguity* vs. *ambiguity*; cost uncertainty, *certain* vs. *uncertain costs*). This design served to test under which conditions participants' 3PP in the 3PPG setting better predicted the intervention behavior against the embezzlement staged in the lab.

Participants

From the sample of 164 participants recruited for Study 1.1 in Chapter 1, only 144 completed the second lab session. We excluded data of 36 participants based on preregistered suspicion checks about the staged embezzlement (for further details, see below). Thus, our final sample size was 108 (83% women; age range from 18 to 30, $M = 22.43$, $SD = 2.49$).

Procedure

Session 1 – 3PPG. In Session 1, participants completed the 3PPG within a battery of different personality questionnaires. As a reminder, we implemented a strategy vector method (e.g., Oxoby & McLeish, 2004). Thus, participants made decisions in the different roles (Person A, B, and C). When deciding as Person C, participants made punishment and compensation decisions that were conditional on seven different possible distributions from Person A – i.e., "Given that Person A transfers [0 to 6] experimental currency units (ECUs) to Person B, how many ECUs do you wish to *deduct* from Person A's / *add* to Person B's endowment?". For further details about the 3PPG, see the procedure of Study 1.1 in Chapter 1 (pp. 20-21).

Session 2 – Staged embezzlement. We individually invited participants to the lab to participate in an experiment, ostensibly about "Learning and Emotions". During the study, participants witnessed two experimenters planning and executing the embezzlement of funds from the lab, and we were interested in how participants would react to this norm violation. Trained female researchers and/or research assistants played the different roles, following a standardized script, the phases of which are described in detail below.

We informed participants that they would be video-recorded during the session. Only after receiving explicit permission through a signed declaration of consent was video material collected. Two cameras, integrated in the computers used for some of the experimental tasks, caught the participants' facial and body expressions, while participants witnessed the embezzlement seated at the experimenter table and while they worked on the different computer-based tasks. We situated a third camera opposite to the computers to ensure a broader angle of the room in which the embezzlement took place.

Phase 1 – Arrival at the lab: Once a participant (henceforth P) arrived at the lab, an experimenter (henceforth E1) asked them to wait and sit in a lab room. Then, a confederate in the role of a second participant (henceforth C) joined and E1 asked her to take a seat. A second experimenter (henceforth, E2) informed P and C about the experimental procedure and the video recording.

Phase 2 – Determining the norm: The project leader (henceforth PL) entered the room, bringing receipts and a cashbox for the payment of participants. The PL reminded E1 and E2 that participants would either receive monetary compensation or receive course credits instead. The PL explicitly highlighted that participants who received the monetary compensation needed to sign the receipt, and that not doing so would be at the expense of the lab funds. This could be overheard by P and C, and it was done to ensure that participants interpreted the subsequent actions of E1 and E2 as a norm violation. Next, the PL informed P and C that they would receive their additional payoff of the 3PPG personally from her at the end of the session. This provided participants with an additional opportunity to intervene against the forthcoming embezzlement by means of reporting it to the PL.

Phase 3 – Compensation: E1 subsequently asked P to enter a second lab room, where the computers and cameras were installed, while C and E2 stayed in the first lab room. In line with the PL's instructions, E1 handed over the payment and asked P to sign the receipt. Then, E1 instructed P to begin with a first questionnaire – which assessed demographic data and emotions – and, after completing it, to remain seated at the experimenter table until C finalised the same questionnaire.

Phase 4 – Planning the norm violation: While P waited, E2 entered the room alone and kneeled down next to E1. Whispering, but sufficiently loud for P to hear, E2 suggested to E1 to fake signatures in the receipt of participants who came for course credits, as if they were actually paid, so that E1 and E2 could

keep the money for themselves. E1 reacted hesitantly at first, but then she stated that they could gain a total of 3.000€ euros and eventually agreed to the proposal. Then, E2 left the room to pick up C.

Phase 5 – Execution of norm violation: Once E2 left, E1 faked a signature on the receipt sheet that P had used earlier. Next, E1 took €30 from the cashbox and put them into her pocket. Two minutes later, E2 brought C to the room and told E1 that C requested course credits as compensation.

Phase 6 – Learning task: E1 placed the course credit sheet on C's workspace and asked P to take a seat at the adjacent computer. E1 instructed P and C to complete an ostensible learning task and some emotion-related measures. Then, E1 and E2 left the room, and P and C remained alone during the completion of the different tasks. After finishing, P and C answered several questions about the learning task and an open-ended suspicion check (i.e., "In your opinion, what was this investigation about?"). Shortly after P and C had finished, E1 entered the room and asked C to go to the other room. Then, E1 requested P to go to the PL's office to receive the additional payoff from the 3PPG of Session 1.

Phase 7 – Project Leader's office: P entered the PL's office alone, where the PL asked P whether everything was fine during the experiment. Then, the PL asked P for their participant code to check their corresponding payoff. Afterwards, the PL handed P a questionnaire with some control questions and asked them to return to the lab while she collected the money.

Phase 8 – Control questions and debriefing: Back in the lab, P filled in the questionnaire that they had received from the PL. The PL arrived at the lab with the receipt and the money and announced to P that the study was over. Then, the PL fully debriefed P about the bogus situation and the actual purpose of the study. If P had intervened in any early phase, the debriefing would have followed immediately. P had the opportunity to ask any question through a structured interview, and they were provided with a declaration of consent through which they could withdraw their conformity with the use of responses and video material collected during the session. Lastly, P filled in a retrospective self-evaluation questionnaire, which included suspicion checks and post-hoc measures of personal perceptions of the situation.

Pre-test

We collected data from an independent sample of 62 participants (61.3% women; age range from 18 to 62, $M = 24.05$, $SD = 6.56$). Participants read a written vignette describing Phases 2 to 5 and answered several items assessing their personal impression of the embezzlement situation. Specifically, participants reported the *perceived severity* (i.e., "I would find it severe if such situation happen in real life") and the *perceived immorality* of the situation (i.e., "I perceive the behavior of the experimenters as immoral"), as well as whether they found the experimenters' behavior justified (i.e., "The experimenters' behavior is justified"). Every item used response options from 0 (*does not apply at all*) to 5 (*applies completely).*

We computed one-sample t-tests comparing the mean levels with the midpoint of the scale (i.e., 2.5). These indicated that participants perceived the situation as considerably severe, $M = 4.28$, $SD = 0.85$, $t(59) = 16.337$, $p < .001$, $d = 2.11$, 95% CI [1.65, 2.56], and immoral, $M = 4.66$, $SD = 0.70$, $t(60) = 23.903$, $p < .001$, $d = 3.06$, 95% CI [2.46, 3.66]. Furthermore, participants found the experimenters' behavior highly unjustified, $M = 0.44$, $SD = 0.81$, $t(60) = -19.918$, $p < .001$, $d = -2.55$, 95% CI [-3.07, -2.00]. Four additional items assessed how difficult and potentially risky the participants found it to intervene in the described situation (e.g., "I think it would be risky to intervene in the situation", $\alpha = .83$). After computing an average index with the four items, we compared the mean level of risk with 0 in an additional one-sample t-test, which indicated that participants perceived intervention as moderately risky, $M = 2.54$, $SD = 1.32$, $t(60) = 15.081$, $p < .001$, $d = -1.93$, 95% CI [1.50, 2.35].

Taken together, the pretest suggested that the chosen embezzlement situation was an appropriate setup for our research purposes. First, participants clearly identified the embezzlement as a severe, immoral, and unjustified norm violation. Second, they perceived that an intervention against it was moderately risky behavior, which ensured that people perceived the intervention as costly, but not to the extent of completely discouraging participants from intervening.

Measures

Session 1.

3PP. We measured the sum of ECUs that participants deducted as Person C in those decisions that implied a reaction to an unequal split of a 10-ECU endowment (i.e., Person A sent [0, 1, 2, 3, or 4] ECUs to Person B). However, as we will explain later for the measure of intervention behavior, we computed a dichotomous measure of 3PP in each condition in order to compare the participants' 3PP in the 3PPG with the behavioral coding of their intervention behavior against the staged embezzlement. This dichotomous measure captured whether participants punished at least 1 ECU (1) or not (0) across the same decisions mentioned above. We excluded the decisions entailing a fair split (i.e., Person A sent [5 or 6] ECUs) because we assumed that these would not be perceived generally as norm violations (Erkut et al., 2015; Krupka & Weber, 2013). If participants did not report more than one decision in a round, their data were not included in the analyses for that round.

Session 2.

Video ratings of intervention behavior. Two independent raters, previously trained and blind to the tested hypotheses, watched the video material and rated the participants' reactions to the embezzlement. To facilitate the procedure, the video material was divided into different sequences, corresponding to the aforementioned phases (i.e., phases 4, 5, 6, and 8). Since phase 7 took place in a different room where video recording was not available, we used the PL's report as a measure of intervention behavior. We used an

ordinal 5-point scale to capture the strength of the intervention against the norm transgression. For each phase, the raters (or the PL, in phase 7) assessed whether P had not intervened (0), had questioned or commented on E1 and E2's actions without identifying them as wrong or immoral (1), had questioned or commented on E1 and E2's actions identifying them as wrong or immoral (2), had urged others to act (e.g., E1 to put the money back or C to intervene; 3), or had taken action themselves (e.g., by taking the receipt, ending the study, or reporting the situation to the PL; 4). Across the different phases, the level of agreement between the two raters was high overall (97.01%). For the 23 cases in which the two raters disagreed, we asked a third independent rater to provide a third rating. When the third rater's score coincided with the score provided by one of the other two raters, we then used it as final score. Otherwise, we consulted a fourth rater to reach agreement.

As we preregistered, we computed a dichotomous measure of intervention behavior from the ordinal scale used for the video ratings. This measure captured whether participants showed any sign of intervention throughout the aforementioned phases (1 = Intervention) or not (0 = No intervention). The reason for dichotomizing the ordinal scale, as we did with the continuous measure of 3PP, corresponded to a matter of divergence between these two measures. Although both of them arguably capture intervention/punishment intensity, the ordinal scale of intervention behavior additionally assessed qualitatively different types of reactions (e.g., identifying the norm violation vs. morally condemning it). This difference between measures could affect the expected relationship between them. Thus, we decided to use their dichotomized versions to ensure that both similarly captured the decision to react (intervene/punish) or not against a norm violation (embezzlement/unfair monetary distribution).

Retrospective measure of intervention behavior. In addition to the video ratings, the trained researchers involved in the staged embezzlement (i.e., E1, E2, C, and PL) made *in situ* retrospective assessments of the participants' reactions of those stages in which they were present, using the same ordinal scale described above. Similarly to our measure from the video ratings, we dichotomized the ordinal scale in people who showed any sign of intervention across phases (1 = Intervention) and those who did not (0 = No intervention).

Suspicion checks. To ensure that participants perceived the situation as credible, we considered two preregistered suspicion checks.

The first check were the answers to the open-ended question presented in Phase 6 (i.e., "In your opinion, what was this investigation about?"). We asked two additional independent raters to code whether participants expressed no doubts about the actual goal of the study (0), whether it was unclear whether they expressed potential doubts (1), or whether they expressed clear doubts by identifying the embezzlement as part of the study (2).

The second check were the potential doubts that participants could have expressed verbally or non-verbally while the situation was taking place. The two raters who watched the video material coded whether participants expressed potential doubts during the embezzlement. They used a similar scale as for the open-ended question (i.e., 0 – no doubts, 1 – unclear doubts, 2 – clear doubts).

We excluded data from 36 participants who expressed clear doubts in either of the suspicion checks (i.e., 4 in the open-ended question, 21 in the video ratings, and 11 in both).

Perception of embezzlement and risk of intervention. We further assessed whether participants properly perceived the embezzlement situation as a norm violation and a potential intervention against it as a risky behavior. We used different post-hoc items included in the questionnaire participants completed after the debriefing. Similar to our pretest, we used one item to assess the *perceived severity* of the norm violation (i.e., "How severe do you find it when such behavior takes place in reality?", 0 – Not at all, 5 - Extremely). To assess how immoral participants perceived the situation, we used six items intended to capture different moral concerns based on the five moral foundations proposed by Graham et al. (2011; e.g., "The experimenters behaved disloyally"), as well as liberty/oppression (i.e., "The behavior of the experimenters restrict other's freedom"). Finally, one item assessed the *perceived risk* of intervening ("Was or would an intervention have been associated with risks or inconveniences for you?" 0 – No, 1 – Yes).

Results

Perception of embezzlement and risk of intervention

We checked whether participants perceived the embezzlement as severe by testing whether their mean ratings were higher than a test value of 3, considering that each scale ranged from 0 to 5. Participants perceived the situation as highly severe, $M = 4.28$, $SD = 0.81$, $t(107) = 16.46$, $p < .001$, $d = 1.58$, 95% CI [1.15, 2.02]. Moreover, participants perceived the embezzlement to be a violation of care, $M = 3.69$, $SD = 1.15$, $t(107) = 6.20$, $p < .001$, $d = 0.60$, 95% CI [0.21, 0.99], fairness, $M = 4.36$, $SD = 0.96$, $t(107) = 14.71$, $p < .001$, $d = 1.42$, 95% CI [0.99, 1.84], loyalty, $M = 4.41$, $SD = 0.99$, $t(105) = 14.57$, $p < .001$, $d = 1.42$, 95% CI [0.99, 1.85], and authority, $M = 4.18$, $SD = 1.01$, $t(107) = 12.07$, $p < .001$, $d = 1.16$, 95% CI [0.74, 1.57]. Yet, they did not consider that the embezzlement entailed a violation of sanctity, $M = 2.71$, $SD = 1.74$, $t(107) = -1.714$, $p = .089$, $d = -0.17$, 95% CI [-0.55, 0.22], or liberty, $M = 2.39$, $SD = 1.58$, $t(106) = -3.99$, $p < .001$, $d = -0.39$, 95% CI [-0.77, 0.00].

With regard to the perceived risk of the intervention, a binomial test indicated that participants were more likely to consider that the intervention was or would be associated with risks for themselves (Yes, 62.26% vs. No 37.73%), and that this probability was significantly higher than 50% ($p = .015$, 95% CI [52.33%, 71.50%]).

3PP and intervention behavior

To test whether the participants' 3PP in the 3PPG predicted their intervention behavior against the staged embezzlement, we fit a multilevel logistic regression model using the R package *lme4* (version 1.1-25, Bates et al., 2015; R version 4.0.3). The model included the dichotomous measure of 3PP (0 = No punishment, 1 = Punishment) as fixed factor. Furthermore, we entered three dummy-coded variables and their interactions with 3PP to assess the effect of 3PP across experimental conditions (Dummy 1: 0 = *no ambiguity* + *no uncertainty*, 1 = *no ambiguity* + *uncertainty;* Dummy 2: 0 = *no ambiguity* + *no uncertainty*, 1 = *ambiguity* + *no uncertainty*; Dummy 3: 0 = *no ambiguity* + *no uncertainty*, 1 = *ambiguity* + *uncertainty*). For our hypothesis, the crucial interaction term was the Dummy 3 x 3PP, which allowed us to test whether 3PP significantly predicted intervention behavior when both ambiguity of the norm violation and cost uncertainty were present

Table 3.1

Multilevel logistic regression models with dichotomous dependent measures of intervention behavior (video ratings and retrospective measures) in Study 3.1.

	Video ratings			Retrospective measure		
Parameter	**Odds ratio** [95% CI]	***z***	***p***	**Odds ratio** [95% CI]	***z***	***p***
Fixed effects						
Constant	0 [0, 0]	-2.88	.004	0 [0, 0]	-3.12	.002
3PP	0.76 [0, 12152]	-0.06	.955	0.78 [0, 5297]	-0.06	.956
Dummy 1	0.90 [0, 69480]	-0.02	.985	0.89 [0, 26298]	-0.02	.982
Dummy 2	0.88 [0, 38514]	-0.02	.982	0.77 [0, 16912]	-0.05	.959
Dummy 3	0.89 [0, 52474]	-0.02	.983	0.77 [0, 23161]	-0.05	.960
3PP x Dummy 1	1.14 [0, 342063]	0.02	.984	1.16 [0,116917]	0.02	.980
3PP x Dummy 2	1.15 [0, 253802]	0.02	.982	1.39 [0, 130053]	0.06	.955
3PP x Dummy 3	1.14 [0, 308449]	0.02	.984	1.38 [0, 162101]	0.05	.957
Random effects						
σ^2	3.29			3.29		
$\tau_{00\ ID}$	4250.17			4079.36		
ICC_{ID}	1.00			1.00		
N_{ID}	106			108		
Observations	421			429		
Marginal R^2 / Conditional R^2	0.000 / 0.999			0.000 / 0.999		

Note. 3PP = Costly third-party punishment, σ^2 = Residual variance; $\tau_{00\ ID}$ = Variance of the intercept; ICC_{ID} = Intraclass correlation coefficient; N_{ID} = Total number of individuals.

(vs. absent). The interactions with Dummies 1 and 2 were included in order to check whether the effect of 3PP was already significant with the unique introduction of ambiguity or cost uncertainty. The model further included the participants' ID as a random factor to account for the within-subject variance. As dependent variables, we used the dichotomous measure of intervention (0 = No intervention, 1 = Intervention) based on the video ratings, as well as the dichotomous measure based on the retrospective reports by the different confederates. The results of both models are summarized in Table 3.1.

Both models indicated that the participants' 3PP did not predict their intervention behavior against the staged embezzlement, independently of the degree of ambiguity of the norm violation or cost uncertainty introduced in the 3PPG.

Study 3.2

In Study 3.2, we used the same embezzlement setting as in Study 3.1, but as a hypothetical norm violation described in a vignette format. The use of vignettes has the methodological advantage of offering high control over the situational information people receive when deciding to intervene. In this study, we used different types of vignettes (i.e., written vs. video vignettes), which provided perfect situational information about the norm violation (i.e., embezzlement). Thus, we hypothesized that people's 3PP in the 3PPG would better predict their intervention intentions when the 3PPG did not incorporate ambiguity of the norm violation in its design. Informed by the null effect of cost uncertainty on 3PP observed in Study 1.1 (see Chapter 1), we did not consider this factor as influencing the extent to which 3PP predicted intervention intentions.

We further explored whether the different vignette formats (written vs. video vignettes) affected our main findings. Without an a priori hypothesis, we tested whether the vignette format influenced people's intervention intentions and, if so, whether it moderated the relationship of 3PP with intervention intentions.

We preregistered our hypothesis, analysis plan, procedure and a log file including every measure in the study in the OSF (https://osf.io/etgq9). For a matter of conciseness, some details regarding the experimental design, the procedure and measures that were already described in Chapters 1 and 2 will not be repeated here.

Method

Design

The 3PPG in this study followed the same 2x2 within-subject design as in Study 3.1.

In the second session, however, we further manipulated between-subjects the *vignette format* (written vs. video vignettes).

Participants

From the sample of 226 participants recruited for Study 1.2 in Chapter 1, only 215 completed the second session. We excluded data of 3 based on preregistered concentration checks associated with the vignettes about the embezzlement (for further details, see below). Thus, our final sample was 212 (74% women; age range from 18 to 68, $M = 23.26$, $SD = 5.59$).

Procedure

The procedure of Study 3.2 was identical to Study 3.1, the only difference being that in Session 2 we presented participants with a vignette describing the different phases of the embezzlement situation staged in Study 3.1.

We randomly assigned half of the sample to watch a video vignette, which displayed the embezzlement from the perspective of the participant. The other half read a written transcription of the situation. In both cases, participants subsequently answered similar control items to the ones used in Study 1 to evaluate how participants perceived the described situation. Next, we assessed intervention intentions. We sequentially asked participants to report how they would react in each of the different phases after the embezzlement had been committed (i.e., Phases 4 to 8, see Study 3.1). Lastly, participants answered data-quality and concentration checks to ensure that they had paid attention to the situation described in the vignette.

Measures

Session 1 – 3PPG. The measures in Session 1 were identical to those in Study 3.1.

Session 2 – Vignettes of the embezzlement.

Intervention intention. To report their intervention intentions, participants used the same ordinal 5-point scale we used in Study 3.1 for the video ratings (0 – no intervention; 1 – question or comment on E1 and E2's actions without identifying them as wrong or immoral; 2 – question or comment on E1 and E2's actions identifying them as wrong or immoral; 3 – urging others to act, e.g., E1 to put the money back or C to intervene; 4 – taking action themselves, e.g., by taking the receipt, ending the study, or reporting the situation to the PL). As in Study 3.1, we dichotomized this scale into participants who reported any intervention intention throughout the different phases (1 = Intervention) and participants who did not (0 = No intervention).

Concentration checks. We used two preregistered concentration checks to ensure that participants had paid sufficient attention to the vignette and that they had put themselves into the embezzlement situation.

The first consisted of four items, which assessed the participants' attention to details described in the vignettes (e.g., "In the described/showed situation, I switched between rooms"). These items were the same for the written and the video vignette. We excluded participants who wrongly answered more than one of these items.

The second check was one further item, through which participants self-reported the effort they made in order to put themselves into the described situation (i.e., "I have made my effort to put myself in the described/showed situation"). We excluded participants who responded 1 or 2 in a scale from 1 (Totally disagree) to 6 (Totally agree).

In total, we excluded data from three participants, two of whom failed in the first attention items, and one who failed in the self-reporting effort item.

Perception of norm violation and risk of intervention. As in Study 3.1, we used one item to assess the *perceived severity* of the norm violation (e.g., "How severe do you find it when such behavior takes place in reality?", 0 – Not at all, 5 – Extremely). We further measured how immoral participants perceived the situation with the same six items intended to capture moral concerns on the five moral foundations proposed by Graham et al. (2011; e.g., "The experimenters behaved disloyally"), as well as liberty/oppression (i.e., "The behavior of the experimenters restricts the freedom of others"). Finally, as in Study 3.1, we used a binary item to assess the *perceived risk* of intervening ("Was or would an intervention have been associated with risks or inconveniences for you?" 0 – No, 1 – Yes). Yet, we further used an average index of perceived difficulty and risk from the four items included in the pretest (e.g., "I think it would be risky to intervene in the situation", $\alpha = .84$).

Results

Perception of norm violation and risk of intervention

We tested whether participants perceived the situation as severe and immoral by comparing their mean ratings with a test value of 3, considering that each scale ranged from 0 to 5. The results are summarized in Table 3.2. Participants perceived the situation as highly severe and immoral. Specifically, participants evaluated the embezzlement as a violation of care, fairness, loyalty, and authority, but, once again, they did not consider that the embezzlement entailed a violation of sanctity or liberty. Descriptively, these results were similar across participants who read the written vignette and those who watched the video vignette.

Regarding the perceived risk of the intervention, a binomial test indicated that participants from both groups were more likely to consider that the intervention would be associated with risks for themselves (Yes, 72.17% vs. No 27.83%) and that this probability was significantly higher than 50% ($p < .001$, 95% CI

[65.62%, 78.09%]). Using the continuous measure of perceived risk/difficulty to intervene, we observed the average level to differ significantly from 0. In fact, it was moderately high, considering that the mean exceeded the midpoint of the scale for both groups (i.e., from 0 to 5; see Table 3.2).

Table 3.2

Mean level comparisons of perceptions of the norm violation and risk of the intervention for the total sample and the different subsamples in Study 3.2.

		M	*SD*	*t*	*df*	*p*	*d* [95% CI]
Severity	Total Sample	4.50	0.83	26.25	211	<.001	1.80 [1.48, 2.12]
	Written vignette	4.50	0.81	19.12	105	<.001	1.86 [1.40, 2.32]
	Video vignette	4.51	0.86	17.97	105	<.001	1.75 [1.29, 2.20]
Immoral	Total Sample	4.73	0.64	39.10	211	<.001	2.69 [2.31, 3.06]
	Written vignette	4.75	0.59	30.64	105	<.001	2.98 [2.42, 3.53]
	Video vignette	4.72	0.70	25.24	105	<.001	2.45 [1.94, 2.96]
Justified (R)	Total Sample	4.67	0.77	31.49	208	<.001	2.18 [1.83, 2.52]
	Written vignette	4.78	0.51	35.40	104	<.001	3.45 [2.84, 4.07]
	Video vignette	4.58	0.95	16.89	103	<.001	1.66 [1.21, 2.11]
Care (R)	Total Sample	3.94	1.21	11.31	211	<.001	0.77 [0.50, 1.06]
	Written vignette	3.99	1.06	9.58	105	<.001	0.93 [0.53, 1.34]
	Video vignette	3.89	1.34	6.81	105	<.001	0.66 [0.27, 1.06]
Fairness (R)	Total Sample	4.61	0.82	28.70	211	<.001	1.97 [1.64, 2.30]
	Written vignette	4.66	0.60	28.50	105	<.001	2.77 [2.23, 3.31]
	Video vignette	4.56	0.99	16.25	105	<.001	1.58 [1.14, 2.02]
Loyalty (R)	Total Sample	4.49	1.00	21.54	211	<.001	1.48 [1.17, 1.78]
	Written vignette	4.48	0.91	16.80	105	<.001	1.63 [1.19, 2.08]
	Video vignette	4.49	1.10	13.98	105	<.001	1.36 [0.93, 1.79]
Authority (R)	Total Sample	4.08	1.29	12.15	211	<.001	0.83 [0.55, 1.12]
	Written vignette	3.93	1.36	7.06	105	<.001	0.69 [0.29, 1.08]
	Video vignette	4.22	1.20	10.41	105	<.001	1.01 [0.60, 1.42]
Sanctity (R)	Total Sample	2.35	1.71	-5.55	211	<.001	-0.38 [-0.65, -0.11]
	Written vignette	2.38	1.70	-3.77	105	<.001	-0.37 [-0.75, 0.02]
	Video vignette	2.32	1.73	-4.05	105	<.001	-0.39 [-0.78, -0.00]
Liberty (R)	Total Sample	2.92	1.64	-0.67	211	.503	-0.05 [-0.32, 0.22]
	Written vignette	2.86	1.53	-0.95	105	0.342	-0.09 [-0.48, 0.29]
	Video vignette	2.99	1.74	-0.06	105	0.956	-0.01 [-0.39, 0.38]
Perceived risk/difficulty	Total Sample	3.34	1.24	39.26	211	<.001	2.70 [2.32, 3.07]
	Written vignette	3.42	1.23	28.76	105	<.001	2.79 [2.25, 3.34]
	Video vignette	3.25	1.25	26.76	105	<.001	2.60 [2.08, 3.12]

Note. All one-sample t-tests used 3 as a test value, considering that every scale ranged from 0 to 5, except for perceived risk/difficulty, where the test value was 0.

3PP and intervention intentions

Table 3.3 summarizes the frequencies in which we observed each type of intervention intention across the different phases, as well as the maximum that participants showed across phases. The intervention intentions of participants from the written-vignette (M = 3.51, SD = 1.02) and the video-vignette groups (M = 3.51, SD = 1.02) did not significantly differ, $t(210) = -0.33$, $p = .738$, $d = -0.05$, 95% CI [-0.32, 0.22].

Table 3.3

Frequencies of intervention intention and maximum intervention intention across phases of the embezzlement.

	Intervention intention				
Phases of the vignette	*0*	*1*	*2*	*3*	*4*
Phase 4 – The two experimenters talk in front of you.	63 (30%)	56 (26%)	34 (16%)	40 (19%)	17 (8%)
Phase 5 – The experimenter puts the money into her pocket.	38 (18%)	52 (25%)	47 (22%)	45 (21%)	25 (12%)
Phase 6 – The other participant (C) receives the course credit.	83 (39%)	60 (28%)	29 (14%)	20 (9%)	19 (9%)
Phase 6 – You and the other participant (C) are alone working on the different tasks.	34 (16%)	109 (51%)	48 (23%)	14 (7%)	7 (3%)
Phase 6 – The experimenter comes back into the room after the tasks.	86 (41%)	47 (22%)	32 (15%)	33 (16%)	13 (6%)
Phase 7 – You are at the Project Leader's office.	38 (18%)	114 (54%)	38 (18%)	22 (10%)	0 (0%)
Phase 8 – The study is over.	46 (22%)	95 (45%)	41 (19%)	29 (14%)	0 (0%)
Maximum intervention intention across phases	2 (1%)	42 (20%)	55 (26%)	77 (36%)	36 (17%)

Note. Ordinal scale for intervention intention (*0* – no intervention; *1* – question or comment on E1 and E2's actions without identifying them as wrong or immoral; *2* – question or comment on E1 and E2's actions identifying them as wrong or immoral; *3* – urging others to act, e.g., E1 to put the money back or C to intervene; *4* – taking action themselves, e.g., by taking the receipt, ending the study, or reporting the situation to the PL).

We preregistered to use a multilevel logistic regression to examine whether the participants' 3PP in the 3PPG predicted their intervention intentions. However, we experienced model convergence problems, probably due to the low variability in the dichotomous measure of intervention intentions (i.e., only 2 participants reported not intervening across all different phases of the embezzlement). Thus, we decided to use the ordinal scale of intervention intentions and to explore its bivariate correlations with the respective continuous measure of 3PP across the rounds of the 3PPG that included ambiguity of the norm violation and those that did not.

We found that people's intentions to intervene against the embezzlement negatively correlated with their 3PP under no ambiguity, $r = -.14$, $t(210) = -1.99$, $p = .048$, 95% CI [-.27, .00], while under ambiguity this

correlation was not significant, $r = -.13$, $t(210) = -1.87$, $p = .063$, 95% CI [-.26, .00]. We tested the difference between these two dependent correlations using Meng et al.'s (1992) z-test, which suggested that the correlations did not significantly differ from each other, $\Delta r = -.008$, $z = -.14$, $p = .892$, 95% CI [-.12, .11].

If we split the data based on the vignette's format, we observed that the video vignette showed higher negative correlations – under no ambiguity, $r = -.18$, $t(104) = -1.91$, $p = .059$, 95% CI [-.36, .01]; under ambiguity, $r = -.19$, $t(104) = -1.93$, $p = .056$, 95% CI [-.36, .00] – than the written vignette – under no ambiguity, $r = -.10$, $t(104) = -1.00$, $p = .319$, 95% CI [-.28, .09]; under ambiguity, $r = -.09$, $t(104) = -0.93$, $p = .356$, 95% CI [-.28, .10]. Yet, no correlation was statistically significant.

General discussion

Previous lab work on the prevalence of 3PP and its role in the maintenance of social norms and human cooperation has been argued to lack external validity (Baumard, 2010; Guala, 2012). From a theoretical and phenomenological perspective, it is therefore important to provide evidence about the generalizability of lab research on 3PP and, in particular, about the extent to which individual differences in 3PP observed in the lab extrapolate to the field. In other words, does the 3PP people show in economic games (e.g., 3PPG; Fehr & Fischbacher, 2004) converge with how people react to actual norm violations in the field? In an attempt to address this question, we examined the relationship between people's 3PP in the 3PPG and their intervention behavior (and intentions) in a field-like situation of a norm violation (i.e., an embezzlement of lab funds). Critically, we tested whether the introduction of ambiguity of the norm violation and cost uncertainty in the 3PPG – factors that potentially characterize situations of norm violation in the field – enhanced the targeted relationship. 3PP did not predict people's intervention behavior against the embezzlement staged in the lab (Study 3.1). Furthermore, 3PP did, if at all, correlate negatively with people's intervention intentions when we described the embezzlement in a vignette format (Study 3.2). Critically, our findings did not differ when considering 3PP under ambiguity of the norm violation and cost uncertainty. Overall, the present research did not support the generalizability of the 3PP in the 3PPG to a field-like situation, which suggests that the interpretation of previous lab findings on 3PP should be more nuanced. Despite the null results, the present research offers relevant considerations regarding the external validity of research on 3PP.

From a phenomenological perspective, the first consideration refers to the psychological and situational factors underlying the discrepancies in 3PP between the lab and the field. The scenario that the 3PPG represents likely differs from many real-life contexts in which third parties address the perpetrator of a norm violation (e.g., littering in public spaces; Balafoutas et al., 2016). In our case, we took one initial step by introducing into the 3PPG the conditions of imperfect information in which third parties arguably react against norm violations outside the lab. Specifically, we considered the ambiguity of the norm violation and

the cost uncertainty as two moderating factors that could influence 3PP in the field, but which have been neglected in previous lab studies using the 3PPG as an experimental paradigm. Yet, the introduction of ambiguity and cost uncertainty in the 3PPG did not improve the convergence of people's 3PP in the 3PPG with their intervention behavior (or intervention intentions) against the embezzlement. This may suggest that further situational factors still created discrepancies between these two contexts that differentially influenced how people decided to react to the respective norm violation.

In the 3PPG, for example, the norm violation entails an unfair monetary distribution that directly affects the payoff of a present victim, whereas in the embezzlement there was no victim (or at least there was none present). One could argue that participants could perceive the lab or the institution where the study was conducted as the direct victim of the embezzlement. However, punishing on behalf of a person versus on behalf of an institutional entity could already involve different psychological mechanisms that led third parties to react differently (e.g., lower perceived interdependence; Pedersen et al., 2020). Another situational difference is that the norm violation in the 3PPG consists of not *sharing* money *with someone*, whereas the embezzlement entailed *taking* money *from someone*. Previous work has shown that the perceived social norms regulating these two types of behavior differ, even when the actual payoff of these actions is equivalent (Krupka & Weber, 2013), which could once again explain discrepancies in the third parties' reactions.

We may further examine certain methodological features of the 3PPG that distance the latter from the psychological conditions in which third parties decide to address norm violations in the field (Pedersen et al., 2013, 2018). For instance, the 3PPG generally offers third parties with one behavioral option (i.e., punishing the dictator), which is in stark contrast with the behavioral repertoire that third parties arguably have in most real-life situations (in the embezzlement situation, directly confronting the experimenters, but also reporting the violation to the other participant or to the project leader). Previous lab experiments actually show that, when other behavioral options are available (e.g., compensating the victim, rewarding the perpetrator for good behavior), third parties prefer those over punitive reactions; these experiments also show that, consequently, 3PP rates decrease (Heffner & FeldmanHall, 2019; Lotz, Okimoto, et al., 2011; Pedersen et al., 2013; van Doorn et al., 2018). Consistently with these findings, our studies showed that levels of third-party compensation were higher than levels of 3PP (see Supplementary Material). In a similar direction, some work suggests that more indirect forms of punishment (e.g., gossiping, social exclusion, or withholding help in the future) are more common in the field than direct punishment, as 3PP is generally operationalized in the 3PPG (i.e., direct confrontation with the perpetrator; Balafoutas et al., 2016; Molho et al., 2020). A second methodological feature that threatens the external validity of the 3PPG is the use of the strategy method, which is also present in our studies. Despite the main advantage of the strategy method (i.e., obtaining the complete strategic decision-making profile of each individual across different hypothetical scenarios), it forces third parties to anticipate their emotional reaction to hypothetical norm violations

(Pedersen et al., 2013). This decision-making mindset, often termed "cold", is clearly far off from the "hot", emotionally driven reactions that have been argued to underlie the third parties' direct confrontations with the perpetrators of real norm violations (Ginther et al., 2021; Gummerum et al., 2016; Sasse et al., 2020; Seip et al., 2009). In summary, the aforementioned features of the 3PPG frame the structural and psychological circumstances under which third parties decide to engage in 3PP differently from most situations in the field. We argue that this would limit the external validity of the 3PP that we usually observe in the 3PPG, since the determinants of 3PP in one and the other context differ.

A second consideration, derived from the previous one, links up with the theoretical implications of the lack of external validity of the 3PPG, which our findings arguably suggest. When we measure 3PP in the lab via the 3PPG, what exactly are we measuring? In his critical review of the literature on punishment research, Guala (2012) offers a useful distinction to answer this question. He differentiates between the "wide" and the "narrow" interpretation of punishment research. The *wide* interpretation refers to the conception of 3PP as a critical reinforcing mechanism of social norms (e.g., Fehr & Fischbacher, 2004). This wide interpretation necessarily assumes that the 3PP observed in the 3PPG extrapolates to a varied range of norm-violation situations to maintain the different social norms regulating people's behavior. This does not imply that the prevalence of 3PP we observe in the lab corresponds one-to-one to the prevalence of 3PP in the field. Yet, it does entail that the behavioral disposition that third parties show when they punish in the 3PPG should transfer or generalize to other norm-violation situations outside the lab. Under this wide interpretation, our findings question whether the 3PP observed in the 3PPG measure this disposition that can be generalized to other norm-violation situations in the field (e.g., an embezzlement), because our participants did not behave similarly across these two settings. Therefore, the present research should invite researchers to revise the notion that 3PP is a critical sanctioning mechanism underlying the maintenance of social norms.

On the other hand, the *narrow* interpretation of punishment research reduces the scope of what can be inferred from lab experiments on 3PP to what these actually show, i.e., that people are willing to incur costs to impose a penalty on perpetrators of unfair distributions of monetary resources under specific lab conditions (some of them described above). Under this narrow interpretation, our results would not offer an appropriate test of external validity, given that, as we previously discussed, the embezzlement situation differs in various ways from the situational and psychological structure of the 3PPG. In other words, we did not test whether the 3PP observed in the 3PPG generalizes to a structurally equivalent situation in the field (e.g., a friend not sharing with another friend the extra meal that he or she won in a restaurant lottery ticket). Although the narrow interpretation might be less impactful for the global understanding of the regulatory mechanisms of social norms, we agree with Guala (2012) that it is nevertheless critically informative. The acknowledgement that the 3PP we observe in the 3PPG is sensitive to an extensive list of structural and

psychological moderators – from the type of norm violation 3PP addresses to the presence or absence of a victim – may give us an idea of how sensitive 3PP is in the field. Our approach of introducing two of these situational moderators into the 3PPG (i.e., ambiguity of the norm violation and cost uncertainty) turned out insufficient in this respect, plausibly due to the remaining moderators that distanced the 3PPG from the embezzlement situation. However, the lack of relationship we observed between the 3PP in the 3PPG and the intervention (intentions) against the embezzlement exemplifies how specific the behavior assessed through the 3PPG seems to be.

Where should future research go from here? Once it is recognized that the results from lab experiments on 3PP have limitations regarding their external validity, researchers should keep extending the investigation of 3PP in a more varied set of contexts and situational circumstances (e.g., different norm violations; Ohtsubo et al., 2010; Pedersen et al., 2018). This will facilitate the identification of further situational and psychological moderators, but also the accumulation of sufficient evidence to proceed with efforts of aggregation. If the experimental paradigms used in lab studies to assess 3PP suffer from situational specificity, the aggregation of findings across different situations may provide a more reliable and externally valid measure of 3PP (Epstein, 1983). These efforts of aggregation include lab studies that assess 3PP across various situational moderators and structures (for an example of aggregation with other economic games, see Baumert et al., 2014). Moreover, further critical and meta-analytic reviews of the literature will also be extremely useful to provide more accurate estimations of when and how 3PP is likely to occur in the lab. In parallel, researchers should continue evaluating 3PP in the field, as this real-life assessment will always nourish and inspire lab research and its aggregation attempts (e.g., Balafoutas & Nikiforakis, 2012; Molho et al., 2020). Only when we gather sufficient aggregated evidence to ignore the situational specificity of 3PP will we be able to abandon its narrow interpretation and discuss its wider implications for the system of social norms.

Supplementary material and open practices

The supplementary material of this chapter is available in the following OSF repository: https://osf.io/bwr8a/. This repository further includes the data, analysis code, codebook and research materials to reproduce the reported results. Note that in the repository, Studies 3.1 and 3.2 are labelled differently (i.e., Study 1 and Study 2, respectively).

Chapter 4

Governmental distancing rules and normative change at the start of the COVID-19 pandemic in Germany

To contain the COVID-19 pandemic, governments worldwide implemented physical-distancing rules. Theoretical and empirical work suggests that such rules should inform people's norm perceptions, personal attitudes and willingness to intervene against others' behavior. We ask whether these rules affected the mentioned elements of the normative system and their interrelation, and tested these effects regarding physical-contact (vs. physical-distancing) behavior in a natural experimental design in Germany. As expected, once governmental rules were implemented, the perceived prevalence of physical-contact behavior decreased. Unexpectedly, there were no changes in perceived social appropriateness or the willingness to intervene. Paradoxically, personal attitudes toward physical-contact behavior became more positive. Furthermore, perceived social appropriateness and personal attitudes independently predicted people's willingness to intervene, but not more so after the rules were implemented. We conclude that governmental rules may prompt the perception of behavioral change, but their contribution to processes of normative change may be less straightforward than theoretically proposed.

Chapter 4 is based on: Toribio-Flórez, D.*, Fahrenwaldt, A.*, Saße, J., & Baumert, A. (under review). Governmental distancing rules and normative change at the start of the COVID-19 pandemic in Germany.
* The first two authors contributed equally to the referred manuscript and they agreed to share joint first authorship.

Social norms have important functions in a society. They serve as decision-making heuristics for social behavior as they reflect (often implicit) common rules (Bicchieri, 2005; Elster, 1989; Reno et al., 1993). As such, they facilitate cooperative relationships among the society's members (Ohtsuki & Iwasa, 2006; Thøgersen, 2008). Critically, they are subject to change, where old norms are revised and new norms emerge.

One important source of such change can be *institutional signals*, which are public rules, penalties or rewards provided by institutions serving the society's government, education or organization (Tankard & Paluck, 2016). Particularly if individuals consider an institution to be legitimate, they may perceive that signals from this institution carry information about social norms (Licht, 2008) and likely about changes therein. To test this notion, research has addressed, for example, how Supreme Court rulings affected perceptions of social support for gay marriage (Tankard & Paluck, 2017). However, previous theorizing and research did not take systematic account of how institutional signals might influence the different elements of what we term a *normative system*. We argue that a normative system can consist at least of four elements: (a) individuals' perceptions of a behavior's prevalence in society and (b) their perceptions of the social appropriateness generally ascribed to this behavior; (c) individuals' personal attitudes regarding this behavior, and (d) their willingness to intervene against others' behavior (by verbally or non-verbally expressing their disapproval).

In the present research, we investigate how institutional signals, in the form of governmental rules, can affect each of these elements (i.e., a-d) of a normative system (RQ1). We also ask whether they can strengthen the relationship between these elements. Specifically, we argue that the willingness to intervene against others' behavior is a neglected, yet important, element of normative systems. Thus, we focus on the effect of governmental rules on the relationship of norm perceptions and personal attitudes with people's willingness to intervene (RQ2). To answer these questions, we use the context of the beginning of the COVID-19 pandemic in Germany to investigate how the introduction of country-wide behavioral restrictions by the German government affected the normative system related to physical-contact behavior. To that end, we surveyed samples of the German general population in three waves, one before and two after the implementation of governmental physical-distancing rules in Germany on March 22nd 2020.

Normative systems and the effect of institutional signals

Social norms describe accepted and typical behaviors in social groups (Bicchieri, 2005). Importantly, the individual group members are recipients as well as contributors of normative information (Tankard & Paluck, 2016). They can infer their group's norms from the perceived prevalence of a behavior among their fellow group members (a.k.a. perceived *descriptive norms*) or via perceptions of others' explicitly expressed attitudes regarding the social appropriateness of that behavior (a.k.a. perceived *injunctive norms*). At the same time, individuals can also contribute normative information by sharing their own personal attitudes regarding the behavior or by showing willingness to intervene against others' behaviors. With these contributions,

individuals may even be able to elicit norm compliance in others (e.g., Blanton et al., 2001; Fehr & Fischbacher, 2004). All these elements seem to be crucial in the development and retention of social norms, thus constituting what we call a normative system.

In line with previous theorizing, we posit that normative change brought about by institutional signals could be reflected in changes in the aforementioned elements. Specifically, if individuals use signals from legitimate societal institutions as indicators of their group's social norms (e.g., Licht, 2008; McAdams, 2000), then their perceptions of a behavior's prevalence and social appropriateness should change in the signaled direction (Tankard & Paluck, 2016). Similarly, theories on the expressive function of the law propose that legal institutional signals can trigger the internalization of the signaled norm, changing individuals' personal attitudes, as well as willingness to intervene against potential norm deviations (Carbonara et al., 2008).

However, the proposed effects of institutional signals on normative systems are debated amongst theorists (Rosenberg, 2008; Schacter, 2008), which is also reflected in mixed empirical findings. For instance, while some studies indicate that institutional signals change perceptions of injunctive and descriptive norms (Tankard & Paluck, 2017) as well as personal attitudes (Kotsadam & Jakobsson, 2011), other research does not show such changes (e.g., Soss & Schram, 2007; note that research on this topic is rather scarce as most studies focus on effects on norm-compliant behavior, e.g., Huber, et al., 2018 and Khatapoush & Hallfors, 2004, including case-studies during the COVID-19 pandemic, e.g., Casoria et al., 2020 and Götz et al. 2020).

Notably, some studies offered insights into potential boundary conditions under which the effect of institutional signals is likely observable. Studies have suggested that, for an institutional signal to be effective, individuals should perceive that the institution represents and serves their social group well (Hogg & Reid, 2006), that many others are exposed to the same signal (Arias, 2014), and that individuals are directly affected by the signal's consequences (Kotsadam & Jakobsson, 2011). Lastly, an important factor determining whether an institutional signal brings about normative change could be the degree of conflict between the newly signaled norm and previous social norm perceptions as well as personal attitudes. Specifically, an institutional signal may not change perceptions of social norms (Tankard & Paluck, 2016), personal attitudes (Andrighetto et al., 2010; Carbonara et al., 2008) or interventions against deviations (Kahan, 2000) in the intended direction if it strongly conflicts with previous social norms.

The normative system during the eruption of the COVID-19 pandemic

To our knowledge, no previous study has investigated the effect of institutional signals on these four elements of the normative system simultaneously. We set out to fill this gap by investigating whether the governmental physical-distancing rules, introduced during the beginning of the COVID-19 pandemic in Germany, affected the until-then established normative system for physical-contact behavior.

In the beginning of March 2020, SARS-CoV-2 spread in Germany, with the first COVID-19-related death registered on March 9th, similarly to other European countries. To contain the spread of the virus, societies had to adjust social behavior quickly. Governments worldwide tried to promote this behavioral adaptation by employing institutional signals in the form of physical-distancing rules. In Germany, the government decided to implement such nationwide rules, jointly with the federal state ministers, on March 22nd 2020. Globally, these institutional signals were crucial in limiting the viral spread (Fazio et al., 2021; López & Rodó, 2020; Moosa, 2020). We asked whether these signals also reconfigured the normative system regarding physical-contact behavior.

The context of our case-study was particularly interesting for addressing this question, because it fulfilled some of the situational prerequisites for effective normative change as suggested in the literature mentioned earlier. Specifically, the German public generally perceived the signaling institution (i.e., the German government) as trustworthy and legitimate (Murtin et al., 2018); the signal (i.e. the introduction of physical-distancing rules) was broadcasted on all possible media channels and likely received by the majority of citizens; and virtually everyone was affected by the physical-distancing rules. Given these situational characteristics, we expect effects of the institutional signal on all four elements of the normative system regarding physical-contact behavior.

However, another important aspect of the COVID-19 context may have been the degree of conflict between the institutional signal and preexisting norms and attitudes. Under normal circumstances, personal freedom of mobility is a valued individual right in Germany and physical contact was completely appropriate before the beginning of the pandemic. Therefore, the new rules stood in stark contrast to preexisting social norms. This may have limited their power to trigger normative change (Tremewan & Vostroknutov, 2020).

At the same time, the start of the COVID-19 pandemic provided individuals with new and contradictory information, for example in the form of incongruent media coverage (Aerzteblatt, 2020; Boberg et al., 2020). As a result, this context had the potential to elicit uncertainty in individuals' perceptions of the social norms regarding physical-contact behavior as well as in their personal attitudes towards this type of behavior and their willingness to intervene against it (Bohner & Dickel, 2011; Lapinski & Rimal, 2005; Merguei et al., 2021). Such uncertainty could render the elements of a normative system more malleable, which should facilitate normative change by institutional signals.

Taking into account the specific circumstances of the COVID-19 pandemic in Germany, we thus expect to observe effects of the governmental rules on the four elements of the normative system, and hypothesize:

After the introduction of governmental physical-distancing rules (vs. before), people perceived behavior entailing a risk of physical contact as less prevalent (**H1a**) and less socially appropriate (**H1b**), held

more negative attitudes towards it (**H1c**), and showed a higher willingness to intervene against this type of behavior (**H1d**).

To explore the specificity of the effects of the institutional signal, we also investigate how the institutional signal affected the four elements of the normative system for behavior involving the use of public spaces without a risk of physical contact (henceforth, *physical-distancing* behavior). Considering that physical-contact behavior was more likely to entail the spread of COVID-19 and therefore to be tackled by even mild rules, we hypothesize:

In comparison with physical-distancing behavior, people perceived physical-contact behavior as less prevalent (H2a) and less socially appropriate (H2b); and they reported more negative attitudes (H2c) as well as a higher willingness to intervene against such behavior (H2d), irrespective of the implementation of governmental rules.[5]

Institutional signals and the relationship between elements of the normative system

Notably, the different elements of the normative system are intertwined (e.g., Terry & Hogg, 1996). We specifically focus on the relationships between the willingness to intervene and the other aforementioned elements. The former may play a key role in attaining stability or change in the normative system. Specifically, the willingness to intervene may help to stabilize the normative system whenever it implies punishing norm-deviant behavior, but it may also reinforce the normative change if individuals are willing to intervene in line with newly emerging or changing norms (Almenberg et al., 2010; Blanton et al., 2001; Cushman, 2015; Fehr & Fischbacher, 2004).

In line with results from prior research (e.g., Deitch-Stackhouse et al., 2015), we hypothesize that individuals' willingness to intervene against physical-contact behavior will be predicted by its perceived social appropriateness and personal attitudes towards it. We also explore[6] whether perceptions of the prevalence of physical-contact behavior will predict individuals' willingness to intervene against it (e.g., Lemay et al., 2019, Traxler & Winter, 2012).

[5] It is important to note that the specific behavioral rules to be introduced by the government were unknown when planning the study. Mild rules only prohibiting physical-contact behavior were conceivable (and eventually opted for). However, the introduction of a nationwide lockdown following the example of other European countries (e.g., Italy, Spain), that prohibited physical-contact as well as physical-distancing behavior, was also possible. Given this uncertainty when designing the study and the expectation that only physical-contact behavior would be affected in either scenario, H2a-d were not preregistered.

[6] The implementation of a full lockdown seemed possible when we designed the study. We expected that such a scenario would drastically reduce the variance of the perceived prevalence of physical-contact behavior, thereby reducing its value as a predictor of the willingness to intervene. Thus, we only explored whether people's perceived prevalence of physical-contact behavior predicted their willingness to intervene against it and how this changed over time, instead of including it as part of our preregistered hypotheses.

More importantly, we argue that, beyond their effects on the single elements of the normative system, institutional signals may strengthen the relationship between individuals' willingness to intervene and their norm perceptions and personal attitudes, respectively. As we described earlier, at the beginning of the COVID-19 pandemic, individuals might have been uncertain about their personal attitudes and about the changing norms regarding physical-contact behavior. Accordingly, these elements of the normative system might have been less motivationally relevant for whether individuals were willing to intervene and express disapproval of others' physical-contact behavior. In this situation, a multitude of other factors may have been more important for individuals' reactions to others' physical-contact behavior. However, the institutional signal posed by the introduction of governmental rules likely facilitated to think about the behavior in question and to form personal attitudes as well as norm perceptions towards it (Chu et al., 2021; Prosser et al., 2020), making individuals feel more comfortable in basing their willingness to intervene on their personal attitudes and norm perceptions.

Crucially, the proposed moderating effect of institutional signals on the relationship of norm perceptions and personal attitudes with the willingness to intervene has not been tested yet by previous research. In line with our reasoning, we hypothesize:

After the introduction of governmental rules, people's perceived social appropriateness and personal attitudes regarding physical-contact behavior should independently predict the willingness to intervene against physical-contact behavior. These associations should be stronger after (as opposed to before) the introduction of governmental rules (H3).

Research overview

We tested our preregistered hypotheses (https://aspredicted.org/blind.php?x=zw5r4e) in a natural experiment following a pre-post-design. Around March 22nd, 2020, when the German government, jointly with the federal state ministers, decided to implement nationwide physical-distancing rules (for the complete set of implemented rules, see the Supplementary Material), we recruited three samples of participants at different time points. We recruited the first sample on March 20th (i.e., two days before the introduction of governmental rules; T1); the second one on March 24th (i.e., two days after the introduction of governmental rules; T2); and the third one on March 31st (i.e., T3). We included this last sample to examine effects over a longer time period.

Method

Participants

We preregistered to collect data from 500 participants per time point, based on availability of resources and time constraints. We anticipated the exclusion of a substantial part of data based on

preregistered exclusion criteria. Therefore, we recruited a total sample of 1,786 participants from a German online panel, who were compensated with 0.50€ for their participation.

After excluding data from participants with incomplete responses or with duplicated IP addresses, our sample consisted of 1,639 participants. We also excluded data from participants whose completion time was 1.5 times above or below the interquartile range (n = 103). Furthermore, we excluded data from residents of the federal states Bavaria and Saarland (n = 234), because these states had already implemented strict governmental physical-distancing rules on March 20th (i.e., T1). Our final sample consisted of 1,302 participants. Post-hoc power-determination analyses based on Monte Carlo simulations indicated that this sample size would allow us to detect small to moderate effects sizes (i.e., β = [.05, .30], d = .22) for every hypothesis with at least 90% statistical power (for further details, see the Supplementary Material).

Table 4.1 presents demographic information from each time point. We provide statistical comparisons between the different samples in the Supplementary Material.

Procedure and measures

We programmed our study in Qualtrics as an online survey (Qualtrics, Provo, UT). At every time point, after providing informed consent, participants were presented with six vignettes describing different kinds of behavior that varied in the extent to which they entailed physical contact (see below for more details). We presented the vignettes three times, together with items assessing the participants' perceptions of (a) the prevalence and (b) the social appropriateness of the described behavior in one block; (c) their personal attitudes toward the behavior in another block; and (d) their' willingness to intervene against it in a third block. To avoid carry-over effects between these measures, we randomized the order of presentation of these three blocks across participants. Also, within each block, the order of the six vignettes was random.

Of the six vignettes, three described behavior which respected *physical distance* (e.g., "Person A goes out on the street to walk to the supermarket, but on the way, he/she stops in front of a friend's house to say 'Hello' from some distance."), whereas the other three vignettes portrayed behavior entailing an apparent risk of *physical contact* (e.g., "Person A has his/her kids at home. As the kids get bored, Person A decides to take them to the closest park, where Person A expects to encounter some neighbors."). Factor analyses supported the distinction between these two clusters of vignettes across the dependent measures a-d (see the Supplementary Material).

To assess (a) the perceived prevalence of each behavior, we asked participants *how likely they believed it* was *that most other people would behave* in the way described in each vignette (1 – Very unlikely, 7 – Very likely). For (b) social appropriateness, we measured the participants' beliefs about *how appropriate, correct, or moral most other people* would *find* the respective behavior, whereas for (c) personal attitudes, the question referred to *how appropriate, correct, or moral* participants *personally found* each behavior (1 – Very inappropriate, 7 – Very

appropriate). Finally, we assessed (d) participants' willingness to intervene with three items. Each item captured different types of reactions toward potential norm transgressions: direct intervention ("I would tell the person directly that his/her behavior is inappropriate"), indirect intervention ("I would tell other people that this person is behaving inappropriately"), and anger expression ("I would be angry with this person and I would let them know"; 1 – Very unlikely, 7 – Very likely). We computed an average score per vignette (Cronbach's $\alpha s \geq .82$).

Once participants had completed the three blocks of measures, they filled in further questionnaires unrelated to the set of vignettes. These questionnaires included demographic information (i.e., gender, age, education, and political orientation) and other measures (e.g., participants' perceived ambiguity of behavioral norms and their dispositional justice sensitivity; see Supplementary Material for the complete list of measures).

Table 4.1

Demographic information of participants across different time points.

	Total sample	**T1**	**T2**	**T3**
N	1,302	440	447	415
Age				
M	50.09	47.72	50.14	52.54
SD	15.79	15.35	15.15	16.58
Range	[18-88]	[18-87]	[18-88]	[18-81]
Gender				
% Females	41.94%	38.18%	40.04%	47.95%
% Males	57.60%	60.91%	59.96%	51.57%
% Diverse	0.31%	0.46%	0%	0.48%
Education				
No education	0.38%	0.45%	0.22%	0.48%
Secondary school, school years 5 to 9 (*Hauptschule*)	5.68%	5.45%	6.94%	4.58%
Secondary school, school years 5 to 10 (*Realschule/mittlere Reife*)	17.67%	17.05%	18.79%	17.11%
Highschool / University entrance diploma (*Abitur/Fachhochschulreife*)	18.28%	21.14%	13.87%	20.00%
Professional training (*Lehre/Berufsausbildung*)	27.42%	26.14%	27.29%	28.92%
University degree (*Hochschulabschluss*)	27.96%	27.27%	31.10%	25.30%
Doctoral degree (*Promotion*)	2.15%	1.82%	1.12%	3.61%
Political orientation				
M	44.82	44.66	43.47	46.43
SD	17.79	18.92	17.04	17.22

Note. *N* = Number of observations, *M* = Observed mean, *SD* = Standard deviation, T = Time point. *Range* = Minimum and maximum observed response.
Political orientation was measured on a continuous scale from 1 (extreme left) to 100 (extreme right).

Results

Governmental rules and elements of the normative system

To test H1a-d, we conducted four separate mixed-effects regression analyses. The data clustered repeated measures within subjects and subjects within time points. In each model, we regressed the dependent variable on the fixed factor *Time* (dummy-coded, with T1 coded 0 as reference category), while controlling for subject and vignettes random effects. Furthermore, the model included *Behavior* (dummy-coded: 0 = physical-contact; 1 = physical-distancing) as a fixed factor (H2a-d), and its interaction with Time. The results of the model for each dependent variable are summarized in Table 4.2, whereas Figure 4.1 depicts mean levels across time points and types of behavior.

We observed that, after the introduction of governmental rules (T2 vs. T1), people perceived physical-contact behavior to be (a) significantly less prevalent, yet (b) similarly socially appropriate. Furthermore, people held (c) significantly more positive attitudes toward physical-contact behavior. We did not find differences in (d) people's willingness to intervene against it. We observed comparable differences between T3 and T1, while T3 and T2 did not significantly differ regarding any of the measures a-d ($p \geq .131$). Thus, the effects observed two days after the introduction of governmental measures seemed stable a week later.

When comparing physical-contact with physical-distancing behavior (H2), people perceived the former as significantly (a) less prevalent and (b) less socially appropriate, whereas they held (c) more negative attitudes towards it. Moreover, people reported significantly (d) higher willingness to intervene against physical-contact than against physical-distancing behavior.

Exploring whether the implementation of governmental rules (T2 vs. T1) differentially affected the perception of physical-distancing versus physical-contact behavior, we found that this was indeed the case for perceived (a) prevalence and (b) social appropriateness of these types of behavior, as evidenced by the significant Time x Behavior interactions. In contrast to physical-contact behavior, the introduction of governmental rules led people to perceive physical-distancing behavior as significantly more prevalent and significantly more socially appropriate. The effect of the governmental rules on (c) personal attitudes and (d) willingness to intervene did not significantly differ across types of behavior, as shown by the non-significant Time x Behavior interactions. Like for physical contact, people held significantly more positive attitudes toward physical-distancing behavior at T2 compared to T1, while their willingness to intervene against each type of behavior did not change.

Table 4.2

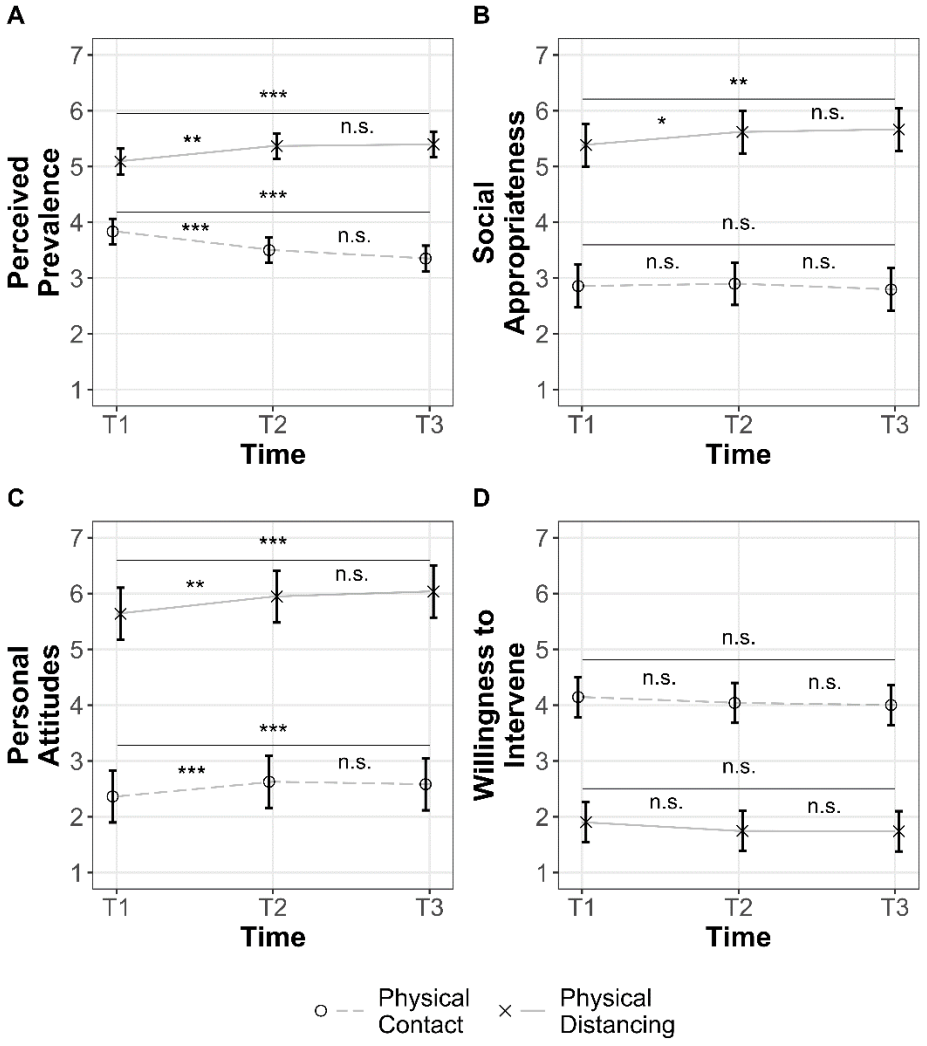
Mixed-effects models examining the effect of time on perceptions of prevalence and social appropriateness, personal attitudes, and willingness to intervene for physical-contact and physical-distance behavior.

	Perceived prevalence			Social appropriateness			Personal attitudes			Willingness to intervene		
Fixed effects	***b*** [95% CI]	***t***	***p***	***b*** [95% CI]	***t***	***p***	***b*** [95% CI]	***t***	***p***	***b*** [95% CI]	***t***	***p***
Constant	3.84 [3.61, 4.07]	32.76	**< .001**	2.86 [2.48, 3.24]	14.69	**< .001**	2.36 [1.90, 2.83]	9.94	**< .001**	4.15 [3.79, 4.50]	22.66	**< .001**
Time [T2]	-0.33 [-0.48, -0.18]	-4.31	**< .001**	0.04 [-0.12, 0.20]	0.52	.605	0.26 [0.12, 0.40]	3.65	**< .001**	-0.11 [-0.27, 0.06]	-1.22	.224
Time [T3]	-0.48 [-0.64, -0.33]	-6.15	**< .001**	-0.06 [-0.22, 0.10]	-0.76	.449	0.22 [0.07, 0.36]	2.97	**.003**	-0.14 [-0.32, 0.03]	-1.62	.105
Behavior [Physical-distancing]	1.25 [0.95, 1.56]	8.01	**< .001**	2.53 [2.00, 3.05]	9.35	**< .001**	3.27 [2.62, 3.93]	9.83	**< .001**	-2.24 [-2.73, -1.75]	-9.00	**< .001**
Time [T2] x Behavior [Physical-distancing]	0.61 [0.46, 0.76]	7.76	**< .001**	0.19 [0.03, 0.36]	2.29	**.022**	0.04 [-0.11, 0.19]	0.58	0.562	-0.05 [-0.19, 0.09]	-0.72	.470
Time [T3] x Behavior [Physical-distancing]	0.79 [0.63, 0.95]	9.88	**< .001**	0.34 [0.18, 0.51]	4.02	**< .001**	0.18 [0.03, 0.33]	2.32	**.020**	-0.02 [-0.16, 0.12]	-0.27	.784
Random effects												
σ^2	2.06			2.34			1.93			1.70		
$\tau_{00\ Subject}$	0.64			0.66			0.51			1.09		
$\tau_{00\ Vignette}$	0.03			0.10			0.16			0.09		
ICC	.25			.25			.26			.41		
$N_{Subject}$	1302			1302			1302			1302		
$N_{Vignette}$	6			6			6			6		
Observations	7811			7812			7812			7812		
Marginal / Conditional R^2	.219 / .412			.371 / .526			.520 / .644			.309 / .592		

Note. σ^2 = Residual variance; τ_{00} = Variance of the intercept; ICC = Intraclass correlation coefficient; N = Number of observations per level.

Figure 4.1

Levels of perceived prevalence (A), social appropriateness (B), personal attitudes (C), and willingness to intervene (D) for physical-contact and physical-distancing behavior before (T1) and after (T2, T3) the introduction of physical-distancing rules by the German government.



Note. * $p < .05$., ** $p < .01$., *** $p < .001$. Error bars represent 95% CIs.

Social Appropriateness, Personal Attitudes and Willingness to Intervene

For testing H3, we regressed willingness to intervene against physical-contact behavior on the fixed effects of *Social Appropriateness*, *Personal Attitudes*, and *Time*, together with the Social Appropriateness x Time and Personal Attitudes x Time interactions. The model accounted for subject and vignette random effects (see Table 4.3). As expected, we found significant effects of Social Appropriateness, $F(1,3680) = 52.274$ $p < .001$, $\eta_p^2 = .01$ and Personal Attitudes on people's willingness to intervene, $F(1,3763) = 305.931$, $p < .001$, $\eta_p^2 = .08$, indicating that the less socially appropriate people perceived physical-contact behavior to be, and the less positive attitudes they held towards it, the more willing they were to intervene against it. However, against

our hypothesis, these effects did not significantly vary across time points, as indicated by the non-significant interactions of Social Appropriateness x Time, $F(2,3660) = 2.179$, $p = .113$, $\eta_p^2 = .001$ and Personal Attitudes x Time, $F(2,3690 = 0.239$, $p = .787$, $\eta_p^2 = .000$. The main effect of Time was not significant either, $F(2,2950) = 1.379$, $p = .252$, $\eta_p^2 = .001$.

Extending the previous model, we estimated an additional, more comprehensive model, the goal of which was twofold: first, to explore whether the perceived prevalence of a type of behavior also predicted the willingness to intervene; and second, to check whether the observed effects of social appropriateness and personal attitudes differed between physical-contact and physical-distancing behavior. Thus, this second model included the fixed factors *Perceived Prevalence*, *Social Appropriateness*, *Personal Attitudes*, *Time* and *Behavior*, and the two-way and three-way interactions with Time and Behavior as moderators. The model controlled for subject and vignette random effects. We display the results in Table 4.3 and Figure 4.2. We observed significant main effects of Social Appropriateness, $F(1,7248) = 121.931$, $p < .001$, $\eta_p^2 = .017$, and Personal Attitudes, $F(1,7233) = 967.306$, $p < .001$, $\eta_p^2 = .118$, as well as a main effect of Behavior, $F(1,21) = 96.715$ $p < .001$, $\eta_p^2 = .823$ Perceived Prevalence did not have a significant effect on the willingness to intervene, $F(1,7321) = 1.585$, $p = .208$, $\eta_p^2 = .000$. Furthermore, we found a significant interaction between Personal Attitudes and Behavior, $F(1,7152) = 55.696$, $p < .001$, $\eta_p^2 = .008$, indicating that Personal Attitudes predicted the willingness to intervene significantly better for physical-contact than for physical-distancing behavior. We also found a significant interaction between Personal Attitudes and Time, $F(2,7254) = 4.915$, $p = .007$, $\eta_p^2 = .001$, indicating that the relationship between Personal Attitudes and the willingness to intervene was significantly weaker at T2 than at T1, but this relationship did not significantly differ when comparing T1 and T3. No other two-way or three-way interactions were significant.

Table 4.3

Pre-registered (Model 1) and exploratory (Model 2) mixed-effects models examining the fixed effects of personal attitudes, social appropriateness and perceived prevalence of behavior on willingness to intervene as a function of time and behavior.

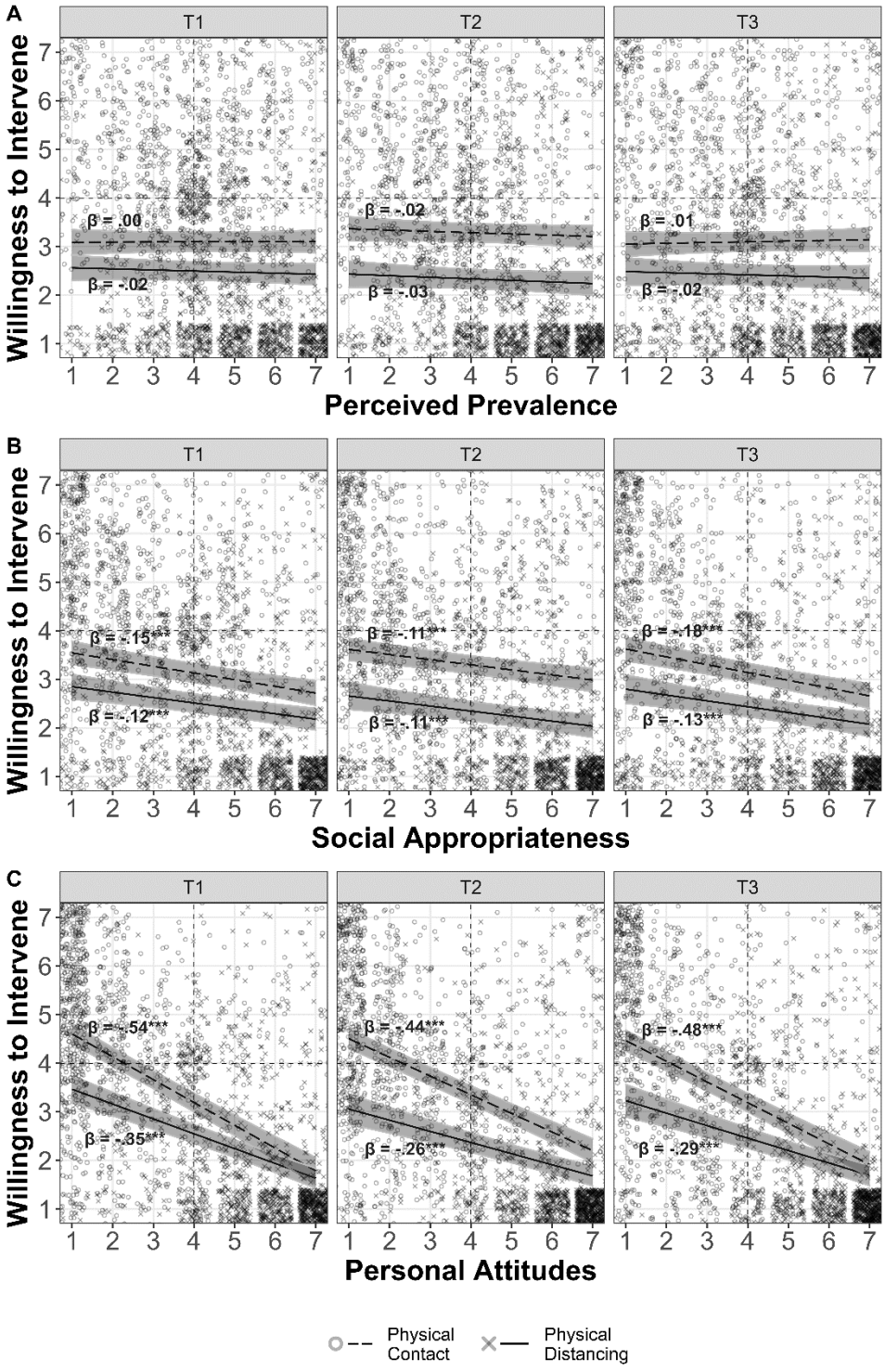
	Model 1			**Model 2**		
Fixed effects	***b*** [95% CI]	***t***	***p***	***b*** [95% CI]	***t***	***p***
Constant	5.21 [4.87, 5.55]	30.09	< .001	5.64 [5.39, 5.89]	43.75	< .001
Personal attitudes	-0.31 [-0.37, -0.24]	-9.91	< .001	-0.47 [-0.52, -0.42]	-18.15	< .001
Social appropriateness	-0.12 [-0.17, -0.07]	-4.31	< .001	-0.14 [-0.19, -0.09]	-5.33	< .001
Perceived prevalence				0.00 [-0.05, 0.05]	0.11	0.915
Time [T2]	-0.24 [-0.54, 0.06]	-1.54	.123	-0.19 [-0.49, 0.10]	-1.29	.196
Time [T3]	-0.04 [-0.34, 0.26]	-0.27	.783	-0.14[-0.43, 0.15]	-0.93	.353

Behavior [Physical-distancing]				-1.30 [-1.71, -0.90]	-6.33	< .001
Personal attitudes x Time [T2]	0.03 [-0.05, 0.10]	0.67	.501	0.09 [0.02, 0.15]	2.52	.012
Personal attitudes x Time [T3]	0.01 [-0.07, 0.09]	0.25	.802	0.04 [-0.03, 0.11]	1.25	.212
Social appropriateness x Time [T2]	0.05 [-0.02, 0.12]	1.33	.182	0.03 [-0.04, 0.10]	0.92	.355
Social appropriateness x Time [T3]	-0.02 [-0.10, 0.05]	-0.63	.526	-0.02 [-0.10, 0.05]	-0.61	.542
Perceived prevalence x Time [T2]				-0.03 [-0.10, 0.04]	-0.81	.420
Perceived prevalence x Time [T3]				0.01 [-0.06, 0.08]	0.34	.735
Personal attitudes x Behavior [Physical-distancing]				0.16 [0.09, 0.24]	4.53	< .001
Social appropriateness x Behavior [Physical-distancing]				0.03 [-0.04, 0.10]	0.75	.452
Perceived prevalence x Behavior [Physical-distancing]				-0.02 [-0.09, 0.04]	-0.72	.470
Time [T2] x Behavior [Physical-distancing]				-0.28 [-0.80, 0.24]	-1.05	.295
Time [T3] x Behavior [Physical-distancing]				-0.11 [-0.64, 0.43]	-0.39	.699
Personal attitudes x Time [T2] x Behavior [Physical-distancing]				-0.01 [-0.11, 0.09]	-0.19	.850
Personal attitudes x Time [T3] x Behavior [Physical-distancing]				0.00 [-0.10, 0.11]	0.09	.929
Social appropriateness x Time [T2] x Behavior [Physical-distancing]				-0.03 [-0.13, 0.07]	-0.49	.622
Social appropriateness x Time [T3] x Behavior [Physical-distancing]				0.01 [-0.09, 0.12]	0.26	.793
Perceived prevalence x Time [T2] x Behavior [Physical-distancing]				0.02 [-0.08, 0.12]	0.37	.711
Perceived prevalence x Time [T3] x Behavior [Physical-distancing]				-0.01 [-0.12, 0.09]	-0.25	.802
Random effects						
σ^2	1.10			1.29		
$\tau_{00\ Subject}$	1.92			1.14		
$\tau_{00\ Vignette}$	0.05			0.01		
ICC	.64			.47		
$N_{Subject}$	1302			1302		
$N_{Vignette}$	3			6		
Observations	3906			7811		
Marginal / Conditional R^2	.115 / .683			.426 / .697		

Note. σ^2 = Residual variance; τ_{00} = Variance of the intercept; ICC = Intraclass correlation coefficient; N = Number of observations per level.

Figure 4.2

Jittered raw data and estimated relationship of perceived prevalence (A), social appropriateness (B), and personal attitudes (C) with willingness to intervene for physical-contact and physical-distancing behavior before (T1) and after (T2, T3) the introduction of physical-distancing rules by the German government.



Note. * $p < .05$., ** $p < .01$., *** $p < .001$. Band widths represent 95% CIs.

General discussion

In the context of the eruption of the COVID-19 pandemic, we investigated whether the introduction of physical-distancing rules by the German government affected four elements of the normative system for physical-contact behavior, namely its perceived prevalence and social appropriateness, personal attitudes towards it, and the willingness to intervene against it. Furthermore, we asked whether institutional signals would strengthen the relationship of perceived social appropriateness and personal attitudes with the willingness to intervene against others' physical-contact behavior. The observed pattern of results does not support a uniform effect of the governmental rules. We find a decrease in the perceived prevalence, but not in the social appropriateness, of physical-contact behavior, and, paradoxically, more positive personal attitudes towards it after the introduction of governmental rules. Moreover, these rules did not strengthen, but rather weakened the relationships between elements. We discuss below how this pattern of results may be explained, at least in part, by specific contextual characteristics.

Predicted and paradoxical effects of institutional signals

Focusing first on the perceptions of social norms, the introduction of governmental rules yielded the expected reduction in prevalence perceptions of physical-contact behavior (i.e., descriptive norms; H1a). This finding is in line with the objective reduction in mobility in Germany from T1 to T2 (Schlosser et al., 2020). However, unexpectedly, the governmental rules did not affect perceptions of the behavior's social appropriateness (i.e., injunctive norms; H1b). It is noteworthy that these ratings were already low at T1, which speaks against the possibility that strongly contrasting, preexisting norms (e.g., freedom of mobility) hindered institutional signals from being effective (an idea we mentioned in the introduction). A more plausible explanation may be that the signaling effect of governmental rules occurred before these rules were officially introduced. Indeed, two days before rule implementation at the national level (i.e., at our T1), two federal states already introduced curfew-like behavioral rules (Connor, 2020) and the possibility of a nationwide curfew was discussed in the daily news (e.g., ARD-aktuell, 2020). Consequently, individuals from the remaining federal states may have anticipated drastic nationwide physical-distancing rules at T1 and updated their perceived social appropriateness of physical contact accordingly. Together with the fact that the rules eventually implemented at the national level were much milder than the forecasted full lockdown, this may explain why we did not observe a reduction in perceived social appropriateness from T1 to T2.

We speculate that the divergence between our results for perceptions of the social appropriateness and the prevalence of physical-contact behavior results from a change in beliefs regarding the gap between others' professed attitudes or intentions and their actual behavior (e.g., Gollwitzer et al., 2009). Specifically, while participants at T1 likely suspected that others did not practice what they preached (regarding keeping

physical distance), the possibility that the government enforced the newly proclaimed distancing-rules at T2 could have led participants to assume that others would change their behavior.

The effect of anticipating stricter rules than eventually implemented also could have affected individuals' personal attitudes (H1c), since we observed that personally, many participants found physical contact rather inappropriate before the governmental rules were introduced. In contrast, at T2, when the rules were implemented, individuals reported significantly more positive personal attitudes toward physical-contact behavior. Thus, the governmental rules seemed to have had a paradoxical effect on personal attitudes. One way to explain this finding is that, when the physical-distancing rules became reality, this could have increased awareness of their potential personal implications (e.g., unemployment, bankruptcies, mental-health issues; Armbruster & Klotzbücher, 2020; Coibion et al., 2020; Jones et al., 2020). Therefore, in addition to the possibility that participants anticipated stricter rules than those eventually implemented, the relaxation of personal attitudes toward physical-contact behavior at T2 also could have been self-serving. This notion has been discussed by others (Rosman et al., 2020) and it aligns with some of our exploratory results, showing that the participants' support of the application of a full curfew significantly decreased from T1 to T2 (see the Supplementary Material).

Lastly, we had expected that the introduction of explicit governmental rules would foster the willingness to intervene against others' physical-contact behavior (H1d). However, these rules did not affect participants' willingness to intervene. One explanation for these findings may be that people delegated the responsibility to enforce the new rules to the state, given that legal sanctions can crowd out norm enforcement by individuals (Kube & Traxler, 2011). Indeed, one of our exploratory measures supported this argument by showing that participants at T2 and T3 (vs. T1) were less likely to have intervened against a behavior they had found inappropriate in the last 2-3 days (see the Supplementary Material).

Institutional signals and the differentiation between types of behavior

Our study design allowed us to delineate differences between behavior that was prohibited by governmental rules (i.e., physical-contact) and behavior that was still explicitly allowed (i.e., physical-distancing). As expected, the former was consistently perceived as less prevalent and less socially appropriate, and it evoked less positive attitudes as well as a higher willingness to intervene than the latter at any time point (H2a-d). Crucially, the introduction of governmental rules seemed to be effective in further differentiating the two types of behavior. While the perceived prevalence and social appropriateness was already quite high and personal attitudes were positive for physical-distancing behavior at T1 (and significantly higher than those reported for physical-contact behavior), these ratings increased even further at T2. For perceived prevalence and social appropriateness, this meant that governmental rules yielded a significant increase in the differentiation between the two types of behavior. Together, these results suggest

that, instead of playing an informative role regarding prohibited behavior, the introduction of governmental rules was successful at signaling that mobility with physical distance was still allowed (see also Bundesregierung, 2020).

Institutional signals did not strengthen the relationships between elements

Following our second main research question, we also investigated the relationships among elements of the normative system, with a special focus on unique predictors of the willingness to intervene against physical-contact behavior. In line with our hypothesis, we found that perceived social appropriateness and personal attitudes independently predicted the willingness to intervene, with personal attitudes explaining more variance than perceived social appropriateness.

However, we were especially interested in whether governmental rules affected this predictive relationship. Specifically, we asked whether such rules could strengthen the relationship of perceived social appropriateness and personal attitudes with the willingness to intervene, and thus, act as a potential accelerator of normative change. Our results indicate that the governmental rules did not moderate the relationship between perceived social appropriateness and the willingness to intervene against physical-contact behavior, but we observed an interesting effect on the relevance of personal attitudes. Unexpectedly, instead of strengthening the predictive power of personal attitudes, the introduction of governmental rules weakened it. The data descriptively suggest that this effect may be driven by those participants at T2 who reported more positive personal attitudes and yet, did not reduce their willingness to intervene accordingly (compared to those at T1). Overall, these results indicate that, prior the introduction of governmental rules, people relied more on their attitudes when deciding whether they were willing to intervene or not. However, after the introduction of those rules, people decided less in line with their own attitudes when asked about their willingness to intervene against physical-contact behavior. Thus, the results contradict the notion that the governmental rules helped people to attain clearer attitudes that would then be more relevant for their willingness to intervene.

Contributions, limitations and future research

Our results add to the scarce and inconsistent literature on the effects of institutional signals on systems of social norms. Taken together, they suggest that institutional signals may have differential effects on the elements of normative systems.

First of all, the onset of the COVID-19 pandemic in Germany offered a unique opportunity for testing whether institutional signals, in the form of governmental rules, affect normative systems. At the same time, our results could be specific to the contextual characteristics. In particular, the context allowed people to anticipate the governmental rules so that these rules could have exerted a normative effect prior to their implementation (for different findings, see Casoria et al., 2020). Future research should try to disentangle

such anticipation effects from effects of the actual implementation on the different elements of the normative system.

Strikingly, we saw paradoxical effects of governmental rules on personal attitudes. These findings resonate with the notion that norm interventions aimed at reducing a specific (e.g., selfish) behavior may end up increasing it by unintentionally providing "normalizing" information, which individuals may then use in a self-serving manner (e.g., Bicchieri & Ganegoda, 2017). Thus, our results add to this strain of literature by showing that such effects may not only affect individuals' behavior, but also their attitudes. Future studies could investigate how tailoring institutional signals to the status-quo of the normative system (i.e. the preexisting norm perceptions and personal attitudes), changes their effectiveness and whether this helps to avoid the paradoxical effects we observed.

Moreover, our results indicate that institutional signals, in the form of governmental rules, may help to differentiate between behavior that the rules prohibit and behavior that is still allowed. In this vein, it could be interesting to broaden the idea of normative change, such that it pertains not only to change in social norms for a single behavior, but to change in norms of differentiation between behaviors. Future research could specify and test such a theory advancement.

Crucially, our findings support the results from previous studies showing a greater relevance of personal attitudes compared to perceived social appropriateness for people's willingness to intervene (Deitch-Stackhouse et al., 2015). Interestingly, while the relationship between norm perceptions and norm-compliant behavior is broadly documented (e.g., Miller & Prentice, 2016), even in the context of COVID-19 (e.g., Tunçgenç et al., 2021), less so is their relationship with interventions against others' behavior. Most importantly, we are the first to address how the predictive power of norms and attitudes for the willingness to intervene could change due to institutional signals. Future research should further investigate this predictive relationship and its susceptibility to situational factors. Specifically, it could be interesting to include more direct measures of the hypothesized psychological processes (i.e. lowered uncertainty and increased confidence regarding the prevailing norms and personal attitudes). In our case, we attempted to measure such changes in uncertainty, but we did not observe any changes over time (for a description of these secondary results, see the Supplementary Material). At this point, it is unclear whether this was due to the employed items not being fully suitable for our purpose, or whether the underlying mechanism is different from what we assumed.

Lastly, our employment of a natural manipulation carried the advantage of comparatively high ecological validity. However, it also came with some disadvantages, primarily a reduced experimental control of the unfolding of events. Therefore, the most important limitation of our study could be posed by its timeframe. For a more detailed examination of the unfolding of normative change processes over time, the

study would have benefitted from including more measurements, especially before T1, but also after T3. However, considering the unprecedented speed at which events occurred when this study was conducted, we decided to focus on examining a relatively short time period (i.e., ± two days from the implementation of governmental rules). This decision aimed at reducing potential noise or confounds affecting the effects of interest. Future research could try to track changes over a longer period of time while closely recording any relevant developments. For example, in the time after we conducted this study, compliance with physical-distancing rules seemed to ebbed away (von Haaren, 2020), people increased their mobility again (Schlosser et al., 2020) and Germany has already faced new waves of the pandemic (tagesschau, 2021). More than one year after the start of the pandemic, even ongoing mobility restrictions cannot fully regulate the infection rates (Rieger & Wang, 2020), possibly because many people transgress the rules in private (Robert Koch Institut, 2021). Together, this might indicate that the implemented governmental rules did not deeply influence the normative system, as our study also suggests.

Conclusion

The physical-distancing rules introduced by the German government at the start of the COVID-19 pandemic did not exert the expected monolithic effects on the system of social norms regarding physical-contact behavior. Nonetheless, our findings are informative for broadening the theoretical understanding of the effect of institutional signals on the different elements of normative systems. The introduction of governmental rules changed people's perceived prevalence of the regulated behavior (i.e., descriptive norms), but not its perceived social appropriateness (i.e., injunctive norms). Crucially, the governmental rules also exerted paradoxical effects by relaxing people's personal attitudes. We attributed this effect to a plausible self-serving response in addition to the possibility that the implemented rules were milder than anticipated. Furthermore, the institutional signal unexpectedly weakened the relationship between personal attitudes and the willingness to intervene, which could indicate that the assumed mechanism was not at work. Lastly, governmental rules appeared to be helpful for differentiating between prohibited and allowed types of behavior. To conclude, the present findings illustrate that the dynamics of normative change are rather complex, and the role institutional signals play herein may be less clear than theoretically proposed (e.g., Licht, 2008; Tankard & Paluck, 2016). The mechanisms underlying normative change may have to be defined separately, and future research should work towards a better understanding of these complex patterns.

Supplementary material and open practices

The supplementary material of this chapter is available in the following OSF repository: https://osf.io/59k7d/. This repository further includes the data, analysis code, codebook and research materials to reproduce the reported results.

General Discussion

The violations of social norms often trigger punitive reactions from uninvolved third parties or bystanders, who, despite being unaffected by the norm violation, directly address the perpetrator while assuming personal costs. The personal costs of this type of punishment behavior can go from the investment of time and resources in addressing the perpetrator, to further physical, social or economic risks that third parties may suffer (e.g., a violent counter reaction by the perpetrator, social ostracism, getting fired). Since this often-termed *costly third-party punishment* (3PP) does not entail personal benefits but immediate costs for third parties, some researchers have focused on studying its underlying individual motivations (Delton & Krasnow, 2017; Gummerum et al., 2016; Jordan et al., 2016; Tan & Xiao, 2018). Taking a broader perspective, at the group or the society level, others have investigated and theorized about 3PP as a fundamental mechanism in the reinforcement and maintenance of social norms (Chen et al., 2020; Egas & Riedl, 2008; Fehr & Fischbacher, 2004).

However, how prevalent is 3PP as a personally costly yet societally beneficial behavior? Almost two decades ago, the first steps of experimental research on 3PP offered an answer to this straightforward question. Under controlled and standardized conditions in the lab, roughly 60% of third parties exerted financially costly 3PP (Fehr & Fischbacher, 2004). These controlled and standardized conditions corresponded to the *third-party punishment game* (3PPG), an economic game introduced by Fehr and Fischbacher (2004) and used in many lab studies on 3PP (e.g., Balafoutas et al., 2014; Lotz et al., 2011; Marlowe et al., 2008). In the 3PPG, participants playing as third parties can impose costs on others who made unfair monetary distributions between themselves and a second party. The rates of 3PP in this experimental setup has been replicated in cross-cultural samples, despite some observed differences between cultures (Henrich et al., 2006). Yet, when researchers have turned to assess 3PP as a reaction to daily norm violations outside the lab, such as littering in public spaces, the frequency of these third-party punitive reactions is drastically lower (Balafoutas et al., 2016; Balafoutas & Nikiforakis, 2012; Winter & Zhang, 2018).

As I argued in the introduction of this dissertation, this discrepancy between lab and field studies in the observed levels of 3PP brings about a series of conceptual and methodological questions regarding how empirical research has approached this behavior. The first question refers to the situational boundary conditions that may hinder 3PP in the field, but which lab experiments may have overlooked. Considering these boundary conditions in experimental setups should reduce the differences between lab and field findings, and thus, reduce the chance of overestimating the prevalence of 3PP based on what is observed in the lab. The second, related question concerns the external validity, or generalizability, of lab findings on 3PP. If lab studies indeed overlooked critical situational boundaries of 3PP, it is unlikely that the levels of 3PP we observe in the lab resemble what occurs in the field. At the individual level, this implies that the way in which people behave in lab experiments differs from how they behave in the field, given the different situational and psychological factors involved (Levitt & List, 2007). Hence, one may argue that the consideration of these

boundary conditions within lab experiments increases the external validity of the latter. And finally, researchers have made important theoretical claims about the role of 3PP in the maintenance of social norms (Fehr & Fischbacher, 2004). Given that these claims were mainly based on evidence from lab studies, does the discrepancy with field evidence invite to revisit these claims or, to the contrary, can we argue that 3PP is indeed a relevant element for the system of social norms?

In the following discussion, I will delve into how the research I conducted during my PhD contributed to answering each of these questions, which aspects remain unclear or unanswered, and which implications this research agenda has for future research on 3PP.

Searching for situational boundaries of 3PP

As a response to the raised concerns about the discrepancies in punishment behavior between the lab and the field (Guala, 2012), researchers urged about the need to assess this type of behavior under conditions of informational noise and uncertainty (see Bereby-Meyer, 2012; Van Lange et al., 2012). Recently published work has made an explicit call for addressing this gap in the literature (Wu et al., 2021), which stresses the importance of this potential situational factor for 3PP and reminds that the gap is still unfilled.

As with other types of behavior, the decision to engage into 3PP of norm violations in daily life situations is often based on incoherent, noisy or incomplete information that individuals receive from their social environment. This imperfect information, which in this dissertation I termed *situational ambiguity*, may hamper the interpretation of the norm violation as such. Given that the interpretation of the perpetrator's behavior as a norm violation is a necessary requirement for the third parties' decision-making (Baumert et al., 2013; Osswald et al., 2010), I argued that situational ambiguity is critical for understanding when 3PP occurs.

In the introduction of this dissertation, I used the example of a young man who seemed not to wear a mask properly in a public space during a global pandemic. His physical position or the possibility of him suffering from a clinical condition that justified why he did not wear a mask could be sufficient to hamper the third parties' interpretation of his behavior as a norm violation. In both cases, the ambiguity about whether his behavior represented a norm violation or not could lead third parties to experience concerns about the appropriateness of 3PP (for similar arguments, see Wu et al., 2021). I argued that third parties may refrain from punishing (in the case of the example, directly reprimanding the young man) when considering the possibility of punishing unfairly or unjustifiably; that is to commit a false positive or *type I error* in the identification of the norm violation (Grechenig et al., 2010). Specifically, third parties may avoid committing these type I errors due to the anticipation of individual regret and moral concerns, or reputational costs as previous research suggests (de Kwaadsteniet et al., 2019).

Literature in willful ignorance and moral wiggle room (Dana et al., 2007; Haisley & Weber, 2010) offers a complementary explanation regarding what occurs to 3PP under ambiguity of the norm violation.

People who aim to avoid incurring the costs of 3PP may exploit this ambiguity to justify their passiveness and thus elude any associated reputational risk (Kriss et al., 2016; Stüber, 2020). Thus, the hindering effect of ambiguity on 3PP could be partly explained by a second concern about avoiding costs.

In Chapter 1, I reported findings from a set of six lab studies using the experimental setting of the 3PPG, which demonstrated that the ambiguity of the norm violation is indeed a critical boundary condition of 3PP. The studies consistently showed that third parties reduced their punishment behavior when facing an ambiguous norm violation. With regard to the underlying concerns driving this effect (i.e., type I errors and cost avoidance), I followed two evaluative approaches.

The first approach was to examine the role of interindividual differences in Justice Sensitivity (JS; Baumert & Schmitt, 2016), a dispositional variable capturing people's readiness to perceive and react against injustice and assume costs in order to restore justice. I aimed to test whether third parties' JS from the perspective of a perpetrator who inflicts injustice on others (i.e., Perpetrator JS) and the perspective of an uninvolved observer (i.e., Observer JS) respectively captured the concerns about type I errors and cost avoidance. In four out of six studies, I found Observer JS, but not Perpetrator JS, to moderate the effect of ambiguity of the norm violation on 3PP. More concretely, third parties with high Observer JS reduced more their 3PP under ambiguity than those with low Observer JS. If one considers that people with high Observer JS are less prone to act less selfishly (Edele et al., 2013; Lotz, Baumert, et al., 2011) even when the situation provides an excuse for it (Lotz et al., 2013), the more plausible explanation is that, under ambiguity, they experienced type I error concerns. However, the mixed results across studies and other alternative, yet less plausible explanations for this interaction effect invite to interpret these results cautiously. For example, under ambiguity, people with high Observer JS could have struggled more in the identification of the norm violation. This is possible, yet unlikely, given that they should be more prone to perceive ambiguous situations of unfairness, as previous research suggests (Baumert & Schmitt, 2009).

A second complementary approach was to examine whether third parties resolve the ambiguity of the norm violation when they have the opportunity to do so and how this determines their subsequent punishment decision. For example, one may approach the young man in the bus to verify whether he is indeed wearing a mask properly. In two studies, some third parties had this option (and thus, the opportunity to learn whether their punishment would be unfair or unjustified). Within this group, those who resolved the ambiguity of the norm violation (even when this entailed an additional cost) reported to be more concerned about punishing unfairly and less concerned about avoiding costs than those who willfully kept the situation ambiguous. The subsequent 3PP of those who resolved the ambiguity was similar or even higher than the 3PP I observed under no ambiguity. These findings support that ambiguity was indeed a critical obstacle hindering the punishment decision of these third parties, who otherwise would be more inclined to address a norm violation.

Still, I found some indications that cost avoidance was a second motivation under ambiguity. First, people with low Observer JS also reduced their 3PP under ambiguity, although this result did not replicate across all studies. Second, those who chose to remain ignorant when given the opportunity to resolve the ambiguity expressed higher concerns about avoiding costs than those who resolved the ambiguity. Perhaps, this is not sufficiently strong evidence to make the claim that ambiguity was exploited with the intention of avoiding costs, but it certainly suggests that for some third parties costs avoidance could be an important consideration when facing an ambiguous norm violation.

Taken together, this dissertation provided a **first answer** regarding the situational boundaries of 3PP, showing that the ambiguity of the norm violation reduces punishment. Moreover, Chapter 1 offered evidence about one plausible psychological mechanism underlying the effect of ambiguity, namely the concerns of third parties about punishing unfairly.

In the field, the situational ambiguity that characterizes situations of norm violations also relates to the information that third parties have about the costs associated to 3PP (Van Lange et al., 2012). While in the lab, third parties know the exact costs that they incur when punishing the perpetrator of a norm violation, in the field, these costs might be unknown or more difficult to anticipate. Thus, a second situational boundary that could explain the discrepancies between findings from lab and field experiments on 3PP is this cost uncertainty.

Based on the assumption that third parties would overestimate the likelihood of high costs, as people do with other extreme life events (Barberis, 2013; Lichtenstein et al., 1978; Rozin & Royzman, 2001; Tversky & Kahneman, 1992), I tested within the setting of the 3PPG whether cost uncertainty decreased 3PP. In the three lab studies reported in Chapter 2, I did not observe this to be the case. Yet, the manipulation I used for investigating this effect was perhaps suboptimal based on the proposed theoretical assumption, which could explain why I obtained inconclusive results. In particular, the uncertain costs I introduced in the 3PPG were perhaps not sufficiently high to resemble the extreme costs, the probability of which third parties may overestimate in real-life situations of norm violations. Thus, it is possible that under low uncertain costs, third parties disregard the level of cost uncertainty (Bombardini & Trebbi, 2012), and therefore, their decision to engage into 3PP remains unaffected.

Still, researchers should not disregard the potential role of cost uncertainty as a boundary condition of 3PP. There is supportive empirical evidence to assume that the lack of information about the involved costs may discourage third parties from addressing norm violations in the field. First, punishment behavior decreases when its costs increase relatively to its impact on the perpetrator (Egas & Riedl, 2008). Second, people tend to attribute more subjective probability to extreme costs, although these are objectively unlikely

(Barberis, 2013; Rozin & Royzman, 2001). Thus, when third parties are uncertain about the costs of 3PP, they may refrain from punishing to avoid the (subjectively high) probability of incurring an extremely high cost.

In short, it remains **unanswered** whether cost uncertainty should also be considered a situational boundary condition of 3PP. Chapter 2 did not provide evidence for an effect of costs uncertainty on 3PP. Rather than discarding this hypothesis, I argue that the lack of support may be explained by methodological shortcomings, given that there are theoretical reasons to assume that cost uncertainty should affect 3PP in the field.

On the issue of the external validity of research on 3PP

As I introduced earlier, the discrepancies between 3PP in the lab and in the field may highlight a problem of external validity or generalizability of findings from lab studies using the 3PPG. However, the consideration of situational boundaries that hinder 3PP within the 3PPG should provide an opportunity to resemble the situational and psychological conditions under which third parties decide to punish in the field, and in turn, to increase the external validity of the 3PPG (Levitt & List, 2007).

Based on my investigation of the effects of ambiguity of the norm violation and cost uncertainty on 3PP, in Chapter 3, I addressed whether the introduction of these two potential situational boundaries in the 3PPG affected its external validity. Specifically, I tested whether people's 3PP under ambiguity and cost uncertainty in the 3PPG better predicted their intervention (or intervention intention) against a norm violation different from the violation of fairness commonly used in the 3PPG (i.e., an embezzlement of lab funds). I did not find people's 3PP to predict their intervention behavior or their intention to intervene against the embezzlement, irrespective of whether the 3PPG included ambiguity of the norm violation or cost uncertainty.

These results suggest that the external validity of the 3PPG is limited, especially if we theoretically assume that 3PP is a fundamental mechanism for the maintenance of social norms and human cooperation (e.g., Fehr & Fischbacher, 2004). If this was the case and 3PP was a widespread reinforcing element of social norms, the same behavioral disposition that third parties show to punish the unfair distribution of money in the 3PPG should generalize to the violation of other social norms outside the lab, such as an embezzlement of lab funds. To the contrary, I observed that people's punishment behavior in the 3PPG does not resemble their (intended) reactions to a different norm violation in a field-like situation, even when considering similar situational factors (e.g., ambiguity of the norm violation) that arguably affect punishment behavior in these two different contexts.

Thus, the findings from Chapter 3 invite to consider a much narrower interpretation of the findings from lab studies using the 3PPG. A narrower interpretation entails to acknowledge the situational specificity

of the 3PPG, and therefore, that the behavior observed in it would only generalize to situationally equivalent situations. The situational specificity of the 3PPG may refer to the specific norm violation (i.e., unfair distribution of money) and three-people structure (i.e., perpetrator, victim, third party) that it commonly presents, as well as the unique form of behavioral reaction that third parties can generally exert in this experimental paradigm (i.e., direct punishment). These and other situational characteristics of the 3PPG drastically differ from many real-life situations in which other social norm are violated (e.g., physical harm), in the number of people involved (e.g., absence of victim), or the multiple other ways in which third parties can react (e.g., withholding future help; Balafoutas, Nikiforakis, et al., 2014). If the 3PP people show in the 3PPG does not share any common interindividual variance with how they behave in other real-life situations, as I reported in Chapter 3, a straightforward conclusion is that the lack of relationship is due to the situational moderators that differentially configure people's behavior in each context (i.e., the 3PPG and the field).

However, this conclusion does not imply that studying 3PP through lab experiments is totally uninformative (Guala, 2012). In fact, lab studies might be the only means through which researchers can empirically study the effect of certain situational factors (e.g., Lewisch et al., 2015). Yet, researchers should be aware that when they assess 3PP through the 3PPG, they might be measuring a very specific type of behavior that hardly resembles how people behave in the field and that, therefore, they might be overestimating its prevalence and its superordinate role within the system of social norms. This rather pessimistic prospect should improve over time, when further research provide a broader picture of the situational boundaries shaping 3PP in both the lab and the field, and when efforts of aggregation through critical literature reviews (e.g., Guala, 2012), multi-method approaches (e.g., Balliet et al., 2022) and meta-analyses take place.

Chapter 3 offers a **second answer**, in this case, to the question of the external validity of lab research on 3PP. If we theoretically assume that 3PP functions as a fundamental mechanism for the reinforcement of social norms, research using the 3PPG has limited external validity (even when considering situational boundary conditions, such as ambiguity of the norm violation). It may thus be necessary to acknowledge the situational specificities of the 3PPG in a narrow sense and to focus future research efforts in the identification of situational moderators and the aggregation of findings across situations, in the lab and in the field, to distil commonalities.

3PP and the system of social norms

If we turn to assess the theoretical assumption that 3PP is a fundamental reinforcing element of social norms (Fehr & Fischbacher, 2004), an ideal scenario is a context of normative change. When groups or societies modify their social norms, the external contingencies that define these norms (Legros & Cislaghi, 2020; Young, 2015) should congruently change to guarantee the success in the transition to the new

normative framework. Thus, a change in people's inclination to punish violations of the new norms would be indicative of the reinforcing role that 3PP may play within the system of social norms.

In this dissertation, I used the context of the outbreak of the Covid-19 pandemic to analyze this notion. In Chapter 4, I aimed to examine the effects of the governmental rules of physical distancing implemented in Germany on the perception of social norms and, importantly for the present matter, on people's willingness to intervene against the violation of these new norms (i.e., against behavior entailing the risk of physical contact). As an institutional signal, these governmental rules could arguably trigger a normative change (Tankard & Paluck, 2016, 2017), which, according to the previous rationale, should be accompanied by a change in people's willingness to intervene.

In a natural quasi-experimental study, I assessed people's perceptions of social norms and their willingness to intervene against counter-normative behavior before and after the introduction of the governmental rules. The study did not show that the targeted institutional signal changed the perceptions of social norms, probably because the process of normative change had already begun prior to our data collection with the anticipation of this institutional signal or with other past signals (e.g., from international health authorities, media coverage, other international governments, etc.). Hence, it is not surprising that people's willingness to intervene did not change either in the assessed time window. Yet, what the study did show was that that people's personal attitudes – i.e., what people individually think is (in-)appropriate – and their perceptions of the injunctive norms – i.e., what people think others think is (in-)appropriate –, but *not* their perceptions of the descriptive norms – i.e., whether people think others behave (in-)appropriately –, predicted their willingness to intervene. More concretely, the more inappropriate a behavior was personally and socially perceived, the more willing people were to intervene against it. Due to the correlational nature of the study, it is not possible to establish a causal relationship. However, the findings already indicate that people's 3PP (or, at least, their willingness to engage in it) is related to and congruent with their perception of the social norms, and therefore, that 3PP is an important element within the system of social norms.

What remains unclear is whether 3PP contributes to the reinforcement of people's perceptions of social norms in the field. This could have been inferred from our study if the government rules had exerted a strong influence on people's perceptions of social norms. In this scenario, the observation that people's willingness to intervene had accompanied this normative change could have supported the notion that 3PP functioned as a reinforcement of the new norms. Recent work already offers evidence from lab studies suggesting that 3PP, in comparison to second-party punishment, exerts a stronger influence on other people's perception of injunctive and descriptive norms (Chen et al., 2020). Yet, as I remarked before with regard to other lab findings, the literature still requires further evidence from the field and across a wider range of norm violation situations to ascertain that 3PP indeed plays this reinforcing role within the system of social norms.

In summary, Chapter 4 provides evidence for the connection between 3PP and the perception of social norms in a natural setting of presumed normative change. However, it remains **unanswered** whether 3PP functions as a reinforcing mechanism of social norms.

Concluding remarks and future directions

It is difficult to deny the existence and role of mechanisms of informal social punishment – including 3PP – in regulating social behavior in our daily lives. However, it is unclear how widespread these are and how much weight they have in the maintenance of the system of social norms. Researchers from multiple fields have accomplished many advancements in the last two decades to address these two questions (e.g., Balafoutas, Nikiforakis, et al., 2014; Bond, 2019; Egas & Riedl, 2008; Fehr & Fischbacher, 2004; Fowler, 2005). The present dissertation contributes to this progress by showing that levels of 3PP decrease when there is situational ambiguity affecting the interpretation of the norm violation (Chapter 1). Furthermore, in a natural setting, research in this dissertation indicates that people's intentions to exert 3PP are directly linked to their personal and social perception of the respective social norm (Chapter 4), which highlights the theorized relationship of 3PP with the system of social norms.

Still, the heavy reliance on lab experiments compromises the scope and advancement of research on 3PP. The discrepancies between findings from lab and field studies cannot be ignored and should invite us to take a step back to revisit conceptually and methodologically what we measure when using experimental settings like the 3PPG. It is unquestionable that the 3PPG and other standardized games provide the rigor and experimental control that field data usually lacks, but every conclusion derived from these experimental settings should be further examined outside the lab. This would not only allow to have better estimations of the prevalence of 3PP – and punishment behavior more generally –, but also to be analytically sensitive to potential situational moderators of this type of behavior that may have been neglected in lab experiments (with the investigated ambiguity of the norm violation as an example). In this direction, the present dissertation provides an assessment of the external validity of the 3PPG (Chapter 3). This assessment suggests that the 3PPG offers limited generalizability to situations that differ in their situational structure and the norm violation involved (e.g., an embezzlement). Other research has indicated that 3PP may occur in multiple and very different ways (i.e., direct vs. indirect types of punishment), each of them potentially affected by different situational factors and potentially fulfilling different roles within the system of social norms (Balliet et al., 2022; Molho et al., 2020).

From here, research in 3PP should invest efforts in two different fronts. The first front is conceptual and relates to theory specification. The identification of the situational determinants of 3PP (and its different types) should lead to a more complex explanatory model of punishment behavior and its relationship with the

system of social norms. This should improve the detail of the fundamental theoretical propositions and auxiliary assumptions about under which circumstances punishment behavior may be expected. As for the second front, this is methodological and refers to the aforementioned effort of aggregation. Researchers should assess 3PP across different experimental paradigms, under different situational moderators and, outside the lab, across a wider ranges of daily-life contexts, like some methodologies already offer (e.g., ambulatory assessment). Moreover, systematic and meta-analytic reviews of the literature may also benefit this goal of aggregation. As a result, the field will gain in better estimations of the prevalence of 3PP and its actual role regarding the reinforcement of social norms.

References

Aerzteblatt. (2020, February 14). *Sars-CoV-2 könnte einer schweren Grippewelle gleichkommen*. https://www.aerzteblatt.de/nachrichten/109390/Sars-CoV-2-koennte-einer-schweren-Grippewelle-gleichkommen

Almenberg, J., Dreber, A., Apicella, C., & Rand, D. G. (2010). *Third party reward and punishment: Group size, efficiency and public goods*. SSRN. https://papers.ssrn.com/abstract=1715305

Anderson, C. M., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, *54*(1), 1–24. https://doi.org/10.1016/j.geb.2004.08.007

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

Andrighetto, G., Villatoro, D., & Conte, R. (2010). Norm internalization in artificial societies. *AI Communications*, *23*(4), 325–339. https://doi.org/10.3233/AIC-2010-0477

ARD-aktuell. (2020, June 22). *Corona-Ausbruch bei Tönnies 1331 Infizierte—Vorerst kein Lockdown*. https://www.tagesschau.de/inland/toennies-coronainfektionen-guetersloh-105.html

Arias, E. (2014). *Media, common knowledge, and violence against women: A field experiment on norms change in Mexico*. Paper presented at the annual meeting of the American Political Science Association. Washington, DC. https://ucema.edu.ar/conferencias/download/2013/06.25AE.pdf

Armbruster, S., & Klotzbücher, V. (2020). *Lost in lockdown? COVID-19, social distancing, and mental health in Germany*. EconStor. http://hdl.handle.net/10419/218885

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters*, *122*(2), 308–310. https://doi.org/10.1016/j.econlet.2013.11.028

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: A natural field experiment. *European Economic Review*, *56*(8), 1773–1785. https://doi.org/10.1016/j.euroecorev.2012.09.008

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

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*(1), 13327. https://doi.org/10.1038/ncomms13327

Balliet, D., Molho, C., Columbus, S., & Dores Cruz, T. D. (2022). Prosocial and punishment behaviors in everyday life. *Current Opinion in Psychology*, *43*, 278–283. https://doi.org/10.1016/j.copsyc.2021.08.015

Barberis, N. (2013). The psychology of tail events: Progress and challenges. *American Economic Review*, *103*(3), 611–616. https://doi.org/10.1257/aer.103.3.611

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

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, *9*(2), 171–192. https://doi.org/10.1007/s11299-010-0079-9

Baumert, A., Beierlein, C., Schmitt, M., Kemper, C. J., Kovaleva, A., Liebig, S., & Rammstedt, B. (2014). Measuring four perspectives of justice sensitivity with two items each. *Journal of Personality Assessment*, *96*(3), 380–390. https://doi.org/10.1080/00223891.2013.836526

Baumert, A., Gollwitzer, M., Staubach, M., & Schmitt, M. (2011). Justice sensitivity and the processing of justice-related information. *European Journal of Personality*, *25*(5), 386–397. https://doi.org/10.1002/per.800

Baumert, A., Halmburger, A., & Schmitt, M. (2013). Interventions against norm violations: Dispositional determinants of self-reported and real moral courage. *Personality and Social Psychology Bulletin*, *39*(8), 1053–1068. https://doi.org/10.1177/0146167213490032

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*, *30*(3), 178–192. https://doi.org/10.1027/1015-5759/a000183

Baumert, A., & Schmitt, M. (2009). Justice-sensitive interpretations of ambiguous situations. *Australian Journal of Psychology*, *61*(1), 6–12. https://doi.org/10.1080/00049530802607597

Baumert, A., & Schmitt, M. (2016). Justice sensitivity. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (pp. 161–180). Springer New York. https://doi.org/10.1007/978-1-4939-3216-0_9

Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field?—Evidence from donations. *Experimental Economics*, *11*(3), 268–281. https://doi.org/10.1007/s10683-007-9192-y

Bereby-Meyer, Y. (2012). Reciprocity and uncertainty. *Behavioral and Brain Sciences*, *35*(1), 18–19. https://doi.org/10.1017/S0140525X11001178

Berger, J., & Hevenstone, D. (2016). Norm enforcement in the city revisited: An international field experiment of altruistic punishment, norm maintenance, and broken windows. *Rationality and Society*, *28*(3), 299–319. https://doi.org/10.1177/1043463116634035

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511616037

Bicchieri, C., & Ganegoda, D. (2017). Determinants of corruption: A sociopsychological analysis. In P. Nichols & D. C. Robertson (Eds.), *Thinking about bribery: Neuroscience, moral cognition and the psychology of bribery* (pp. 179–205). Cambridge University Press. https://doi.org/10.1017/9781316450765.008

Blanton, H., Stuart, A. E., & Van den Eijnden, R. J. J. M. (2001). An introduction to deviance-regulation theory: The effect of behavioral norms on message framing. *Personality and Social Psychology Bulletin*, *27*(7), 848–858. https://doi.org/10.1177/0146167201277007

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). *Pandemic populism: Facebook pages of alternative news media and the corona crisis – A computational content analysis.* ArXiv. http://arxiv.org/abs/2004.02566

Bohner, G., & Dickel, N. (2011). Attitudes and attitude change. *Annual Review of Psychology*, *62*(1), 391–417. https://doi.org/10.1146/annurev.psych.121208.131609

Bombardini, M., & Trebbi, F. (2012). Risk aversion and expected utility theory: An experiment with large and small stakes. *Journal of the European Economic Association*, *10*(6), 1348–1399. https://doi.org/10.1111/j.1542-4774.2012.01086.x

Bond, R. M. (2019). Low-cost, high-impact altruistic punishment promotes cooperation cascades in human social networks. *Scientific Reports*, *9*(1), 2061. https://doi.org/10.1038/s41598-018-38323-7

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535. https://doi.org/10.1073/pnas.0630443100

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: A first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375–398. https://doi.org/10.1007/s10683-011-9272-x

Brauer, M., & Chekroun, P. (2005). The relationship between perceived violation of social norms and social control: Situational factors influencing the reaction to deviance. *Journal of Applied Social Psychology*, *35*(7), 1519–1539. https://doi.org/10.1111/j.1559-1816.2005.tb02182.x

Bundesregierung. (2020, March 22). *Besprechung der Bundeskanzlerin mit den Regierungschefinnen und Regierungschefs der Länder. Die Bundeskanzlerin und die Regierungschefinnen und Regierungschefs der Länder fassen am 22. März 2020 folgenden Beschluss.* https://www.bundesregierung.de/breg-de/themen/coronavirus/besprechung-der-bundeskanzlerin-mit-den-regierungschefinnen-und-regierungschefs-der-laender-1733248

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Carbonara, E., Parisi, F., & von Wangenheim, G. (2008). Lawmakers as norm entrepreneurs. *Review of Law & Economics*, *4*(3). https://doi.org/10.2202/1555-5879.1320

Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior & Organization*, *62*(4), 522–542. https://doi.org/10.1016/j.jebo.2005.05.004

Casoria, F., Galeotti, F., & Villeval, M. C. (2020). *Perceived social norm and behavior quickly adjusted to legal changes during the Covid-19 pandemic in France*. SSRN. https://doi.org/10.2139/ssrn.3670895

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PLoS ONE*, *15*(3), e0229510. https://doi.org/10.1371/journal.pone.0229510

Chu, H., Yang, J. Z., & Liu, S. (2021). Not my pandemic: Solution aversion and the polarized public perception of Covid-19. *Science Communication*, *43*(4), 508–528. https://doi.org/10.1177/10755470211022020

Chung, A., & Rimal, R. N. (2016). Social norms: A review. *Review of Communication Research*, *4*(1), 1–28. https://doi.org/10.12840/issn.2255-4165.2016.04.01.008

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology, Vols. 1-2, 4th ed* (pp. 151–192). McGraw-Hill.

Coibion, O., Gorodnichenko, Y., & Weber, M. (2020). *Labor markets during the COVID-19 crisis: A preliminary view* (NBER Working Paper No. w27017). National Bureau of Economic Research. https://doi.org/10.3386/w27017

Connor, R. (2020, March 20). German states move closer to near-total lockdowns. *Deutsche Welle*. https://p.dw.com/p/3ZoCo

Crandall, C. S., Miller, J. M., & White, M. H. (2018). Changing norms following the 2016 U.S. presidential election: The Trump effect on prejudice. *Social Psychological and Personality Science*, *9*(2), 186–192. https://doi.org/10.1177/1948550617750735

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, *10*(2), 117–133. https://doi.org/10.1111/phc3.12192

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, *84*, 103800. https://doi.org/10.1016/j.jesp.2019.03.011

Deitch-Stackhouse, J., Kenneavy, K., Thayer, R., Berkowitz, A., & Mascari, J. (2015). The influence of social norms on advancement through bystander stages for preventing interpersonal violence. *Violence Against Women*, *21*(10), 1284–1307. https://doi.org/10.1177/1077801215592720

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

Dimitroff, S. J., Harrod, E. G., Smith, K. E., Faig, K. E., Decety, J., & Norman, G. J. (2020). Third-party punishment following observed social rejection. *Emotion*, *20*(4), 713–720. https://doi.org/10.1037/emo0000607

Edele, A., Dziobek, I., & Keller, M. (2013). Explaining altruistic sharing in the dictator game: The role of affective empathy, cognitive empathy, and justice sensitivity. *Learning and Individual Differences*, *24*, 96–102. https://doi.org/10.1016/j.lindif.2012.12.020

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878. https://doi.org/10.1098/rspb.2007.1558

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99–117. https://doi.org/10.1257/jep.3.4.99

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360–392. https://doi.org/10.1111/j.1467-6494.1983.tb00338.x

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., Aldashev, A., Andersson, P. A., Andrighetto, G., Anum, A., Arikan, G., Aycan, Z., Bagherian, F., Barrera, D., Basnight-Brown, D., Batkeyev, B., Belaus, A., Berezina, E., Björnstjerna, M., … Van Lange, P. A. M. (2021). Perceptions

of the appropriate response to norm violation in 57 societies. *Nature Communications*, *12*(1), 1481. https://doi.org/10.1038/s41467-021-21602-9

Erkut, H., Nosenzo, D., & Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, *130*, 28–31. https://doi.org/10.1016/j.econlet.2015.02.021

Fazio, R. H., Ruisch, B. C., Moore, C. A., Granados Samayoa, J. A., Boggs, S. T., & Ladanyi, J. T. (2021). Social distancing decreases an individual's likelihood of contracting COVID-19. *Proceedings of the National Academy of Sciences*, *118*(8), e2023131118. https://doi.org/10.1073/pnas.2023131118

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., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications*, *5*, 5306. https://doi.org/10.1038/ncomms6306

Fetchenhauer, D., & Huang, X. (2004). Justice sensitivity and distributive decisions in experimental games. *Personality and Individual Differences*, *36*(5), 1015–1029. https://doi.org/10.1016/S0191-8869(03)00197-1

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, *102*(19), 7047–7049. https://doi.org/10.1073/pnas.0500938102

Galizzi, M. M., & Navarro-Martinez, D. (2018). On the external validity of social preference games: A systematic lab-field study. *Management Science*, *65*(3), 976–1002. https://doi.org/10.1287/mnsc.2017.2908

Ginther, M. R., Hartsough, L. E. S., & Marois, R. (2021). Moral outrage drives the interaction of harm and culpable intent in third-party punishment decisions. *Emotion*. https://doi.org/10.1037/emo0000950

Gollwitzer, P. M., Sheeran, P., Michalski, V., & Seifert, A. E. (2009). When intentions go public: Does social reality widen the intention-behavior gap? *Psychological Science*, *20*(5), 612–618. https://doi.org/10.1111/j.1467-9280.2009.02336.x

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Grechenig, K., Nicklisch, A., & Thöni, C. (2010). Punishment despite reasonable doubt: A public goods experiment with sanctions under uncertainty. *Journal of Empirical Legal Studies*, *7*(4), 847–867. https://doi.org/10.1111/j.1740-1461.2010.01197.x

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–15. https://doi.org/10.1017/S0140525X11000069

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (2016). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology*, *65*, 94–104. https://doi.org/10.1016/j.jesp.2016.04.004

Haisley, E. C., & Weber, R. A. (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior*, *68*(2), 614–625. https://doi.org/10.1016/j.geb.2009.08.002

Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific Reports*, *9*(1), 13219. https://doi.org/10.1038/s41598-019-49680-2

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Hogg, M. A., & Reid, S. A. (2006). Social identity, self-categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30. https://doi.org/10.1111/j.1468-2885.2006.00003.x

Janssen, M. A., & Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. *Journal of Theoretical Biology*, *254*(3), 541–545. https://doi.org/10.1016/j.jtbi.2008.06.017

Jones, L., Palumbo, D., & Brown, D. (2020, June 30). Coronavirus: A visual guide to the economic impact. *BBC News*. https://www.bbc.com/news/business-51706225

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/10.1038/nature16981

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences*, *111*(35), 12710–12715. https://doi.org/10.1073/pnas.1402280111

Jordan, J., & Kteily, N. (2020). *Punitive but discriminating: Reputation fuels ambiguously-deserved punishment but also sensitivity to moral nuance*. PsyArXiv. https://doi.org/10.31234/osf.io/97nhj

Kahan, D. M. (2000). Gentle nudges vs. hard shoves: Solving the sticky norms problem. *The University of Chicago Law Review*, *67*(3), 607. https://doi.org/10.2307/1600336

Karmali, F., Kawakami, K., & Page-Gould, E. (2017). He said what? Physiological and cognitive responses to imagining and witnessing outgroup racism. *Journal of Experimental Psychology: General*, *146*(8), 1073–1085. https://doi.org/10.1037/xge0000304

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting Affective and behavioral responses to racism. *Science*, *323*(5911), 276–278. https://doi.org/10.1126/science.1164951

Kotsadam, A., & Jakobsson, N. (2011). Do laws affect attitudes? An assessment of the Norwegian prostitution law using longitudinal data. *International Review of Law and Economics*, *31*(2), 103–115. https://doi.org/10.1016/j.irle.2011.03.001

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

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior & Organization*, *128*, 159–177. https://doi.org/10.1016/j.jebo.2016.05.017

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524. https://doi.org/10.1111/jeea.12006

Kube, S., & Traxler, C. (2011). The interaction of legal and social norm enforcement. *Journal of Public Economic Theory*, *13*(5), 639–660. https://doi.org/10.1111/j.1467-9779.2011.01515.x

Kurzban, R., Descioli, P., & Obrien, 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

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science*, *15*(1), 62–80. https://doi.org/10.1177/1745691619866455

Leibbrandt, A., & López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization*, *84*(3), 753–766. https://doi.org/10.1016/j.jebo.2012.09.018

Levitt, S. D., & List, J. A. (2007). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économique*, *40*(2), 347–370. https://doi.org/10.1111/j.1365-2966.2007.00412.x

Lewisch, P., Ottone, S., & Ponzano, F. (2015). Third-party punishment under judicial review: An economic experiment on the effects of a two-tier punishment system. *Review of Law & Economics*, *11*(2). https://doi.org/10.1515/rle-2015-0018

Licht, A. N. (2008). Social norms and the law: Why peoples obey the law. *Review of Law & Economics*, *4*(3), 715–750. https://doi.org/10.2202/1555-5879.1232

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 551–578. https://doi.org/10.1037/0278-7393.4.6.551

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32. https://doi.org/10.1037/rev0000074

López, L., & Rodó, X. (2020). The end of social confinement and COVID-19 re-emergence risk. *Nature Human Behaviour*, *4*(7), 746–755. https://doi.org/10.1038/s41562-020-0908-8

Lotz, S., Baumert, A., Schlösser, T., Gresser, F., & Fetchenhauer, D. (2011). Individual differences in third-party interventions: How justice sensitivity shapes altruistic punishment. *Negotiation and Conflict Management Research*, *4*(4), 297–313. https://doi.org/10.1111/j.1750-4716.2011.00084.x

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology*, *47*(2), 477–480. https://doi.org/10.1016/j.jesp.2010.10.004

Lotz, S., Schlösser, T., Cain, D. M., & Fetchenhauer, D. (2013). The (in)stability of social preferences: Using justice sensitivity to predict when altruism collapses. *Journal of Economic Behavior & Organization*, *93*, 141–148. https://doi.org/10.1016/j.jebo.2013.07.012

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1634), 587–592. https://doi.org/10.1098/rspb.2007.1517

McAdams, R. H. (2000). An attitudinal theory of expressive law. *Oregon Law Review*, *79*(2), 339–390.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10. https://doi.org/10.1016/j.cognition.2014.08.013

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Merguei, N., Strobel, M., & Vostroknutov, A. (2021). *Moral opportunism and excess in punishment decisions* [Unpublished Manuscript]. Maastricht University. http://www.vostroknutov.com/pdfs/MSV-00.pdf

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, *67*(1), 339–361. https://doi.org/10.1146/annurev-psych-010814-015013

Molho, C., Tybur, J. M., Van Lange, P. A. M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, *11*(1), 3432. https://doi.org/10.1038/s41467-020-17286-2

Moosa, I. A. (2020). The effectiveness of social distancing in containing Covid-19. *Applied Economics*, *52*(58), 6292–6305. https://doi.org/10.1080/00036846.2020.1789061

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781. https://doi.org/10.2139/ssrn.1804189

Murtin, F., Fleischer, L., Siegerink, V., Aassve, A., Algan, Y., Boarini, R., González, S., Lonti, Z., Grimalda, G., Vallve, R. H., Kim, S., Lee, D., Putterman, L., & Smith, C. (2018). *Trust and its determinants: Evidence from the Trustlab experiment* (OECD Statistics Working Papers No. 2018/02). OECD. https://doi.org/10.1787/18152031

Niesta Kayser, D., Greitemeyer, T., Fischer, P., & Frey, D. (2010). Why mood affects help giving, but not moral courage: Comparing two types of prosocial behaviour. *European Journal of Social Psychology*, *40*(7), 1136–1157. https://doi.org/10.1002/ejsp.717

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, *31*(4), 259–264. https://doi.org/10.1016/j.evolhumbehav.2009.12.007

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Osswald, S., Greitemeyer, T., Fischer, P., & Frey, D. (2010). What is moral courage? Definition, explication, and classification of a complex construct. In C. L. S. Pury & S. J. Lopez (Eds.), *The psychology of courage: Modern research on an ancient virtue.* (pp. 149–164). American Psychological Association. https://doi.org/10.1037/12168-008

Oxoby, R. J., & McLeish, K. N. (2004). Sequential decision and strategy vector methods in ultimatum bargaining: Evidence on the strength of other-regarding behavior. *Economics Letters*, *84*(3), 399–405. https://doi.org/10.1016/j.econlet.2004.03.011

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), 20122723–20122723. 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*, *147*(4), 514–544. https://doi.org/10.1037/xge0000410

Pedersen, E. J., McAuliffe, W. H. B., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M. E. (2020). When and why do third parties punish outside of the lab? A cross-cultural recall study. *Social Psychological and Personality Science*, *11*(6), 846–853. https://doi.org/10.1177/1948550619884565

Prosser, A. M. B., Judge, M., Bolderdijk, J. W., Blackwood, L., & Kurz, T. (2020). 'Distancers' and 'non-distancers'? The potential social psychological impact of moralizing COVID-19 mitigating practices on sustained behaviour change. *British Journal of Social Psychology*, *59*(3), 653–662. https://doi.org/10.1111/bjso.12399

Qualtrics. (2019). *Qualtrics*. Provo, UT. https://www.qualtrics.com

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104–112. https://doi.org/10.1037/0022-3514.64.1.104

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences*, *109*(37), 14824–14829. https://doi.org/10.1073/pnas.1203179109

Rieger, M. O., & Wang, M. (2020). Secret erosion of the "lockdown"? Patterns in daily activities during the SARS-Cov2 pandemics around the world. *Review of Behavioral Economics*, *7*(3), 223–235. https://doi.org/10.1561/105.00000124

Robert Koch Institut. (2021). *Täglicher Lagebericht des RKI zur Coronavirus-Krankheit-2019 (COVID-19)*. Robert Koch Institut. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Maerz_2021/2021-03-26-de.pdf?__blob=publicationFile

Rosenberg, G. N. (2008). The hollow hope: Can courts bring about social change? In *The Hollow Hope*. University of Chicago Press. https://doi.org/10.7208/9780226726687

Rosman, T., Chasiotis, A., Kerwer, M., Steinmetz, H., Wedderhoff, O., Betsch, C., & Bosnjak, M. (2020). *Will COVID-19-related economic worries superimpose the health worries, reducing acceptance of social distancing measures? A prospective pre-registered study*. PsychArchives. https://doi.org/10.23668/psycharchives.3005

Rothmund, T., Baumert, A., & Zinkernagel, A. (2014). The German "Wutbürger": How justice sensitivity accounts for individual differences in political engagement. *Social Justice Research*, *27*(1), 24–44. https://doi.org/10.1007/s11211-014-0202-x

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sasse, J., Halmburger, A., & Baumert, A. (2020). The functions of anger in moral courage—Insights from a behavioral study. *Emotion*. https://doi.org/10.1037/emo0000906

Schacter, J. S. (2008). Courts and the politics of backlash: Marriage equality litigation, then and now. *Southern California Law Review*, *82*(6), 1153–1224.

Schlosser, F., Maier, B. F., Hinrichs, D., Zachariae, A., & Brockmann, D. (2020). *COVID-19 lockdown induces structural changes in mobility networks—Implication for mitigating disease dynamics*. ArXiv. http://arxiv.org/abs/2007.01583

Schmitt, M., Baumert, A., Gollwitzer, M., & Maes, J. (2010). The justice sensitivity inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. *Social Justice Research*, *23*(2–3), 211–238. https://doi.org/10.1007/s11211-010-0115-2

Seip, E. C., van Dijk, W. W., & Rotteveel, M. (2009). On hotheads and dirty harries: The primacy of anger in altruistic punishment. *Annals of the New York Academy of Sciences*, *1167*(1), 190–196. https://doi.org/10.1111/j.1749-6632.2009.04503.x

Soss, J., & Schram, S. F. (2007). A public transformed? Welfare reform as policy feedback. *American Political Science Review*, *101*(1), 111–127. https://doi.org/10.1017/S0003055407070049

Stüber, R. (2020). The benefit of the doubt: Willful ignorance and altruistic punishment. *Experimental Economics*, *23*(3), 848–872. https://doi.org/10.1007/s10683-019-09633-y

Szekely, A., Lipari, F., Antonioni, A., Paolucci, M., Sánchez, A., Tummolini, L., & Andrighetto, G. (2021). Evidence from a long-term experiment that collective risks change social norms and promote cooperation. *Nature Communications*, *12*(1), 5452. https://doi.org/10.1038/s41467-021-25734-w

tagesschau. (2021, March 17). *Corona-Pandemie: Die perfekte dritte Welle*. tagesschau.de. https://www.tagesschau.de/faktenfinder/corona-dritte-welle-101.html

Tan, F., & Xiao, E. (2018). Third-party punishment: Retribution or deterrence? *Journal of Economic Psychology*, *67*, 34–46. https://doi.org/10.1016/j.joep.2018.03.003

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, *10*(1), 181–211. https://doi.org/10.1111/sipr.12022

Tankard, M. E., & Paluck, E. L. (2017). The effect of a Supreme Court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, *28*(9), 1334–1344. https://doi.org/10.1177/0956797617709594

Terry, D. J., & Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. *Personality and Social Psychology Bulletin*, *22*(8), 776–793. https://doi.org/10.1177/0146167296228002

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. E. (2021). Economic games: An introduction and guide for research. *Collabra: Psychology*, *7*(1), 19004. https://doi.org/10.1525/collabra.19004

Thøgersen, J. (2008). Social norms and cooperation in real-life social dilemmas. *Journal of Economic Psychology*, *29*(4), 458–472. https://doi.org/10.1016/j.joep.2007.12.004

Tremewan, J., & Vostroknutov, A. (2020). *An informational framework for studying social norms: An extended version* [Unpublished Manuscript]. University of Auckland. http://www.vostroknutov.com/pdfs/tv-normsframework.pdf

Tunçgenç, B., El Zein, M., Sulik, J., Newson, M., Zhao, Y., Dezecache, G., & Deroy, O. (2021). Social influence matters: We follow pandemic guidelines most when our close circle does. *British Journal of Psychology*, *112*(3), 763–780. https://doi.org/10.1111/bjop.12491

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2018). An exploration of third parties' preference for compensation over punishment: Six experimental demonstrations. *Theory and Decision*, *85*(3–4), 333–351. https://doi.org/10.1007/s11238-018-9665-9

van Lange, P. a M., Agnew, C. R., Harinck, F., & Steemers, G. E. M. (1997). From game theory to real life: How social value orientation affects willingness to sacrifice in ongoing close relationships. *Journal of Personality and Social Psychology*, *73*, 1330–1344. https://doi.org/10.1037/0022-3514.73.6.1330

Van Lange, P. A. M., Balliet, D. P., & IJzerman, H. (2012). What we need is theory of human cooperation (and meta-analysis) to bridge the gap between the lab and the wild. *Behavioral and Brain Sciences*, *35*(1), 41–42. https://doi.org/10.1017/S0140525X11000872

von Haaren, F. (2020, July 6). Warum so viele Menschen die Corona-Regeln verletzen. *Welt*. https://www.welt.de/politik/deutschland/plus209058199/Verstoesse-gegen-Corona-Regeln-Menschen-lassen-alle-Grenzen-fallen.html

von Neumann, J., & Morgenstern, O. (2007). *Theory of Games and Economic*. Princeton University Press. https://doi.org/10.1515/9781400829460

Winter, F., & Zhang, N. (2018). Social norm enforcement in ethnically diverse communities. *Proceedings of the National Academy of Sciences*, *115*(11), 2722–2727. https://doi.org/10.1073/pnas.1718309115

Wu, J., Luan, S., & Raihani, N. (2021). Reward, punishment, and prosocial behavior: Recent developments and implications. *Current Opinion in Psychology*, S2352250X2100172X. https://doi.org/10.1016/j.copsyc.2021.09.003

Wu, J. J., Zhang, B. Y., Zhou, Z. X., He, Q. Q., Zheng, X. D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, *106*(41), 17448–17451. https://doi.org/10.1073/pnas.0905918106

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116. https://doi.org/10.1037/0022-3514.51.1.110

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*(1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322

Acknowledgements

I always portrayed the pursuit of a PhD as a very hard endeavor, and now that I look back, I confirm my suspicion. Focusing your professional life – and investing part of your personal life – on investigating a single topic for several years is indeed a quite demanding, exhausting and, at times, very frustrating task. This was especially the case under the circumstances in which we all have lived during the last two years of global pandemic. It did not make it easier to be deprived of one of the most important sources of inspiration and motivation, namely, the social support from my peers, friends and family. For my scientific peers, this does not come at surprise; they all have probably suffered similar experiences, given how much we depend on social contact to keep our brains and work ongoing. However, for those less familiar with how scientific research works, it may sound cliché. Reason why I would like to stress that doing science is nothing but a constant exchange of ideas that rarely come from a single mind. And in a social science like the one I am proudly part of, this exchange of ideas very often nourishes from daily conversations with friends, relatives, flat mates, etc. In my case, they all played a huge role in offering that kick of inspiration and motivation that I sometimes missed. Thus, I would like to dedicate this section to acknowledge the people who, directly or indirectly, pushed me professionally and emotionally to go through my PhD and develop the present dissertation. I could not have accomplished it without the (situationally) limited, yet incredibly valuable support from all of them.

I will start with the person that granted me the opportunity to initiate this journey. Anna, I am so grateful for your supervision and guidance during these four years. From day one, I felt trusted and free to make my own decisions, which I initially feared (in the end, one does not get rid of insecurities and impostor syndromes that easily) but which I highly appreciated later on. You have also been super encouraging and supportive with regard to my side projects and the reinforcement of my professional network. In retrospect, these have proved to be extremely valuable resources that you contributed to ensure. Yet, you have not only been a great supervisor, but also a mentor to rely on a personal level. Sharing with you my up and downs was never a problem and really helped me to gain perspective with some of the issues I encountered throughout my PhD. So, for these and many other reasons I may forget to mention, thank you very much! I really hope we keep in touch and, who knows, even stay in active collaboration for further projects!

I would also like to thank Julia for her support. In the presence or absence of Anna, you have always been a reference point when I had to deal with technical and professional decisions. Your guidance and feedback during these four years have been extremely valuable, proof of which are the different chapters of the present dissertation. I hope we can cross paths in the future, this being in Germany, the Netherlands, the UK, or perhaps, Italy!

Aya and Fiona. What to say about you two, my dear *colleagues*? Despite your persistent impression about my frustration with your endless chatting at the office, you should know how much I appreciated those

conversations about society, politics and research that we frequently got (and hopefully, will get) involved into. Those conversations that did not only reframe and improve my projects more than once, but also that never left untouched my viewpoints and sensitivities, forcing me to rethink and delve into my never-ending ambivalent tendencies. Aya, I do not want to repeat words I already shared with you after you left. I just want to add that you are one of these few people in life that passes by and revolutionize (I thoroughly choose this verb) your way of thinking and feeling about almost everything. I really need to thank you for making me become a better version of myself. I hope that this colleague-becomes-friendship relationship ignores the borders between us and lasts long in time. Fiona, it took me some time to get used to your critical style, but now I genuinely appreciate it as one of your many virtues. As a colleague, I think that your frankness and transparency are incredibly valuable for doing research and our group has always benefited from them. As a friend, I think that both frankness and transparency gain even more value, because they are very helpful (at least, they were for me) to bring clarity to daily personal issues and decisions. I cannot be more grateful to have met you and I hope we stay in close touch!

I would also like to thank the rest of the Moral Courage research group for contributing to create a great work environment of collaboration and personal appreciation. Mengyao, I have to thank you for your professional advice and (together with Chad) for your always-welcoming evenings at your place, this being for amazing hotpots, dumplings or board game sessions! Niklas, it has been great to have you as a colleague and to collaborate with you in both research and lecturing because I definitely learnt many things from you in the process. I wish you the best for the rest of your PhD! I would also like to highlight the indispensable work of David, Gabriela, Isa, Amelie, Lucy, Andreas, Marie, Silja and Ezra, the diligent research assistants that often helped me to carry out my research projects.

Moreover, I would like to extend my acknowledgements to my network of external collaborators, who allowed me to expand my research interests and experiences to other domains during my PhD. Among them, I would like to thank Frenk and Iris for giving me the push to get my first paper published and, more broadly, for all their support in my early academic career; the members of the Open Science group of the Max Planck PhDnet for being a very good example of what motivated and organized researchers from different fields can achieve; and finally, my colleagues Eva and Josh from the EASP Summer School, it was a pleasure to develop our small project together!

I am lucky to have built some friendships over time, both in Bonn and beyond, that helped me to find a better balance between work and life to survive these four years. First, my flat mate and friend Manon, who had to deal with my emotional rollercoaster and, simultaneously, with her own PhD. You managed both very successfully! And regarding your PhD, you are almost there girl! Wrap up this year and break free from that lab! Sham, thank you for always listening and for validating my fears and concerns. I will always admire

your way of spicing reality from that personal look of yours, which reflects in both your personality and your artistic talent. Maj-Britt, you were one of the first people that came to introduce herself when I joined the first summer school and I really appreciated it. I wish we had more time for climbing and continuing with our regular chitchats, but we should now focus on our next professional steps! Best of luck in Konstanz! To the group from Eiffel Strasse (Rupert, Ronja, Charlotte and Paula), a big THANK YOU for "adopting" me and being such a nice bunch of people, I am glad to have you around for existential chats, working and climbing sessions followed by spontaneous dinners. I am also thankful for my Amsterdam crew, with whom I gave my very first steps in research! Robin, Sam, Rabia, Maaike, Adam, Sanne and Simon, I miss you all guys! Especially my two caballeros, who have been an indispensable pillar for the last 6 years of my life… Thank you all for being there guys! And Anna (Luiken), I still remember that very sad afternoon when you left me at the bus stop direction to Schiphol, but four years later, I still feel you close and available for whatever I need. Thank you too for holding me when I need it!

And last but not least, I need to show my most sincere gratitude to the people who, from Spain, still felt close and shared my good and bad moments during this 4-year journey. Por supuesto, a mi madre y a mi padre, quienes siempre me han mostrado todo su apoyo en todas mis decisiones, y sin los que no podría haber dado los pasos que me han llevado hasta aquí. Mis chicas Alicia, Sara, Amaya y Ruth, quienes a pesar de los años que pasan, hace tiempo que siento como incondicionales y a las que debo mucho. Juls, por ser una desertora cómplice, una increíble mentora profesional, pero ante todo una buena amiga. Y finalmente Carlos, Rubén y Víctor, por recordarme desde la distancia mis raíces, llenarme de nostalgia, e ilusionarme cada vez que coincidimos en algún lugar del mundo.