UNIVERSITÄT ZU KÖLN

HOW SOCIAL NORMS AND ETHICS SHAPE ECONOMIC BEHAVIOR

- Three Essays in Experimental Economics -

Inauguraldissertation
zur Erlangung des Doktorgrades
der Wirtschafts-und Sozialwissenschaftlichen Faktultät

Vorgelegt von
CAROLINE STEIN

BERLIN
2023

Referent: Prof. Dr. Bernd Irlenbusch

Korreferent: Prof. Dr. Matthias Sutter

ACKNOWLEDGMENTS

During the past years, I have met and worked with various people who consistently supported and inspired me, for which I feel deeply grateful.

First of all, I would like to thank my advisor Bernd Irlenbusch from whom I learned a lot about conducting experimental research and who supported me throughout my time as a doctoral student.

I also want to express my gratitude to the members of my committee Matthias Sutter who accepted to be my second advisor and helped me with insightful comments and encouragement, especially during the work on the paper with my co-author Anna Untertrifaller. And also, Gönül Dogan, head of my thesis committee, with whom I had very insightful talks on research.

My very special thanks goes to Ernesto Reuben, whom I met at the beginning of my PhD, and who became my co-author and an excellent teacher for me. Our collaboration has been very inspiring and encouraging for me.

I am also thankful to my co-author Anna Untertrifaller with whom I spent many hours of fruitful discussions where we learned so much from one another.

My sincere thanks goes to Agne Kajackaite, whom I visited at the WZB Berlin and who supported me with her valuable feedback and consistent encouragement.

I am also thankful for my dear colleagues for their very helpful comments on my work, or for great discussions on research topics. Among others, those people are Anja Bodenschatz, Sebastian Butschek, Mira Fischer, Anastasia Danilov, Robert Stüber, Marina Schröder, Jarid Zimmermann.

I gratefully acknowledge financial support from the Cologne Graduate School (CGS), the Center for Social and Economic Behavior (C-SEB), and German Science Foundation (DFG) through the Research Unit Design and Behavior (FOR 1371).

I dedicate this thesis to my family, whom I love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

In philosophical dialectic thinking, the source of knowledge is the consciousness about oneself. Yet at the end of his critical analysis of the pure reason, Kant concludes that it is not the pure reason but the practical reason by setting moral laws that has a transcendental force and stimulates the motivation of human-being (Kant, 1787, KrV A795-A833). Taking this school of thought into consideration, intrinsic motivation for ethical and prosocial behavior does not only seem to be one driver for human behavior but might be considered essential for our fundamental perception and thinking. However, the extent to which intrinsic ethical motivations drive behavior is ultimately an empirical question.

In my thesis, I explore the intrinsic motivation of individuals to act according to ethical rules and social norms in economically relevant situations. While the former I define as rules that derive from a theory of ethics and apply to individuals independent from a social group, the latter describes the rules as they prevail within society through shared beliefs about what others think is right (Elster, 2015; Bicchieri, 2005). The two may often coincide as they aim for the same concepts as fairness or equality.

In economics, the interest in social norms has come along with the puzzle about its existence (Ostrom et al., 1992; Fehr and Fischbacher, 2004a). Why would people act against their personal profit even in situations among strangers to comply with a social norm, e.g. cooperation? The research on social norms in behavioral and experimental economics started with this question. Thus there is a large body of research showing that punishment is a considerably effective tool for enforcing social norms (e.g. Fehr and Gächter, 2000; Fehr and Fischbacher, 2003, 2004b; Balafoutas et al., 2014a; Gürerk et al., 2006; Sutter et al., 2010). Recently, another question has become of increasing interest in economic research, which is how social norms evolve and can be shaped (e.g. Bicchieri et al., 2021; Dimant and Gesche, 2021; Gächter et al., 2017; Ali and Bénabou, 2020; Bursztyn et al., 2020).

Through the lens of this recent research on the dynamics of social norms, I investigate punishment in chapter two. Specifically, I analyze how the experience of punishment can help shape a norm. Chapter three also deals with the question of norm enforcement but looks at a different tool, which is more relevant in an organizational context: whistle-

blowing. In chapter four, the topic moves from social norms to ethical decision-making and its consequences on performance.

In this thesis, I use laboratory and online experiments as the methodological toolbox for my empirical analysis. This method is a widely accepted practice in behavioral economics, which allows for observing behavior in a completely controlled setting (Kagel and Roth, 2020).

In chapter 2, I study in an online-experiment how the experience of punishment affects individual norm perceptions and behavior in situations where no punishment exists. The literature on punishment mainly focuses on its function to shift the incentive structure and make behavior not compliant with the norm unprofitable. However, in this paper, I focus on the norm signaling function as it has been proposed by scholars, particularly in law and economics (Sunstein, 1996). I argue that the experience of peer punishment can serve as a signal that the exhibited behavior is considered less appropriate than the punished subject thought it is. This could help update the normative perception. I assume this to happen, especially for behavior where no major agreement prevails over its inappropriateness or appropriateness but where some ambiguity exists.

I use a simple investment game with costly punishment and run two treatments to elicit trustees' behavior and norm perceptions separately. In the behavior treatment, subjects participate in the investment game for three rounds, whereas only in the second round can the investor punish a trustee for low returns. In the norm treatment, subjects also participate in an investment game with punishment in the second round. However, in contrast to the behavior treatment, subjects evaluate the social appropriateness of each return option in a hypothetical investment game without punishment in rounds one and three.

As a first result, I find clear differences in the agreement levels on the appropriateness of various return behaviors. As expected, returning nothing or returning an amount that results in an equal split exhibit large agreement over its inappropriateness or appropriateness, respectively. Secondly, the results suggest that individual normative perception correlates with actual behavior. Subjects who find returning low amounts more appropriate than the average return lower amounts in the second round. When it comes to the

question of how the social norm perceptions change, it seems that the subjects shift their normative perceptions to a more favorable evaluation towards lower returns, which I call an erosion of norms over rounds. However, this effect is significantly smaller for those subjects who were punished. The results indicate that the experience of punishment can not help elevate the social norm by making people perceive norm deviations as more inappropriate than before, but it instead seems to prevent norm erosion. This finding is consistent with the behavior treatment results where I also find a decrease in returns. The reduction in return is more minor for punishment. However, this difference is not significant.

With this chapter, I mainly contribute to experimental economic research on how the communication of a social norm interacts with the effectiveness of punishment (Andrighetto et al., 2013; Bicchieri et al., 2021; Dimant and Gesche, 2021). These studies show that providing additional social information about the norm can increase the effectiveness of punishment either through a direct positive effect on norm compliance or through the perception of punishment. I draw on this literature but flip the relationship of interest and thereby consider the social norm perception endogenous.

In chapter 3, which is joint work with Bernd Irlenbusch and Ernesto Reuben, we investigate the effectiveness of external whistleblowing regimes when internal whistleblowing is also possible. This study is motivated by the debates among policymakers and other societal stakeholders about which whistleblowing channels should be supported and protected by the law. Overall, it has been shown that employee whistleblowing is a highly effective tool to disclose corporate misconduct (Dyck et al., 2010; Call et al., 2018; Stubben and Welch, 2020). But employees usually have two ways to go. Employees who witness corporate misconduct typically have two options, they either go to an external institution, such as a prosecuting agency, or they approach someone within the organization (Near and Miceli, 1985). Yet the success of external whistleblowing is observable as it usually becomes public compared to internal whistleblowing, which is resolved within the company. Hence, it is unclear how effective internal whistleblowing can be on its own and how much an external whistleblowing channel can add to this. Furthermore, as external whistleblowing becomes public, it often comes with more severe consequences for the company. Therefore, we investigate to what extent these consequences might affect the effectiveness of external whistleblowing.

We design and conduct a laboratory experiment to study the effectiveness of internal and external whistleblowing by members of an organization. Specifically, we ask if the provision of an external channel can help to increase disclosure and sanctioning of lying even if it comes with more severe consequences for the organization. In the experiment, we formed groups of team members and public members. The team members were to perform a real-effort task and then state their performance to the public members. They could either state it truthfully or overstate their performance, which would increase their own income to the harm of the public members. In the next stage, team members can observe the decision of another team member about whether he stated the performance truthfully or not. In case of an untruthful statement, the observing team member can report this misconduct either to another team member, internal whistleblowing, or to the public, external whistleblowing. This is the baseline treatment where internal and external whistleblowing exists, and only the recipients' incentives to accept the report between the two channels are different. We compare this baseline treatment to two other treatments with varying consequences for the team members. As a control treatment, we have a scenario where team members can only report to another team member.

As a result, we find that the external channel indeed increases sanctioning. However, it does not seem to encourage more people to blow the whistle. Instead the external channel turns out to be more efficient in sanctioning. This is not only driven by employees in the organization who lie and therefore do not want to sanction other members who misbehaved. We also find a non-negligible fraction of honest subjects who abstain from whistleblowing and from going forward with reports for prosecuting other members.

With this chapter, we contribute to a growing strand of experimental economic research on whistleblowing, exploiting the methodological advantage that all misconduct can be observed, which is not possible in the field (Reuben and Stephenson, 2013; Butler et al., 2019; Schmolke and Utikal, 2018; Muehlheusser et al., 2020). In contrast to these studies that only provide one option to report - usually to an external party or a computer - we analyze a situation where two different channels of whistleblowing are available.

Whereas in chapters two and three, the focus lies on the enforcement of social norms and the prevalence of norm enforcement within social groups, the next chapter instead

focuses on the individual ethical decision-making and how this affects performance.

Chapter 4 is joint work with Anna Untertrifaller, where we analyze whether being responsible for an ethical or unethical work environment affects workers' performance. This study is motivated by combining two strands of experimental literature on intrinsic motivation and performance. On the one hand, there is evidence that people perform better when they consider their job meaningful (Ariely et al., 2008; Chandler and Kapelner, 2013) or it has a pro-social mission (Fehrler and Kosfeld, 2014; Tonin and Vlassopoulos, 2015; Carpenter and Gong, 2016; Charness et al., 2016; Kajackaite and Sliwka, 2017; Cassar, 2018). In these studies, the positive effect on performance is explained by selection: people who have preferences for such a pro-social mission choose this job. However, on the other hand, there is also experimental research suggesting that responsibility and autonomy over a decision can increase performance (Charness, 2000; Babcock et al., 2015; Falk and Kosfeld, 2006; Bartling and Fischbacher, 2011). Therefore, we ask if the performance increase for a pro-social or ethical job environment could not only come from the selection but also from the fact that the people feel responsible for their ethical behavior and thereby more committed to the task they do (Deci, 1971; Ryan and Deci, 2000; Gagné and Deci, 2005).

In a laboratory real-effort experiment, we use a specific randomization technique that allows us to separate the responsibility effect from a possible selection effect. For this, we assign subjects either the role of a worker or an employer and form groups of two. The worker has to eventually perform a task, about which both players can decide. Specifically, both players can overstate the piece rate, which is beneficial for both of them. However, apart from the aspect that the overstatement is untruthful, we also made clear that it is against the rules. Thus a violation of this rule we call unethical. Only after both players decide, will a random choice implement either the decision of the employer or the worker. Hence, we elicited the worker's preferences before the decision took place. When the employer's decision was implemented, we call it No Responsibility because the worker performs a task according to a piece rate he did not choose. He knows this and also knows whether or not the employer overstated. In case the worker acts according to the piece-rate he decided for, we call this situation Responsibility. We suggest a theoretical framework that use a standard utility function and adds ethical costs and intrinsic motivation from ethical responsibility. From that, we the hypothesis that under

a truthful piece-rate, workers who prefer a truthful piece-rate would exert higher effort if their own decision was implemented (Responsibility) compared to a situation where the decision of the employer was implemented (No Responsibility).

We find that workers who prefer to work under a truthfully stated piece-rate (an ethical work environment) perform better if they are also responsible for it, compared to a situation imposed on them. We do not find this positive incentive effect of responsibility for workers that prefer an unethical work environment. Moreover, we observe that if an unethical environment is imposed, workers who prefer an ethical environment perform worse than those who are aligned with the environment.

The results from the three experimental studies I present in this thesis show that people seem to be influenced by others' behavior, as being punished can have a sustaining effect on social norms. Moreover, we learn that some people are willing to report an observed wrongdoing irrespective of the expected success of the whistleblowing action. Internal reporting seems less effective in sanctioning because not only do those who committed a wrongdoing refrain from letting whistleblowing through but even subjects who act norm compliant. Lastly, I show that ethical decision-making can also positively affect performance when people can bear responsibility. Hence this thesis contributes to a better understanding of the dynamics of social norm enforcement and the motivational effect of acting ethically. In the following chapters, the experimental studies are presented in detail.

CHAPTER 2
TEACHING RIGHT FROM WRONG: THE ROLE OF PUNISHMENT IN NORMATIVE UPDATING

2.1 Introduction

People punish others for norm violations that harm the punisher herself or a third party. Experimental economic and psychological research finds extensive evidence for this phenomenon in the field and in the lab (e.g. Fehr and Gächter, 2000; Balafoutas et al., 2014b). This well studied observation is essential to understanding basic concepts of norm enforcement and the evolution of norm-enforcing institutions (Gürerk et al., 2006; Sutter et al., 2010). In the presence of potential punishment people usually act more compliant to the norm - even in situations with complete strangers.(e.g. Lergetporer et al., 2014; Fehr and Gächter, 2002)

In this field, most studies model punishment as an exogenous tool that shifts the incentive structure (Becker, 1968; Ostrom et al., 1992; Sigmund, 2007; Fehr and Gächter, 2002). Under the threat of punishment, people expect to suffer a high payoff reduction if they deviate, and they would act compliant because it is the most profitable. However, there is a large body of research, primarily in the field of law and economics, arguing that punishment not only comes with a material cost but also conveys a signal about the normative appropriateness of an action (Sunstein, 1996; Tyran and Feld, 2006; Masclet et al., 2003; Galbiati and Vertova, 2008). It thereby emphasizes the norm-signaling function of law and punishment, which may also affect norm compliance, especially in situations where uncertainty exists about the prevailing social norm (Benabou and Tirole, 2011).

In a recent study, Galbiati et al. (2021) take up the idea of the norm-signaling function and study how laws can affect the perception of the social norm. Specifically, the authors study how past lockdown policies during the Covid-19 pandemic drastically changed the perception of the social norm of social distancing. However, the focus lies on the influence of laws and thus on the institutional level of punishment. I draw on this research but with a different focus. Specifically, I am interested in norm enforcement among peers and ask if the experience of punishment by another person could also affect the perceived social norm.

To illustrate the research question, let us take another example from the Covid-19 pandemic. Public health organizations have heavily promoted proper hand-washing behavior as a simple but considerable contribution every individual ought to make to help mitigate the spread of the virus. However, a proper practice costs people some effort and is not always observable by others. A person might deviate from the recommended practice and not do it carefully enough, but think it still is appropriate because it is something. If someone sees it and punishes this deviating behavior, this may help the person update the belief about what is commonly considered appropriate. In other words, being punished could change the evaluation of how accepted a deviation of the social norm is, which might in turn, create a higher cost for this deviating behavior. Therefore, the punishment could help update and adjust how a person evaluates a behavior and thereby supports the internalization of a norm.

In the present paper, the main research question is whether the experience of being punished influences how the person perceives the norm and behaves in a situation where punishment is not possible. Specifically, I am interested in how punishment could change the evaluation of an action that lies in a range of normative ambiguity and, thus, for which no explicit agreement prevails over its appropriateness or inappropriateness. To study the research question, I use an online-experiment. The main reason for this methodological choice is that it allows me to measure the perceived social norm and behavior separately. These two variables of interest usually come together in the field and are hard to disentangle. Furthermore, the experimental study delivers observations within subjects directly before and after punishment is experienced, which helps to study the causal relationship between punishment and social norms.

Specifically, I use a simple investment game with weak punishment and run two treatments to elicit behavior and norm perceptions separately on a stranger matching protocol. In the behavior treatment, subjects participate in the investment game for three rounds, whereas only in the second round can the investor punish a trustee. In the norm treatment, subjects also participate in an investment game with punishment in the second round. However, in rounds one and three, subjects evaluate the social appropriateness of each return option in a hypothetical investment game without punishment using an incentivization similar to Krupka and Weber (2013).

As the main result, I conclude that punishment seems to have a sustaining effect on the evaluation of actions for which normative ambiguity prevails. Overall, I find that trustees significantly shift their perception of the social norm towards a more favorable evaluation of deviations from the social norm. However, the deteriorating effect seems to be only prevailing for those trustees who did not receive a punishment. Controlling for the return behavior in round 2, I find that the experience of punishment significantly diminishes the shift to a more favorable evaluation of norm deviations. Hence, it appears that the experience of not being punished encourages the trustee to consider norm deviations more appropriate than before. At the same time, the experience of punishment prevents this erosion of the norm. As for the behavior treatment, the results hint at a similar conclusion. I also find a significant decrease in returns from round 1 to round 3 in the behavior treatment for trustees who were punished or not. However, while deteriorating shift appears to be more pronounced for trustees who were not punished, the difference is not significant.

This paper contributes to existing evidence and conceptual thinking about norm enforcement for two main reasons. First, the results of this study unfold how punishment could affect norm perceptions and behavior even beyond the situation where it actually takes place. Second, in contrast to other experimental papers on social norm perceptions (Dimant and Gesche, 2021; Bicchieri et al., 2021; Reuben and Riedl, 2013) I not only look at the aggregated evaluations and how they might change or be different between groups. In this paper, I also focus on the individual norm perception to analyze norm change within subjects. Specifically, I use two different measures. I look at the switching point, which indicates the first action that a subject considers appropriate among all decision options put in an order. And I look at how appropriately a subject evaluates its own behavior in round 2. These two measures together may give a compound picture of how norms can change on an individual level.

2.2 Related literature

This paper is related to various strands of literature. First, it contributes to experimental research on the interaction between social norm communication and peer punishment. Andrighetto et al. (2013) study a public good game and combine the material threat of punishment with an additional norm-signaling component. They show that this interac-

tion can increase and sustain pro-social behavior more than punishment alone. While the authors find a positive effect of the norm-signal on the effectiveness of punishment, they consider the social norm information as a separate component not relative to punishment.

Bicchieri et al. (2021) recognize the norm-signaling function punishment holds itself. As the experimental design in the present paper, the authors use a trust game with weak punishment such that the incentive effect is not sufficient to change behavior. The authors concentrate, therefore, on the norm-signaling function of punishment and ask how additional information about the norm could change the perception of the legitimacy of the punishment. They find that weak punishment combined with normative information about what people think is appropriate to do can increase compliance compared to no punishment. I draw on this research described above but flip the causal relationship of interest. While in those studies, the authors use the normative signal as an exogenous variation to investigate the effect of norm information on the effectiveness of punishment, in the present paper, I focus on the question of how punishment can affect the social norm perception.

Therefore, this study relates to the experimental literature on interventions to change individual normative perceptions. In a recent experimental study, Dimant and Gesche (2021) investigate whether the provision of norm information changes the norm perception of the action to be punished and could increase punishment. For this, they look at the behavior and normative evaluation separately and find that, indeed, information about the prevailing social norm increases the amount of punishment. This finding is consistent with their observation that the norm information also lowers the average perception of the norm violation. Furthermore, they show that the norm information has a more substantial effect on norm perception in an environment of conflicting social norms and thus where the evaluation of norm violations is rather ambiguous.

In another study by Gächter et al. (2017), the authors examine how the observed behavior of the peers affects the individual norm perception. In a sequential three-player dictator game, they elicit the behavior and the normative perception of the behavior of the second dictator in two separate experiments. They find that if observable, the first dictator's choice has a significant influence on the perception of the social appropriateness of norm deviating behavior. Hence, this study shows that the observation of

a single person and not only the information about a broader group can heavily affect norm perceptions. Furthermore, this effect is larger for actions where normative ambiguity prevails. This result is consistent with the findings of the present paper that the experience of punishment by another person influences the norm perceptions of behavior for which no explicit normative agreement exists. However, in those previous studies, the comparison remains between subjects, and does not consider how normative perceptions may actually evolve within subjects.

As for the within-subject comparison of norm-related behavior and perceptions, this paper also relates to a strand of literature on spill-over effects of enforcement regimes on behavior and normative perceptions (see Galizzi and Whitmarsh (2019) for an overview). Peysakhovich and Rand (2016) show in their study on habit formation that subjects that were repeatedly exposed to an incentive environment that encourages cooperation, keep behaving more cooperative in other subsequent games. Engl et al. (2021) present evidence of spill-over effects of enforcement institutions on simultaneous behaviors in unregulated domains. Another strand of literature investigates how the exposure to norm enforcement institutions changes cooperative behavior through rational learning (Galbiati et al., 2019; Acemoglu and Jackson, 2017). In this paper, I refer to this literature and adopt the reasoning of a rational updating of beliefs about the social norm. Specifically, I hypothesize that subjects, who received a punishment for an action for which normative ambiguity exists, may infer new information and thus update their normative perception towards a less favorable evaluation.

2.3 Experimental Design and Procedures

General Structure of the Experiment

In this paper, I study whether the experience of being punished can make people change their perception of the social norm and can thereby affect their subsequent behavior in the same domain of action. I use an experimental design by which I can compare outcomes within subjects.

Figure 2.1 provides an overview of the experiment. In short, the experiment consists of three rounds. In round 1, subjects participate in a basic investment game as a variant of the Trust Game (Berg et al., 1995). In round 2, the investment game is repeated. However, in addition to the basic setting, the investor has the opportunity to costly punish

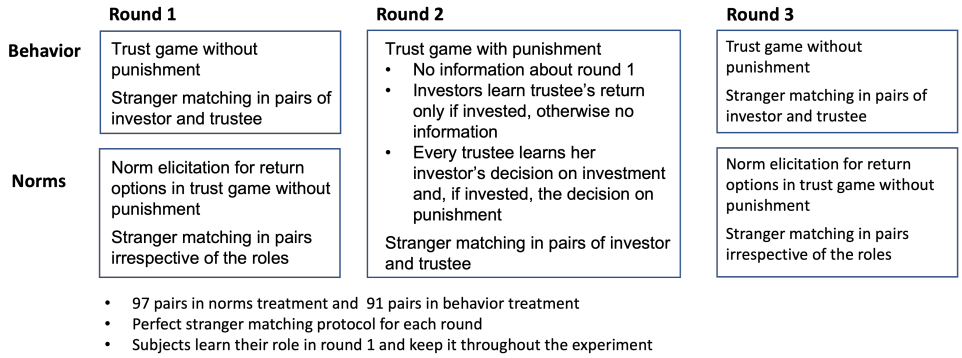


Figure 2.1. Overview of the experiment by treatment and rounds

the trustee. In the third round, the subjects will play the basic version of the investment game as in round 1. This structure thus allows me to observe decision outcomes before and after trustees may have experienced a punishment.

Specifically, I am interested in two different outcomes - the trustees' perception of the social norm regarding the return behavior and the trustee's actual return decision. Therefore, I look at these two outcomes separately, and divide the study into two treatments: Norms treatment and Behavior treatment. In the *Norms treatment*, the subjects assess how socially appropriate they think are the trustee's return options in the presented investment game. The norm elicitation happens in rounds one and three. In round 2, the subjects actually take part in the investment game with punishment and, respectively, to their role, they decide on return, investment or punishment.[1] In the *Behavior treatment*, subjects make decisions on investment and return not only in round 2 but also in rounds one and three.

Prior to round 1, participants are randomly assigned the role of an investor or the role of a trustee. They keep this role throughout the session and are re-matched for each new round on a perfect stranger matching protocol. Hence, the experimental setup makes sure that subjects would never interact with another subject in this session more than once. This is also clearly stated in the instruction.

Behavior treatment

In round 1, one investor and one trustee[2] are randomly matched into a pair. Both subjects receive an endowment of 15 tokens. The basic game consists of two decisions: investment

[1]This is necessary because the trustees shall be exposed to an actual punishment as I study its effect on the subsequent norm perception.

[2]In the following, I refer to the investor in the feminine form and the trustee in the masculine form.

decision and return decision.

Investment decision: The investor decides whether or not to send 10 tokens of her endowment to the trustee, $i \in \{0, 1\}$. In contrast to a standard investment game, this is a binary decision.[3] Thus, the investor decides whether or not to invest. If the investor sends money, her 10 tokens will be multiplied by four and the trustee will receive 40 tokens.

Return decision: The trustee chooses how much he would return to the investor, where $r \in \{0, 5, 10, 15, 20, 25, 30\}$ before he learns about the investor's decision. Thus the trustee always decides how much to return. However, the return decision comes into effect only if the corresponding investor chooses to invest.

Based on the subjects' decisions in each group j the earnings are as follows:

$$\Pi^j_{Investor} = 15 - i^j * (10 - r^j). \tag{2.1}$$

$$\Pi^j_{Trustee} = 15 + i^j * (40 - r^j). \tag{2.2}$$

At the end of round 1, participants neither learn their payoff nor receive feedback about their partner's decisions. Instead, they will learn about the decisions in this round at the end of the session.

In round 2, investors and trustees are again randomly matched into pairs. A second round of the investment game takes place. Again, the subjects make - according to their role - a decision on investment or return as in round 1. If the investor decides to invest (i = 1) in this round, she will get the opportunity to punish her trustee.

Punishment decision: If the investor decides to invest, i = 1, she learns how much her trustee decides to return. After that, she makes a binary decision on whether or not to punish her trustee, $p \in \{0, 1\}$. If the investor chooses to punish, a random draw, $d \in \{0, 1\}$, determines the outcomes. Specifically, with a 50 percent probability, the trustee's payoff will be reduced by a fixed amount of 20 tokens, and the investor will pay a fee of 5 tokens. In the other 50 percent of the cases, the trustee does not suffer a payoff reduction, and the investor will not pay a fee.

[3] I restrict the investor's decision space to a binary choice because my focus lies on the social norm perception of the return behavior. However, the return might be dependent on the investment choice made before. To keep it simple for the participants in the experiment, I decided to have a binary choice here.

After all decisions were made in group j, the subjects are informed about their own earnings in this round. Based on the subjects' decision in each group j the earnings are as follows:

$$\Pi^{j}_{Investor} = 15 - i^{j} * (10 - r^{j} + (d|i^{j} = 1, p^{j} = 1) * (p^{j}|i^{j} = 1) * 5) \tag{2.3}$$

$$\Pi^{j}_{Trustee} = 15 + i^{j} * (40 - r^{j} - (d|i^{j} = 1, p^{j} = 1) * (p^{j}|i^{j} = 1) * 20). \tag{2.4}$$

As the trustees receive feedback about their earnings, they are also informed about the investment decision of their corresponding investor. If the investor made an investment, the trustee learns whether he is punished. As for the investors, the feedback depends on their investment decision. Investors who invest already learn the partner's return decision before they decide to punish or not. Those investors who decide not to invest (i = 0) will not be informed about the return decision before the end of the session.

In round 3, the subjects are matched into groups of one investor and one trustee again. After that, they make the same decisions as in round 1. At the end of this round, the participants receive feedback about the decisions of their respective partners of each round.

Norms treatment

In round 1 of this treatment, subjects face the identical decision situation as in round 1 in the behavior treatment. The difference is that the subjects do not take part in this game. Instead, I ask them to evaluate how socially appropriate they find each possible return option.

The participants start with the same instructions as for the behavior treatment in round 1. However, in addition to the instructions from the behavior treatment, the participants learn that they will not make an actual decision in this round but state normative evaluations. Furthermore, they learn that the decision situation will also become relevant for them in the next round. More concretely, the subjects learn their role at the beginning of the session. They can check their role in the header of each screen. Thus, they will make their evaluation in expectation to act in the role of a trustee or a investor in a subsequent round.

As for the normative evaluation in round 1, the subjects are informed that they shall

evaluate every return option provided to the trustee in the return decision. Specifically, subjects can assign an evaluation item from the following range to each choice option separately: very socially appropriate, rather socially appropriate, rather socially inappropriate and very socially inappropriate. For this, they find all return options listed [4] on the screen. In the instructions, I refer to the definition by Krupka and Weber (2013) and describe the concept of socially appropriate as "a well-accepted behavior, which is generally viewed as an action that ought to be done". Furthermore, in the instructions, I describe a socially inappropriate behavior "as an action that is generally considered unacceptable and for which you might receive very angry reactions".

The norm elicitation is incentivized after Krupka and Weber (2013)[5] and works as follows. Subjects are - randomly and irrespective of their roles - matched into a pair. The subjects are endowed with an amount of 15 tokens. Additional to the endowment, they can earn a bonus of 15 tokens. The bonus depends on their evaluations and the answers of the matched partner. Specifically, for each pair a random draw selects one out of the seven return options the subjects are to evaluate. The subjects will receive the bonus only if they make the same evaluation for this selected return option as their matched partner. Otherwise, they receive no bonus in this round. Like in the behavior elicitation group, participants do not receive feedback about the other participants' decisions in this round. They are informed that they will learn the matched partners' choices at the end of the session.

In round 2, one investor and one trustee are randomly matched into a pair. At the beginning, the participants learn that they are to make actual decisions in this round. They face the same decision situation as participants from the behavior treatment. Thus, the basic game consists of the investment decision and return decision. In addition, investors who decided to make an investment in the first place also make the punishment decision. The instructions for this round are identical to the instructions for round 2 in the behavior treatment.

In round 3, the subjects are again randomly matched into pairs. As in round 1, the matching is independent of the assigned roles. The subjects shall evaluate the possible

[4]For the evaluation, all return options are listed in order from return zero to return 30.

[5]A recent study by Fallucchi and Nosenzo (2021) shows that using a coordination game to elicit the norm is a valid technique and robust against other focal points that are not social norms.

return options of a trustee again in a basic investment game without punishment, as they did in round 1.[6]

Experimental Procedures

I ran an online-experiment with 376 participants in total. Specifically, I collected data of 91 pairs in the behavior treatment and 97 pairs in the norms treatment. It was an interactive study with direct feedback. Hence, the experimental sessions took place in predetermined time slots as announced in the invitation e-mails. Within the session, I used time limits which I made very clear at the beginning in the general instructions. Specifically, after round 1 and round 2, subjects had to answer questions to ensure a solid understanding of the structure of the proceeding game.[7] For these questionnaires I provided a time limit that was three times higher than the average time used for the answers. If a subject was time outed, it dropped out the whole session. Likewise, for the decision screens, I set a time limit. For each decision, subjects had five minutes, which is several times higher than the average decision time, and should make sure not to put pressure on the subjects. If the subjects did not make the decision in this time frame, they also dropped out the game.

I programmed the experiment in Otree (Chen et al., 2016) and used the online-survey subject pool of the Cologne Laboratory for Economic Research of the University of Cologne. Participants earned 1 Euro for 5 tokens in the game. The show up fee was 2.5 Euros. They were paid out only one round that was randomly selected for each session. On average, subjects earned 10 Euros.

[6]When all three rounds are completed, the participants filled out a short survey about their potential motives to punish, irrespective of their roles. Specifically, I provide them with a concrete situation as it could have happened in the investment game, in which they are an investor, and the corresponding trustee returned five tokens. I then give them eight statements describing a specific motive to punish (anger as emotional reaction to the low return, (2) inequality aversion (3) moral obligation (4) joy of power (5) reciprocity (6) norm-signaling) and ask how relevant this motive would be (in a range of very relevant, somewhat relevant, somewhat irrelevant, very irrelevant) to their personal decision whether or not to punish in such a case. However, due to the scope of this paper, I did not further analyze the data for this paper.

[7]I gave examples to clarify from solving this questionnaire what the equal split action would be in this game.

2.4 Hypotheses

In economic research, punishment has been mainly investigated as a tool to enforce norms by putting a (monetary or non-monetary) cost on non-compliant behavior. However, punishment might be effective also because it signals a norm. Particularly, for people who behave in a way that is neither commonly agreed to be inappropriate nor considered socially appropriate, being punished could help them adjust their normative evaluation of the behavior. Hence, in the following, I will focus on this argument and discuss hypotheses primarily for the norms treatment. First of all, I expect punishment to prevail for both treatments, which provide the basis of the experiment.[8] So is punishment necessary to prevail in both treatments to proceed with the following argument.

Norms treatment

Recalling the experimental setup of the norms treatment, in round 1, all subjects - irrespective of their assigned role - are to evaluate each return option of a trustee in a range between very socially inappropriate and very socially appropriate. I expect here a large agreement over the social appropriateness of returning 25 points and, as well, agreement over the inappropriateness of returning zero or five as purely selfish action.[9] Likewise, there should be more heterogeneity in the normative evaluations for returns in the range between 10 and 20. These amounts would cover the investment costs for the investor but would not implement payoff equity. Thus it is a deviation from the social norm of equity[10] which is not purely selfish. This behavior might be assessed differently by the subjects and thus create normative ambiguity, as already observed by various empirical studies in the context of a dictator game (Krupka and Weber, 2013; Gächter et al., 2017; Dimant and Gesche, 2021; Reuben and van Winden, 2010).

[8] There is no monetary incentive for investors to punish. Nonetheless, there are various non-pecuniary motives such as preferences for fairness and equity (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), negative reciprocity (Reuben and Van Winden, 2008; Brandts and Solà, 2001) or emotional satisfaction (Bosman and Van Winden, 2002; Hopfensitz and Reuben, 2009) that should cause a significant amount of punishment even if the incentive function is not in place, see also (Kosfeld et al., 2009; Engel, 2014). All these behavioral explanations mutually rely on the assumption that the investor punishes because she expected a higher return and thus derives utility punishing the misbehavior(Xiao, 2018).

[9] At least in this setting, it could not be justified by any social norm or fairness principle. For example, Almås et al. (2020) show that there are different norms on redistribution and tolerance for inequality prevailing across cultures. However, this finding emerges only when merit is also a factor.

[10] Payoff equity is a well-established social norm in behavior economics (Krupka and Weber, 2013; Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999)

In round 2, trustees make a return decision, for which they gave evaluations of appropriateness in the previous round. Hence, I expect the return decisions to be correlated with the normative evaluations elicited in round 1. Specifically, I draw on a growing body of empirical research (Bicchieri, 2005; Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016; Fehr and Schurtenberger, 2018; d'Adda et al., 2020; Bicchieri et al., 2021) suggesting that a large proportion of subjects have a desire to act in a socially appropriate way. Hence, if one subject believes that deviating from the social norm is more appropriate than others, she would also suffer a smaller intrinsic utility cost from returning lower amounts.

To illustrate the intuition I refer in the following to the utility function presented by Krupka and Weber (2013)[11]:

$$U(x) = y_x + \alpha * n_x, \tag{2.5}$$

where y_x denotes the utility a person derives from the monetary gain of action x, n_x ranges between $\{-1, 1\}$ and it describes the individual evaluation of social appropriateness of the action x . Furthermore, α with ≥ 0, denotes how much an individual values the normative appropriateness of an action.

If $\alpha > 0$, the utility function implies that subjects derive utility from their own behavior the more socially appropriate they perceive it.[12] Hence, given an action x and ceteris paribus, those who evaluate x more socially appropriate than others would also derive more utility. Therefore, one might expect that subjects who perceive low returns as more socially appropriate than others will be, on average, more inclined to return low amounts.[13] Translating this expected behavior to the present experiment, I hypothesize that trustees who return low amounts perceive, on average, low returns more appropriate in round 1 than trustees who returned a higher amount. In contrast, subjects with $\alpha = 0$

[11] There are various other approaches to incorporate normative considerations e.g. Bénabou and Tirole (2006) and a recent model by Fehr and Schurtenberger (2018). I use the model by Krupka and Weber (2013). It fits best to this present setting because I look at the social norm perception on an individual level and also need to capture the assessment of deviations from the social norm that might be more ambiguous.

[12] However, the causal relationship could come from both directions. While there exist attempts to identify the causal effect of norms on behavior by endogenously varying the context (Krupka and Weber, 2013), the information on the norm (Fehr and Schurtenberger, 2018; Bicchieri et al., 2021), there also exists evidence for a causal effect in the reverse direction, for example on motivated belief to justify behavior (Bicchieri et al., 2022).

[13] I assume y_x and n_x to be independently distributed.

only value the monetary gain of an action. Thus those subjects will most likely choose to return nothing, and their evaluation of the social appropriateness should not be correlated.[14]

Hypothesis 1 (Norms treatment): *On average, trustees who return amounts lower than the equal split in round 2, evaluate low returns more appropriate than others in round 1.*

In the next step, I turn to my primary question of how the experience of punishment in the second round might affect a subject's norm perception. As argued above, in this present setup, punishment conveys a signal of the norm, which possible selfish motives of the investor cannot undermine.[15] Hence, punishment transmits normative information about how others think about deviating from a social norm by serving as a signal. Moreover, this information might be different from what the receiving trustee believed. According to Bayes' rule, such new normative information would make the trustee update his belief about how appropriate a norm deviation is. Hence, a trustee who initially assumed that his behavior might be considered appropriate infers new information about the norm from the experienced punishment, which is that the behavior is less appropriate than believed. Consequently, being punished would shift the normative evaluation n_x of a given action x downwards.

However, a trustee infers new information from the punishment only if, in the first place, he believed his behavior was at least somewhat appropriate and thus did not expect the punishment. Therefore, the return in round 2 must be in a range where normative ambiguity and disagreement over its evaluation exist. In contrast, if the trustee sent a return of zero (or five), he would hardly think this is appropriate behavior.[16] The

[14]Indeed, one could argue here that pro-social preferences are independent of the preference to adhere social norms. Thus, even if a subject does not care about a prevailing social norm, she might still act pro-socially. However, I follow existing empirical evidence that social norms and pro-social preferences are strongly linked to each other (Kimbrough and Vostroknutov, 2016; Kölle and Quercia, 2021) and can either be explained by social or self-image concerns (Falk, 2021).

[15]Studies show that cooperation significantly decreases when punishment can serve selfish motives (Fehr and Rockenbach, 2003; Houser et al., 2008). This evidence suggests that monetary incentives can interfere with the normative aspect of punishment. Thus the normative aspect of punishment seems to be overshadowed.

[16]It is possible to assume an entirely altruistic investor would prefer a trustee taking everything. However, this type exists, if anything, very rarely but may be indeed assumed by a trustee to justify one's selfishness. In that specific case, also zero returners might update their norm perception.

punishment would not give him new information. Likewise, a trustee who returned 25 and thus equally split, which can be considered an unambiguously appropriate action, would not learn from being punished either. He might instead consider it a malevolent act from the investor.[17] Therefore, a change in the individual norm perception through belief updating would primarily happen for those who sent a return between 10 and 20 points in round 2.

Hypothesis 2 (Norms treatment): *Trustees who were punished in round 2, will subsequently evaluate low returns as less appropriate than in round 1.*

Behavior treatment

Putting together hypotheses 1 and 2, the prediction about how the experience of punishment affects behavior is straight forward. On the one hand, I argued that I expect trustees to be more likely to update their norm perceptions if they were punished for a return r $\in \{10, 15, 20\}$. On the other hand, I hypothesize that many subjects derive utility from the adherence to a social norm such that their behavior is to some extent correlated with their norm perception. Consequently, trustees who experience punishment in round 2 for a return in a range of some normative ambiguity will update their belief on how appropriate this action is. Through the norm-updating, the subject will decrease the normative evaluation, n_x, of the behavior he was punished for. This, in turn, will decrease the utility from taking that action or even render it costly (when the evaluation changes from a $n_x > 0$ to $n_x < 0$). Hence, a decrease in n_x might make a subject change the behavior and thus return a higher amount.

Specifically, I expect that trustees who experience punishment in round 2 will be, on average, more likely to change their behavior according to their updated beliefs about the norm.

Hypothesis 3 (Behavior treatment): *A trustee who was punished in round 2, will be more likely to increase the return from round 1 to round 3 as compared to a trustee who was not punished.*

[17] For public good settings, evidence has been found that antisocial punishment also exists, which means that some people would punish others for high contributions (Herrmann et al., 2008). While the motivations for this behavior have not yet been fully understood, it is not clear whether this will also prevail in investment game settings.

2.5 Results

In this section, I analyze whether the experience of punishment can help people update their evaluation of how socially appropriate deviations from the norm are. Therefore, the primary focus lies on the norms treatment and the question of how individual norm evaluations prevail before and after the experience of punishment.[18] First, I look at the aggregated level of norm perceptions to see how the norm initially prevails in round 1.

Norms treatment - Prevalence of norms and correlation with behavior

Table 2.1 depicts the relative frequencies for how appropriate the given actions were evaluated by all subjects[19] distinguished between round 1 (in panel A) and round 3 (in panel B). It shall illustrate how the overall perception of the social norm prevails. For now, I focus on the evaluations made in round 1 (panel A). The table presents a large majority of subjects who evaluate returns of zero (98 percent) and five (81 percent) as "very socially inappropriate". Likewise, among all options, the return of 25 is evaluated as "very socially appropriate", with the largest majority of 82 percent of the subjects. Whereas for the options in between, we find more heterogeneity in evaluations. Hence, there seems to be more consensus on the normative perception towards giving nothing or splitting equally.

From the frequencies of normative evaluations from Table 2.1, I compute the percentage (levels) of agreement that is to be expected for each return. Figure 2.2 depicts these agreement levels, and it illustrates the clear consensuses about the inappropriateness of returning zero and the social appropriateness of returning 25. In contrast, there seems to be much more disagreement for those returns that are in the range between 10 and 20. As for the returns 10 and 15, more than half of the matched pairs would disagree. Moreover, if an agreement exists, it is not clear on what evaluation.

In order to test for differences between the agreement levels as shown in Figure 2.2, I use a permutation test (Holt and Sullivan, 2021; Kennedy, 1995). This standard non-parametric test relies on the idea that the null distribution can be built from the empirical sample by shuffling the observational data across the comparison groups. Hence, the null

[18] In both treatments, we find large fractions of punishment. Specifically, among the trustees who received an investment, 43 percent in the norms treatment and 40 percent in the behavior treatment were eventually punished.

[19] In the appendix, Table 2.7 presents the same aggregated values in round 1 and 4 but only for trustees.

Table 2.1. Relative frequencies of evaluations about the appropriateness for each return

Note: Evaluations of all subjects independent of their roles in round 1 and round 3. The evaluation items are described as follows: (- -) very socially inappropriate, (-) somewhat socially inappropriate, (+) somewhat socially appropriate, (++) very socially appropriate. I assign to each item a number from 0 (- -) to 3 (++) and can, therefore, take the mean values.

	Panel A: Overall evaluations in round 1					**Panel B:** Overall evaluations in round 3				
Action	**Mean**	- -	-	+	++	**Mean**	- -	-	+	++
Give 0	0.03	**98%**	1%	0%	1%	0.03	**98%**	2%	0%	1%
Give 5	0.22	**81%**	17%	2%	1%	0.16	**85%**	15%	0%	1%
Give 10	0.79	29%	**63%**	6%	1%	0.74	34%	**60%**	5%	2%
Give 15	1.37	6%	**54%**	38%	3%	1.28	11%	**53%**	34%	3%
Give 20	2.02	1%	11%	**74%**	14%	1.92	2%	14%	**74%**	10%
Give 25	2.79	1%	3%	14%	**82%**	2.85	1%	2%	11%	**87%**
Give 30	2.68	2%	5%	19%	**75%**	2.70	2%	5%	16%	**78%**

hypothesis is that there is no systematic difference between the two groups. In order to calculate the test statistics and null distribution, I use a re-sampling method.[20] Therefore, I use the individual evaluations of each subject from round 1. Specifically, for each return option, I randomly draw two subjects (disregarding the role) and check whether or not they would agree on how they evaluate the appropriateness of that return. Doing this, I generate a binary variable indicating whether or not a pair would agree on a specific return option. For each return, I simulate values of ten thousand pairs. I define the fraction of pairs that agreed as the agreement level from these samples.

As for the permutation test, I shuffle this binary variable on agreement across the two return options that are of comparison. From this resulting set of agreement values, I randomly draw two samples of 100 simulated pairs, calculate the agreement levels for these two samples and the test statistic[21]. I do this one thousand times and thereby build a null distribution. Likewise, I create the sampling distribution of the test statistic without shuffling across return levels but keep them fixed. The p-value is then given by the fraction of events, where the test statistic from the null distribution is more extreme than from the distribution with fixed return levels as they prevail in the observed data. Figure 2.2 presents significant differences between the agreement levels of the return op-

[20] I do not take the average agreement as they prevail from the matched pairs in the experiment for two reasons. First, the values would strongly rely on this one random matching realization. Second, I matched the pairs only once in the experiment to check for agreement on all return levels (but paid out for one option). These individual fixed effects could confound the analysis.

[21] As a test statistic, I use the difference in agreement levels.

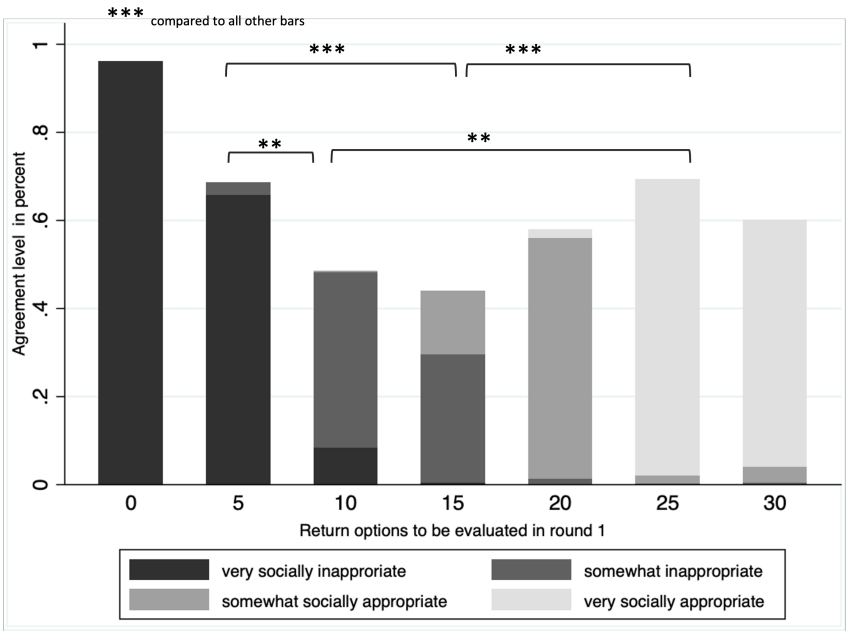


Figure 2.2. Average agreement (in percent) over return options in round 1

Note: This figure presents fractions of agreements I obtain from the evaluation frequencies of Table 2.1. For statistical testing, I use a two-sided permutation test with a re-sampling technique as described in the text. *** and ** denote statistical significance at 1% and 5%.

tions. The marked agreement over returning zero differs significantly from the agreement levels over the other returns. Furthermore, the figure confirms that the agreement over returning 25 is significantly different from the agreement over the returns 15 and 10.[22]

Result 1 *In round 1, there is a large agreement on the appropriateness of returning 25 and on the inappropriateness of returning nothing among all subjects. I find significantly more agreement over the normative evaluation of the returns 0, 5, and 25 than for the returns 10 and 15.*

As a second question, I ask whether the general normative beliefs as they prevail in round 1 change on an aggregate level in round 3. Panel B in Table 2.1 depicts the relative frequencies of evaluations from all subjects for each return in round 3. However, I do not find significant differences when comparing the evaluations between round 1 and round 3. In Table 2.2 and Table 2.3, I make the same illustration but narrow the perspective to only trustees who were punished (Table 2) and those who were not punished but received an investment in round 2. They show that for neither of these groups the social norm changes.

[22]If I use a p-value from a one-sided test, I find that the agreement level for 25 is significantly higher than for return 20, which would also comply with the directed hypothesis of more agreement for the equal split.

Table 2.2. Relative frequencies of evaluations about the appropriateness for each return for trustees who were punished in round 2

Note: Evaluations of the respective subjects in round 1 and round 3. The evaluation items are described as follows: (- -) very socially inappropriate, (-) somewhat socially inappropriate, (+) somewhat socially appropriate, (++) very socially appropriate. I assign to each item a number from 0 (- -) to 3 (++) and can, therefore, take the mean values.

	Panel A: Evaluations of **punished trustees** in round 1					**Panel B:** Evaluations of **punished trustees** in round 3				
Action	**Mean**	- -	-	+	++	**Mean**	- -	-	+	++
Give 0	0.08	**97%**	0%	0%	3%	0.11	**94%**	3%	0%	3%
Give 5	0.31	**78%**	17%	3%	3%	0.28	**78%**	19%	0%	3%
Give 10	0.81	31%	**61%**	6%	3%	0.92	22%	**67%**	8%	3%
Give 15	1.42	8%	**44%**	**44%**	3%	1.42	6%	**50%**	42%	3%
Give 20	1.97	0%	11%	**81%**	8%	1.97	0%	11%	**81%**	8%
Give 25	2.75	0%	3%	19%	**78%**	2.78	0%	0%	22%	**78%**
Give 30	2.72	3%	3%	14%	**81%**	2.75	0%	6%	14%	**81%**

On aggregate levels, nothing seems to have changed in the overall perception of the social norm. However, the potential effect of punishment on the individual norm perception might diminish as they cancel out across subjects.[23] For this reason, I analyze how the individual norm perception is affected by an experience of punishment and control for the return in round 2. Specifically, I use two different measures that shall describe the individual norm perception and allow for a comparison between rounds.

First, I look at a switching point, which I define as the the first return among all options in the range from 0 to 30 that an individual evaluates as an appropriate action. For example, an individual evaluates returns of zero and five as very inappropriate, 10 as rather inappropriate, and 15 as rather appropriate. In that case, the individual considers all actions below 15 inappropriate and switches at a return of 15 points to appropriate, so 15 would be the switching point.[24] As a second measure, I look at how trustees evaluate the return option they choose in round 2.

[23] Subjects who returned high amounts and were punished (not punished) might change their perception differently than those who returned low amounts and were punished (not punished).

[24] There is, indeed, the possibility that an individual exhibits multiple switching points, which may hint at an incoherent evaluation scheme. As that pattern implies that the individual considers one return appropriate and another higher return inappropriate again. However, there are only very few such cases, which means that the evaluations are overall coherent and consistent. In case of two switching points, I choose the lower amount.

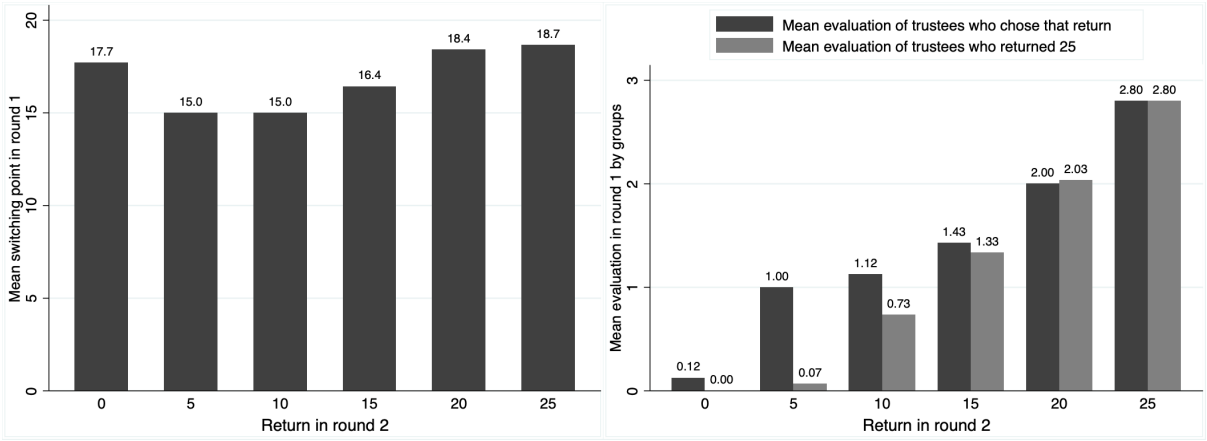
Table 2.3. Relative frequencies of evaluations about the appropriateness for each return for trustees who received an investment and were not punished in round 2

Note: Evaluations of respective subjects in round 1 and round 3. The evaluation items are described as follows: (- -) very socially inappropriate, (-) somewhat socially inappropriate, (+) somewhat socially appropriate, (++) very socially appropriate. I assign to each item a number from 0 (- -) to 3 (++) and can, therefore, take the mean values.

	Panel A: Evaluations of **not punished trustees** in round 1					**Panel B:** Evaluations of **not punished trustees** in round 3				
Action	**Mean**	- -	-	+	++	**Mean**	- -	-	+	++
Give 0	0.02	**98%**	2%	0%	0%	0.0	**100%**	0%	0%	0%
Give 5	0.13	**87%**	13%	0%	0%	0.15	**85%**	15%	0%	0%
Give 10	0.85	23%	**70%**	8%	0%	0.83	25%	**68%**	8%	0%
Give 15	1.38	6%	**53%**	**40%**	2%	1.30	9%	**53%**	36%	2%
Give 20	2.08	0%	11%	**70%**	19%	2.04	0%	11%	**74%**	15%
Give 25	2.75	0%	6%	13%	**81%**	2.91	0%	2%	6%	**92%**
Give 30	2.64	2%	6%	19%	**74%**	2.68	2%	4%	19%	**75%**

Before studying the dynamics between round 1 and round 3, I check if the trustees' behavior in round 2 is to some extent correlated with the normative perception from round 1 as argued in hypothesis 1. In Figure 2.3, the left panel presents the average switching point in round 1 for the trustees sorted by the returns they sent in round 2. It shows that the average switching point increases with the return sent in round 2, suggesting that trustees who consider higher returns to be appropriate, also return higher amounts in round 2. As argued in the hypotheses section, the increase seems more pronounced for trustees who return an amount greater than zero.

Using Pearson correlation between the return in round 2 and the switching point from round 1, I find a moderate positive relationship that is marginally significant ($p = 0.08$) if I include zero returners but highly significant ($p = 0.003$) if I exclude those subjects. A similar picture provides on the right panel in Figure 2.3. For each return option, the Figure compares the mean evaluation in round 1 of those trustees who chose that specific option with trustees who chose to return 25 points. Specifically, those who return lower amounts seem to consider this choice more appropriate than trustees who did an equal split. Using a t-test with the null hypothesis that the self-evaluations are not systematically different from the mean taking from subjects who returned 25 token, we find a marginally significant difference ($p=0.08$).



(a). Switching points **(b). Self-evaluation**

Figure 2.3. Two different norm measures in round 1 grouped by the return sent in round 2

Note: This figure presents average values of the individual norm perception of all trustees in round 1 conditioned on their returns in round 2. The individual norm perception is measured in two ways. In Panel A it displays the individual switching points from round 1, which are defined as the amount of return at which an individual starts to consider an action appropriate. In Panel B, it shows how socially appropriate subjects evaluated in round 1 a behavior they eventually exhibited in round2. Both measures are presented over the returns from round 2. As for calculations, I assigned numbers from 0 to 3 to each evaluation item: very socially inappropriate = 0, somewhat socially inappropriate = 1, somewhat socially appropriate = 2, very socially appropriate = 3.

Result 2 *The return choices in round 2 seem to be correlated with the trustees' norm perceptions in round 1. The threshold from which onward returns are considered socially appropriate (switching point) is lower for trustees who return low amounts in round 2.*

Norms treatment - The effect of punishment on norm perceptions

In the next step, I will look at the differences in the norm perception between round 1 and round 3 to see how the experience of punishment may induce a change. For this comparison, I restrict the sample to those trustees matched with an investor who invested in round 2 to ensure that the information the trustees receive only varies in the punishment experience.[25]

In a signrank test, I compare the individual norm perception between round 1 and 3. Taking the self-evaluation as a measure, I find, on average, that trustees evaluate a return option that equals the amount they sent in round 2, more appropriate in round 3

[25]Otherwise, in the group of trustees who were not punished, I could not distinguish between not being punished because the investor did not invest or because the investor chose not to punish. The information of no investment might affect norm perceptions as well and thus could interfere with the effect of punishment.

Table 2.4. Effect of Punishment on individual norm perceptions

Note: For statistical testing, I use a OLS regression. ***, **, and * denote statistical significance at 1%, 5%, and 10%. Both OLS models use as a dependent variable the difference between round 3 and round 1 for the two measures A) self-evaluation and B) switching point. In model (1), I describe the regression with overall punishment as a regressor. In model (2), punishment indicates the case where the trustee was punished and the punishment was implemented. The sample is restricted to trustees whose matched investor invested in round 2. P-values for the marginal average effect of punishment are presented in brackets below.

	A) Self evaluation		B) Switching Point	
	(1)	(2)	(1)	(2)
Marginal average effect of punishment:	-0.21*	-0.30*	0.90	1.75*
	(0.090)	(0.052)	(0.116)	(0.036)
Marginal effect of punishment at return:				
0	0.06	0.10	0.50	0.06
5	-0.03	-0.03	0.64	0.63
10	-0.12	-0.15	0.77	1.20
15	-0.21*	-0.28*	0.90	1.76*
20	-0.31**	-0.40**	1.04	2.33**
Observations	89	77	89	77

than they did in round 1 (p= 0.03). If I now split the sample into those trustees who experienced punishment and those who were not punished, the significant effect on the norm perception over rounds only prevails for those not being punished (p<0.01)[26]. If I use the switching point as a measure for individual norm perception, I do not find significant effects for either group.

Table 2.4 depicts results of an OLS regression where I use the differences in the two different measures of individual norm perception from round 1 and 3 as dependent variables. For the regression model, I include the binary variable punishment as a regressor, control for the return from round 2 and include an interaction term between these two variables. I add this latter component as I hypothesize that the effect of punishment strongly depends on the return a subjects is punished for. Furthermore, the control for return is important as the groups of punished and not punished subjects are systematically different in this dimension (subjects who returned low amounts are more punished than those with high amounts [27]).

[26] I find the same result when I exclude trustees who have returned zero in round 2 and, therefore, might be less likely to change their norm perception as described in the hypothesis part.

[27] See in the Appendix, Table 2.6 presents raw data on the prevalence of punishment by returns in round 2.

I use two different models (1) and (2), where I distinguish between two concepts of punishment. Even if I focus on the effect of the norm-signalling function of punishment, the actual implementation of the punishment may still be an important factor in this context. For instance, it could be possible that the payoff reduction will make the punishment more salient to the trustee and, therefore, amplify the effect.[28] Hence I look at punishment in two different ways. In the model (1), I take into account every punishment decision where p = 1, irrespective of its implementation (d=0 or d=1). In addition, in the model (2), I only include punishment if it was implemented[29] (d=1) in order to see whether the effect of punishment is more pronounced in that case.[30]

As for better interpretation of the overall effects of punishment and the effect conditioned on return in round 2, I refer in the Table 2.4 to marginal average effects and marginal effects for each return. The coefficients presented in this table refer to the effect on the differences between round 3 and round 1. Hence they present the difference in the change of normative perceptions (the self-evaluation or the switching point).

In the left panel, we find the results for the regression models where the change in self-evaluation is the dependent variable. Hence in that case we look the change in the evaluation of one's own action between round 3 and round 1. If, for instance, the change is positive this means that subjects would perceive the action more favorable than before. In the table we find significant marginal average effects of punishment on the change in the evaluation of one's own behavior (p=0.09 and p=0.05). Thus the negative coefficients in both models suggest that subjects who were punished increase the social appropriateness of their behavior, they exhibited in round 2, from round 1 to round 3 less pronounced than those subjects who were not punished. This result shows for both models whether or not I restrict punishment to implemented punishment. This first result aligns with the hypothesis (2) in that it suggests an alleviating effect of being punished on normative evaluation. However, it goes somewhat in another direction than expected. The experience of punishment does not seem to make the people evaluate norm deviations

[28] Experimental studies show that the payoff reduction seems to have a separate effect on norm compliance (Xiao, 2018; Masclet et al., 2003).

[29] Consequently, I discard those observations where punishment was intended but not implemented. These are 12 observations in total.

[30] Due to power, I can not disentangle the effect between the two types of punishment, though.

less appropriate. Instead, trustees who were not punished seem to update their belief, such that they evaluate their behavior more appropriate than initially. It might be that the experience of being not punished serves as a signal relative to the norm deviation and makes people evaluate a deviation of the social norm more favorably.[31] Whereas the experience of punishment prevents people from eroding their norm perception.Moreover, this effect becomes stronger for higher returns. Therefore, it seems that only for returns that are not too strongly deviating from the norm, and thus where ambiguity exists, the experience of not being punished serves as a signal of a norm.

The right panel of Table 2.4 depicts the marginal effects of the regression of the difference of switching points as a dependent variable. As the coefficients for punishment are positive, this hints in the same direction. It appears that especially for high returners, who yet deviate from the equal split, the experience of punishment increases the switching point to a higher return. Respectively, trustees who were not punished lower the amount for which they consider behavior appropriate. However, the effect is only significant for model (2) and thus only for punishment where an actual payoff reduction is implemented (p=0.036).

Result 3 *Whereas overall, subjects seem to perceive lower returns more favorable in round 3 than in round 1, subjects who experienced punishment in round 2 seem to do less so.*

Behavior treatment

From the previous results, I concluded that subjects who experienced punishment do not seem to evaluate their behavior less appropriately than before. Instead, subjects who experienced no punishment tend to evaluate their behavior more appropriately than they did initially.

In this part, I focus on the behavior treatment and look if observed norm perception dynamics manifest in behavior. However, considering the previous results, one might also expect behavioral differences that are slightly different from what I hypothesized.

[31] However, learning could also exist in a setting without feedback where no punishment option exists. In an experimental study, Weber (2003) shows that in a competitive repeated guessing game learning takes place even in a situation where subjects do not experience outcomes from previous rounds. Hence, in the context of norms learning might also happen without the experience of the normative signals by others. To test if the erosion of the norm really comes from the signal of no punishment, one would need to check how the normative perceptions evolve over the three rounds when no punishment option exists in round 2 and if it is systematically different from the case where subjects were not punished.

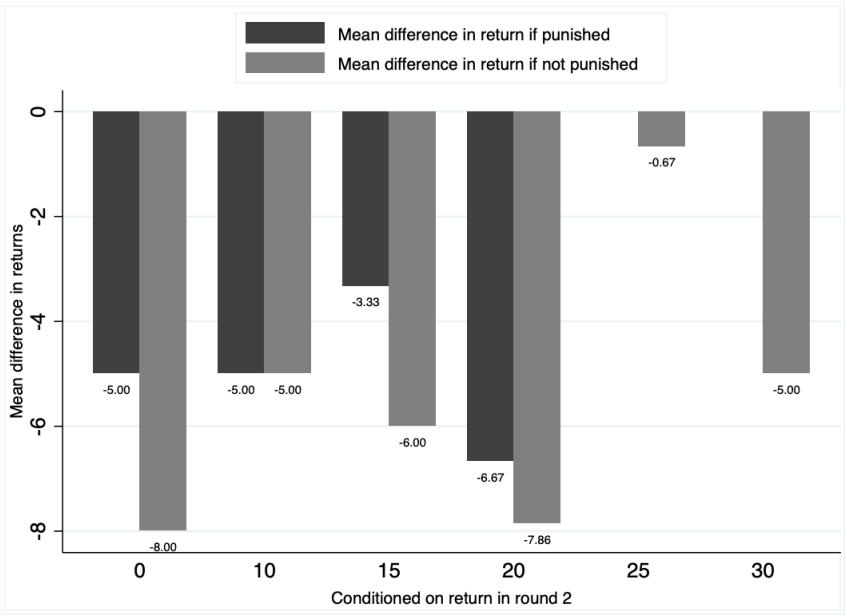


Figure 2.4. Mean difference in returns from round 1 to round 3 conditioned on return in round 2

Note: This figure presents the average difference in returns of trustees from the behavior treatment. The return differences are grouped by the return behavior in round 2 and distinguishes between punished and not punished trustees. I restricted the sample to trustees whose investors invested in round 2.

As the punishment prevents trustees from considering a norm deviation as more appropriate than before, punishment might therefore rather lead to a smaller reduction in return. Figure 2.4 presents at least a tendency for such behavior. Conditioned on the return sent in round 2, the Figure depicts the differences in return between round 3 and round 1 distinguished by punished and no punished trustees. The reduction in return is smaller for punishment. However, even if I find a significant decrease in returns from round 1 to round 3 for both groups (signrank, $p<0.001$ in both groups), this change is not significantly different between punished and not punished subjects.

Thus we see - if anything - a decrease in behavior that is less severe for those being punished. This observation aligns with the previous finding that those who were not punished feel instead encouraged to be more selfish in the next round. Overall, this would show that punishment is not necessarily there to install a norm but can help not let deteriorate a norm, so it will be necessary to send the signal several times to keep people reminded.

2.6 Conclusion

The results presented in this paper shed new light on the norm-signaling function of punishment. In contrast to existing empirical research that analyzes how punishment interacts with social information as an additional component, I take a step back and only focus on the message sent by the punishment itself. I can show that the experience of being punished seems to affect the normative perception of actions where no clear agreement over the appropriateness or inappropriateness prevails. However, the results indicate that the experience of punishment can not help elevate the social norm by making people perceive norm deviations as more inappropriate than before, but it instead seems to prevent norm erosion. Specifically, I find that people who were not punished perceive norm deviations more appropriate, whereas those who were punished significantly differ from this tendency and do not change their normative perceptions. This finding leads to the following three insights, which provide room for further investigation on the evolution and enforcement of social norms on an individual level.

First, this result gives reason to assume that the inaction in a situation where punishment is possible can send a message and thereby approve tacitly the observed behavior. Especially for actions where normative ambiguity exists, subjects who deviate from the social norm but did not receive a punishment might even overweight this signal and use it for a motivated belief update. Hence, at least in the setting that I investigate, the social norm appears to be a fragile construct that needs to uphold constantly. While punishment may serve as a tool to sustain a norm, the evidence presented does not indicate its effectiveness in shaping the norm or even steering it to another behavior as perceived appropriate.

The previous argument leads to a the second thought. In the literature on norms and enforcement of pro-social behavior, punishment is considered most widely a crucial and outstanding mechanism to be in place and explain the sustaining prevalence of social norms. However, whereas sanctions by institutions or peers have been shown to be effective in enforcing an existing norm, they might not explain the origin or change of norms. Although this study shows that punishment as a marginal intervention by one individual could sustain another subject's norm perceptions, it does not seem to have the power to give new inspiration to shift it. Hence, it might be the next step to study how other forms of interaction on an individual level have another function in inspiring people to change their normative perceptions. Thus, referring to existing literature on the effect of

role models and voluntary leadership (Rivas and Sutter, 2011), it might be powerful to see someone else in the same situation behaving differently and setting a new standard that might also translate to a new perception of the norm.

Third, the result might provide a new narrative on why people punish at all. In light of its relevance explaining the motivation to cooperate among strangers, it seems relevant to explain the motivation for punishment. However, there is no compelling answer why especially in stranger matching situations, people would incur a cost to punish someone else if she does not expect to meet this punished person in the future again. This paper presents results indicating that punishment is effective beyond its actual implementation. Specifically, I find that punishment prevents norm erosion in situations where no enforcement is possible. People might take this into account when they make their punishment decision.

2.7 Appendix A: Additional Figures and Tables

Table 2.5. Raw data on decisions for investors and trustees per round for each treatment

	Behavior treatment			*Norms treatment*		
	Invest	Return	Punish	Invest	Return	Punish
Round 1	0.78	15.1				
Round 2	0.74	12.7	0.3	0.92	15.4	0.37
Round 3	0.54	10.6				

Table 2.6. Frequencies of return decisions and punished returns by treatment

	Return options for trustee						
	0	5	10	15	20	25	30
Behavior treatment							
total	33	0	8	10	16	23	1
of this punished	20	0	1	3	3	0	0
Norms treatment							
total	25	1	8	7	20	36	0
of this punished	5	1	4	6	4	1	0

Table 2.7. Relative frequencies of evaluations about the appropriateness for each return for trustees

Note: Evaluations of all subjects independent of their roles in round 1 and round 3. The evaluation items are described as follows: (- -) very socially inappropriate, (-) somewhat socially inappropriate, (+) somewhat socially appropriate, (++) very socially appropriate. I assign to each item a number from 0 (- -) to 3 (++) and can, therefore, take the mean values.

Panel A: Overall evaluations in round 1						**Panel B: Overall evaluations in round 3**				
Action	**Mean**	- -	-	+	++	**Mean**	- -	-	+	++
Give 0	0.04	**98%**	1%	0%	1%	0.04	**98%**	1%	0%	1%
Give 5	0.20	**84%**	14%	1%	1%	0.19	**84%**	15%	0%	1%
Give 10	0.80	28%	**65%**	6%	1%	0.82	27%	**65%**	7%	1%
Give 15	1.36	7%	**52%**	39%	2%	1.32	8%	**54%**	36%	2%
Give 20	2.01	0%	12%	**74%**	13%	1.97	0%	14%	**74%**	11%
Give 25	2.76	0%	4%	15%	**80%**	2.85	0%	1%	13%	**86%**
Give 30	2.63	2%	5%	20%	**73%**	2.68	1%	5%	19%	**75%**

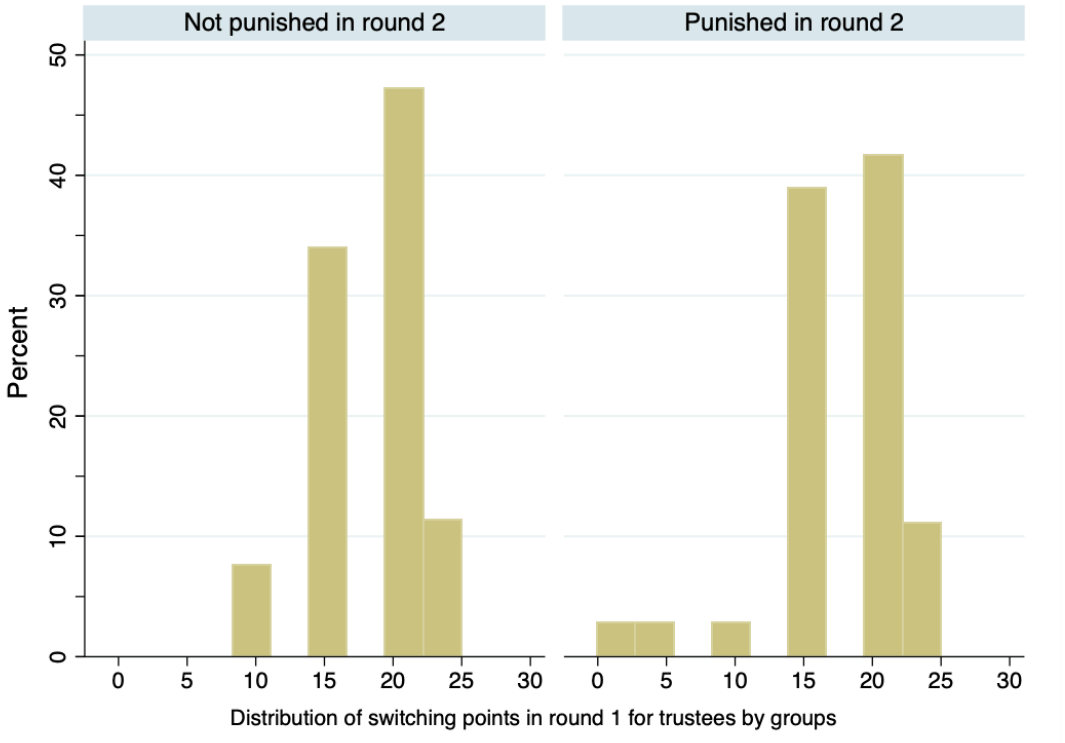


Figure 2.5. Distribution of switching points in round 1 of trustees

Note: This figure presents the frequencies of switching points among the groups of those subjects who were punished in round 2 and those who were not punished. The sample is restricted to trustees who received an investment in round 2. The switching point is defined as the threshold taken across the return options in an order at which a subject first evaluates an option appropriate.

2.8 Appendix B: Instructions and Questionnaire

Instructions[32]

Consent

We welcome you to this online study.

You will participate in a **study on economic decision behavior** and will be asked to make some decisions.

During today's session, you will **interact with other participants** via the computer. Thus, it may be the case that you respond to a decision made by another participant or that another participant is asked to respond to your decision. Therefore, we ask that you **remain focused on this study** throughout the duration of this session.

For your participation in this study, you will receive a **fixed amount of 2,50 Euro**. In addition, **you can earn additional money**.

During the study, your income will be calculated in tokens. At the end of the study, your income will be converted, where **5 tokens are worth one Euro**. The payment of your income will take place via Paypal. You will receive your money within 24 hours.

Your participation will be completely anonymous. This means that the participants you interact with during the study will not know anything about your identity. We will also ensure that your choices are never linked to your personal information. For the payout via Paypal, the request for your contact details will be made by the Cologne Laboratory for Economic Research (CLER). Your personal data can thus not be linked to your decisions.

Your **participation is completely voluntary** and you can withdraw from the study at any time without giving reasons. However, in order for us to use your decision data

[32]Note: I only attach here the instructions for the Norms treatment, because it is a longer version of the instructions I provided to the Behavior treatment. I used the exact wording except for the introduction in Part 1 and 3, where I explained that they will make the decisions (instead of evaluating them). Furthermore I explain the matching in these two parts differently as they are matched with a partner of a different role in part 1 and 3. The evaluation task is excluded. Instructions translated from German. German instructions and instructions for Behavior treatment available upon request.

for research, it is necessary for you to complete all parts of this study. If you decide to withdraw from participation during the study, you will receive 2.50 Euros for your appearance, but not any other amounts you may have earned during the study.

To ensure that the study can be completed by the end of the specified time period of this session, **we use time limits**.

This means that you will always see a specific amount of time when you are asked to make decisions or provide answers. It is **important that you always make your decision or submit a response in the allotted time**. The **time is very generous so that you can make decisions or give answers without rushing**. If you exceed the time limit, we will have to exclude you from further participation (you will still receive 2.50 euros for your participation).

By going to the next screen, I agree to participate in this study.

General information about the experiment

At the beginning of the study, all participants are randomly assigned to one of **two possible roles: Player A or Player B**. You will be informed via the screen which role you have been assigned to. You will take the **same role throughout the study**.

This study is divided into **three parts**. First, we will inform you about Part 1. Only when Part 1 is completed will you receive the information for the following Part 2. The same is true for Part 3, which you will not be informed about until you have completed Part 2.

After all three parts of this study are completed, **one of the three parts will be randomly selected** . **Your payment will be based on the income you have earned in this one - randomly selected - part**.

You will first receive detailed information about part 1. Please click on "Next" to continue.

Role assignment

You are player A / player B .

You have been assigned to another participant in this study with whom you will interact only in Part 1.

On the next page you will get more information about part 1. To do this, please click on "NEXT".

Detailed information about Part 1

We provide you with **a decision situation in which participants from another session actually found themselves**. Here, player A and player B have interacted with each other.

In Part 1, we ask you how socially appropriate you would rate the various courses of action between which Player B can choose. **Your income** in this Part 1 will **depend on your own evaluation and the evaluation of others in this session**.

Note that the decision situation now presented will be relevant not only for this first part but also for the second part. More specifically, **in the next part, you will find yourself in a similar situation**, making a decision **in your role as player A/ player B**.

Below you will learn more about the decision situation. After that, we will explain how you can earn money with your evaluation.

Decision situation

At the beginning of Part 1, groups of two are formed again, with one player A randomly assigned to one player B at a time.

Player A receives an **endowment of 15 tokens**. **Player B** also receives an **endowment of 15 tokens**.

Player A's decision

Player A decides whether to invest or not to invest:

- **Invest:** Player A invests **10 tokens from his 15 tokens endowment**. These 10 tokens are multiplied by 4, and **40 tokens are distributed to player B**. Player B can then give a share of these 40 tokens to Player A.
- **Not invest:** Player A does not invest. Therefore, **no tokens are distributed to player B**. Player B cannot give any tokens to player A.

Player B's decision

Player B **decides how much** of his 40 tokens he would **give to Player A**, in case Player A has actually invested. For this, player B can choose between the following seven options:

I give to player A:
0 tokens
5 tokens
10 tokens
15 token
20 token
25 token
30 token

Please note that **every player B makes this decision**. However, the decision is **only implemented if the assigned player A** has actually **invested** 10 tokens beforehand, and player B thus has 40 tokens at his disposal.

How the other player has decided, both find out only at the end of the study after all parts have been completed.

Income of players A and B

Calculation of income

1. If player A invests:

Income of player A = 15 token (initial endowment) - 10 token (investment) + x token (return of player B)

income of player B = 15 token (initial endowment) + 4*10 token (invested by player A) - x token (return of player B)

2. If player A does not invest:

Income of player A = 15 tokens (initial endowment)

Income of player B = 15 tokens (initial endowment)

Calculation of income

Player B returns:	**Player A does not invest**		**Player A does invest**	
	Player A	Player B	Player A	Player B
0 tokens	15	15	5	55
5 tokens	15	15	10	50
10 tokens	15	15	15	45
15 tokens	15	15	20	40
20 tokens	15	15	25	35
25 tokens	15	15	30	30
30 tokens	15	15	35	25

Your evaluation task

In this Part 1, we will ask you to **evaluate all seven possible choice options** that Player B can choose between.

For this, we will ask you the following question: Imagine that player B is in the described decision situation and decides how many tokens he wants to give to player A out of the 40 tokens. How socially appropriate do you think player B's behavior is?

For your evaluation, you can choose from the **following scale**: 'very socially appropriate (=3)', 'rather socially appropriate (=2)', 'rather socially inappropriate (=1)', and 'very socially inappropriate (=0)'.

'Socially appropriate' should be understood as a **generally accepted behavior** that describes **what one should do**. In contrast, **socially inappropriate behavior** can be thought of as an action that is generally **considered unacceptable** and to which one might **expect angry reactions**.

Your evaluation (just for illustration on this screen)

Player B returns:	Very socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Very socially appropriate
... 0 tokens to player A	0	1	2	3
... 5 tokens to player A	0	1	2	3
... 10 tokens to player A	0	1	2	3
... 15 tokens to player A	0	1	2	3
... 20 tokens to player A	0	1	2	3
... 25 tokens to player A	0	1	2	3
... 30 tokens to player A	0	1	2	3

Your income

Please note that for one of the parts, the tokens will be converted into euros **(1 Euro = 5 tokens)** and the income will be paid to you.

In this part, your income is composed of a **fixed amount of 15 tokens plus a possible bonus of 15 tokens**

Whether you receive the additional bonus **depends on your evaluation and that of another participant** in this session. For this, you were randomly assigned to a participant from your session.

From the seven choice options of player B, which you are to evaluate, **one is chosen at random**.

We then compare how you and your assigned participant rated this action option. **If both evaluations match**, you and your assigned participant will **each receive 15 tokens**.

Note that you could be assigned to any participant from this session - regardless of their role. **To maximize your chances of receiving a bonus, you should rate each action option as you think most participants in their session do.**

Detailed information about Part 2

You will **make a decision yourself** in this part 2. You will find yourself in a **similar decision situation as described in part 1**. Here, first player A makes the decision whether to invest and player B how much he would give away. So you can read here again exactly the same information about these two decisions.

In the **next tab on your screen** you will find the **information about the additional decision of player A**.

Recap: Decision situation part 1

Comment: participants find here the same texts as provided above in the part 'Decision situation'.

Information about the additional decision of player A

In case player A has invested 10 tokens, player A will be informed how many tokens player B has sent back. **Player A can make an additional decision as a reaction** if he wants **to punish player B**. On the other hand, player A who has not invested, will not be informed about player B's decision. For this reason, **only player A who has invested, can make an additional decision**.

Player A can choose between the following options in this case:

- **Punish**: Player A pays 1 token, and in return, player B's income in part 2 is reduced by 10 tokens.
- **Not Punish**: Player A will not incur any costs. Player B will not have anything deducted from his earnings.

If player A punishes player B, a random number generator decides whether the punishment will actually be enforced. That is, **with a probability of 50 percent**, the penalty will be enforced, **player B will actually be deducted 10 tokens**, and **player A will incur the cost of 1 token**. Otherwise, no tokens will be deducted from player B, and there will be no cost to player A.

In any case, player B will be informed about player A's decision(s). That is, Player B is notified of one of the following cases at the end of Part 2:

Player A has not invested.
OR
Player A has invested and has not punished Player B.
OR
Player A has invested and has punished Player B. The penalty was enforced.
OR
Player A has invested and punished player B. The penalty was not enforced.

Income table - depending on the decisions made by player A and player B.
Please note that for one of the parts, the tokens will be converted into Euros (1 Euro = 5 tokens), and the income will be paid to you.

Player B returns	**Player A does not invest**		**Player A does invest**			
	Player A	Player B	Punishes AND punishment implemented		Does not punish OR punishment not implemented	
			Player A	Player B	Player A	Player B
0 tokens	15	15	4	45	5	55
5 tokens	15	15	9	40	10	50
10 tokens	15	15	14	35	15	45
15 tokens	15	15	19	30	20	40
20 tokens	15	15	24	25	25	35
25 tokens	15	15	29	20	30	30
30 tokens	15	15	34	15	35	25

Detailed information about Part 3

In this part, you find yourself in **the identical decision situation as described in part 1**. Again, as in part 1- , **we ask you how socially appropriate you would rate the possible choice options** between which player B can choose.

Please note that your income in this part depends on your own evaluation and the evaluation of others in this session.

You can therefore reread here **exactly the same information** about these two decisions **as in Part 1**.

Recap: Decision situation part 1

Comment: participants find here the same texts as provided above in the parts 'Decision situation' and 'Your evaluation task'

Questionnaire

Questionnaire after the information about Part 1

Before you can start with Part 1, we ask you to answer the following comprehension questions.

Please press 'Next' below if you want to submit your answers. If you have pressed 'Next' and **one or more answers are incorrect**, you will be notified and can **just try again**.

Tip I: You can press the 'Evaluation Task' or 'Decision Situation' tab at the top of your screen to scroll back and look up the information.

Questions for the evaluation task

1.Your evaluation of Player B's choice options is compared to the evaluation of your assigned participant. If your score and the other participant's score for a randomly chosen action option match, you both receive a bonus of what amount each?

Questions about the decision situation of players A and B

2. How many tokens does player A receive as endowment?
3. How many tokens does player B receive as endowment?
4. Player A can send either 10 tokens or none to player B.
5. If player A decides to send 10 tokens, how many tokens will be distributed to player B?

Tip II: It is best to look at the income table for players A and B for the following questions.

Income table - depending on the decisions made by player A and player B

6. *Comment: Income calculations for player A and B for various scenarios. These questions were randomized to avoid priming*

If player A invests, and player B sends back 5 tokens, what is the income of...

Questionnaire after the information about Part 2

Before you can start with part 2, we ask you to answer the following questionnaire. Please press 'Next' below if you want to submit your answers. If you have pressed 'Next' and one or more answers are incorrect, you will be notified and can just try again.

1. If player A decided to punish player B and the punishment is not enforced, does player A still bear the cost of 1 token?

2. Player B will only know if Player A intended to punish him if the punishment is enforced.

3. If player A decides to punish player B and the punishment is enforced: How many tokens will be deducted from player B?

4. Player A decides to punish Player B after Player B has not sent any tokens back. If the punishment is enforced, what is the income of players A and B.

5. Player A decides not to punish player B after player B has not sent any tokens back. What is the income of players A and B.

Questionnaire after the information about Part 3

Comment: No questionnaire for Part 3.

CHAPTER 3
WHISTLING IN THE WIND?
THE EFFECT OF THE EXTERNAL WHISTLEBLOWING REGIME ON INTERNAL REPORTING

3.1 Introduction

Whistleblowing by employees has become an increasingly recognized instrument by policymakers for uncovering breaches of law in organizations.[33] Yet, there still exists a large disagreement on which channels of whistleblowing should be protected by the law. In fact, whistleblowers can usually choose between the internal and the external channel. This means they can either share information about a wrongdoing within the organization, i.e., with the supervisor or a designated compliance officer, or they directly approach an external institution such as a prosecuting authority or the media. And there is a trade-off between these two options. On the one hand, the external channel is considered reliable in taking action against the wrongdoing. Whereas for the internal channel, the incentives for recipients of a whistleblowing report to prosecute and sanction are less clear. People within the organization might themselves be involved in wrongful activities or they just consider the report snitching other colleagues. On the other hand, the external channel may incur more damage for the organization. External whistleblowing usually takes away the control over information and makes the allegations public, which could cause an enormous reputational loss.[34]

Thus, among policymakers, the question arises how effective internal whistleblowing can be on its own and to which extent an external whistleblowing regime can add to or even interfere with the self-regulatory function of an organization. In the course of the drafting

[33] As a legislative response to corporate scandals and in the aftermath of the financial crisis, in the US the Sarbanes-Oxley Act in 2002 (https://www.investor.gov/introduction-investing/investing-basics/role-sec/laws-govern-securities-industrysox2002) and the Dodd-Frank Act 2010 (https://www.govinfo.gov/content/pkg/PLAW-111publ203/pdf/PLAW-111publ203.pdf) were enacted to better protect and promote whistleblowing. In 2019 also the EU launched a Directive (https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L1937) to provide a comprehensive protection scheme for whistleblowers.

[34] See, for example, the case of Oxfam where whistleblowers have disclosed cases of sexual exploitation: "This also changed the public perception of overseas charity workers and damaged the charity sector as a whole" (See https://www.europeanceo.com/business-and-management/whistle-while-you-work-the-benefits-of-corporate-whistleblowing/). Furthermore, empirical evidence suggests that whereas for external whistleblowing cases long term negative effects on a company's performance prevail (Bowen et al., 2010), this seems not to be the case for internal whistleblowing (Stubben and Welch, 2020).

period of the EU Directive on the protection of whistleblowers, for example, there was indeed a substantial debate about which whistleblowing to protect and whether to treat both channels as identical options for the eligibility of protection.[35] In fact, policymakers can easier design the potential consequences of external whistleblowing for the organization, while this is much less the case for the consequences of internal whistleblowing.

In this paper, we take first empirical steps in this debate and ask how the provision of an external whistleblowing channel affects the overall rate of reported and sanctioned wrongdoing when internal whistleblowing is also possible. Furthermore, we are interested in whether the effectiveness of internal and external channels depends on the possible consequences of the external channel for the organization. For this, we study the effectiveness of various external whistleblowing regimes by comparing the aggregate outcomes of disclosed and sanctioned wrongdoing. In addition, we ask by whom the different channels are actually used. Therefore, we look at the individual strategies to see to which extent honest behavior correlates with the usage of each whistleblowing channel and the acceptance of internal whistleblowing. We also investigate whether moral preferences for loyalty or fairness could explain the choice for the whistleblowing channel.

To be able to exogenously vary the possible consequences of external whistleblowing we design and conduct a laboratory experiment. It allows us to observe the full prevalence of misbehavior as well as the actual fractions of lying disclosed and punished by internal or external whistleblowing. In contrast, in the field, misbehavior is usually difficult or even impossible to observe if not disclosed. The same is true for the act and the consequences of internal whistleblowing since it usually occurs outside public surveillance. We form experimental groups in which a subject is either assigned the role of the public or a firm member. The firm members report their individual performance from a previous task to the public. They have an individual benefit from over-reporting, which is, however, socially costly because it harms the public. Within each firm, the members observe each others' lying and can decide if they want to blow the whistle. For this, they can share the information either with someone in the firm or with the two external public members.

[35] See https://www.politico.eu/article/whistleblower-protection-rules-held-up-reporting-clash-european-parliament/ from 28 February 2019. A similar debate arose in the course of the enactment of the Dodd-Frank Act in 2010. The law's focus on the external channel for reward provision raised deep concerns that it could undermine self-regulatory competencies of a company (e.g. https://www.businessroundtable.org/archive/resources/business-roundtable-letter-to-sec-on-whistleblower-provisions).

Thus, they have the choice between internal and external whistleblowing. The reported firm members will be punished and the public compensated, only if the subjects, who are informed about the lies, eventually decide to prosecute. In this basic setup, two channels are available, which solely differ in the recipient of the information about the lie. It serves as our primary treatment (EXTERNAL). As a control treatment we look at a situation where team members can only report internally (NoEXTERNAL). Furthermore, we introduce two more treatments to see how the consequences of external whistleblowing may affect the overall effectiveness of fraud disclosure. These treatments are based on our primary treatment EXTERNAL. Specifically, in one treatment (DAMAGE), a successful act of external whistleblowing comes with damage to the firm, which is a small cost for every member of the firm, whether she lied or not. In another treatment (INVESTIGATION) a successful external whistleblowing act follows a firm-wide investigation and leads to a punishment of every member who lied.

We assume at least two different motives that may drive whistleblowing behavior. First, subjects will whistleblow because they value the outcome of a successful report, which is a redistribution from the wrongdoer back to the public. Second, subjects may feel a "warm glow" about their own altruistic and potentially truth-correcting act, and thereby derive a private gain from the action of whistleblowing itself, which is independent from the success of whistleblowing. We therefore expect that whistleblowing already occurs in a situation where only the internal channel exists. If an external channel is available, whistleblowers might switch from the internal to the external channel. In addition, there might be subjects willing to blow the whistle externally but not internally. The negative externalities of external whistleblowing in the treatments DAMAGE might lead some whistleblowers to switch to the internal channel or fully abstain from whistleblowing. The same might be even more the case in INVESTIGATION where the existence of an external channel could lead to even more sanctions.

As a first finding, we see that if external whistleblowing is possible the fraction of lies that is sanctioned from the number of committed lies increases from 8 in NoEXTERNAL to almost 30 percent in EXTERNAL. Thus, the provision of an external whistleblowing channel clearly makes a difference and increases the effectiveness of whistleblowing overall. The result could be driven by at least two effects. A higher fraction of sanctioned lying may result from a higher acceptance rate by the recipients of the whistleblowing

reports. This is indeed the case. Whereas over the internal channel, slightly more than 20 percent of the whistleblowing requests follow consequences, external whistleblowing almost certainly leads to punishment. Another explanation that could drive the result is that if an external channel exists the total number of people who blow the whistle increases because people expect a better chance to implement sanctions. This, however, is not the case. We do not find a significant difference in the overall total reporting (external and internal) across the two treatments. Instead, we observe almost a complete shift from the internal to the external channel, which is eventually more effective. Whistleblowers seem to care about the actual consequences of their action but, at the same, time are also motivated by the "warm glow" of their altruistic action such that a more effective channel will not encourage more people to whistleblow.

Furthermore, we find that the share of whistleblowers significantly drops in DAMAGE compared to EXTERNAL. We see that more people use the internal channel in DAMAGE and the total amount of sanctioned wrongdoing decreases compared to EXTERNAL. In contrast, in treatment INVESTIGATION, the amount of whistleblowers does not change compared to EXTERNAL. However, we see that internal reporting in INVESTIGATION is more effective compared to EXTERNAL. This can be explained by the fact that in INVESTIGATION more people that lied themselves are willing to accept requests compared to EXTERNAL.

Overall, whereas in previous experimental research, external whistleblowing is analyzed as the only option for disclosing observed misconduct, we add the internal reporting channel in our study. Thus, our experimental design mirrors a decision situation employees usually face. We find that the external channel can indeed increase the amount of disclosed and sanctioned corporate fraud. However, the higher effectiveness does not seem to motivate more people to report. Thus, it might be the same set of people who would report internally and those who report externally if a more effective external whistleblowing channel becomes available. This finding suggests that even if the provision of an external channel cannot help increase the number of whistleblowers, external channels nevertheless seem desirable if the internal systems are ineffective in sanctioning fraud. Furthermore, if the whistleblowing procedures are well designed our findings suggest that external whistleblowing may even help to strengthen the self-regulatory forces within a company.

The remainder of the paper is organized as follows. In section 2 we give a short review of related literature, section 3 describes the experimental design and procedures, section 4 presents empirical finds and section 5 concludes.

3.2 Related literature

There is a long strand of non-experimental research in social sciences investigating whistleblowing. Near and Miceli (1985); Miceli and Near (1984, 1996, 2002); Miceli et al. (2008) use survey data and, thereby, provide valuable insights in how individual and organizational traits correlate with the decision to blow the whistle.[36] Whereas a prominent and commonly accepted definition of whistleblowing by Near and Miceli (1985)[37] includes internal and external whistleblowing, the distinction between the channels has so far played a minor role.

In fact, research in the field has mostly focused on external whistleblowing, because it is easier to observe. Thus, in a study using field data, Dyck et al. (2010) analyze a large number of corporate fraud cases in the US in order to assess the actual prevalence of external whistleblowing in companies. The authors show that whistleblowing by employees is prevalent and even play a key role. Besides of its role in disclosing misconduct, Call et al. (2018) ask if external whistleblowing can help increase the financial outcome of prosecution. For this, the authors look at a large US data set on enforcement actions with and without whistleblower involvement. As a result they find that the presence of whistleblowing is associated with higher enforcement outcomes, e.g. larger monetary sanctions for the targeted firm, higher prison sentences for targeted employees and quicker starts of enforcement actions. Thus, employee whistleblowing is a highly relevant source that can even make the process faster and more effective. Considering that employees usually do not have any monetary benefit from whistleblowing [38] but may even be exposed

[36] It is difficult, however, to identify causal effects on whistleblowing from survey data. For example, the authors find that among those respondents who observed wrongdoing, non-whistleblowers reported to have gathered less evidence than whistleblowers. The authors add for considerations that this correlation might be explained by a post-decisional justification and more research for causal identification should be done. (Near and Miceli, 1985)

[37] "[...]the disclosure by organization members (former or current) of illegal, immoral, or illegitimate practices under the control of their employers, to persons or organizations that may be able to effect action." (p.4)

[38] An exception is the Whistleblower Program by the Securities and Exchange Commission (SEC) in the US. The Commission is authorized by Congress to provide monetary awards to eligible individuals who

to retaliation, the question arises of what employees' motivation for blowing the whistle is.

However, even if external whistleblowing is observable, it is still difficult to study in the field because the data are usually restricted only to those cases where whistleblowing actually happened. In contrast, in the laboratory, the total amount of wrongdoing and therefore also the fraction of the cases where people remain silent can be observed. Due to this methodological advantage, there is a growing experimental literature studying the determinants of whistleblowing motivation of employees.[39] In one of the first experimental papers, Reuben and Stephenson (2013) find that people seem intrinsically motivated to report an observed wrongdoing. Butler et al. (2019) investigate how monetary incentives could affect the willingness to blow the whistle in varying environments of public scrutiny. Similar to our setting, they form two different kinds of groups indicated either as firm or public[40], where wrongdoing can happen in a firm which harms a corresponding public. Overall, they find that monetary rewards help increase whistleblowing, and do not seem to crowd out intrinsic motivation arising from the expected public approval. Schmolke and Utikal (2018) also study the effect of monetary incentives on whistleblowing for varying consequences of the observed wrongdoing. They find that monetary rewards increase whistleblowing in situations where employees observe misconduct they profit from. If the observed wrongdoing harms the potential whistleblower, whistleblowing rates are already very high even without monetary incentives.

In another recent paper by Muehlheusser et al. (2020), the authors study the effectiveness of whistleblower protection measures. In their setting, they also consider an employee who could observe and report a wrongdoing committed by her employer, which harms an uninvolved third party. The observing employee could make the report to someone who has a monetary incentive to investigate and a successful investigation leads to a compensation for the third party. This feature is similar to the external channel in the experiment in the present paper. In contrast to our setting, however, in their design fraudulent reports are possible. The results suggest that a protection regime increases

come forward with high-quality original information that leads to a Commission enforcement action in which over 1,000,000 dollars in sanctions is ordered. The range for awards is between 10 percent and 30 percent of the money collected (https://www.sec.gov/whistleblower).

[39]There is another strand of experimental literature on whistleblowing research (Abbink and Wu, 2017; Apesteguia et al., 2007; Bigoni et al., 2012; Feltovich and Hamaguchi, 2016; Hinloopen and Soetevent, 2008; Buckenmaier et al., 2018), which rather focuses on whistleblowing as a strategic tool in context of collusion and leniency programs.

[40]The authors use that framing in their experiment, while we use neutral language.

the number of successful investigations against wrongdoers and decreases misbehavior in the first place, but only if it specifically targets non-fraudulent whistleblowers.

Apart from situational factors, there are a few other studies focusing on personal traits and individual valuation of moral principles. Waytz et al. (2013) find that conflicting norms of fairness and loyalty create a dilemma for potential whistleblowers, which makes the decision to blow the whistle in the end depend on the individual's weighing of these two norms. Similarly, Bartuli et al. (2016) find that the personal valuation of fairness is highly correlated with whistleblowing.

All those studies described above (except Reuben and Stephenson (2013)) have in common that they look at a hierarchical structure between the potential whistleblower and the wrongdoer. This describes a situation in which the employee can only observe and report misconduct by her supervisor, but is not herself exposed to the possibility of doing wrong. In contrast, we study a situation where the potential whistleblower also has an incentive to commit wrongdoing and, therefore, might find herself in a different situation to judge and report another's misbehavior. We argue that such a situation is important to address because it is common among organizational members and allows us to study how one's own compliant behavior is potentially correlated with whistleblowing and the willingness to process other's whistleblowing requests.

Furthermore, the review of experimental literature shows that research has primarily focused on external whistleblowing. In the studies mentioned above, the employee could report either to an external person with a strong incentive to implement sanctions or to the computer, which always sanctions.[41] Hence, In our paper, we not only introduce a second channel but also we vary the expected outcome of whistleblowing. Therefor, we contribute to the experimental literature on the motivation of whistleblowing by investigating if people care about the effectiveness of the whistleblowing channel and how it eventually affects the whistleblowing decision. We find that even if people almost fully

[41] There also exists experimental research on internal reporting. For example, Carpenter et al. (2018) study how reporting one another for shirking can improve team performance. This behavior is, however, usually not considered as internal whistleblowing, because the observed wrongdoing does not entail law violations and, therefore, it is not of relevance for public enforcement authorities. Butler et al. (2019) argue that their results can be reinterpreted for internal whistleblowing. This seems feasible, however, only when the probability of prosecution of an report is comparable between channels, which we doubt is the case for many internal channels. Furthermore, in our setting, we allow for both channels and do not assume them to be interchangeable.

switch to a more effective channel - if provided - it does not seem to encourage more people to whistleblow.

So far, there is only little empirical research on internal reporting, which we define as providing information to someone within the same company. In the field, it is quite difficult to observe because even if it is successful, it usually happens outside the public. In a recent study by Stubben and Welch (2020), however, the authors try to fill this gap by using a large data set on summary information about internal reports from the largest provider of reporting systems for companies. Primarily, the authors present results on the actual prevalence of internal reporting and its characteristics. They show that it varies substantially across firms and industries. Furthermore, they find a negative correlation between the number of internal reports processed in a company and the amount of dollars spent on state fines or material lawsuits. Thus, their results suggest that internal reporting happens and can help the companies mitigate the damage and solve the problems internally. We draw on these insights from the field, and investigate the effectiveness of internal whistleblowing alone and compare it with different external whistleblowing regimes in a controlled experimental setting.

3.3 Experimental Design and Procedures

3.3.1 Experimental Design

We randomly match participants in groups of eight. Six participants in each group are randomly assigned the role of a team member and the remaining two are assigned the role of the public[42]. Participants remain in the same role and in the same group matching throughout the entire experiment. The experiment consists of two parts. In the first part, all team members perform a real effort task for five minutes. Specifically, they are asked to type in random sequences of letters and numbers that are displayed on the screen and renewed every time they click on a 'next' button. They receive a point for each sequence they type in correctly. After the team members have completed the task, they are privately informed about how many sequences they succeeded to type in correctly. The number of points determines the payoff of each team member individually. Specifically, team members earn 9 Euros if they obtained 21 or more points, 6 Euros

[42]In the instructions, we use the framing of team member and non-team member, which we call public members in the paper.

otherwise. Public members do not perform the task and earn a fixed amount of 6 Euros. In the second part, there are three different decision stages, which may affect the earnings.

Lying stage

In stage one, team members state their earnings from the previous task to the public, i.e., they can state either 6 Euros or 9 Euros. Importantly, they know that the statement does not need to match the actual earnings, i.e., they can lie. The final payoff team members receive is based on the earnings they stated. Thus, all team members that earned 6 Euros in the first part have an incentive to state the higher payoff of 9 Euros which we will denote as a lie in the following.[43] At the same time, stating higher earnings incurs social cost on the public. Specifically, if a team member lies and states an earning of 9 Euros instead of the true 6 Euros, this increases her payoff by 3 Euros but reduces the payoff of each of the two corresponding public members by 2 Euros. Lying is, therefore, welfare-decreasing in this setting.

Whistleblowing stage

In stage two, each team member i observes one other team member j in her group, where $i \neq j$, and makes a decision whether or not to blow the whistle and accuse j to have lied in case j stated an earning of 9 Euros to the public. To elicit behavior, we use a hypothetical scenario and ask each team member i what she would do if team member j had lied before she eventually learns the actual decision of team member j . This allows us to observe every team member's decision, because it is independent from the actual lying behavior in stage one.[44] Only if team member j actually stated a higher payoff, does the reporting decision of team member i become effective. This excludes the possibility of fraudulent reports, which we make clear in the instructions.

We introduce four treatments where we vary (i) the provision of the external channel and (ii) the consequences that come along with external whistleblowing. In a baseline treatment NoEXTERNAL, only the internal whistleblowing channel is provided to the

[43] In order to avoid selection we set the threshold for receiving the 9 Euro in part one very high so that no participant actually reached it. Therefore, every team member had an incentive to lie.

[44] There is a methodological debate on eliciting decisions by such a method i.e., there are concerns that the results from so called cold decisions might systematically differ from hot decisions that are made in the actual situation. For our case, we think the advantage of this method outweighs the concerns, as our results would be dependent on the amount of lying in the first stage. Furthermore a review on this methodological debate summarizes that an elicitation of cold decisions does if anything underestimate an effect (Brandts and Charness (2011)).

team members in the whisteblowing stage. Specifically, team member i can decide if she wants to share the information about potential lying of team member j to two other team members, k and l, where $k, l \neq j$ and $k, l \neq i$. Thus, in the instructions, we make clear that a team member will never receive a report against herself and a whistleblower will never send a report to herself. For an internal whistleblowing report the reporting team member i pays a small cost of 10 Cents. This cost only becomes effective if the reported team member j actually lied. Thus, team members only pay if their report is valid. Internal whistleblowing is only successful and therefore leads to consequences, when both team members k and l decide to accept sanctions in the subsequent stage. If this is true, team member j will be sanctioned and the public members compensated. Specifically, team member j must pay back the cost of in total 4 Euros. This is the cost she imposed on the public members by overstating her payoff in the lying stage. Therefore, the consequences merely imply a redistribution, which has no impact on welfare (if we disregard the marginal cost of whistleblowing itself).

In our treatment EXTERNAL, we add the external channel. Thus, in the whistleblowing stage, team members can choose between whistleblowing to two other team members (internal channel) as described in NoEXTERNAL, or report to the two public members (external channel). For each whistleblowing report - either internal or external - the reporting team member i pays a small cost of 10 Cents, which again becomes effective only if team member j actually lied. Furthermore, team members can choose if they want to use only one of the two channels or both sequentially. More concretely, each team member i has the possibility to first report to her other team members, and - only if this does not lead to sanctions - report to the public members. The sequential choice option is restricted to this order. If a team member i uses both channels to report the lying of her team member j, she must pay 20 Cents. External whistleblowing is only successful and leads to consequences, if both public members decide to prosecute in the subsequent sanctioning stage. For internal whistleblowing two team members must accept for sanctioning to become effective as described in the treatment NoEXTNAL. In treatment EXTERNAL, the consequences of whistleblowing are the same no matter which channel is used and they are identical to the consequences of whistleblowing in NoEXTERNAL. The team member j will be sanctioned and the public members compensated. Thus, the team member j must pay back 2 Euros to each public member. The consequences are again welfare neutral (if we disregard the marginal cost of whistleblowing itself).

In treatment DAMAGE, a potential whistleblower can choose between the same choice options as in treatment EXTERNAL. In this treatment, only the consequences of external whistleblowing differ in one dimension. Specifically, we want to mirror a situation, where whistleblowing might lead to organization-wide damage for the firm and thus for all employees independent from their own behavior. Thus, if a team member is sanctioned for lying by external whistleblowing, she must pay back the money to the public as in treatment EXTERNAL. But in addition to that, every member of the team has to pay a fine of 50 Cents. This additional damage occurs for each successful external whistleblowing against a team member. Thus, in this case, external whistleblowing is welfare decreasing. In treatment INVESTIGATION, the choices are again the same as in treatment EXTERNAL. Here, we want to simulate the common feature of whistleblowing, i.e., initiation of a company-wide investigation. Therefore, in this treatment successful external whistleblowing not only causes a sanction for the lie of the one team member j that was reported but for everyone in the team who lied and thereby damaged the public. Therefore, in this case public members are fully compensated for their loss and will get their fixed payoff of 6 Euros. In contrast to DAMAGE, however, the consequences of external whistleblowing are not welfare decreasing because it is only a redistribution of costs for all lying within a team. Team members that made a correct statement about their earnings are not affected.

Accepting stage

In the third stage, depending on the treatment only team members (in NoEXTERNAL) or both team members and public members (in all other treatments) make a decision if they want to accept sanctions for team member who lied and were reported to them. As we do for whistleblowing, we again elicit the decision to accept by hypothetical question. This means, when the participants make their decision, they do not know if there was actually a team member reported to them. Participants make one decision that then applies to all reports they receive [45]. As for the team members, instructions make clear that they will never make a decision about a report that concerns themselves. They know that, only if both corresponding participants choose to accept sanctions for the reported team members, the respective consequences will follow. Public members have a direct material incentive to accept, because they will be compensated for the loss they suffer

[45] Team members can receive at maximum two whistleblowing reports by other team members. Instead, public members can receive at maximum six reports if all members of a team lied and blew the whistle.

from lying by a team member. In contrast, for team members the decision to accept does not affect their own payoff. They will make a decision about internal whistleblowing, which only affects the person reported and the public.

At the end, we use a standard psychological questionnaire to elicit individual moral foundations (Graham et al., 2011).[46]

3.3.2 Experimental Procedures

We ran the experiment in the CLER laboratory of the University of Cologne in October 2017 and June 2018. In total, we recruited 368 subjects and conducted three sessions per treatment. Specifically, we collected 88 independent individual strategies for the two treatments NoEXTERNAL and INVESTIGATION, respectively, and 96 independent individual strategies for the treatments EXTERNAL and DAMAGE, respectively.

Instructions were common knowledge and read out aloud before each session. Only after part one was completed, instructions for part two were handed out. In the instructions we used the framing of team members and non-team members but we did not use words like "whistleblowing". The study was fully anonymous, which we made clear in the beginning of each session.

We paid a show-up fee of 10 Euros in order to compensate for possible negative payoffs public members could end up with after the second part. We made clear in the instructions that everyone would receive at least 4 Euros in this session, which corresponded to the regular show-up fee of the CLER. The experiment lasted for approximately 90 min and participants earned on average 16 Euros including the show-up fee. To program our experiment, we used the Z-tree software (Fischbacher, 2007).

3.4 Hypotheses

In this section, we discuss different motives why people are willing to report wrongful behavior by others even if it is costly. Based on these motivations, we derive predictions

[46] This questionnaire is derived from the moral foundation theory proposed by Haidt and Joseph (2004), which relies on an intuitionist approach (Haidt, 2001) and claims that people's conception of a moral behavior can be pinned down to a set of moral foundations: care, fairness, loyalty, authority, sanctity. Individuals are assumed, however, to prioritize foundations in different ways and thus derive different moral principles and beliefs.

about how the provision of an external reporting channel affects internal whistleblowing and the effectiveness of whistleblowing in general. For simplicity, we first consider the whistleblowing decision as an independent choice and will later discuss how it might correlate with the subjects' honesty and acceptance behavior.[47]

In our settings, there is a good chance that the material profits from lying are higher than the expected loss of being sanctioned. This is especially true in the treatments NoEXTERNAL, EXTERNAL and DAMAGE as a sanction would require reporting by other members of their team. Yet, there is strong empirical evidence that considerable many subjects seem to derive utility from viewing oneself as an honest person (Bénabou and Tirole, 2016; Abeler et al., 2019; Bénabou et al., 2019). Thus, in all treatments we expect significantly many team members to be motivated by a desire to support an honest self-image. Therefore, in the treatments NoEXTERNAL, EXTERNAL and DAMAGE we expect relatively constant levels of honest behavior. In contrast, there might be a higher fraction of honest behavior in INVESTIGATION since in this treatment, only one team member is necessary to sanction all team-members. The threat of being punished becomes more likely and could prevent also some selfish agents from lying.

After the decision whether to lie or not, subjects make a whistleblowing decision. While whistleblowing is monetarily costly, subjects might derive intrinsic utility from this action. On the one hand, subjects might feel a *warm glow* about their own good intention to punish and rectify justice (Andreoni, 1989, 1990).[48] This private gain should come into effect irrespective of the actual consequences of one's action. As in the experiment, the expected costs - at least for internal whistleblowing - are the same in all treatments, this would lead to the same amounts of whistleblowing across the treatments.

Not only the warm glow but also should the actual consequences of whistleblowing have

[47] If we take the strategy space and apply a standard selfish agent model there are two salient symmetric pure-strategy equilibria that apply to all of our treatments. First, subjects can coordinate on lying, no whistleblowing, no acceptance. Second, subjects could coordinate on not lying, whistleblowing via the internal channel and acceptance by team-members. In the treatments with an external channel coordination on not lying, whistleblowing via the external channel and acceptance by the public also constitutes an equilibrium since subjects with the role public have an incentive to always accept.

[48] Ouss and Peysakhovich (2015) suggest a similar motivation also for punishers, who would receive a "cold glow" utility by reducing payoffs of subjects who violated social norms. In our setting, we could think of both "cold and warm glow" as whistleblowing not only leads to punishment but also increases the payoff of the public.

an effect on utility (Fehr and Gächter, 2002; Fehr and Fischbacher, 2003)[49]. At least for punishment behavior, studies show that indeed the effectiveness of punishment has a significant influence (Casari, 2005; Egas and Riedl, 2008). Hence, with regard to this type of motivation subjects derive utility from whistleblowing, only if it eventually leads to consequences.[50] Yet whistleblowing does not necessarily succeed such that the probability of success should be a decisive factor for the whistleblowing behavior. And in this regard, the internal channel and the external channel are different. Whereas the recipients of an external report have a monetary incentive to accept, this is not the case for internal whistleblowing. Therefore, the expected probability that a whistleblowing succeeds should be higher for the external whistleblowing channel.[51] Therefore, we hypothesize in EXTERNAL a clear shift to external whistleblowing. This will in turn also result in a more effective prosecution and, therefore, lead to an increase in the rate of punishment compared to NoEXTERNAL. It is not clear though if we also see an increase in overall whistleblowing. This would require subjects that do not take any private gain from their action but only care of the consequences, and would, therefore only make a report if the chances are high enough that it will be implemented.

Hypothesis 1 *Compared to NoEXTERNAL, in EXTERNAL, whistleblowers shift to the external channel. The fraction of lying that is actually sanctioned, therefore, increases.*

Let us now look at the two other treatments with an external channel where the consequences for successful external whistleblowing are different compared to internal whistleblowing. In DAMAGE, a successful external whistleblowing will make all team members of the firm suffer a small additional fine. This implies that the expected costs for external whistleblowing are higher than for using the internal channel. Depending on whether this extra costs outweigh the utility gain from a more effective channel, some subjects who would have reported externally in treatment EXTERNAL are expected to shift to the

[49] In many aspects, altruistic punishment is strongly related to whistleblowing. A major difference could be seen in that in the case of whistleblowing the person who observes the misbehavior is not in power to implement a punishment but depends on a third party.

[50] The underlying mechanisms may comprise other-regarding preferences for fairness and altruism (Rabin, 1993; Fehr and Schmidt, 1999; Fowler et al., 2005; Bolton and Ockenfels, 2000), preference for the enforcement of a social norm (De Quervain et al., 2004), or emotional satisfaction (Bosman and Van Winden, 2002; Hopfensitz and Reuben, 2009) from the punishment. In all these explanations, the eventual implementation of punishment is necessary.

[51] Note that we do not vary the probability exogenously but assume that the team members will correctly anticipate the different acceptance behavior between team members and public members. We elicited their expectations and they confirm our assumption.

internal whistleblowing or fully abstain from whistleblowing in DAMAGE. As discussed above, internal whistleblowing is less likely to result in sanctioning of lies.

Hypothesis 2 *In* DAMAGE *as compared to* EXTERNAL*, the fraction of sanctioned lies decreases.*

In INVESTIGATION, the prosecution of all members in a team who lied will follow a successful external whistleblowing. Thus, external whistleblowing by one team member could lead to the sanctioning of all wrongdoing in a team. If whistleblowing would be mainly motivated by the expected success in punishing and correcting damage, agents would derive greater utility from whistleblowing in the INVESTIGATION treatment than in the EXTERNAL treatment[52]. We would expect to see an increase in external whistleblowing and, therefore, an increase in the fraction of lying disclosed.[53]

Hypothesis 3 *In* INVESTIGATION*, the usage of the whistleblowing channel as well as total reporting increase as compared to* EXTERNAL*.*

We will now briefly discuss, how the decisions might correlate with one another within subjects. The decisions to whistleblow or accept a report do not come with a material gain so that they must be rather driven by intrinsic motivation such as self-image as described above in context of honesty. Thus, we assume this intrinsic gain will more likely apply to those subjects who make an honest decision before. Otherwise they would want to punish a behavior they did themselves. Psychological research suggests, however, that people are concerned about their own behavior to be consistent (Cialdini, 1983; Guadagno and Cialdini, 2010). Furthermore, one could also think of a more general preference of honesty which not only implies the own honest behavior but which might make some people also reluctant to see other people's dishonesty. Thus, following these lines of argument, we would find a positive correlation between honest behavior and whistleblowing, as well as

[52]This only applies to honest whistleblowers, as they will not punish themselves in INVESTIGATION. However, we do not assume a significant fraction of dishonest whistleblowers as we will argue in the following subsection on the correlation between decisions on the individual level.

[53]This situation might be considered similar to a bystander problem (Darley and Latané (1968)) where diffusion of responsibility might decrease the amount of whistleblowing (Choo et al. (2019)). Since in case of multiple lying cases, there are also multiple team members who could whistleblow externally and thereby sanction all lying at once. However, in this, treatment, team members are uncertain about their pivotality, because there is a fair chance that the potential whistleblower is the only one who observe a lie and is pivotal for the sanctioning. Therefore, we do not think diffusion of responsibility can be applied.

between honesty and acceptance behavior.

Hypothesis 4 *On an individual level we expect, that whistleblowing and the acceptance of whistleblowing reports is highly correlated with honest behavior.*

3.5 Results

First we present a comparative analysis of the effectiveness of internal and external whistleblowing in disclosing and sanctioning lying across treatments. After that we will have a closer look at the predominant strategies used by the team members. Specifically, we show how honest behavior is correlated with reporting as well as accepting, and we look at the moral foundations as a possible determinant of whistleblowing.

3.5.1 Resampling and permutation test

For our analysis of the effectiveness of whisteblowing regimes, we primarily focus on the aggregated outcomes on a team level such as the fractions of lying reported through whistleblowing and fractions of sanctioned lying. These outcomes, however, not only depend on the individuals' decisions but also on the random matching in the experiment. For instance, if a team member decides to whistleblow, this report only counts as whistleblowing if the matched team member has lied. In order to eliminate possible confounding effects of the random matching, we need to separate our analysis from the very realization of the matching in our sample. Hence, we use a re-sampling method. Specifically, we take the strategies of all subjects and randomly rematch them into a team. For this team we can then generate the fractions of reported and sanctioned lying.[54] We let the permutation program perform this procedure ten thousand times for each treatment and thus receive an average value of reported and sanctioned lying on a team level as presented in the Figures 3.1 and 3.2.

[54]In more detail, for each treatment, separately, we keep the team members' strategies, which imply for each subject a set of three individual decisions, i.e., on lying, whistleblowing and accepting, respectively. From this subsample, we randomly draw (without replacement) individual strategies of six team members. We then aggregate outcomes as if these team members had played together in a team. Furthermore, for each team that we simulated as described above we also randomly draw (without replacement) two public members from the pool of subjects in the respective treatment that are assigned the role of a public member. For them we do not see much variation in acceptance behavior since almost everyone independent from the treatment accepts potential sanctions (See Table 3.3 in Appendix A).

As for statistical inference, we also use the simulated distribution of team outcomes as a basis for testing instead of the actually realized outcomes in our sample.[55] Specifically, the null hypothesis in our case is that the team outcomes are randomly allocated between treatments. We, therefore, take the team outcomes from the simulation and shuffle them across two treatments that are of comparison. From this distribution we then take ten thousand pseudo samples of eight simulated teams (as is the sample size of each treatment) and calculate our test statistic[56], and, therefore, compute the null distribution of our test. For the reasons we explained above, we do not compare a single sample mean against the null distribution but also a distribution of test statistics, that we compute in the same way as the null distribution but with fixed treatment labels. The p-value is then given by the fraction of cases, where the test statistic value generated under the null hypothesis is at least as extreme as the value we simulate with fixed treatments as they prevail in our observed sample.

3.5.2 Lies and sanctioned lies

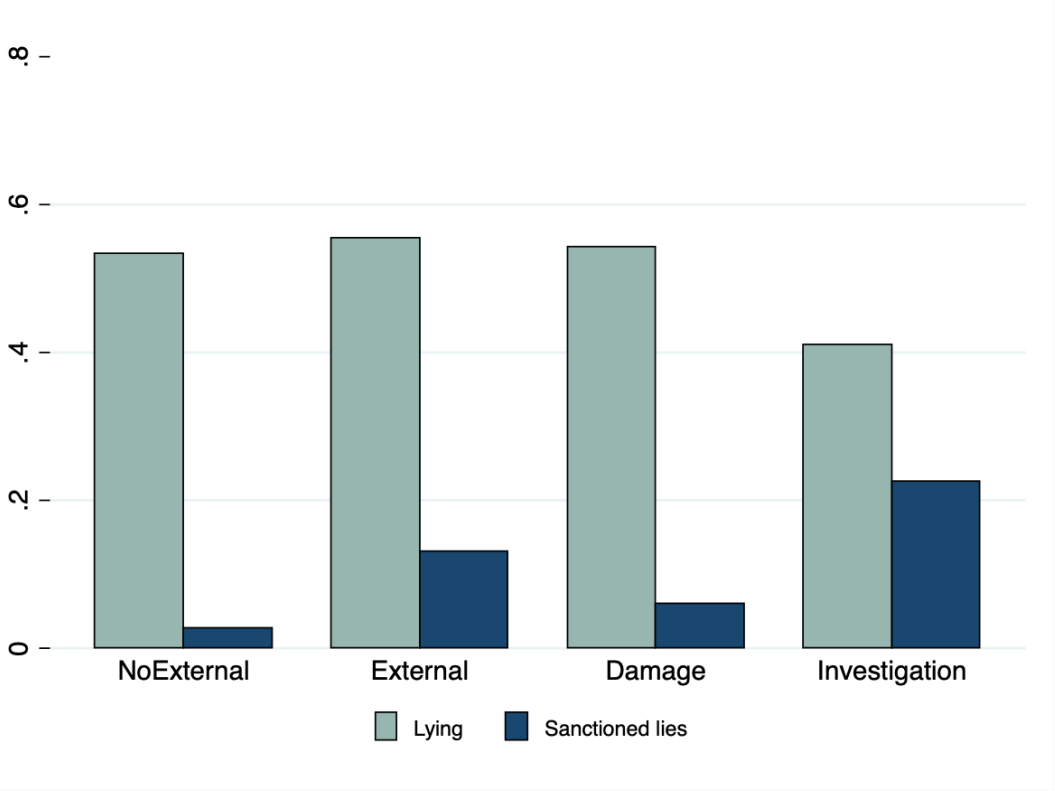
The light bars in Figure 3.1 depict the fraction of liars in each treatment. The first three light bars look very similar and present a level of lying at about 50 percent. Only in INVESTIGATION we see a drop down to 41 percent, which is, however, not significantly different compared to EXTERNAL with 53 percent ($p = 0.13$). Whereas lying behavior does not seem to vary much over treatments, we do find differences in the fractions of sanctioned lies.

The dark bars in Figure 3.1 show the fractions of lies that are actually sanctioned according to the resampling exercise described above. If we concentrate on the two left dark bars, we find that the provision of an external whistleblowing channel increases the fraction of sanctioned team members from 2 percent with only internal whistleblowing to 13 percent in a regime where also the external channel exists ($p = 0.03$). Furthermore, the last three dark bars allow us to see how varying the consequences of external whistle-

[55] For this, we apply a two-sided permutation test. This is a standard non-parametric test, which relies on re-sampling and comes with the advantage that it does not make assumptions on an underlying sampling distribution (Holt and Sullivan, 2021; Kennedy, 1995). The basic idea is that the null distribution can be derived from the empirical sample data itself through re-sampling or, more concretely, shuffling the treatment labels. Thus, the null hypothesis assumes a random allocation of the treatment and, therefore, no treatment effect.

[56] As a test statistic we use the difference in sample outcomes. For instance, for the amount of reported lying in a sample as an outcome, we compute this for two simulated samples and take the difference.

Figure 3.1. Fractions of team members who lie and are sanctioned



Note: The figure presents the fractions of team members who lie and the fractions of team members who are sanctioned separately for each treatment. The fractions of sanctioned lies are obtained through resampling.

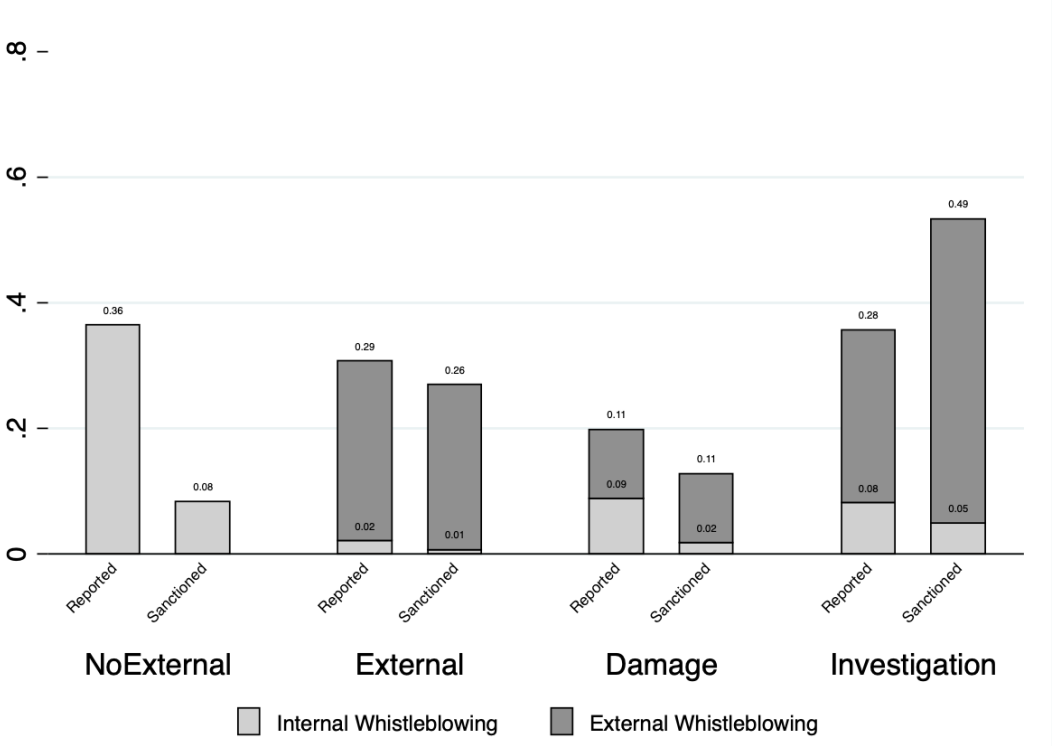
blowing on the team affects the fraction of sanctioned lies. Compared to EXTERNAL we see that sanctioned lies go up to 23 percent in INVESTIGATION and down to 6 percent in DAMAGE. However, both comparisons are not statistically significantly different (p=0.37 and p=0.19, respectively).

3.5.3 Reported lies and sanctioned lies

In a next step, we take a closer look at why a larger fraction of lies is sanctioned when an additional external channel exists. On the one hand, whistleblowers might switch to the external channel as it is more effective, and, therefore, lies might be sanctioned more often. On the other hand the external channel might also increase the amount of reporting. Due to its higher chances to successfully implement sanctions, the external channel might encourage more people to make a report than in case where only internal reporting is possible.

Figure 3.2 presents the fractions of lies that are reported and sanctioned in each treatment.

Figure 3.2. Fractions of lies that are reported and sanctioned by whistleblowing channel and by treatments



Note: The figure presents the fractions of lies that are reported and the fraction of lies that are sanctioned separately for both whistleblowing channels and for each treatment. The fractions of sanctioned lies are obtained through resampling.

Again, we first concentrate on the comparison between the treatments EXTERNAL and NoEXTERNAL to see if the introduction of an external channel may increase the chance of a lie to be reported and sanctioned. We observe that in NoEXTERNAL 36 percent of lies are reported, whereas in EXTERNAL we have a slightly lower 31 percent ($p = 0.66$). This finding suggests that the external reporting channel seems to barely affect the likelihood of getting reported. If we now look at the right bars of the treatments EXTERNAL and NoEXTERNAL in Figure 3.2, we see how many of those team members who lie are eventually sanctioned. It seems that the effectiveness of reporting increases tremendously. Almost every lie that is reported is eventually sanctioned in treatment EXTERNAL, whereas in the NoEXTERNAL it is only about 20 percent of the reports that lead to sanctions ($p = 0.012$). A marked shift from internal to external whistleblowing explains this difference. It seems that if provided with the possibility to report externally people use it and blow the whistle externally. This suggests that people correctly anticipate the external to be the more effective channel. However, the introduction of an external channel does not seem to motivate more people to report an observed lie.

Result 1 *In* EXTERNAL *whistleblowing almost fully shifts to the external channel but the total amount of whistleblowing does not change as compared to* NoEXTERNAL.

As a next step, we turn to the treatments DAMAGE and INVESTIGATION and compare their outcomes with EXTERNAL. In DAMAGE we find a drop in the share of lies reported through external whistleblowing (from 29 percent in EXTERNAL down to 11 percent in DAMAGE, $p = 0.09$). This finding is not surprising as in this treatment an external report comes with an additional cost of damage for everyone in the team. However, people who now refrain from the external whistleblowing do not seem to necessarily use the internal channel. In comparison to EXTERNAL, the fraction of internal whistleblowing is slightly higher but not significantly though ($p = 0.67$). Although the total share of reported lies is not different from EXTERNAL, the internal channel does not seem to catch up with the drop in the external channel. This could suggest that some people who would have reported in EXTERNAL may fully refrain from reporting in DAMAGE.

Result 2 *In* DAMAGE*, the share of lies disclosed through external whistleblowing significantly drops compared to* EXTERNAL. *The total amount of whistleblowing does not change significantly.*

In INVESTIGATION, a fraction of 36 percent of lies is reported which is not significantly different from the corresponding fraction of 31 percent in EXTERNAL ($p = 0.76$). Despite the fact that in this treatment external whistleblowing can fight wrongdoing in a team more effectively, it does not seem to encourage more people to report. Turning to the sanctioned lies, we see that the fraction of sanctioned lies increases from 27 percent in EXTERNAL to 54 percent in INVESTIGATION. However, this difference is not statistically significant either ($p = 0.21$). Furthermore, we see that the fraction of sanctioned lies exceeds the amount of lying that is reported. This results from the team-wide investigation, which means that all lying in a team is disclosed if only one team member blows the external whistle.

Result 3 *A team-wide investigation following external whistleblowing neither affects the fraction of reported lies nor the fraction of sanctioned lies.*

Table 3.1. Frequencies of strategies in each treatment

Strategies...	No External	External	Damage	Investigation
...without Whistleblowing				
[Lying- No Whistleblowing - No Accept]	**48.48**	**40.28**	**38.89**	**24.24***** $_{++}$
[Lying- No Whistleblowing- Accept]	1.52	12.50**	13.89***	12.12**
[No Lying- No Whistleblowing- No Accept]	7.58	12.50	**23.61**** $_{+}$	16.67
[No Lying- No Whistleblowing- Accept]	10.61	8.33	6.94	15.15
...with Whistleblowing				
	Internal	Internal External	Internal External	Internal External
[No Lying- Whistleblowing- Accept]	**28.79**	1.39*** **18.01**	5.56$_{+++}$ 8.33*	3.03$_{+++}$ **22.73**
Other	3.03 –	0.00 6.94	1.39 1.39$_{+}$	1.52 4.55
Observations	66	72	72	66

Note: For statistical testing we use a Chi squared test for proportions. ***, **, and * denote statistical significance at 1%, 5%, and 10% compared to NoEXTERNAL. $^{+++}$, $^{++}$, and $^{+}$ denote statistical significance at 1%, 5%, and 10% compared to EXTERNAL. We compare strategies with external whistleblowing only among those treatments that provide such a channel. (EXTERNAL, DAMAGE, INVESTIGATION). For comparing strategies that include internal reporting we consider all treatments.

We will now dig deeper and look at team members' entire strategies to see how honest behavior correlates with reporting and acceptance across treatments.

3.5.4 Team members' strategies

Table 3.1 presents the relative frequencies of the predominantly chosen strategies. We emphasize the frequencies of the first two most common strategies for each treatment.[57] It shows that across all treatments the most prominent strategy is to lie, not to report and not to accept. Hence, we see partial coordination on a dishonest equilibrium, which is efficient from the perspective of the team members. Furthermore, we see that even though this strategy is predominant in all treatments there are differences in its prevalence. Whereas we find relatively constant levels of about 40 percent in NoEXTERNAL, EXTERNAL and DAMAGE, the frequency drops significantly to only a quarter of the team members in INVESTIGATION. Among the team-members who do not report, we find a significant increase in the share of those who are honest (Chi-squared test, INVESTIGATION vs. EXTERNAL, $p = 0.06$).[58]

As a second commonly used strategy overall, we observe the combination of not lying, reporting and accepting, which is in line with an honest equilibrium strategy. With levels of a total of 20 to 30 percent, it turns out to be a fairly common strategy in almost every treatment except for DAMAGE. Furthermore, in case of reporting we do not see much heterogeneity in strategies used as we do for the case without reporting. So it seems that reporting comes with honest behavior and acceptance in almost every case. And, indeed, using Pearson correlation we find a significant positive relationship between honest behavior and reporting as well as for accepting behavior and reporting in every treatment.

Looking deeper into this particular strategy, Table 3.1 also divides reporting into the two possible channels, internal and external, and presents the frequencies for each. For the treatments, where external whistleblowing is an option, we observe a marked shift to the external channel. Yet, in DAMAGE the external channel is used less frequently compared to EXTERNAL and INVESTIGATION.[59] This finding goes clearly in line with our results from the aggregated outcomes, where we find that lying reported through external whistleblowing drops in DAMAGE. Instead, people rather seem to be honest but neither report nor do they accept a report. This strategy is the second most prevalent in

[57] We provide the raw data for each decision separately by treatment in Table 3.3 (see in Appendix A of this chapter).

[58] INVESTIGATION vs. NoEXTERNAL p= 0.05, INVESTIGATION vs. DAMAGE $p = 0.30$.

[59] For the sake of simplicity we count subjects who decide for sequential reporting as external whistleblowers.

DAMAGE and significantly more often used than in EXTERNAL.

We now turn to the acceptance behavior of team members. The two most commonly used strategies over all treatments display a consistent behavior in respect to the acceptance decision: Subjects, who lie and do not report, do not accept sanctions for others, whereas honest subjects, who report, tend to accept sanctions. When we use Pearson correlation over all strategies we find a significant positive relationship between honest behavior and acceptance across all treatments (except for DAMAGE). However, in Table 3.1 we also see fractions of people that would not accept sanctions for others. In particular in DAMAGE, no lying, not reporting and not accepting elevates to a very common strategy. But also in every other treatment this strategy is used by a fraction of people that is significantly greater than zero. We can, therefore, conjuncture that the effectiveness of the internal channel being much lower compared to external whistleblowing might primarily result from dishonest team members that do not accept sanctions. But another aspect plays a role as well. There seems to be a fraction of subjects that are honest themselves but refuse any engagement in prosecuting other tea-members, even if accepting another subjects report is costless.

Result 4 *Across all treatments, the dominant behavior is either the dishonest strategy (lie, not whistleblow, not accept) or the honest strategy (not lie, whistleblow, accept). Thus, we see a high correlation between honest behavior and both whistleblowing and acceptance. However, we do find a non-negligible fraction of subjects who are honest but abstain from whistleblowing and even from acceptance.*

3.5.5 Moral foundations

In the previous section, we looked at the most common strategies and we conjectured that reporting is most often observed from honest people. In this section, we, therefore, focus on those subjects that did not lie and ask how individual moral foundations could possibly explain their reporting decision and, in particular, the choice for one of the two reporting channels. For this, we restrict the data to those treatments where all reporting channels are provided and use the decision, whether a subject reports and which channel she chooses, as dependent variable. Due to the nature of our variable of interest, which assigns discrete values to the three possible outcomes - not reporting, internal reporting and external reporting - without ordinal information, we use a multinomial probit re-

Table 3.2. Determinants of Reporting Decision - Moral foundations

	Average Marginal Effects	Standard errors
Fairness		
No Whistleblowing	-0.136***	0.043
Internal Whistleblowing	0.006	0.025
External Whistleblowing	0.130***	0.040
Loyalty		
No Whistleblowing	0.036	0.045
Internal Whistleblowing	0.034	0.021
External Whistleblowing	-0.070*	0.039
Number of observations	104	

Note: Results from a multinomial probit regression with whistleblowing decision as dependent variable with three possible outcomes (no whistleblowing, internal whistleblowing and external whistleblowing). We count sequential reporting as external whistleblowing. The observations are restricted to honest subjects in the treatments EXTERNAL, DAMAGE and INVESTIGATION. We use robust standard errors. ***, **, and * denote statistical significance at 1%, 5%, and 10%, respectively.

gression model. We regress our dependent variable on an indicator for treatment[60] and on dimensions from the moral foundation questionnaire Graham et al. (2011). Specifically, for our model presented in Table 3.2 we only choose the two dimensions loyalty and fairness as regressors thereby referring to previous psychological research which identifies the loyalty-fairness assessment as an important determinant of whistleblowing (Waytz (2016); Waytz et al. (2013)).[61] More concretely, we summarize the responses to the six respective questions for each dimension and use the mean values standardized over all treatments as the independent variables.

Table 3.2 depicts the average marginal effects of these two dimensions and robust standard errors from a multinomial probit regression model as described above. The results suggest that to some extent moral considerations can explain the decision on whether to report an observed wrongdoing and which channel to use. The upper rows show how the fairness perception influences on average the probability that a person chooses one of the three alternatives. It suggests that if a person weights fairness one standard devia-

[60] We do not find an indication for an effect of the treatment variation on the elicitation procedure as we find no correlation between treatment and the elicited responses for each moral dimension.

[61] When we include all dimensions into the model, we do, indeed, not find other dimensions to be significant other than loyalty and fairness.

tion more this would on average decrease the probability that she will fully refrain from reporting by 13.6 percentage points. Furthermore, a greater emphasis on fairness also seems to increase the probability that a person chooses external whistleblowing.

In the last three rows, we see on the other hand how loyalty might affect the reporting decision. We find that a higher valuation of loyalty as a moral principle would make a person slightly more likely to abstain from external whistleblowing. It has, however no influence on the choice for the internal channel or no reporting, respectively.

3.6 Conclusion

In this paper, we study the effectiveness of two different whistleblowing channels. Specifically, we ask to which extent the introduction of an external channel with varying negative externalities for the organization can help increase the frequency of whistleblowing and the rates of disclosed and punished wrongdoing.

Our results suggest that the external channel does not encourage more people to whistleblow as we find no difference in the number of whistleblowing attempts between treatments. However, we see that if an external channel is provided, whistleblowers almost fully shift to the supposedly more effective channel. From this result we conjecture that people seem to be motivated by both the "warm glow" they feel about their altruistic act of whistleblowing but they also seem to care about the actual outcome and use the channel that is more effective. We, indeed, see an increase in sanctioning when external whistleblowing is possible. Thus, our findings suggest that the external channel can help increase the rate of detection and punishment, even if the negative externalities for the company become more severe. The observation that the internal channel is clearly less effective cannot only be due to liars who do not want to punish behavior of other members. In fact, we see a significant fraction of honest subjects who do not whistleblow and do not accept whistleblowing reports, even if the latter would not cost them anything. We find that this fraction is significantly higher when an external channel is provided that could be damaging for the organization.

From a policy perspective where the aim is to establish high rates of disclosures and sanctioning, the provision of an external whistleblowing channel seems to be an effective instrument. The primary reason for this seems to be that the external channel transmits

information about wrongdoing to people who are more willing to sanction.

From a corporate perspective, it appears to be of very importance to address employees in an organization who are honest themselves to be more engaged, in whistleblowing and accepting behavior internally. Our observation that honest people even abstain from accepting a whistleblowing report suggests that they either do not gain "warm glow" utility, and thus are indifferent or it might even incur a cost on them. Whistleblowing and accepting other whistleblowers reports could be still widely considered to be snitching or illoyal (Reuben and Stephenson, 2013; Waytz et al., 2013). Indeed, we see that among honest subjects the personal evaluation of fairness and loyalty as important moral foundations are correlated with the decision to whistleblow. Consequently, organizations should create an environment and culture that finds whistleblowers responsive and raise awareness of whistleblowing being an overall accepted tool to fight internal wrongdoing. In fact, this is already happening in organizations, which conduct internal compliance campaigns to raise awareness and focus on a culture that would point to misbehavior rather than ignore it. Yet, to this point, the efforts overall do not seem to have reached the goal. At least, surveys suggests that still a large fraction of employees observe misbehavior at the workplace which is not addressed by the company. For instance, in a global survey on ethics at the workplace 2020, about one third of the employees indicate to have observed organizational misconduct. While compared to previous years, the number of respondents who answered to have reported the wrongdoing has increased, also the number of alleged retaliation is at a peak. It gives a glimpse that not just the willingness to whistleblow in an organization is enough for an organization to be effectively self-regulatory, it also needs to appreciate the whistleblowing within the organization.[62] Though, it might be interesting to study how retaliation might differ between the channels, and how this could also affect the effectiveness of the channels.

Hence, we do not claim our results to give a complete picture about the possible determinants for the effectiveness of whistleblowing overall and the functioning of each channel. A growing body of experimental literature unfolds a wide range of factors that seem to have an influence on the whistleblowing decision and its effectiveness. So far, however, they have not been investigated in an environment where two different channels exist. For instance, Butler et al. 2019 show that the public approval subjects expect to

[62]See for the report of the Global Ethics Survey 2020 at: https://www.ethics.org/global-business-ethics-survey/

receive from whistleblowing seems to have a positive effect on the whistleblowing decision. However, this result might only apply for external whistleblowing and could shift even to the contrary for internal whistleblowing for two reasons. Subjects might not expect approval from their colleagues to whistleblow internally but even a negative response. Furthermore, they will not receive a response by the public, because the wrongdoing may be resolved internally. This could additionally motivate subjects to rather use the external channel.

3.7 Appendix A: Additional Tables

Table 3.3. Raw data on decisions for team members and public members across treatments

Note: This figure presents the average values for lying, reporting and accepting decisions.

	NoEXTERNAL	EXTERNAL	DAMAGE	INVESTIGATION
Decisions				
Team members				
Lying	0.530	0.556	0.542	0.410
Reporting				
Internal channel	0.318	0.014	0.069	0.045
External channel	-	0.236	0.069	0.197
Sequential	-	0.014	0.028	0.076
Accepting	0.409	0.417	0.361	0.545
Observations	66	72	72	66
Public members				
Accept	-	0.958	1.00	0.954
Observations	22	24	24	22

3.8 Appendix B: Instructions and Questionnaire

Instructions

Note: The parts in the dashed line frame box, are parts that were presented to No-baseline treatments: EXTERNAL, INVESTIGATION and DAMAGE. As for the latter treatments, the instruction only differ in the part on sanctions. I made clear which text was provided in which treatment.

General information about the experiment

We welcome you to this study.

You are participating in a study about economic decision-making and are therefore asked to make some decisions in the following. You are receiving a starting fee of 10 Euros for your appearance. Please read the instructions carefully to learn how you can earn additional money or get money subtracted based on your decisions and the decisions of the other participants of this study. Please note that you will receive at least 4 Euros at the end of this study.

Your income during the study will be calculated in Euros. At the end of the study, you will receive your payment in cash.

All interactions between you and the other participants in today's study will occur through a computer. Please remain quiet and do not talk to the other participants or contact them in any other way. If you have questions, please raise your hand, and one of us will come to you.

Your participation will be completely anonymous. This means that you will neither get to know the identity of the participants you interact with before the study nor afterward. These participants will also not get to know anything about your identity.

At the beginning of the study, all participants will be allocated to one of two possible roles: **Team member** and **non-team** member. You will be informed on your screen which role you have been assigned to. You will keep this role throughout the whole study.

This study is divided into two parts. First, we will hand out and explain the instructions for part 1. Only after part 1 has been conducted will you receive the instructions for part 2.

You will receive your final income at the end of part 2.

Part 1

In part 1, team members can earn money by finishing the following typing task. The task is correctly typing out a randomly sorted row of numbers and letters. Team members are asked to correctly type out as many rows as possible during 5 minutes. Please note that every single element must be typed out correctly. After you have typed out a row, press the "Next"-button to get a new row with different numbers and letters. You can see how many rows you have already typed out correctly at the bottom part of your screen. Below this text, you can find a screenshot of the typing task.

Preliminary income at the end of part

The income for team members is determined by the number of correctly typed out rows, as you can see in the table below.

Non-team members do not carry out the typing task and receive a fixed amount of 6 Euros.

Correctly typed rows	20 or less	More than 20
Euros	6	9

Please note that the income that you receive at the end of part 1 will be directly transferred to part 2 and is therefore subject to change. You will only get a payout of your final income at the end of part 2.

Part 2

At the beginning of part 2 each team member is randomly allocated to a team of 6 members. In each team the team members are labeled with a number:

Team member 1
Team member 2
Team member 3
Team member 4
Team member 5
Team member 6

Furthermore, each team is assigned two non-team members. The non-team members are also labeled with a number:

Non-team member 1
Non-team member 2

From now on, we will indicate the number of Euros each team member earned in the typing task as the actual output of the team member.

Now assume for the rest of the instructions that you have been labeled as team member 1 of a team:

A short overview of the decisions you will make in part 2, which will be explained in more detail in the following:

1. Reporting: You report your output and can choose between 6 and 9 Euros.

2. Notifying: You decide if you want to notify your team member 2 in case he or she has reported a higher output than his or her actual output

3. Acceptance: You decide if you want to accept a notification that has been sent to you by team member 3 or 4 and the corresponding sanctions of that.

Decision 1: Reporting

First, you are asked to report your output from part 1.

You can either **report 6 or 9 Euros** as your output. Your report is sent to the two non-team members assigned to your team. During this decision, only you know your actual output.

For calculating **your income** at the end of the first decision, only your **reported output** is used – independent of your actual output from part 1.

If your actual output is **6 Euros and you report 9 Euros as your output, both non-team members receive a reduction in payment by 2 Euros.**

At the same time, all other team members 2 – 6 are asked to report their output in the same way.

Decision 2: Notifying

For decision 2 (notifying), **each team member observes** at a given point of time the **actual and the reported output of another team member**. For example, you as team member 1, observe the actual and reported output of team member 5.

Before you observe the actual and reported output of team member 6, you are asked to decide whether you will notify team member 5 in case he or she has reported a higher output.

You must pay 10 Cents for each notification you send. Your notification is only sent if team member 5 has actually reported higher output.

You can choose between the following options:

A: Notify team member 5 to the other team members.

B: Notify team member 5 to the non-team members.

C: Notify team member 5 to the other team members. If this notification is rejected, send a notification to non-team members.

B or D: Do not notify team member 5.

At the same time, all other team members 2-6 make the same decision for the team member they observe. Non-team members do not make this decision.

Decision 3: Acceptance

Based on the decisions that were made in decision 1 and decision 2 notifications are sent to

non-team members and/or

team members.

Notifications to non-team members

Before each non-team member actually receives a notification, he or she is asked to decide whether he or she **wants to accept or decline all potential notifications**. In other words, each non-team member makes **one decision which counts for all notifications** that he or she will actually receive.

Only if **both non-team members decide to accept** notifications, all notifications that are actually sent to non-team members are accepted. Sanctions that are more precisely explained in the next section follow for accepted notifications.

You as a team member make a similar decision.

Notifications to team members

Before you actually receive a notification, you are asked to decide **whether you want to accept or decline all potential notifications**. In other words, **you make one decision which counts for all** notifications that you will actually receive. Each team member can receive two notifications at most.

Please note:

- You will never decide about notifications that concern yourself.
- You will never decide about notifications that concern team members that observe you.

- You will never decide about notifications that were sent by a team member that you observe.

For example, as team member 1 you could receive notifications of your team members 2 and 4.

This means in return that **each notification is sent to two team members**. The **notification is accepted if both corresponding team members decided to accept all notifications they received.**

If you decided to accept all potential notifications that you are receiving, these notifications will in fact only be accepted if the other corresponding team member accepts the notification as well.

At the same time all other team members 2 – 6 make the same decision as you about potential notifications that they can receive.

Sanctions

Accepted notifications can cause sanctions. Sanctions from non-team members differ from sanctions that team members implement.

The exact procedure for the implementation of sanctions

1 Sanctions that non-team members implement

For EXTERNAL and DAMAGE:

For each successful notification, the team member who has been notified will suffer a reduction in payment by 4 Euros. Therefore, each one of the two non-team members receives a payback of 2 Euros. In other words:

- Each team member who has been successfully notified looses 2 + 2 = 4 Euros.
- Each non-team member receives 2 Euros back from each team member that has been notified successfully.

(Only for Damage treatment:
Additionally, you and all other team members get a reduction of 50 Cents for each notification that non-team members accepted.)

For INVESTIGATION:

Suppose non-team members accept at least one notification. In that case, all team members who reported a higher output will suffer a reduction in payment by 4 Euros. Therefore, each one of the two non-team members receives a payback of 2 Euros. In other words:

- All team members that reported a higher output loose 2 + 2 = 4 Euros.
- Each non-team member receives 2 Euros back from each team member that reported a higher output.

The second stage only applies if non-team members accept no notifications.

2 Sanctions that team members implement

For each notification that team members accepted, the team member who was notified suffers a reduction in payment by 4 Euros. Therefore, each one of the two non-team members receives a payback of 2 Euros. In other words:

- Each team member that was notified and whose notification was accepted loses 2+2 = 4 Euros.
- Each non-team member receives 2 Euros back from each team member that was notified and whose notification was accepted.

Final Income at the end of part 2

Your final income at the end of part 2 (the income for team members 2 – 6 is composed the same way

Your final income = your reported output – sanctions – costs for notifications

The final income of non-team members in part 2:

Final income = 6 – (Number of team members that report a higher output) * 2 + (Number of sanctioned team members that have reported a higher output) * 2

Questionnaire

Reporting

1. Team member 1's actual output is 6 Euros. What can Team member 1 report to the non-team members?

a. Only 9 Euros.
b. Only 6 Euros.
c. Either 9 or 6 Euros.

Notifying

2. Each team member can notify at most one other team member.
3. Suppose team member 1 chooses option C and team member 2 actually reported a higher output. If the corresponding team members 3 decided not to accept any potential notifications, team member 1 has to pay:
    a. Nothing
    b. 10 Cents
    c. 20 Cents
4. If a team member notifies another team member, sanctions are automatically implemented.

Acceptance

5. Only non-team members can receive notifications from team members.
6. Each notification is sent to two participants.
7. A notification is accepted only if ... decide/s to accept any potential notifications:
    a. both participants b. at least one participant
8. It can happen that a team member is asked to accept a notification that concerns him or herself.

Sanctions

9. Only if all non-team members either did not receive any notifications or did not accept a notification, sanctions accepted by team members will be implemented.
10. Sanctions implemented by non-team members differ from sanctions implemented by team members.
11. If team members implement sanctions only those team members are sanctioned that reported a higher output than their actual output, that were actually notified, and whose notification was accepted by team members.

Income

12. The income of non-team members can be 6 Euros at most.

13. Suppose team member 2 reported 9 Euros, which is higher than his actual output, and all other team members reported their actual output of 6 Euros. Furthermore, team member 1 decides to notify team member 2 to non-team members if team member 2 reported a higher output than his or her actual output. As this is the case, a notification is sent to the two non-team members.
Suppose the corresponding both non-team members accept the notification. Please indicate the final income of all team members and non-team members:

Income of team member 1
Income of team member 2
Income of team member 3
Income of team member 4
Income of team member 5
Income of team member 6

Income of non-team member 1
Income of non-team member 2

CHAPTER 4
ETHICAL RESPONSIBILITY AND PERFORMANCE

4.1 Introduction

Financial incentives are a well-studied and widely used mechanism to increase workers' productivity. A long strand of empirical studies shows that monetary compensation is a crucial factor for employees' effort choice (Lazear, 2000; Shearer, 2004; Bandiera et al., 2005). At the same time, it is widely accepted among social scientists and practitioners that this is not the only factor that influences workers' motivation. In the past decades, economists and psychologists have emphasized the role of non-financial incentives and investigated how much the work environment can affect workers' productivity (Deci, 1971; Gneezy and Rustichini, 2000; Ariely et al., 2009).

In fact, several studies involving field and lab experiments show that employees exert more effort when they regard their job as meaningful (Ariely et al., 2008; Chandler and Kapelner, 2013) or when the job has a pro-social mission, i.e., contributes to a public good (Fehrler and Kosfeld, 2014; Tonin and Vlassopoulos, 2015; Carpenter and Gong, 2016; Charness et al., 2016; Kajackaite and Sliwka, 2017; Cassar, 2018).

These results suggest that workers' motivation is higher if their work environment matches their ethical values and social standards. In other words, it appears that only if the work environment is aligned with the workers' pro-social and moral values do non-financial incentives become effective. For firms, however, this can be challenging and costly to implement. It might, therefore, be useful to provide employees with more discretion to shape their work environment in a way which meets their own social and ethical standards. In this regard, it has become frequent over recent years for an increasing number of companies to grant employees more flexibility to shape their work environment, offering workers decision rights to match their ethical and social values at their workplace.

One example is the US online retailer Zappos. The former CEO Tony Hsieh introduced a no-script policy for those who work in the customer service. He argued that employees should be able to "let their true personalities shine during every phone call" by giving them the freedom to interact freely with the customers. Thus, workers can make their own decision whether to consult in a way best for the clients or to sell unnecessary ser-

vices. This is in contrast to the usual practice of firms to provide written scripts where workers are strongly encouraged to sell services by all means. We argue that a worker is more motivated when he is free to choose his own work environment because this gives him a sense of responsibility over his work.

In this paper, we investigate whether being responsible for a work environment that matches one's own ethical values can serve as an incentive to increase performance. Psychological research studies indicate that the feeling to have self-determined one's own behavior and, thus, the feeling of being responsible for an environment plays a decisive role on individuals' intrinsic motivation. It makes the individual feel more coherent with one's self and thus more committed to the action one executes (Deci, 1971; Ryan and Deci, 2000; Gagné and Deci, 2005).

However, there exists so far limited empirical evidence on this question. Fehrler and Kosfeld (2014) find that participants who choose to contribute their produced output to a charity exerted more effort compared to participants who are randomly assigned to contribute to it. The authors explain this effect by self-selection arguing that the choice acts as a vehicle to efficiently target only those workers for whom Corporate Social Responsibility is an effective non-financial incentive. However, the observed increase in performance might also be caused by the choice itself because workers find themselves responsible for their pro-social behavior. It is, indeed, difficult to separate one explanation from the other because they usually coincide: When an individual makes a choice, she acts the way she prefers. Thus, choosing an ethical or pro-social work environment might increase motivation through (i) the fact that the individual acts according to her preferences or (ii) by the feeling of being responsible for it.

In our experiment, we control for selection effects and focus on how being responsible for an ethical or unethical decision affects performance. In particular, we match participants into pairs composed of one worker and one employer. The worker and the employer are each independently assigned a piece-rate for a task the worker eventually has to perform. The payoffs for both - the employer and the worker - are the same and depend on the reported piece-rate and the worker's performance. The employer and the worker are asked to individually report the assigned piece-rate, where they have the incentive to over-report this piece-rate. At the same time, we explicitly declare over-reporting a

violation of the rules of the experiment which is, however, not punished. This creates a trade-off between one's own financial gain and ethical considerations. In the following we will refer to a situation where the worker performs under an over-reported piece-rate, he is not entitled to, as an "unethical" environment. Conversely, we will call the situation an "ethical" environment when a worker and the employer receive a payoff based on the designated piece-rate. Furthermore, we introduce a randomization procedure which implements the reporting decision of either the worker or the employer. This creates our two major "treatment" conditions: *Responsibility* and *NoResponsibility*. In case the worker's decision is implemented we refer to the treatment condition as *Responsibility* and as *NoResponsibility* otherwise.

As a first result we find that given an ethical environment, workers in the *Responsibility* condition perform better than in the *NoResponsibility* condition. This result can still be driven by selection because the *NoResponsibility* case also includes workers who would have preferred to work in an unethical environment. We want to focus on the effect of responsibility only and, therefore, concentrate only on those workers who prefer to work in an ethical work environment. When comparing their performance in the *Responsibility* and the *NoResponsibility* condition, we still find that performance is higher for workers in the *Responsibility* condition. This gives us reason to infer that responsibility has an effect on performance. However, we do not find this effect for an unethical environment. Thus, workers who are responsible for an unethical environment do not seem to be more motivated. Lastly, we find that when workers are asked to perform in an unethical work environment they are not responsible for, ethical workers perform significantly worse than unethical workers. We interpret this result such that workers bear ethical costs in an unethical work environment, even if they are not responsible for it.

With our findings we add to a field of experimental literature studying the relationship between responsibility and effort provision. So far this has only been investigated in gift exchange settings where social preferences explain positive effort choices. Charness (2000), for instance, finds that agents provide more effort when a random procedure determines agents' wages compared to a situation where a third party chooses agents' wages. The author explains this finding by stating that in presence of a random procedure an agent feels solely responsible for final payoffs, while the same agent can shift part

of this responsibility when a third party has determined his wage. Moreover, in a more recent study Charness et al. (2012) show that when employers delegate their wage decision to their workers, workers respond with higher effort. Controlling for other possible explanations such as reciprocity, the authors show that feeling more responsible for the outcome seems to be a driving factor of the effect. Similarly, Falk and Kosfeld (2006) find that when principals restrict the effort choice set of the employees by setting a minimum effort level - and thus also employees' responsibility - employees perform worse than in cases where choice sets are not restricted. Thus, responsibility for final outcomes seems to enhance pro-social motivations and through this channel increases effort. In contrast to these studies, we approach the relationship between responsibility and effort provision from a different angle by asking if being responsible for an ethical work environment creates additional motivation which results in higher performance.[63]

Furthermore, this paper relates to experimental research on shifting of costs for unethical behavior. In a dictator game, Bartling and Fischbacher (2011) find that dictators are punished significantly less if they delegate an unfair decision to a second person instead of making the same unfair decision themselves. Even if delegation by design eliminates the possibility of a fair outcome, Oexl and Grossman (2013) observe that dictators are able to successfully shift the blame for an unfair outcome to a powerless delegee. Whereas in this literature the focus lies on the costs that are incurred by other peoples' punishments, we look at self-image related costs that are incurred by the person who acts in an unethical environment. We find that people seem to incur these ethical costs even if they work in an unethical environment they are not responsible for, which implies that they cannot fully shift these costs to the decision maker.

As a specific feature of our design, we separate the choice from its implementation in order to observe the selection that would have evolved in a natural field setting and control for it. This relates our paper to research that studies the effect of democratic institutions on behavior. This body of literature provides various attempts to disentangle the effect of making a decision from the effect of self-selection (Dal Bó et al., 2010; Sutter

[63]Looking at the effort provision in a work environment after the choice was implemented distinguishes our paper from recent research on delegation, which investigates the motivational effect of having decision rights (Fehr et al. 2012, Bartling et al. 2014). In this literature, the focus lies on the effort people provide to keep the right to choose and thus on motivational gains before the choice is implemented. Adding to that literature, Sloof and von Siemens (2019) introduce an implementation stage but do not focus on the effect of responsibility.

et al., 2010; Dal Bó et al., 2019). In our paper, we use a technique which is similar to that of Dal Bó et al. (2010). In their setting the group members cast a vote on the implementation of an effective sanctioning institution to enforce cooperation but a random mechanism either implements the group's voting outcome or overrides the vote. They find that people cooperate more if their own decision outcome is implemented compared to having the same decision exogenously imposed on them. After controlling for possible other explanations such as information transmission, the authors show that the effect is robust and call it an "endogeneity premium", which we would interpret as a valuation of being responsible. Beside this strand of research, there is one study by Babcock et al. (2015) where the authors argue that the mere act of choosing might be motivating for individuals. In a field experiment, they find that people perform better if they can choose their incentive scheme compared to a situation where they are randomly assigned to it. In contrast to the other studies, Babcock et al. (2015) could rule out selection as an explanation ex post since an almost complete selection into one choice option occurred. In this paper, we can control for selection ex ante and therefore can explicitly observe the workers' types even in case when the choice of the employer was implemented. Besides of that, we look at the effect of choosing when it comes with an ethical dimension and implies a trade-off between financial gains and ethical standards.

The remainder of this paper is organized as follows. Section 2 describes the experimental design and procedure. In Section 3, we introduce a basic framework that shall illustrate our behavioral hypotheses about the expected treatment effects. Section 4 presents the results of the study and Section 5 concludes.

4.2 Experimental Design and Procedures

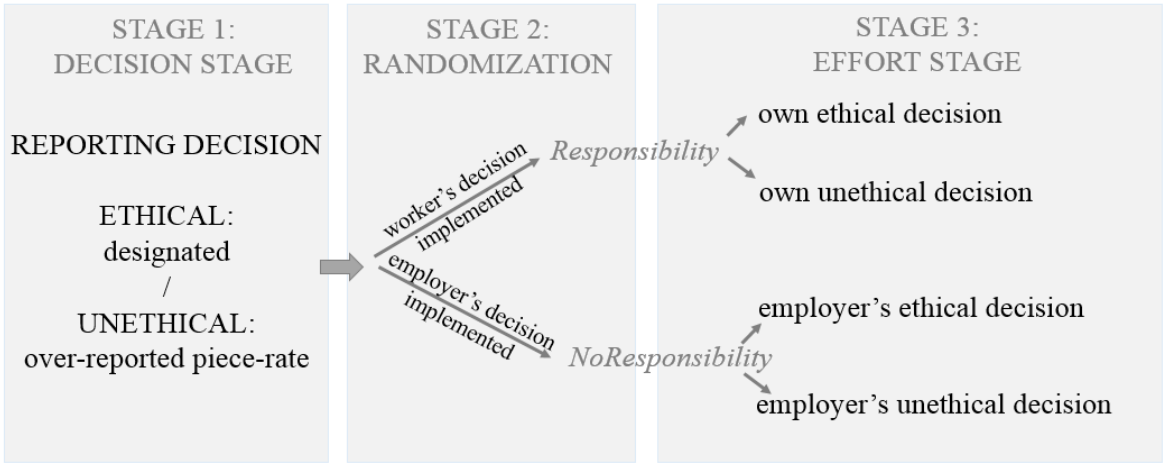
4.2.1 Experimental Design

We run a laboratory experiment where we group subjects into pairs and randomly assign one the role of an employer and the other the role of a worker. The experiment consists of three stages (see Figure 4.1).

Stage 1: Decision stage

In the decision stage, the employer and the worker are independently assigned a piece-

Figure 4.1. Experimental design



rate.[64] They are then asked to simultaneously report this piece-rate, which in the effort stage will be used to calculate the earnings for both, the employer and the worker. More precisely, in the effort stage the worker performs a real effort task which in 50 percent of the cases is remunerated based on his own reported piece-rate and in 50 percent of the cases based on the piece-rate reported by the employer. Earnings are the same for employers and workers. This means that a higher reported piece-rate implies higher earnings for both employer and worker.[65]

For this decision, both players find on their screens two different piece-rate options, namely 5 and 8 points per unit of effort, they can choose from. Thus, employers and workers who were actually assigned a piece-rate of 5 points per unit of effort - which happens in 4 out of 5 cases[66] - have the opportunity to report a piece-rate of 8 points per unit of effort. The experimental instructions explicitly clarify that reporting a higher piece-rate violates the rules of the experiment. Specifically, the instructions say: "Reporting a piece-rate that is different from your designated piece-rate is considered a violation of the rules. If you do so anyway, your earnings will be calculated based on your reported but not the designated piece-rate. You will, therefore, receive a piece-rate which does not

[64] For more detailed information on the experimental procedure, please look into the instructions provided in the Appendix.

[65] The employers' and workers' payoffs are aligned in order to provide them both with the same incentive to over-report. Moreover, we think it is crucial to let the employer also make a decision. If we only had a random procedure simulating an employer's decision, we could not argue that the environment is considered ethical or unethical to the same extent.

[66] Employers and workers have no information regarding the exact distribution of the piece-rates. From the reporting decision on the screen they can infer that both piece-rates, 5 and 8 points per unit of effort, are possible.

correspond to the piece-rate designated for your task".[67] In this sense, we consider the decision to report the designated piece-rate "ethical" and the decision to over-report "unethical" as it violates a stated rule and implies that one earns a higher than designated piece-rate, one is not entitled to.[68] Accordingly, if a worker performs under an ethical decision we consider it an ethical work environment, whereas working under an unethical decision implies an unethical work environment. Please note that only the workers will learn about the nature of the real effort task, which happens only after the decision stage. Therefore, we can rule out that employers or workers condition their reporting decision on their belief about the worker's performance in the effort stage because they will have no information about what the worker will do.

Hence, we will only consider decisions of participants that were assigned a piece-rate of 5 points per unit of effort and, therefore, faced an actual trade-off between a financial gain and the adherence to rules as a potential ethical standard. This means that the ethical decision always yields a piece-rate of 5 points per unit of effort, while the unethical decision yields a piece-rate of 8 points per unit of effort.

Stage 2: Randomization

After both participants reported their individual designated piece-rate, the reported piece-rate of either the employer or the worker is implemented with equal probability. The participants know this procedure beforehand. The randomization device creates our "treatment" conditions. In treatment condition *Responsibility*, the worker performs a task that is compensated based on a piece-rate he has reported himself. Thus, he can work in an ethical or in an unethical environment, but in either case he is responsible for it. In treatment condition *NoResponsibility*, the worker performs a task that is compensated based on a piece-rate the employer has reported. Hence, in that case the worker

[67] Through explicit control questions we ensure that subjects did not expect further consequences from circumventing the stated rule of reporting truthfully.

[68] One could indeed make an objection here as we call an efficiency maximizing behavior unethical. This comes along with two implications. Firstly, outcome-based preferences such as altruism or social preferences would lead to behavior which we define unethical. For the unethical case, we would, if anything, under-estimate the effect of ethical costs opposing the effect of responsibility. However, in a questionnaire after the experiment ended the answers of those participants who reported a higher piece-rate show that nobody explained the behavior with ethical reasons but only with their own financial gain. Secondly, workers who make an ethical decision might want to compensate for the resulting efficiency loss. However, that would imply that workers feel guilty towards an allegedly unethical employer for having made an ethical choice.

acts in an ethical or unethical environment he is not responsible for. In the *NoResponsibility* treatment condition, the worker is informed about the designated and reported piece-rates of the employer. This ensures that the worker - when performing the real effort task - knows whether the employer did over-report the piece-rate or not. In case of *Responsibility* the worker receives no information regarding the employer's choice.

Stage 3: Effort stage

In the effort stage, workers are asked to count the occurrences of "7"s that appear in successive sequences of numbers. The payoffs for the employer and the worker are the same and calculated by the worker's number of correctly solved sequences times the respective reported piece-rate implemented by the random draw. The number of correctly solved tasks will be our measure for performance.[69]

Table 4.1 displays all relevant conditions in which the worker could perform the task eventually. First, we distinguish between *Responsibility* and *NoResponsibility*, which we exogenously vary through the randomization stage and, therefore, call our "treatment conditions". In case of *Responsibility* the worker performs under his own decision. Based on the worker's own decision, he can be responsible either for an ethical or for an unethical work environment. In case of *NoResponsibility* the worker performs under the employer's decision. Thus, he works in an environment he is not responsible for. Second, we distinguish between the workers' types, which reveal endogenously through the workers' decisions in the decision stage. Specifically, we call workers that chose an ethical work environment in the decision stage ethical workers and, respectively, workers who over-reported the piece-rate unethical workers. Therefore, in *NoResponsibility* the worker's type and the work environment chosen by the employer create four possible scenarios in which the worker either performs according to his preferences or not. The worker's preferences are matched with the work environment when (i) an ethical worker performs in an ethical environment, or (ii) an unethical worker performs in an unethical way. Conversely, the worker's preferences are mismatched, if he acts in an environment he would not have chosen himself. At the end of the experiment the employer and the

[69] We use the number of correctly solved tasks as our measure of performance as this measure is incentivized. In Figure 5 in the Appendix we additionally report the results when we use the total number of attempted matrices as our measure of performance. It seems that the effect of responsibility does not reveal in the quantity of work but in the quality of work.

worker receive information regarding the number of correctly solved sequences by the worker, the reported piece-rate, and the earnings of the pair.

Table 4.1. Relevant conditions in which worker possibly performs the task

	NoResponsibility Employer's decision implemented		**Responsibility** Worker's decision implemented
	Ethical worker	**Unethical worker**	
Ethical decision implemented	Worker is *not responsible* for the *ethical decision* and acts *according* to his preferences (matched) N=29	Worker is *not responsible* for the *unethical decision* and does *not* act *according* to his preferences (mismatched) N=12	Ethical worker is *responsible* for the *ethical decision* N=48
Unethical decision implemented	Worker is *not responsible* for the *unethical decision* and does *not* act *according* to his preferences. (mismatched) N=9	Worker is *not responsible* for the *unethical decision* and acts *according* to his preferences. (matched) N=7	Unethical worker is *responsible* for the *unethical decision* N=20

4.2.2 Experimental Procedures

We ran our experiment at the University of Cologne using the software ztree (Fischbacher, 2007) in January and April 2017. The participants earned €1 for 25 points. On average, the participants earned €10.50 including show-up fee.[70] The instructions were common knowledge and read out loud at the beginning of each session. In total, 312 subjects participated in our experiment.[71] We have 68 observations in our treatment condition

[70] Prior to their registration at the Cologne Laboratory for Economic Research, all future participants give their informed consent to voluntarily participate in the experiments.

[71] One observation was dropped due to a participant writing in the questionnaire: "I was tired and, unfortunately, did not know which role I was." This means that this person did not know if his effort would actually count for the payoff calculation, since employers were also asked to perform a real-effort task. However, employers' performance did not count for final payoffs. This was clearly stated in the instructions. As mentioned above, in the following, we will also rule out the observations of those participants who got an actual piece-rate of 8 points per unit of effort assigned and, thus, did not experience a trade-off between a financial gain and the adherence to rules as a potential ethical standard.

Responsibility and 57 observations in *NoResponsibility*. Table 4.1 displays the exact numbers of observations for each condition as described in our experimental design. We preregistered our experiment at AEA RTC Registry (RCT ID: AEARCTR-0001956).

4.3 Conceptual framework

To develop the intuition of the possible effects of being responsible for an ethical or unethical decision, we provide the following framework. First, individuals decide between piece-rate p^H, p^L with $p^H > p^L$ and where the decision for p^H implies over-reporting and creates an unethical work environment. After the workers report a piece-rate $p \in \{p^L, p^H\}$, they choose an effort level e. We suggest that they maximize the following utility function:

$$U_i(e, p, v, \pi_i, \delta_i, \alpha, c_i) = pe - c_i e^2 + r\pi_i e - [rv\delta_i + (1-r)v(1-\alpha)\delta_i]e \qquad (4.1)$$

The utility function is composed of the monetary earnings, pe, minus a standard effort cost function with a convex cost function, ce^2. We extend that function by two additional components. First, we allow for a utility gain from being responsible for a decision, $r\pi_i$, where $\pi_i \in [0, \overline{\pi}]$ is the value of responsibility and where $r \in \{0, 1\}$ denotes the treatment condition. We assume that π_i only applies in a situation where a worker performs in a work environment he is responsible for, which is the case in our treatment condition *Responsibility* ($r = 1$). For simplicity, we assume a constant marginal gain from responsibility. We do not claim that this assumption is generalizable but we assume it is appropriate for our design. In our experiment, the worker can be responsible for the piece-rate which applies for each unit of effort. In this way, the feeling of being responsible for acting ethically or, respectively, unethically applies to every unit of effort.

Second, we introduce an ethical cost, δ_i, where $\delta_i \in [0, \overline{\delta}]$. We assume that this ethical cost applies in situations where the worker achieves an outcome through a violation of a rule $v = 1$, where $p = p^H$. We assume that acting under an unethical decision increases effort costs at a constant marginal rate. This again derives from our experimental design where the unethical decision concerns the piece-rate the worker earns for each unit of effort. Therefore, the rule violation applies to each unit of effort. Due to our experimental design, we also assume ethical costs and effort costs to be uncorrelated. The decision whether to over-report occurs before participants are informed about the nature

of the real effort task. This eliminates the concern that the two costs components might be correlated. To capture the phenomenon of shifting ethical costs, we allow the cost of being unethical to be deduced by a rate of α, where $0 \leq \alpha \leq 1$, in case the unethical decision was made by someone else. In particular, we refer with this additional parameter to previous literature showing that the costs of being unethical or unfair can be (partly) shifted to the person who actually made the decision (Bartling and Fischbacher, 2011; Oexl and Grossman, 2013).

The worker chooses an optimal effort level e^* which is contingent on the condition the worker finds himself (see Table 2). Backward induction reveals the threshold of ethical costs for which individuals are indifferent between choosing a higher piece-rate by making an unethical decision and being honest with the lower piece-rate (see Appendix B of this chapter for the formal derivation of this threshold).

$$\hat{\delta} = p^H - p^L \tag{4.2}$$

To give the intuition, this threshold simply illustrates that individuals are predicted to choose an ethical work environment only if the marginal cost from violating the rule (δ_i) exceeds its marginal benefit such that $\delta_i > \hat{\delta}$.

If we assume all non-standard parameters to be zero, we would find no ethical behavior in the first place. For a standard selfish decision-maker, ethical costs would not apply. Therefore, we would find no ethical behavior at all because only the marginal benefit from over-reporting would be taken into account. Furthermore, we would find no treatment difference in performance because being responsible would not affect effort cost of the worker. We get the same prediction for any form of outcome-based social preference models. Choosing the highest piece-rate by violating the rule would increase efficiency without distorting equity because the payoffs of the employer and worker are aligned.

We will now discuss the predictions that follow if we assume the parameters we additionally introduced to be non-zero. Let us first look at the ethical costs that would apply when a worker performs in an unethical work environment. If we assume δ_i to be distributed over a sufficiently large interval,[72] we would find a fraction of workers that chooses not to over-report and thus make an ethical decision. Furthermore, since the

[72]Specifically, we need to assume here that δ_i is distributed over an interval $[0, \overline{\delta}]$, where $\overline{\delta} \geq \hat{\delta}$.

threshold in (2) does not depend on the worker's effort cost c_i[73] the decision itself to be ethical or unethical would not lead to differences in performance between *Responsibility* or *NoResponsibility* conditions.

In case $\pi_i > 0$ workers are motivated by the mere fact that they work in an environment they are responsible for.[74] Specifically, we would find a higher effort level for ethical workers if they act according to their own ethical decision compared to ethical workers that act according to an ethical decision made by someone else.

Hypothesis 1 *Ethical workers exert more effort when they perform in an ethical work environment they are responsible for than in an ethical work environment that was chosen by the employer.*

For the unethical case, our framework does not give a clear prediction because we have two possible effects coming along with responsibility that go in contrary directions. Comparing unethical workers between treatment conditions, the motivational gain from being responsible might be counteracted by ethical costs that are higher if one is responsible for it. The observable effect of responsibility on performance in an unethical decision depends on how much workers can shift the ethical cost to the employer who made the decision in case of *NoResponsibility*. To be more specific about this possible counteracting effect, let us now consider α which is the degree to which ethical costs can be shifted.

First, one could assume $\alpha = 1$, which simply illustrates the case where workers can fully shift their ethical costs δ_i to the employer in *NoResponsibility*. In contrast to unethical workers acting in an unethical work environment in *NoResponsibility*, unethical workers with *Responsibility* add to their optimal effort provision the component $\frac{\pi_i - \delta_i}{2c_i}$. The direction of the effect of responsibility, therefore, depends on the relative magnitudes of π_i and δ_i. Furthermore, with $\alpha = 1$ we would find no difference in effort provision between ethical and unethical workers who work in an unethical work environment chosen by the employer in *NoResponsibility*.

[73] This follows from the assumption that effort and ethical costs are uncorrelated. As explained above we make this assumption as it follows from our experimental design.

[74] For the sake of simplicity, we do assume in this framework a parameter which is uncorrelated with the type of the decision - whether it is an ethical or unethical decision. Alternatively, one could argue that responsibility has a stronger or any impact only in a situation where the individual made an ethical decision. Responsibility would then only be effective if it is connected to a self-image enhancing action and, thereby, act like a moral boost. As we will find in the results, we cannot rule out this alternative approach.

Table 4.2. Optimal effort levels by condition

Note: This table presents the optimal effort levels in the scenarios with and without responsibility for the implemented choice. We focus on the effect of responsibility on performance. Hence, we compare ethical workers whose ethical preferences match the environment (matched workers), but only the responsibility varies. And we compare unethical workers who work in an unethical work environment between responsibility and no responsibility. To better illustrate the comparison of interest, we split the column NoResponsiblity between ethical worker an unethical worker. In addition, for each ethical type the effort level is encircled if preferences are matched in case of NoResponsibility.

	NoResponsibility Employer's decision implemented, r = 0		**Responsibility** Worker's decision implemented, r = 1
	Ethical worker	**Unethical worker**	
Ethical decision implemented (v=0, p=p^L)	$e^* = \frac{p^L}{2c_i}$	$e^* = \frac{p^L}{2c_i}$	$e^* = \frac{p^L+\pi_i}{2c_i}$
Unethical decision implemented (v=1, p=p^H)	$e^* = \frac{p^H-(1-\alpha)\delta_i}{2c_i}$	$e^* = \frac{p^H-(1-\alpha)\delta_i}{2c_i}$	$e^* = \frac{p^H+\pi_i-\delta_i}{2c_i}$

In our experiment, however, we expect that the worker cannot fully shift ethical costs in case of *NoResponsibility* ($1 > \alpha \geq 0$). Even if the employer is responsible, workers still need to perform in this unethical environment. Therefore, we assume that workers cannot fully dissociate from the unethical behavior. This partial shift of ethical costs would also be consistent with results from Oexl and Grossman (2013) who show that observers would blame individuals who carry out an unethical behavior even if they did not initiate it themselves and were not able to correct it. Assuming only a partial shift of ethical costs in *NoResponsibility*, the costs for an unethical work environment would arise in both conditions in *Responsibility* and *NoResponsibility* but would be lower in *NoResponsibility*. Lower ethical costs in case of *NoResponsibility* could, in turn, offset the motivational gain from being responsible. The compound effect of responsibility, therefore, depends on the relative magnitudes of π_i, δ_i, and α and does not give a clear prediction. Furthermore, a partial shift of ethical costs implies that in the treatment condition *NoResponsbility* unethical workers choose a higher effort level than ethical workers in case the employer chose an unethical work environment. This is because ethical workers reveal a $\delta_i > \hat{\delta}$ and therefore bear higher ethical costs than unethical workers, where $\delta_i < \hat{\delta}$.

Hypothesis 2 *In case the workers act in an unethical work environment they are not responsible for, unethical workers exert higher effort than ethical workers.*

4.4 Results

In the following, we compare the workers' mean performance levels between the different conditions distinguished by treatments and workers' types as described above. We will only compare conditions with the same incentive scheme. We use the non-parametric Mann-Whitney-U test which is suitable for independent observations and small samples. In the first section we will look at aggregated outcomes without controlling for selection. To address the hypotheses we derived from our framework, we will then look at the specific types of workers to see (1) if responsibility itself has an effect on performance and (2) if ethical costs are shifted in case of *NoResponsibility*.

4.4.1 Responsibility versus NoResponsibility in aggregated outcomes

Figure 4.2 compares the mean performance of workers who act under *Responsibility* (dark gray bars) to workers that act under *NoResponsibility* (light gray bars). The left panel illustrates the workers' performance in an ethical work environment, whereas the right panel displays the mean performance when the workers act in an unethical work environment. First, the figure shows that there are participants who prefer an ethical work environment despite the lower financial incentives. Specifically, we find that even 70 percent of the workers do not violate the rules. This suggests that, in our setting, people incur ethical costs from reporting a higher piece-rate.[75] Second, if we pool over treatment conditions and only compare the performance between the ethical and unethical work environment, we see that performance is slightly higher in the unethical case (26.1 versus 28.4 correctly solved sequences). However, the difference in aggregate outcomes is not statistically significant (p=0.1767). With regard to the incentives structure this seems surprising. While participants in the ethical work environment earn 5 points per unit of effort, the earnings correspond to 8 points per unit of effort in the unethical work environment. This suggests that not only monetary incentives but also non-monetary considerations affect performance levels, which we will discuss in the following.

In order to assess the impact of responsibility on performance, we now turn to the left

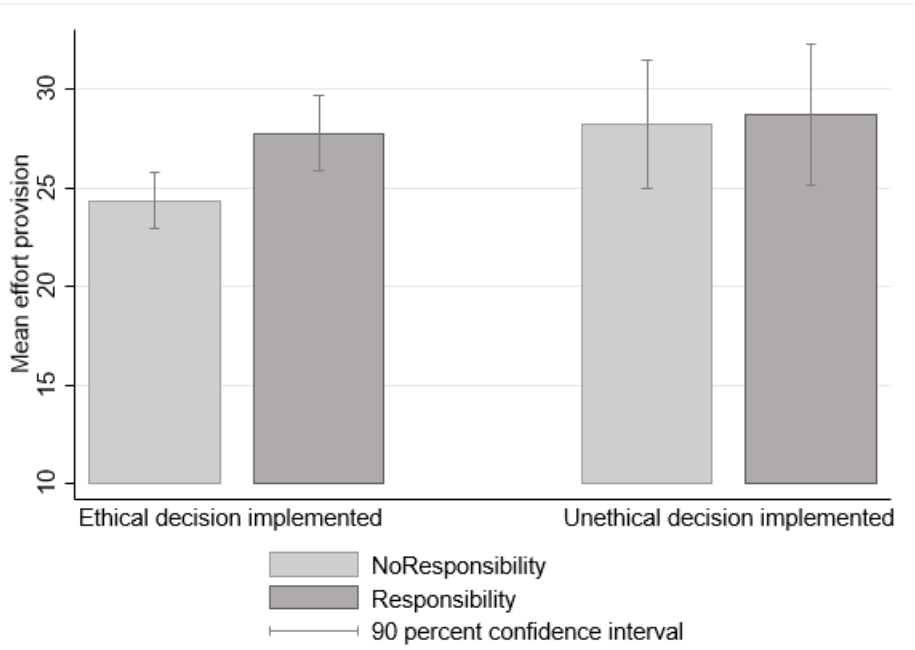
[75] We find the same fraction of employers making an ethical decision.

panel of Figure 4.2 and, therefore, look at the workers' performance in an ethical work environment under low incentives. We find that the performance is higher for workers who are responsible for the ethical decision compared to workers that perform in an ethical work environment chosen by the employer (27.8 versus 24.2 correctly solved sequences). This is an increase in mean performance of about 15% which is economically and statistically significant (p= 0.049). The positive effect of responsibility we find for an ethical environment can be explained by two potential factors. On the one hand, we hypothesize that workers are potentially more motivated if they act in a work environment which they chose and for which they are responsible. On the other hand, the effect could be driven by self-selection. Whereas in the *NoResponsibility* condition the ethical work environment might match or mismatch the worker's preference, workers choosing the ethical work environment in the *Responsibility* condition revealed their preference for it. Thus, the comparison of these two different groups of workers could create a performance difference as, for instance, acting according to one's preferences might be motivating. We want to rule out this latter possible effect as an explanation and, therefore, only compare workers who anyways prefer to work in an ethical environment.

Before we go to the type specific comparison, we have a brief look at the right panel of Figure 4.2. It displays the mean performance levels in case of an unethical work environment. Given such an environment, we do not find any difference in performance between treatment conditions. However, this does not tell us yet how responsibility affects performance because also in this case selection could be going on. In particular, our model shows that in an unethical environment, ethical types bear higher ethical costs and this might lead them to perform worse than unethical types. Therefore, our model anticipates that selection might play a role when an unethical work environment is implemented.

Consequently, we will now proceed with type specific comparisons in order to explore the underlying mechanism that causes the differences in aggregate outcomes.

Figure 4.2. Performance of workers under ethical and unethical decisions by treatment condition



Note: This figure presents the mean performance levels of workers who act under *Responsibility* (dark gray bars) and *NoResponsibility* (light gray bars). It distinguishes between the scenarios of an ethical decision being implemented, where v=0 and $p = p^L$ (left panel) and of unethical decision being implemented, where v=1 and $p = p^H$ (right panel).

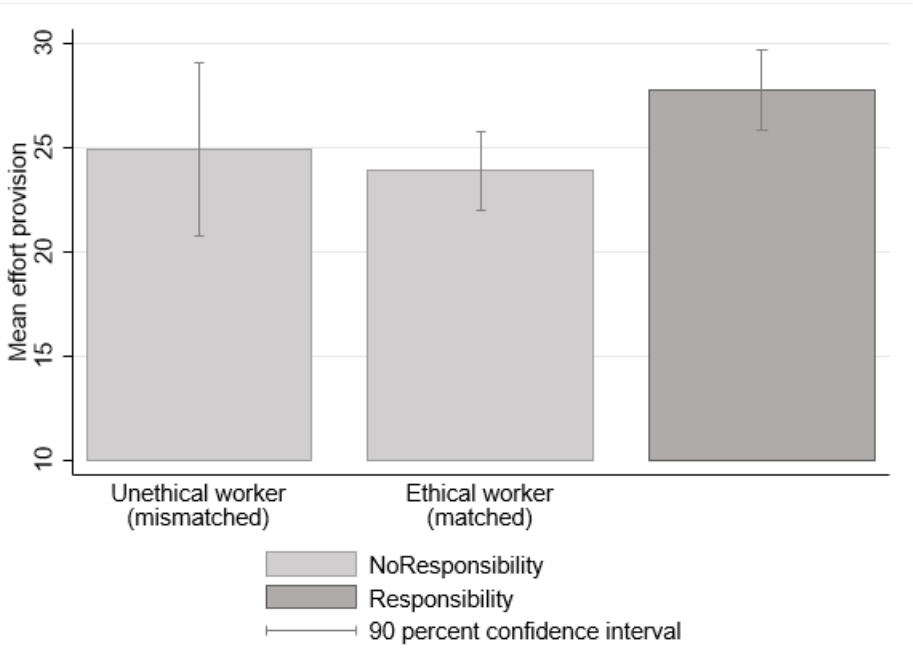
Result 1 *Workers who work in an ethical work environment for which they are responsible perform better than workers who work in an ethical environment chosen by the employer. We do not find this effect for implemented unethical decisions.*

4.4.2 Type specific comparisons

Figure 4.3 displays mean performance of workers that perform in an ethical environment. Similar to Figure 4.2, we also distinguish between the two conditions *Responsibility* and *NoResponsibility*. In addition to this, Figure 4.3 separates the types of workers in the *NoResponsibility* condition. The right light gray bar displays the mean performance of ethical (matched) workers, who thus act according to their preferences. And, respectively, the left light gray bar denotes the mean performance of the unethical (mismatched) workers who would not have chosen that environment they eventually work in.

We can now compare mean performance of the same types between the two different treatment conditions in order to control for possible selection effects. Therefore, we com-

Figure 4.3. Performance with the ethical decision being implemented by treatment conditions and worker's type



Note: This figure presents the mean performance levels of workers, who work in an ethical work environment, where $p = p^L$ and $v = 0$. It compares the mean levels between *Responsibility* (dark grey) and *NoResponsibility* (light grey). Furthermore, for *NoResponsbility*, the figure distinguishes between unethical workers, whose preferences were not matched (mismatched) and ethical workers (matched).

pare ethical workers that work in an ethical environment they are responsible for (the dark gray bar) to ethical workers performing in an ethical work environment under *NoResponsibility* (right light gray bar). We still find that the difference between the dark gray bar and the left light gray bar is statistically significant (27.8 versus 23.9 correctly solved sequences, p-value=0.037). This result provides evidence that the mere fact of being responsible for an ethical work environment seems to increase motivation and, thereby, positively affects performance.[76] We now turn to the left light gray bar which displays mean performance of mismatched unethical workers in the *NoResponsibility* case. When we compare the two light gray bars, we find that unethical types perform as well as ethical types when executing the employer's ethical decision (p-value=0.518).

[76]Figure 4.6 in the Appendix A of this chapter reinforces this finding by displaying the cumulative distribution of effort provision of ethical workers when an ethical work environment is implemented. With *Responsibility* the distribution shifts to the left assessing a positive impact *Responsibility* has on the whole distribution of ethical workers.

Result 2 *In an ethical work environment, ethical workers perform better under Responsibility than under NoResponsibility. This result suggests that the mere fact of being responsible for an ethical work environment positively affects performance.*

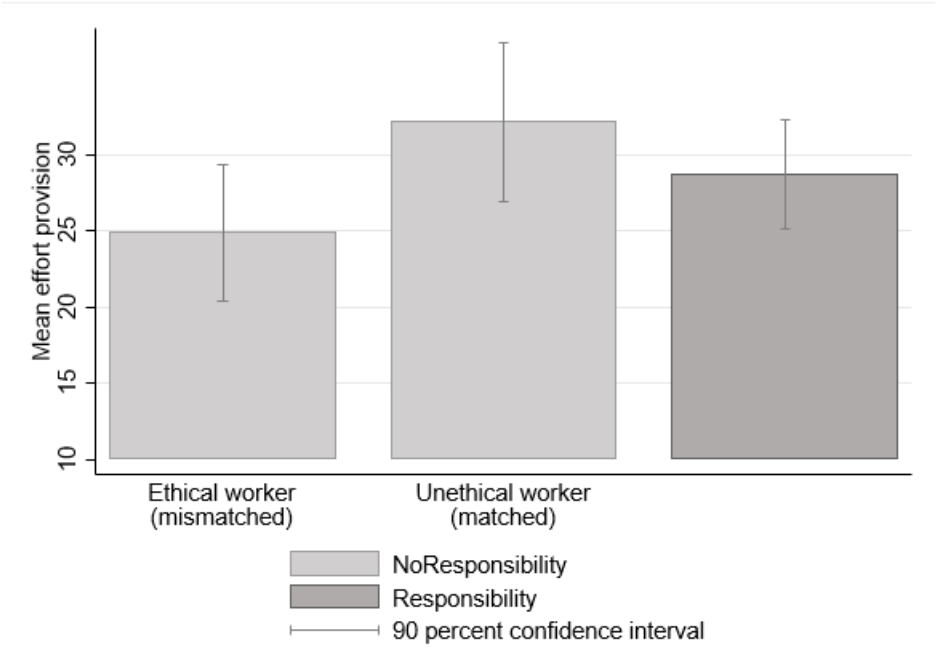
Figure 4.4 shows the mean performance of workers when an unethical decision is implemented. Again, we distinguish between the treatment conditions. The dark gray bar represents mean performance under *Responsibility* and the light gray bars stand for mean performance under *NoResponsibility*. As before, we distinguish between types which are the mismatched ethical workers (left light gray bar) and the matched unethical workers (right light gray bar).

We first look at the workers acting under their own unethical decision (dark gray bar) and compare their mean performance to unethical types working in an unethical work environment chosen by the employer *NoResponsibility* (right light gray bar). The difference in mean performance is, if anything, going in the opposite direction compared to the ethical case which means that under responsibility workers perform worse than under *NoResponsibility*. We, however, do not find a significant difference between treatments (p=.331). Even if we cannot draw a clear conclusion from this result, it still gives some suggestive evidence for the hypothesis that in *NoResponsibility* workers can to some extent shift their ethical costs. Thus, relative to *NoResponsibility* the workers incur higher ethical costs in *Responsibility* which would counteract the motivational gain from responsibility in that treatment condition.

When we compare the two light gray bars we do find a significant difference. Specifically, in an unethical work environment ethical workers perform significantly worse than unethical workers (24.9 versus 32.1 correctly solved sequences, p=.034). This results goes in line with our hypothesis 2 and suggest that different ethical costs ethical and unethical workers bear translate into different performance levels when an unethical work environment is implemented. Furthermore, this result suggests that even in case of *NoResponsibility* workers are not able to fully shift their ethical costs to the employer.

Result 3 *In case of an unethical work environment, there is suggestive evidence that responsibility does not increase performance as much as it does in case of an ethical decision. If the unethical decision of the employer was implemented, we do find that ethical*

Figure 4.4. Performance under an unethical decision by treatment conditions and worker's type



Note: This figure presents the mean performance levels of workers, who work in an unethical work environment, where $p = p^H$ and $v = 1$. It compares the mean levels between *Responsibility* (dark grey) and *NoResponsibility* (light grey). Furthermore, for *NoResponsbility*, the figure distinguishes between unethical workers, whose preferences are aligned with the environment (matched) and ethical workers (mismatched).

workers perform significantly worse than unethical workers.

4.5 Conclusion

We study whether the feeling of being responsible for one's own work environment might serve as an incentive to increase workers' performance. For this purpose, we let workers choose between an ethical or unethical work environment. In the field, having the choice about one's own work environment usually implies not only responsibility but also allows people to sort into environments they prefer. Both mechanisms might affect workers' motivation. Using a specific randomization technique in a laboratory experiment, we can separate the former from the latter possible effect. Specifically, we can compare workers that both act according to their preferences but once with and once without responsibility for this environment.

We find that responsibility for a work environment can serve as an incentive but it

depends on the way the work environment is shaped. In particular, it seems that the incentive effect becomes effective if workers are responsible for an ethical work environment. In this case, it seems that workers perform better than those who also prefer an ethical environment but are not responsible for it. In contrast, those workers who choose an unethical work environment do not respond to the same extent to responsibility. In that case the incentive effect of responsibility seems to be counterbalanced by ethical costs which arise when acting in a way that violates ethical standards. From this finding one might conjecture that responsibility effectively increases performance only for workers with high ethical cost who have a preference for an ethical environment even if it means to sacrifice monetary gains. These types of workers are more likely to be found in organizations with strong ethical and social missions as workers would rather sort into environments they prefer. Those types of organizations often face tight budget constraints. Thus, our finding might be particularly relevant for these organizations as it provides them with a cost-saving tool to improve performance.

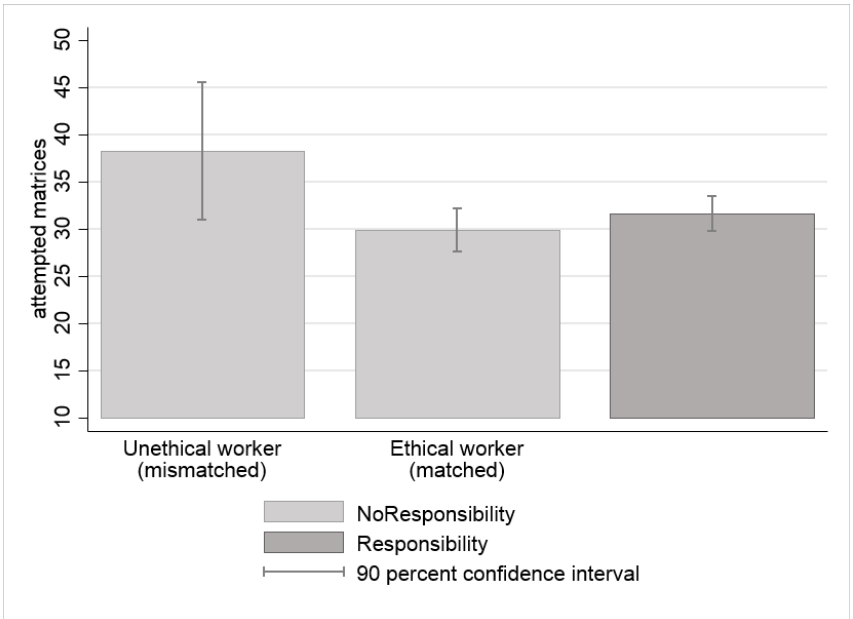
Our results might also have interesting implications for the field of compliance management, which installs monitoring systems to effectively ensure workers' behavior to comply to legal and ethical standards. Our results show that monitoring might come at costs that have not been accounted for so far because it could spoil the positive effect of responsibility on performance. This means that workers would provide less effort when they are forced to act ethically compared to a situation where they autonomously choose to work this way. In this light, apart from the important role of monitoring, a firm might also put emphasis on careful screening for workers with high ethical standards. Screening for employees with high ethical standards and allowing them to actively shape their work environment might not only save monitoring costs but even increase the motivation of workers.

Lastly, our results show that imposing a work environment which does not match workers' ethical standards might have deteriorating effects on their performance. Coming back to our example from the beginning - most of the customer service departments enforce very detailed scripts workers are required to follow exactly and mostly serve to up-sell services by all means. In case this procedure violates a worker's ethical values, we find that this work environment might lead to performance reduction. Specifically, we observe that workers perform worse if they prefer to work in an ethical work environment but are

forced to act against their ethical standards compared to workers who themselves were willing to install an unethical work environment. Our results show that even higher monetary incentives cannot compensate for this mismatch in ethical standards. This might provide further justification for allowing employees to shape their work environment in a way that meets their ethical standards. Only in this case can firms ensure that employees with high ethical standards will be able to unfold their potential fully.

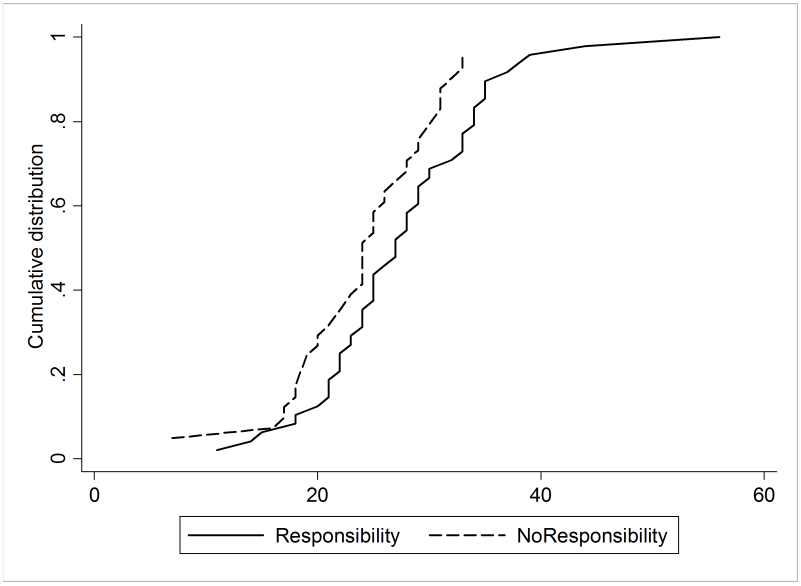
4.6 Appendix A: Additional Figures and Tables

Figure 4.5. Number of attempted matrices with the ethical decision being implemented by treatment conditions and worker's type



Notes: When using the total number of attempted matrices - in contrast to the total number of correctly solved matrices - we find no effect of responsibility on performance. Ethical workers acting under *Responsibility* do not try to solve more matrices compared to ethical workers acting in an ethical work environment imposed by the employer (31.6 vs 32.4; p=0.256, Mann-Whitney-U Test). In this sense, not the quantity of the work but the quality of the work improves in the presence of responsibility.

Figure 4.6. Cumulative distribution of effort provision of ethical workers when the ethical decision is implemented



Notes: Figure 6 displays the cumulative distribution of effort provision of ethical workers in an ethical work environment. The solid line illustrates the *Responsibility* condition, while the dashed line highlights effort provision in the *NoResponsibility* condition. With *Responsibility* the distribution shifts to the right showing that effort provision under *Responsibility* is higher for all effort levels.

4.7 Appendix B: Derivation of effort and threshold for unethical decision

Utility function as proposed in equation (1) :

$$U_i(e, p, v, \pi_i, \delta_i, \alpha, c_i) = pe - c_i e^2 + r\pi_i e - [rv\delta_i + (1-r)v(1-\alpha)\delta_i]e$$

$$\frac{\partial U}{\partial e} = p - 2c_i e + r\pi_i - rv\delta_i - (1-r)v(1-\alpha)\delta_i$$

Derive optimal effort levels by $\Rightarrow \frac{\partial U}{\partial e} = 0$

Ethical decision implemented with Responsibility (r=1, v=0):
$p^L - 2ce + r\pi_i = 0 \Rightarrow e^* = \frac{p^L + \pi_i}{2c_i}$

Ethical decision implemented with NoResponsibility (r=0, v=0):
$p^L - 2ce = 0 \Rightarrow e^* = \frac{p^L}{2c_i}$

Unethical decision implemented with Responsibility (r=1, v=1):
$p^H - 2ce + r\pi_i - \delta_i = 0 \Rightarrow e^* = \frac{p^H + \pi_i - \delta_i}{2c_i}$

Unethical decision implemented with NoResponsibility (r=0, v=1):
$p^H - 2ce - (1-\alpha)\delta_i = 0 \Rightarrow e^* = \frac{p^H - (1-\alpha)\delta_i}{2c_i}$

Backward induction to determine $\hat{\delta}$ for which individuals are indifferent between choosing a higher piece-rate by making an unethical decision and being honest with the lower piece-rate:

$$p^L(\tfrac{p^L+\pi_i}{2c_i}) - c_i(\tfrac{p^L+\pi_i}{2c_i})^2 + \pi_i(\tfrac{p^L+\pi_i}{2c_i}) \geq$$
$$p^H(\tfrac{p^H+\pi_i-\delta_i}{2c_i}) - c_i(\tfrac{p^H+\pi_i-\delta_i}{2c_i})^2 + \pi_i(\tfrac{p^H+\pi_i-\delta_i}{2c_i}) - \delta_i(\tfrac{p^H+\pi_i-\delta_i}{2c_i})$$

$$\tfrac{1}{4c_i}(p^L + \pi_i)^2 \geq \tfrac{1}{4c_i}(p^H + \pi_i - \delta_i)^2$$

$$\delta_i \geq p^H - p^L$$

4.8 Appendix C: Instructions and Questionnaire

Instructions

Note: Instructions translated from German. German instructions available upon request.

General information about the experiment

We welcome you to this experiment. Please read the first page of the instructions and the detailed explanations of this experiment carefully. We will read everything aloud afterwards.

If you read the following explanations carefully, then - depending on your decisions and the decisions of the other participants - you can earn money in addition to the 4 euros that you receive as an entry fee for your participation. It is therefore very important that you read these explanations carefully. If you have any questions, please raise your hand. We will then come to you and answer them.

During the experiment, you are not allowed to talk to the other participants or to use your cell phone.

In this experiment, we do not refer to Euros but to points. So your income will first be calculated in points. These will be converted to euros at the end of the experiment, where:

$$25 \text{ points} = 1 \text{ euro}$$

At the end of today's experiment, we will pay you your earnings from the experiment plus 4 euros in cash for showing up.

Neither before nor after the study will you learn the identity of the participants you interact with during the study. Likewise, these participants will not learn anything about your identity.

In the following pages, we will explain the exact procedure of the experiment.

Detailed information about the experiment

In this experiment, you will be divided into groups of two. One member will take on the role "A" and the other member the role "B". You will find out which role you take before the experiment starts via your computer screen. So if you are a member "A", you will be assigned a member "B". On the other hand, if you are a member "B", you will be assigned a member "A".

Task member "B":

Member "B" must perform one of the five tasks below. For this task "B" has 10 minutes. There are five possible tasks that "B" can perform:

- Task "green",
- Task "blue",
- Task "orange",
- Task "red",
- Task "yellow".

Each of these tasks consists of many small subtasks. Member "B" does not find out exactly what these subtasks consist of and how much he will earn for each subtask until shortly before the task begins. In any case, the more subtasks member "B" solves correctly, the higher the earnings for the group of two.

Decision on the task:

Member "A" and member "B" simultaneously decide which task they think member "B" should perform. With 50% probability, the decision of "A" is implemented and with 50% probability, the decision of "B" is implemented.

To decide which task "B" should perform, an experimenter comes to each participant individually and brings five different colored envelopes. Each envelope represents the task in the corresponding color. All participants can then choose between each of the five envelopes, i.e. the five tasks.

- Task "green",
- Task "blue",
- Task "orange",
- Task "red",

- Task “yellow”.

Keep in mind that at the time of your election you do not know what the chosen task consists of, nor what the associated earning for each subtask is.

Entering the task and the piece rate:

Once you have chosen an envelope and thus a task, you are prompted to open the selected envelope. The envelope contains three pieces of information:

- a password that you must type on the screen to inform the computer of your choice,
- the unit wage intended for the task, which you must click on the screen to allow the computer to calculate your earnings,
- a numerical code that will become relevant later.

Please enter the first two pieces of information on the screen:

1. Make sure that the password is correct. This is the only way to display the selected task.
2. Make sure that you enter the unit wage correctly. The piece rate you specify will be used to calculate your earnings.

Earnings:

Please note: The earnings are the same for both members and consist of the number of subtasks correctly solved by ”B” multiplied by the respective specified piece rate.

Earnings = number ob subtasks solved by ”B” * reported piece-rate.

This means:

- The more subtasks ”B” solves, the higher the earnings.
- The higher the specified piece rate, the higher the earnings.

If you report a piece rate other than the one provided, it will be a violation of the rules just mentioned. Your earnings will be calculated using the piece rate you have reported and not the intended piece rate. Accordingly, you will receive a piece rate that is not

intended for the task.

Random decision:

The random device determines which decision (the chosen task of member "A" or the chosen task of member "B") is selected and thus carried out by "B".

a) If the decision of member "A" is determining:

   - Member "B" learns via the computer the task selected by "A" and the piece rate reported by "A".
   - In addition, "B" receives the envelope chosen by "A". In this envelope, in the third place (after the password and the intended piece rate), there is a numerical code.
   - "B" is requested to enter this numerical code, only then "B" can start processing the task.

   For information: By entering the numerical code, "B" can neither change the task selected by "A" nor the piece rate specified by "A". The entry of the numerical code is solely to ensure that "B" is informed of the choice made by "A", the intended and the reported piece rate wage, and is ready to begin the task.

b) If the decision of member "B" is determining:

   - "B" is requested to enter the numerical code located in his chosen envelope in the third place (after the password and the intended piece rate).
   - Only then "B" can start working on the task.

   By entering the numerical code, "B" cannot change the choice he has made regarding the task, nor his specified piece rate. The entry of the numerical code serves exclusively to ensure that "B" is ready to start with the task.

Task for member "A":

While "B" is performing the task chosen by "A" or "B", "A" is asked to perform another task. This task is not decisive for the payment of the two members. "A" is informed about the task type directly via the screen.

Control questions:

Before starting and being told what role to take and choosing an envelope, all participants are asked to answer some control questions to make sure that all participants have understood the instructions.

Questionnaire

Control questions are displayed only on the screens before the actual experiment starts:

Please state whether the following statements are true or false:

1. Both participants decide which task "B" should perform, but only one of the two decisions (that of "A" or "B") is relevant.
2. Member "B" always performs the task he has chosen.
3. The input of the password and the piece rate is decisive for the chosen task and the earnings.
4. The number of subtasks correctly solved by "B" has no effect on the earnings of "A".
5. The higher the specified piece rate for the selected task, the greater the earnings.

CHAPTER 5
REFERENCES

Abbink, K. and Wu, K. (2017). Reward self-reporting to deter corruption: An experiment on mitigating collusive bribery. *Journal of Economic Behavior & Organization*, 133:256–272.

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.

Ali, S. N. and Bénabou, R. (2020). Image versus information: Changing societal norms and optimal privacy. *American Economic Journal: Microeconomics*, 12(3):116–64.

Almås, I., Cappelen, A. W., and Tungodden, B. (2020). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *Journal of Political Economy*, 128(5):1753–1788.

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of political Economy*, 97(6):1447–1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477.

Andrighetto, G., Brandts, J., Conte, R., Sabater-Mir, J., Solaz, H., and Villatoro, D. (2013). Punish and voice: punishment enhances cooperation when combined with norm-signalling. *PloS one*, 8(6):e64941.

Apesteguia, J., Dufwenberg, M., and Selten, R. (2007). Blowing the whistle. *Economic Theory*, 31(1):143–166.

Ariely, D., Gneezy, U., Loewenstein, G., and Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76(2):451–469.

Ariely, D., Kamenica, E., and Prelec, D. (2008). Man's search for meaning: The case of legos. *Journal of Economic Behavior & Organization*, 67(3-4):671–677.

Babcock, P., Bedard, K., Charness, G., Hartman, J., and Royer, H. (2015). Letting down the team? Social effects of team incentives. *Journal of the European Economic Association*, 13(5):841–870.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2014a). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45):15924–15927.

Balafoutas, L., Nikiforakis, N., and Rockenbach, B. (2014b). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, 111(45):15924–15927.

Bandiera, O., Barankay, I., and Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *The Quarterly Journal of Economics*, 120(3):917–962.

Bartling, B. and Fischbacher, U. (2011). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bartuli, J., Djawadi, B., and Fahr, R. (2016). Business ethics in organizations: An experimental examination of whistleblowing and personality. *available at IZA DP No. 10190*.

Becker, G. S. (1968). Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer.

Bénabou, R., Falk, A., and Tirole, J. (2019). Narratives, imperatives, and moral persuasion. *NBER working paper*.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Benabou, R. and Tirole, J. (2011). Laws and norms. Technical report, National Bureau of Economic Research.

Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–64.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2022). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. *Available at SSRN 3326146*.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? the effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, 188:209–235.

Bigoni, M., Fridolfsson, S.-O., Le Coq, C., and Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *The RAND Journal of Economics*, 43(2):368–390.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 90(1):166–193.

Bosman, R. and Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *The Economic Journal*, 112(476):147–169.

Bowen, R. M., Call, A. C., and Rajgopal, S. (2010). Whistle-blowing: target firm characteristics and economic consequences. *The Accounting Review*, 85(4):1239–1271.

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

Brandts, J. and Solà, C. (2001). Reference points and negative reciprocity in simple sequential

games. *Games and Economic Behavior*, 36(2):138–157.

Buckenmaier, J., Dimant, E., and Mittone, L. (2018). Effects of institutional history and leniency on collusive corruption and tax evasion. *Journal of Economic Behavior & Organization*.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–48.

Butler, J. V., Serra, D., and Spagnolo, G. (2019). Motivating whistleblowers. *Management Science*.

Call, A., Martin, G., Sharp, N., and Jaron, W. (2018). Whistleblowers and outcomes of financial misrepresentation enforcement actions. *Journal of Accounting Research*, 56(1):123–171.

Carpenter, J. and Gong, E. (2016). Motivating agents: How much does the mission matter? *Journal of Labor Economics*, 34(1):211–236.

Carpenter, J., Robbett, A., and Akbar, P. (2018). Profit sharing and peer reporting. *Management Science*, 64(9):4261–4276.

Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics*, 8(2):107–115.

Cassar, L. (2018). Job mission as a substitute for monetary incentives: Benefits and limits. *Management Science*.

Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133.

Charness, G. (2000). Responsibility and effort in an experimental labor market. *Journal of Economic Behavior & Organization*, 42(3):375–384.

Charness, G., Cobo-Reyes, R., Jiménez, N., Lacomba, J. A., and Lagos, F. (2012). The hidden advantage of delegation: Pareto improvements in a gift exchange game. *American Economic Review*, 102(5):2358–79.

Charness, G., Cobo-Reyes, R., and Sanchez, A. (2016). The effect of charitable giving on workers' performance: Experimental evidence. *Journal of Economic Behavior & Organization*, 131:61–74.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Choo, L., Grimm, V., Horváth, G., and Nitta, K. (2019). Whistleblowing and diffusion of responsibility: An experiment. *European Economic Review*, 119:287–301.

Cialdini, R. B. (1983). *Influence: The psychology of persuasion*. HarperCollins New York City.

d'Adda, G., Dufwenberg, M., Passarelli, F., and Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, 124:288–304.

Dal Bó, P., Foster, A., and Kamei, K. (2019). The democracy effect: a weights-based identification strategy. Technical report, National Bureau of Economic Research.

Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5):2205–29.

Darley, J. M. and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of personality and social psychology*, 8(4p1):377.

De Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., et al. (2004). The neural basis of altruistic punishment. *Science*, 305(5688):1254.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, 18(1):105.

Dimant, E. and Gesche, T. (2021). Nudging enforcers: How norm perceptions and motives for lying shape sanctions.

Dyck, A., Morse, A., and Zingales, L. (2010). Who blows the whistle on corporate fraud? *The Journal of Finance*, 65(6):2213–2253.

Egas, M. and 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.

Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.

Engel, C. (2014). Social preferences can make imperfect sanctions work: Evidence from a public good experiment. *Journal of Economic Behavior & Organization*, 108:343–353.

Engl, F., Riedl, A., and Weber, R. (2021). Spillover effects of institutions on cooperative behavior, preferences, and beliefs. *American Economic Journal: Microeconomics*, 13(4):261–99.

Falk, A. (2021). Facing yourself–a note on self-image. *Journal of Economic Behavior & Organization*, 186:724–734.

Falk, A. and Kosfeld, M. (2006). The hidden costs of control. *American Economic Review*, 96(5):1611–1630.

Fallucchi, F. and Nosenzo, D. (2021). The coordinating power of social norms. *Experimental Economics*, pages 1–25.

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

Fehr, E. and Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in cognitive sciences*, 8(4):185–190.

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

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, 90(4):980–994.

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

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–140.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458–468.

Fehrler, S. and Kosfeld, M. (2014). Pro-social missions and worker motivation: An experimental study. *Journal of Economic Behavior & Organization*, 100:99–110.

Feltovich, N. and Hamaguchi, Y. (2016). The effect of whistle-blowing incentives on collusion: An experimental study of leniency programmes. Technical report, mimeo, Monash University.

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

Fowler, J. H., Johnson, T., and Smirnov, O. (2005). Egalitarian motive and altruistic punishment. *Nature*, 433(7021):E1–E1.

Gächter, S., Gerhards, L., and Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, 97:72–86.

Gagné, M. and Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, 26(4):331–362.

Galbiati, R., Henry, E., and Jacquemet, N. (2019). Learning to cooperate in the shadow of the law.

Galbiati, R., Henry, E., Jacquemet, N., and Lobeck, M. (2021). How laws affect the perception of norms: Empirical evidence from the lockdown. *PloS one*, 16(9).

Galbiati, R. and Vertova, P. (2008). Obligations and cooperative behaviour in public good games. *Games and Economic Behavior*, 64(1):146–170.

Galizzi, M. M. and Whitmarsh, L. (2019). How to measure behavioral spillovers: a methodological review and checklist. *Frontiers in psychology*, 10:342.

Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Guadagno, R. E. and Cialdini, R. B. (2010). Preference for consistency and social influence: A review of current research findings. *Social Influence*, 5(3):152–163.

Gürerk, O., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312(5770):108–111.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral

judgment. *Psychological review*, 108(4):814.

Haidt, J. and Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

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

Hinloopen, J. and Soetevent, A. R. (2008). Laboratory evidence on the effectiveness of corporate leniency programs. *The RAND Journal of Economics*, 39(2):607–616.

Holt, C. A. and Sullivan, S. (2021). Permutation tests for experimental data. *Available at SSRN 3957609*.

Hopfensitz, A. and Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, 119(540):1534–1559.

Houser, D., Xiao, E., McCabe, K., and Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, 62(2):509–532.

Kagel, J. H. and Roth, A. E. (2020). *The handbook of experimental economics, volume 2*. Princeton university press.

Kajackaite, A. and Sliwka, D. (2017). Social responsibility and incentives in the lab: Why do agents exert more effort when principals donate? *Journal of Economic Behavior & Organization*, 142:482–493.

Kant, I. (1787). *Kant's gesammelte Schriften. Hrsg. von der königlich preussischen Akademie der Wissenschaften, Berlin: G.Reimer, 1900ff.*

Kennedy, F. E. (1995). Randomization tests in econometrics. *Journal of Business & Economic Statistics*, 13(1):85–94.

Kessler, J. B. and Leider, S. (2012). Norms and contracting. *Management Science*, 58(1):62–77.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

Kölle, F. and Quercia, S. (2021). The influence of empirical and normative expectations on cooperation. *Journal of Economic Behavior & Organization*, 190:691–703.

Kosfeld, M., Okada, A., and Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, 99(4):1335–55.

Krupka, E. L. and 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.

Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Lergetporer, P., Angerer, S., Glätzle-Rützler, D., and Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional

cooperation. *Proceedings of the National Academy of Sciences*, 111(19):6916–6921.

Masclet, D., Noussair, C., Tucker, S., and Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93(1):366–380.

Miceli, M. P. and Near, J. P. (1984). The relationships among beliefs, organizational position, and whistle-blowing status: A discriminant analysis. *Academy of Management journal*, 27(4):687–705.

Miceli, M. P. and Near, J. P. (1996). Whistle-blowing: myth and reality. *Journal of Management*, 22(3):507–526.

Miceli, M. P. and Near, J. P. (2002). What makes whistle-blowers effective? three field studies. *Human Relations*, 55(4):455–479.

Miceli, M. P., Near, J. P., and Dworkin, T. M. (2008). A word to the wise: how managers and policy-makers can encourage employees to report wrongdoing. *Journal of Business Ethics*, 86(3):379–396.

Muehlheusser, G., Roider, A., and Mechtenberg, L. (2020). Whistle-blower protection: Theory and experimental evidence. *European Economic Review*, 126.

Near, J. P. and Miceli, M. P. (1985). Organizational dissidence: The case of whistle-blowing. *Journal of business ethics*, 4(1):1–16.

Oexl, R. and Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics*, 16(3):306–312.

Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American political science Review*, 86(2):404–417.

Ouss, A. and Peysakhovich, A. (2015). When punishment doesn't pay: Cold glow and decisions to punish. *The Journal of Law and Economics*, 58(3):625–655.

Peysakhovich, A. and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302.

Reuben, E. and Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1):122–137.

Reuben, E. and Stephenson, M. (2013). Nobody likes a rat: on the willingness to report lies and the consequences thereof. *Journal of Economic Behavior & Organization*, 93:384–391.

Reuben, E. and Van Winden, F. (2008). Social ties and coordination on negative reciprocity: The role of affect. *Journal of Public Economics*, 92(1-2):34–53.

Reuben, E. and van Winden, F. (2010). Fairness perceptions and prosocial emotions in the power to take. *Journal of Economic Psychology*, 31(6):908–922.

Rivas, M. F. and Sutter, M. (2011). The benefits of voluntary leadership in experimental public goods games. *Economics Letters*, 112(2):176–178.

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78.

Schmolke, K. U. and Utikal, V. (2018). Whistleblowing: Incentives and situational determinants. *Available at SSRN 3198104*.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *The Review of Economic Studies*, 71(2):513–534.

Sigmund, K. (2007). Punish or perish? retaliation and collaboration among humans. *Trends in ecology & evolution*, 22(11):593–600.

Sloof, R. and von Siemens, F. A. (2019). Effective leadership and the allocation and exercise of power in organizations. *The Leadership Quarterly*.

Stubben, S. R. and Welch, K. T. (2020). Evidence on the use and efficacy of internal whistleblowing systems. *Journal of Accounting Research*, 58(2):473–518.

Sunstein, C. R. (1996). On the expressive function of law. *University of Pennsylvania law review*, 144(5):2021–2053.

Sutter, M., Haigner, S., and Kocher, M. G. (2010). Choosing the carrot or the stick? endogenous institutional choice in social dilemma situations. *The Review of Economic Studies*, 77(4):1540–1566.

Tonin, M. and Vlassopoulos, M. (2015). Corporate philanthropy and productivity: Evidence from an online real effort experiment. *Management Science*, 61(8):1795–1811.

Tyran, J.-R. and Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *scandinavian Journal of Economics*, 108(1):135–156.

Waytz, A. (2016). Whistleblowers are Motivated by Moral Reasons Above Monetary Ones.

Waytz, A., Dungan, J., and Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology*, 49(6):1027–1033.

Weber, R. A. (2003). 'learning'with no feedback in a competitive guessing game. *Games and Economic Behavior*, 44(1):134–144.

Xiao, E. (2018). Punishment, social norms, and cooperation. In *Research Handbook on Behavioral Law and Economics*. Edward Elgar Publishing.