WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Robert Stüber

The Benefit of the Doubt: Willful Ignorance and Altruistic Punishment

Discussion Paper

SP II 2019–215

November 2019

WZB Berlin Social Science Center

Research Area
Markets and Choice

Research Unit
Market Behavior

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Discussion papers of the WZB serve to disseminate the research results of work in progress prior to publication to encourage the exchange of ideas and academic debate. Inclusion of a paper in the discussion paper series does not constitute publication and should not limit publication in any other venue. The discussion papers published by the WZB represent the views of the respective author(s) and not of the institute as a whole.

Robert Stüber
The Benefit of the Doubt: Willful Ignorance and Altruistic Punishment

Affiliation of the author:

Robert Stüber
WZB Berlin Social Science Center

Wissenschaftszentrum Berlin für Sozialforschung gGmbH
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Abstract

The Benefit of the Doubt: Willful Ignorance and Altruistic Punishment

by Robert Stüber*

Altruistic punishment is often thought to be a major enforcement mechanism of social norms. I present experimental results from a modified version of the dictator game with third-party punishment, in which third parties can remain ignorant about the choice of the dictator. I find that a substantial fraction of subjects choose not to reveal the dictator's choice and not to punish the dictator. I show that this behavior is in line with the social norms that prevail in a situation of initial ignorance. Remaining ignorant and choosing not to punish is not inappropriate. As a result, altruistic punishment is significantly lower when the dictator's choice is initially hidden. The decrease in altruistic punishment leads to more selfish dictator behavior only if dictators are explicitly informed about the effect of willful ignorance on punishment rates. Hence, in scenarios in which third parties can ignore information and dictators know what this implies, third-party punishment may only ineffectively enforce social norms.

Keywords: Third-party punishment, Willful ignorance, Sorting, Social preference

JEL classification: C91; D01; D63; D83

*E-mail: robert.stueber@wzb.eu.
I am grateful to Philipp Albert, Kai Barron, Björn Bartling, Alexander Cappelen, Marvin Deversi, Urs Fischbacher, Zack Grossman, Rustamdjan Hakimov, Erin Krupka, Dorothea Kübler, Homayoon Moradi, Lisa Spantig, Kristina Strohmaier, Joël van der Weele, Justin Valasek, Roel van Veldhuizen and seminar participants at the ESA World Meeting 2018, the IMEBESS 2018, the SABE/IAREP 2018, the Lisbon Meeting in Economics and Political Science 2018, the theem 2019, and the Berlin Behavioral Economics Workshop for their helpful suggestions. I am also very thankful to Jennifer Rontganger for copy editing.

1 Introduction

A large and influential strand of literature shows that individuals are willing to punish other individuals if they violate social norms, even if the punishment comes at a monetary cost and yields no material gain (e.g., Fehr and Gächter, 2000; Fehr and Fischbacher, 2004; Carpenter, 2007; Carpenter and Matthews, 2012). Some of these studies show that this altruistic punishment of norm violations is even conducted by third parties, whose own economic payoff is unaffected by the norm violation. The willingness to altruistically punish norm violations has been suggested as being one major enforcement mechanism of social norms. In turn, social norms that are enforced by social sanctions are seen as a key driver of cooperation between strangers, individuals' willingness to be generous, and the existence of human societies more generally.[1]

Social preferences are thought to be the reason for third-party punishment: Unaffected third parties punish subjects who violate norms, although it is costly and they receive no material benefit from it. They do so, presumably, because they expect a benefit for others.[2] However, more recent studies emphasize that people willfully ignore information and in turn exploit ambiguities about the consequences of their actions. In a seminal paper by Dana et al. (2007) and a plethora of follow-up studies (Larson and Capra, 2009; Cain and Dana, 2012; Grossman, 2014; Feiler, 2014; van der Weele, 2014; Grossman and van der Weele, 2017; Moradi and Nesterov, 2017), it is shown that dictator game-giving declines when subjects can choose not to reveal how their actions affect a passive recipient's payoff.[3]

It is an open question whether people's tendency to remain ignorant in order to avoid costly moral behavior might also lower their willingness to altruistically punish norm violations. At the same time, one can think of many real-world scenarios in which willful ignorance might impact altruistic punishment. A university professor supervising an exam might prefer to look away rather than check carefully whether a student brought a forbidden cheat sheet, knowing that if she found the cheat sheet she would need to engage in a nerve-wracking discussion with the cheating student and exclude him from writing the exam. A restaurant manager who suspects that one of his waitresses is not sharing her tips with her colleagues as agreed upon, might be reluctant to check, because finding out that the waitress was not sharing her tips would imply the need to confront her, which would have detrimental effects on the working atmosphere and, ultimately, his profits. And, when a firm asks to be

[1] In this vein, the findings on altruistic punishment have shaped research in various fields such as economics, biology, anthropology, psychology, and neuroscience.

[2] Third-party punishment can also be consistent with inequity aversion (Fehr and Schmidt, 1999) or spite (Levine, 1998), see also Leibbrandt and López-Pérez (2012).

[3] In more applied settings, Kandul (2016) and Kajackaite (2015) also document that subjects remain willfully ignorant in order to make selfish decisions.

paid for its services without providing an invoice, people might refrain from asking for the invoice and from checking whether it contains the total amount – even though they would find it unfair if the firm were to evade taxes – because asking might increase the amount one has to pay oneself.

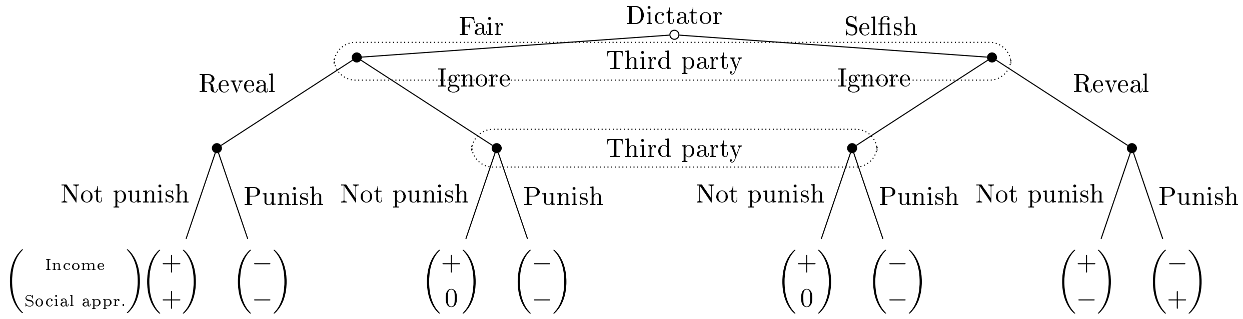
To study the effect of willful ignorance on third-party punishment, I run a laboratory experiment in which I modify the dictator game with third-party punishment ("third-party punishment game"; Fehr and Fischbacher, 2004). The game consists of two stages, each played by a group of three players: a dictator, a recipient, and a third party. In the first stage, the dictator decides between a selfish option, giving him a high payoff and the recipient a low payoff, or a fair option that gives a lower payoff to the dictator but a higher payoff to the recipient. In the second stage, the third party whose payoff is unaffected by the dictator's choice, has the opportunity to punish the dictator. I vary whether the third party always observes the dictator's choice prior to making its decision (baseline treatment) or can choose to reveal it at no cost (hidden information treatment).

I find that a substantial fraction (36%) of third parties avoid learning the choice of the dictator in the hidden information treatment. These third parties act as if the dictator has chosen the fair option. Dictators who choose the fair option are almost never punished. As a result, the fraction of subjects choosing to altruistically punish a selfish dictator is significantly lower when the information about the dictator's behavior is initially unobserved compared to when it is exogenously provided: The frequency of altruistic punishment decreases by 50%. Hence, the possibility to avoid information diminishes third-party punishment. Surprisingly, although this drastic decrease in altruistic punishment significantly alters dictators' payoffs and their payoff-maximizing choice, dictators do not choose the selfish option more often. There is no treatment difference in dictator choices.

In a second step, I investigate the social norms related to third-party punishment using the incentivized norm elicitation method proposed by Krupka and Weber (2013) in a separate experiment. Eliciting social norms is important for two reasons. Firstly, the social norms related to third-party punishment *under full information*, i.e., without moral wiggle room, have not been investigated in the literature before. The most important question in this context is whether punishing a selfish dictator is seen as being prosocial. The results make clear that punishment is indeed seen as the moral action. Secondly, I provide evidence regarding the social norms that prevail in a situation of initial ignorance. I show that the norm prescribes revealing the information about the dictator's choice. Hence, ignorance is not appropriate. I then explore two ways in which the social norms might still be in line with the choices observed. I do not find that it is more or less appropriate to punish a norm violation depending on whether the information about the norm violation was revealed or exogenously given. Contrarily, if a subject chooses to remain ignorant of the dictator's choice, then the social norm prescribes not punishing the dictator. And, little can be gained in terms of appropriateness by revealing the norm violation and punishing. This finding deviates from the result for the dictator game with hidden information about the recipient's payoff and indicates why the possibility to remain ignorant might have a particularly strong impact on altruistic punishment.

In a third step, I explore why some subjects remain ignorant in order to avoid altruistic punishment. I follow two approaches. I first investigate whether the choices observed in the experiment can be predicted on an aggregate level based on monetary payoffs and the measured social norms (see Figure 1). A third party who reveals the information might observe a fair dictator, in which case both monetary and normative incentives are aligned, or a selfish dictator, in which case she faces a trade-off between punishing the dictator (which is costly, but appropriate) or not (which saves on income, but is inappropriate). In contrast, for a third party who remains ignorant, not punishing is income-maximizing and not socially inappropriate. Depending on how strongly a third party weighs monetary incentives compared to adhering to the social norm, she hence might choose to remain ignorant and to not punish. Doing a similar exercise for all possible strategies of the third party, I find that monetary incentives and social norms can well predict the distribution of choices. I then allow for observable heterogeneity between subjects and examine whether different choices can be explained by the subjects being of varying social types. Considering a measure of prosociality, I find that the third parties who reveal a selfish dictator choice and choose to punish are the most prosocial. However, if I analyze punishment choices depending on whether third parties selected into receiving information on the norm violation or were exogenously informed of it, and if I consider a measure of self-image, I do not find evidence of a sorting of types.

Figure 1: Predicting Behavior Based on Monetary Payoffs and Norms



Note: Figure 1 shows a stylized version of the game tree in the hidden information treatment including the income implied by and social appropriateness of the third party's actions.

In a final step, I revisit the finding that dictator behavior does not vary across treatments. Whether dictator behavior can be affected by the treatment variation is crucial, because no treatment difference in dictator behavior implies a null-effect on norm compliance. In two additional treatments, the baseline informed treatment and the hidden information informed treatment, dictators are informed about the proportion of dictators that were punished conditional on their choice (selfish or fair) in the baseline and the hidden information treatment, respectively. I find that when dictators have perfect information on the consequences of willful ignorance for punishment, dictators hold adjusted beliefs and behave significantly more selfishly when their choice is initially hidden.

The first contribution of this study is to show that third-party punishment significantly decreases if the third parties have the possibility to remain ignorant of a potential norm violation. This has important implications for future research and policy. Third-party punishment may not be as effective in enforcing social norms as previously thought, given that in a richer design that allows for avoiding information but leaves everything else unchanged, I no longer find it to be very common. I provide evidence on the social appropriateness of altruistically punishing norm violations supporting the narrative about a socially appropriate punishment, which has been used, but not analyzed, in the literature on third-party punishment so far. I also provide evidence on the social norms that prevail in a situation in which there is initial uncertainty over whether a norm violation has taken place. Based on this, I explain *why* the possibility to remain ignorant reduces altruistic punishment by explaining choices based on norms and monetary incentives. Moreover, I analyze whether behavior under initial ignorance can be explained by individual characteristics (such as a subject's prosociality), which also allows testing theoretical predictions. Finally, I provide evidence showing that whether willful ignorance is likely to affect norm compliance depends on the salience of its effect to the dictators.

The remainder of this paper is structured as follows. In section 2, I briefly discuss the related literature. In section 3, I describe the main experiment and its results. Section 4 analyzes the social norms in punishment behavior. Section 5 is dedicated to explaining choices. In section 6 the effect of providing information about punishment rates on norm compliance is studied. Section 7 concludes.

2 Related Literature

Two existing studies consider moral wiggle room and altruistic punishment. Kriss et al. (2016) analyze the resoluteness of altruistic punishment. In the study, subjects first make a punishment decision and are then asked to report the outcome of a die roll that determines whether the punishment is actually implemented. The study clearly shows that the decisions of third parties to punish norm violations are reluctant, they avoid actually implementing the costly punishment they previously intended.[4] A nice feature of the design of Kriss et al. (2016) is that it allows measuring the effect of moral wiggle room without the experimenter observing which subjects exploit the moral wiggle room. It hence provides a measure of the reluctance to altruistically punish that is not contaminated by the third parties' choices being observed by the experimenter. My study adds to Kriss et al. (2016) in four major respects. First, my finding that third-party punishment decreases through willful ignorance is consistent with the finding in Kriss et al. (2016). Yet, while I analyze a different form of moral wiggle room such that there are a lot of real-life settings captured by the design of this study but not by the design of Kriss et al. (2016), and vice versa, both studies also have very different policy implications.[5] In particular, I analyze scenarios in which the information about a potential norm violation is accessible, but people choose not to access it, while in Kriss et al. (2016) whether a norm violation has taken place is always observed but third parties might revoke their punishment decision. Second, the mechanisms behind both forms of moral wiggle room are likely to differ. While both findings imply that third parties have a preference for not punishing norm violations without clearly signaling this preference, in Kriss et al. (2016) subjects do not seem to be strategic with regard to this: The decision of whether to state an intention to punish is not affected by whether a third party will subsequently have the opportunity to avoid implementing this decision by misreporting the die roll. I demonstrate that third parties *deliberately* do not reveal information that might force them to engage in altruistic punishment. Hence, the sheer possibility that one would punish a norm violation if one observed it, seems to be sufficient for subjects not to feel selfish if they avoid altruistic punishment by being ignorant.[6] Thus, both studies have different implications for theoretically modeling punishment behavior. Third, I provide a quantitive measure on how willful ignorance influences third-party punishment.[7] Fourth, by observing which (types of) individuals remain willfully ignorant, I am able to shed light on the underlying mechanisms driving willful ignorance.

Bartling et al. (2014) analyze how dictators might be able to deter punishment from third parties by remaining deliberately ignorant. The decisive difference between their study

[4] These authors also show that, in contrast, second-party punishment is substantially more resolute.

[5] One message learned from Kriss et al. (2016) is that the rate at which norm violations are altruistically punished will be higher if third parties cannot get out of their punishment decision. This study shows that this rate is higher when third parties are exogenously informed about norm violations.

[6] In Kriss et al. (2016) subjects signal once to themselves that they are willing to punish norm violations.

[7] While providing clean evidence of an effect of moral wiggle room on third-party punishment, the effect cannot be quantified in Kriss et al. (2016), as a preference for being honest may cause third parties to not misreport the outcome of the dice roll.

and mine is that the willful ignorance involves *dictators* forgoing information (how their actions affect others), while I focus on *third parties* remaining ignorant (about the choices of others). Bartling et al. (2014) show that ignorant dictators are punished less than dictators who reveal the consequences of their actions (before implementing them) if their actions lead to an unfair outcome. Hence, combining the findings from their study and mine might suggest that moving from a set-up of full information to a setting where the actors can remain ignorant, the frequency in which norm violations are punished is reduced via two different channels: Dictators who remain ignorant and violate a distributional norm are less often punished *and* third parties who can remain ignorant about the potential violation of a distributional norm punish less often.

My finding that people willfully ignore information to avoid a moral obligation corroborates the finding of Dana et al. (2007) and follow-up studies that people exploit ignorance as a form of moral wiggle room.[8] My results go beyond existing findings by showing that people willfully ignore information about the behavior of others (instead of the outcome of random draws) and that this ignorance changes the amount of altruistic punishment (instead of generous giving), for which the welfare-enhancing effect is uncertain and lies in the future. I hence show that willful ignorance generalizes to a setting that describes a more complex social interaction.[9] By analyzing whether specific types of subjects remain ignorant, my findings bear on studies that explore whether there is a sorting of types in generosity decisions (Dana et al., 2007; Larson and Capra, 2009; Kajackaite, 2015; Grossman and van der Weele, 2017). In this regard my findings also relate to findings showing that people actively avoid situations in which being generous is possible (Dana et al., 2006; DellaVigna et al., 2012; Lazear et al., 2012; Trachtman et al., 2015; Andreoni et al., 2017).

By investigating how willful ignorance influences third-party punishment, my findings address a large body of literature analyzing the robustness of altruistic punishment (e.g., Charness et al., 2008; Egas and Riedl, 2008; Nikiforakis, 2008; Lewisch et al., 2011; Lotz et al., 2011; Nikiforakis and Engelmann, 2011; Balafoutas and Nikiforakis, 2012; Nikiforakis and Mitchell, 2014; Balafoutas et al., 2016; Goeschl and Jarke, 2016).[10] These studies show that the extent of altruistic punishment strongly depends on its design as, for instance, the possibility of reward or retaliation. I can show that even a change in the *information structure* decreases altruistic punishment.

[8] Similarly, van der Weele et al. (2014), Matthey and Regner (2015) and Regner (2018) study the extent to which forms of moral wiggle room other than willful ignorance affect negative and positive reciprocity.

[9] Bartling et al. (2015) and Felgendreher (2018) find that the possibility to avoid information does not have a strong impact on consumption decisions in markets.

[10] Some studies provide evidence that altruistic punishment is affected by the diffusion of responsibility of the dictator or the directness of his decisions (Coffman, 2011; Bartling and Fischbacher, 2012; Oexl and Grossman, 2013).

Finally, my findings also closely relate to previous studies on whether variations in social norms translate into variations in actual behavior and/or investigating norms under initial ignorance (e.g., Krupka and Weber, 2013; Gächter et al., 2013; Gächter et al., 2017; Grossman and van der Weele, 2017; Fehr and Schurtenberger, 2018). I am the first to document a norm for punishing selfish dictators as well as a strong norm for revealing information about whether a norm violation took place. I show that norms can help to explain altruistic punishment both under full information and when individuals have the possibility to remain ignorant.

3 Altruistic Punishment under Willful Ignorance

3.1 Experimental Design and General Procedures

The main experimental game is a modified version of the dictator game with third-party punishment ("third-party punishment game;" Fehr and Fischbacher, 2004). It consists of two stages and three players: a dictator, a recipient, and a third party. In the first stage, the dictator makes a binary decision that affects his income and the income of the recipient. The dictator can either choose option A1 which gives him a high payoff of €6 but a low payoff of €1 to the recipient, or option A2 that gives him a lower payoff of €4 but leaves a higher payoff of €4 to the recipient. For now, I label option A1 the egoistic option and option A2 the fair option.[11] The third party is unaffected by the dictator's decision and is informed that she receives €6 as an endowment. This stage is the same in both treatments.

Dictator chooses	Dictator receives	Recipient receives
A1	€6	€1
A2	€4	€4

The second stage differs between treatments. In the second stage of the baseline treatment the impartial third party immediately observes the dictator's action and can decide to punish the dictator (option C1) or not (option C2). In the second stage of the hidden information treatment, the third party does not observe the choice of the dictator, but can reveal it at no cost. Irrespective of whether the third party reveals the choice of the dictator, she can decide between options C1 and C2. In both the baseline and the hidden information treatment punishing reduces the payoff of the third party by €1 and that of the dictator by

[11] The social norm elicitation in section 4 will suggest that this labeling is justified. Note that option A2 is also more efficient than option A1, as it implies a joint income of €8 in stage 1 for the dictator and the recipient rather than a payoff of €7.

€ 3, but does not affect the payoff of the recipient. Therewith, the third party's payoff is € 5 (€ 6) if she chooses (not) to punish the dictator.

Experimental subjects played the game only once in a between-subject design and in groups of three. The roles were labeled neutrally, that is, the dictator was called "participant A," the recipient "participant B," and the third party "participant C." In order to obtain punishment decisions by two-thirds of the subjects, recipients and third parties played the game under role uncertainty, i.e., they were informed that they were either participant B or participant C and were asked to make their decision as participant C with their true roles assigned ex post. If assigned the role of a third party, their choice was implemented. If assigned the role of a receiver, their choice had no consequence.[12]

To ease comparison with Dana et al. (2007), I adapted several features of their design. First, the "decision-maker" (the dictator in the study of Dana et al.; the third party in this study) decides between an egoistic choice that gives her € 6 or an altruistic choice, giving her € 5. Hence, the costs of the prosocial action and the decision-maker's potential incomes are the same. Second, the status quo of the decision is the same (inaction implies ignorance), which is likely to matter (see Grossman, 2014 and Cox et al., 2017). Third, the framing regarding the revelation decision is almost identical.

Decisions were made anonymously on separated computer terminals. Instructions were provided on screen and with common information within each treatment. To ensure that the subjects understood the game, prior to making their decisions, subjects had to correctly answer an extensive set of control questions. They were also informed about the screens they would see during the game, depending on their own and other subjects' choices. Hence, the third parties in the hidden information treatment knew that if they did not reveal the choice of the dictator, they would avoid learning about his and the recipient's final payoffs, as they would never be informed about them.

The experiment was conducted using z-Tree (Fischbacher, 2007) at the TU-WZB lab in Berlin. The recruitment was done using ORSEE (Greiner, 2015). The experiment was conducted in 10 sessions between December 2017 and February 2018. Two hundred and twenty-two subjects participated and, hence, I observe 148 third-party decisions (thereof 60 in the baseline treament and 88 in the hidden information treatment). About 95% of subjects were students. After the main experimental game was concluded, subjects were shown new instructions described in sections 3.3 and 5.2. On average, each session lasted approximately 42 minutes and the average payment was € 13.43, with a range between € 8

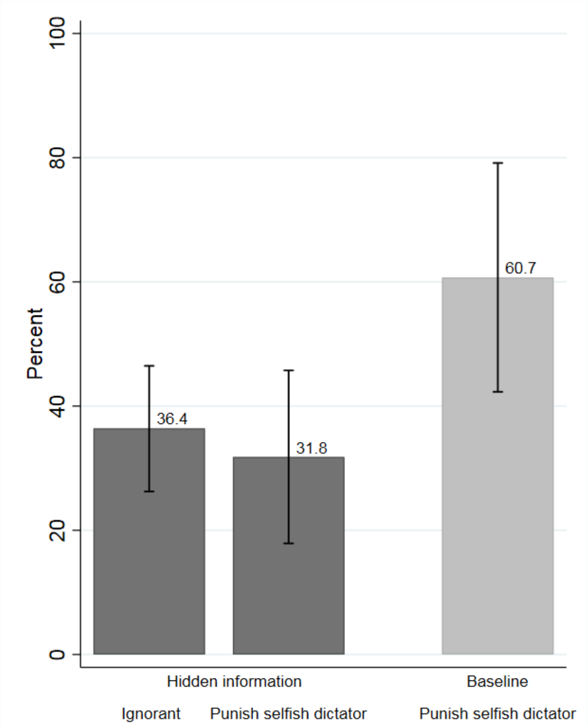
[12] It has been argued that eliciting third parties' punishment decision under role uncertainty does not influence treatment effects (Bartling et al., 2014; Nikiforakis and Mitchell, 2014). If there was an effect, for instance, by increasing the fraction of third parties who reveal the decision of the dictator, any treatment effect should be underestimated.

and €18.

3.2 Third-Party Behavior

Thirty-six percent (or 32 out of 88) of the third parties deliberately remain uninformed and do not reveal the dictator's choice (see the left bar in Figure 2). This fraction is a bit lower than the 44% of dictators remaining ignorant in Dana et al. (2007), but still sizable and even larger than the fraction of dictators remaining ignorant in Moradi and Nesterov (2017) (34%), which replicates Dana et al. (2007) using the same subject pool as I do shortly before my study took place.[13]

Figure 2: Percentage of third parties remaining ignorant and percentage choosing to punish an egoistic dictator by treatment



Note: The figure shows the percentage of third parties who remain ignorant in the hidden information treatment and the percentage choosing to punish an egoistic dictator for the hidden information and the baseline treatment along with 95%-confidence intervals.

Does the fact that people remain ignorant influence the rate at which selfish dictator choices are altruistically punished? Thirty-two percent of the third parties choose to punish an egoistic dictator in the hidden information treatment, but 61% of the third parties choose

[13] Subjects who participated in Moradi and Nesterov (2017) were not invited to participate.

to punish an egoistic dictator in the baseline treatment (see also Figure 2). This reduction of 29 percentage points or 48% is statistically significant at the 5% level (Fisher's exact test (FET), p=0.027).[14] Hence, although about two-thirds of subjects choose to punish altruistically if the information about the egoistic dictator behavior is readily made available, only one-third choose to do so if the information has to be revealed. Thus, the informational structure influences third-party punishment.[15]

The treatment effect is mainly driven by the fact that third parties who remain ignorant almost exclusively choose not to punish the dictator, as can be seen in Table 1. Only 1 out of 32 ignorant third parties choose to punish the dictator. If the dictator behaved altruistically, 1 out of 32 of the third parties choose to punish the dictator in the baseline treatment, while 1 out of 44 choose to punish in this case in the hidden information treatment. Hence, ignorant third parties act as if the dictators have chosen fairly. This finding causes that, if I simultaneously analyze punishment rates for all dictator choices, the fraction of third parties that choose to punish declines from 30% in the baseline treatment to 17% in the hidden information treatment (FET, p=0.073). These results show that people remain ignorant and, by remaining ignorant, they seem to avoid the costly punishment of a norm violation by choosing as if no norm violation had taken place.

Table 1: Fraction of third parties choosing to punish by treatment and dictator choice

	Selfish dictators	Fair dictators	All dictators
Baseline treatment	17/28 (60.71%)	1/32 (3.13%)	18/60 (30.00%)
Hidden information treatment	14/44 (31.81%)	1/44 (2.27%)	15/88 (17.05%)
Ignorant third parties	1/15 (6.67%)	0/17 (0%)	1/32 (3.13%)
Δ	0.027	1.000	0.073

Note: The table displays the fraction (percentage) of third parties choosing to punish by treatment for selfish dictators (column 1), fair dictators (column 2), and all dictators (column 3), as well as the p-value from a two-sided FET for a treatment difference (Δ).

A regression analysis not reported here shows the robustness of these findings. It also

[14] Throughout the study, all reported tests are two-tailed tests.

[15] Note that the observed punishment rate (the fraction of third parties choosing to punish) in the baseline treatment is similar to punishment rates after norm violations in previous studies that use continuous sanctioning measures. In particular, in Fehr and Fischbacher (2004), when dictators do not share equally, about 60% of the third parties choose to engage in some punishing. Equally, Henrich et al. (2006) conducting experiments with subjects from five continents report that, on average, two-thirds of the third parties are willing to punish the dictator if she leaves zero to the recipient. While these studies differ from mine in several respects this is indicative of the fact that my findings are not driven by especially high or low punishment rates in the baseline treatment.

indicates that the treatment effect does not depend on the gender, age, nationality, and semester of the subject.

3.3 Dictator Behavior, Resulting Allocations, and Third-Party Beliefs

Given these results, it is informative to investigate (beliefs about) dictator behavior and welfare effects to answer the following questions: How do dictators behave? Is exogenously providing the information about dictators' choices welfare-enhancing in the short run by inducing more fair dictator choices and higher total payoffs compared to a situation of initial ignorance? How should dictators behave given the observed treatment effect in punishment rates if they were to maximize their expected payoffs? Are differences in third parties' beliefs about dictator behavior able to explain the treatment effect?

Averaging across treatments, the choices of the dictators are very balanced between the selfish (49%) and the fair option (51%). Dictators' behavior is very similar across the two treatments, as 53% of the dictators in the baseline treatment and 50% of the dictators in the hidden information treatment choose the fair option (FET, p=0.816). Combined with the observation that punishment rates are substantially higher in the baseline treatment this causes dictators' payoffs to be significantly higher in the hidden information treatment than in the baseline treatment (diff: €0.52, Mann-Whitney U test (MW-test), p=0.038). Hence, the difference in punishment choices translates into noticeable payoff consequences. As dictators' and recipients' average incomes are not different between treatments, total payoffs are not significantly different between the baseline and the hidden information treatment and there is no significant short-run treatment effect on total welfare (t-test, p=0.421).[16]

From a normative perspective, the following findings emerge: In the baseline treatment, it pays for dictators to be fair. There is a small but significant difference between opting for the fair and the selfish choice (diff.: €0.14, MW-test, p=0.027). Contrarily, in the hidden information treatment, the dictators' average income is substantially higher if they choose to be selfish (diff.:€1.18, MW-test, p=0.016). In this respect it is interesting that dictators' choices do not vary across treatments, although, in expectation, choosing fairly is beneficial in the baseline treatment but going for the big piece of the cake pays in the hidden information treatment.[17] I further investigate this finding in section 6.

[16] Equally, as dictators do not behave differently across treatments and as, conditional on a fair dictator choice, third parties also do not behave differently across treatments, the most efficient outcome (fair dictator, no punishment) is chosen equally often in both treatments (MW-test, p=0.772).

[17] Studying the effect of counter-punishment opportunities on third-party punishment, Balafoutas et al. (2014) find that, although the opportunity to counter-punish reduces punishment, the proportion of norm violations is identical with and without counter-punishment.

After the game was completed, third-party beliefs about the average choices of the dictators were elicited. The third parties were asked to guess the percentage of dictators in the lab that had chosen option A1 (€6 for participant A, €1 for participant B). Substantial monetary incentives to report the beliefs truthfully were provided with a maximum additional payoff of €4.[18]

The average belief about the percentage of dictators choosing selfish is virtually 50% (49.77%).[19] This is of interest, as third parties' willingness to reveal the dictators' choices might depend on the probabilities with which the dictators choose the two options. With this result, the "probability of being in a conflict of interest" is close to the one in Dana et al. (2007), where it is 50% by design.[20]

If these beliefs differ between treatments, the observed treatment effects might not only be driven by changes in the information structure, but also by the fact that, for instance, third parties in the hidden information treatment expect dictators to choose the selfish option more often. To find out whether third-party beliefs about the average dictator choice differ between treatments, I regress the belief about average dictator behavior on a treatment dummy, the actual choice of the dictator of the same group, and the interaction between the two. The coefficient on the hidden information treatment is small and not statistically significantly different from zero (effect size 5.70 percentage points, p=0.352).[21] Thus, the difference in punishment rates between the baseline and the hidden information treatment is driven not by different expectations about dictators' decisions, but by the change in the information structure.

4 Social Norms in Punishment Behavior

To identify the social norms that prevail in third-party punishment, I conducted a second experiment with different experimental subjects using the incentivized elicitation method proposed by Krupka and Weber (2013). In this way I obtain a social appropriateness rating for the actions of the third party for each of the four (six) possible choice combinations of the baseline treatment (hidden information treatment). I also obtain separate appropriate

[18] I assess the beliefs about the average dictator behavior after subjects were informed of the outcome of the game, but as I observe the choices of the dictators I can control for its effect on average beliefs and even allow it to vary between treatments.

[19] Observing a selfish dictator has a strong influence on this belief: The average belief for thirds, who observed a selfish (fair) dictator is 67% (34%) (t-test, p<0.001).

[20] There is mixed evidence on whether changes in this probability of conflicting payoffs affect the rate of information avoidance (see van der Weele, 2014 and Moradi and Nesterov, 2017, but Feiler, 2014).

[21] Looking at the raw data, the treatment difference in beliefs is 7 percentage points and also not statistically significant (t-test, p=0.172).

ratings for the third party's decision to reveal the choice of the dictator or not.

I first investigate the social norms in the baseline treatment. This is helpful for two reasons. On the one hand, considering the social norms in the baseline treatment reveals whether punishing a selfish dictator is indeed considered to be socially desirable. In the analyzed setting punishing the selfish dictator choice might be appropriate either because it is selfish or because it is inefficient (if third parties care about the efficiency after stage 1, but not stage 2), or both. On the other hand, if social norms matter for punishment behavior then both should be correlated. Prior findings show that unfair dictators are often, though not always, punished, while fair dictators are almost never punished. Hence, one might expect a strong norm to not punish a fair dictator and a (weaker) norm to punish a selfish dictator.

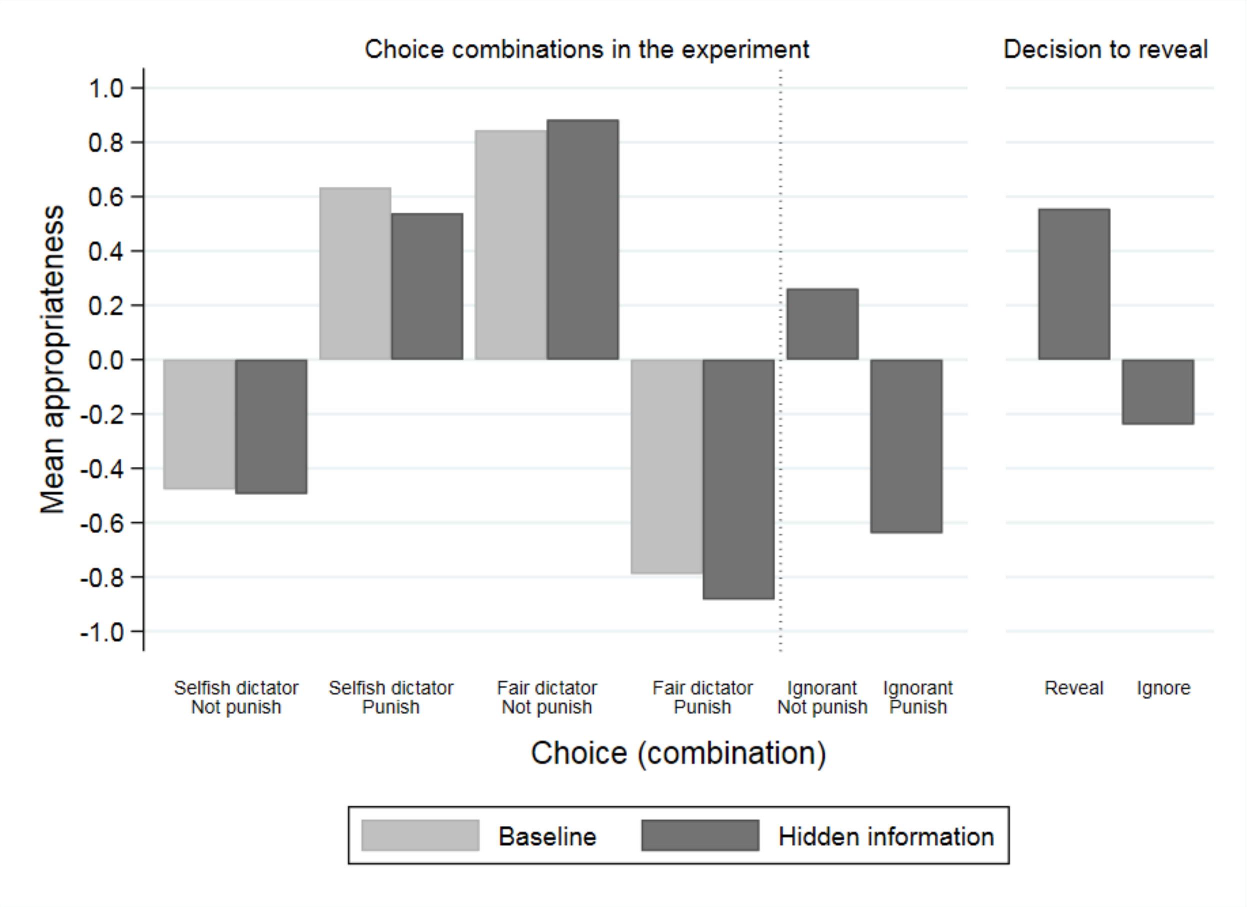
The light gray bars in the left part of Figure 3 display the mean appropriateness ratings for each possible combination of choices (selfish or fair dictator choice on the one hand, and decision to punish or not on the other) in the baseline treatment. The large majority (88%) of subjects think punishing a selfish dictator is very or somewhat appropriate. If the dictator behaved fairly, it is very socially inappropriate to punish him. Hence, as expected, it is very inappropriate to punish a fair dictator and appropriate to punish a selfish dictator. These results indicate that experimental subjects interpret the game as it is intended to be interpreted and find it ethically correct to punish a selfish dictator. At the same time, the social norms vary with individual behavior.[22]

Under hidden information, it is first important to know whether revealing the information about the potentially selfish dictator choice is seen as the ethically correct thing to do. As the third party is revealing the moral choice of another subject in the laboratory, it could in fact be that it is more appropriate to "mind one's own business." The social norm might even prescribe being a trusting person by remaining ignorant.

The right part of Figure 3 provides evidence that this is not the case. It shows the social appropriateness ratings for the information acquisition decision. While it is about somewhat inappropriate not to reveal, it is between somewhat and very appropriate to inform oneself about the dictator's choice (Wilcoxon signed-rank test (WSR-test), $p<0.001$). Hence, the

[22] In fact, I can explore the extent to which the elicited norms can explain behavior by predicting the choice probabilities of the four choices in the baseline treatment based on the social appropriateness of the action and its monetary payoff (see section 5.1 for a thorough discussion). I predict that upon observing a selfish dictator choice punishment will be chosen with a high probability (67%) and not punishing with a corresponding low probability (33%). Upon observing a fair dictator choice, I predict that not punishing will be chosen with a very high probability (99%) and punishment with a corresponding very low probability (1%). Hence, the predicted choice probabilities match the actual fraction of choices closely. At the same time, it makes sense that the social norms do not coincide with behavior, because the social norms elicited with the method proposed in Krupka and Weber (2013) are injunctive norms, that is, norms regarding what individuals "ought" to do and not necessarily what they actually do.

Figure 3: Mean appropriateness ratings by choice (combination) and treatment



Note: To the left, the figure shows the mean appropriateness rating of each choice combination for the baseline treatment (light gray) and the hidden information treatment (dark gray) and to the right the mean appropriateness rating of revealing the dictator's choice or remaining ignorant in the hidden information treatment.

norm prescribes revealing the dictator's choice.

Having established this, the crucial question becomes: What social norms prevail for altruistic punishment with the possibility of remaining ignorant? Two different mechanisms might explain the treatment effect. First, it could be that the fact that the third party does not immediately observe the choice of the dictator makes an egoistic dictator choice more excusable and, thus, punishing the dictator less appropriate. Then, punishing a selfish dictator should be differently appropriated depending on whether the third party is initially informed about the behavior of the dictator or not or more generally, the appropriateness of the same choice combinations should differ between the baseline treatment, in which the information about the choice of the dictator was exogenously provided, and the hidden information treatment, in which the same information was endogenously revealed.

Looking again at the left part of Figure 3, we see that the dark gray bars, which represent the mean appropriateness ratings for the four choice combinations after revealing the dictator's choice in the hidden information treatment, closely resemble the bars for the baseline treatment. The appropriateness does not significantly differ between treatments for any of the choice combinations (WSR-tests, selfish dictator, no punishment: p=0.533; selfish dictator, punishment: p=0.201; fair dictator, no punishment: p=0.664; fair dictator, punishment: p=0.076). Thus, it is not more or less desirable to punish a selfish dictator whether one initially observes his choice or not.

Alternatively, it could be that the social norms prevailing in the experiment are as such that remaining ignorant and choosing not to punish is an appropriate alternative to revealing and punishing. Two empirical questions are of interest in this regard. First, comparing the two choices under ignorance, punishing the dictator or not, the social norm should prescribe not punishing the dictator. Whether or not this holds true is likely to depend on both the third parties' beliefs about dictator behavior and on the way individuals make the trade-off between false negatives and false positives.[23] Second, the difference in appropriateness ratings between remaining ignorant and choosing not to punish and revealing that the dictator was selfish and engaging in altruistic punishment should be limited. At the same time, remaining ignorant and choosing not to punish should be more appropriate than revealing a selfish dictator choice and choosing not to punish it.

I find that if one remains ignorant, the social norm clearly prescribes not punishing the dictator (WSR-test, $p<0.001$). While this is socially appropriate, 93% find it inappropriate to remain ignorant and to punish. Furthermore, I find that not revealing the choice of the dictator and not choosing to punish is more appropriate than revealing that the dictator took the selfish option and not engaging in punishment (WSR-test, $p<0.001$), and revealing a norm violation and punishing is more appropriate than not revealing the choice of the dictator and not punishing (WSR-test, $p<0.001$). However, the difference in the appropriateness of these latter two choice combinations is small and much smaller than between the equivalent choice combinations in the dictator game. And, almost half of the subjects (48%) find it more or equally appropriate to remain ignorant and not to punish than to reveal a norm violation and to punish.

Summing these findings up, it is not that the decision to punish is more or less appropriate

[23] More precisely, if the third parties, on average, believe that dictators chose the selfish option, then it might be more appropriate to punish than to not punish, because the dictator is more likely selfish. In addition, it should matter whether it is more appropriate not to punish a selfish dictator than to punish a fair dictator. When making distributive choices, people seem to avoid false negatives (giving individuals more than they deserve), rather than false positives (giving individuals less than they deserve) (Cappelen et al., 2018). To the degree that these preferences are a reflection of social norms and to the degree to which they can be transferred to punishment decisions, the norm to punish a selfish dictator might be less strong than the norm not to punish a fair dictator.

depending on whether one was exogenously informed about the dictator's choice or chose to inform oneself: If the dictator was egoistic, and one knows this, it is inappropriate not to punish her. In contrast, if one does not observe the choice of the dictator then i) the norm generally prescribes revealing the choice of the dictator, ii) if one does not acquire information about the dictator behavior there is a strong social norm not to punish, and iii) conditional on an egoistic dictator choice, little can be gained from revealing the information, as revealing this dictator choice and punishing is only slightly more socially appropriate than remaining ignorant and not punishing. These observations are unique for altruistic punishment and lead to remaining ignorant being an attractive "outside option."

5 Explaining Punishment Behavior under Willful Ignorance

5.1 Predicting Behavior Based on Monetary Payoffs and Norms

How can we explain the observed behavior in section 3? Having estimated the existing social norms, one possibility is to try to explain behavior based solely on monetary payoffs and social norms. In the spirit of Krupka and Weber (2013) (see their online appendix), I compare the distribution of actual choices with the predicted distribution based on the exogenously given monetary payoffs and the appropriateness ratings elicited in section 4.

To derive a predicted distribution of choices I proceed in two steps. In the first step, I obtain the choice probability of each of the six possible strategies the third party can take in the hidden information treatment. These strategies are "remain ignorant and do not punish," "remain ignorant and punish," "reveal, do not punish a selfish dictator, and do not punish a fair dictator," "reveal, do not punish a selfish dictator, and punish a fair dictator," "reveal, punish a selfish dictator, and do not punish a fair dictator," and "reveal, punish a selfish dictator, and punish a fair dictator." To this end, I derive the expected utility of each of the six strategies based on the strategy's expected monetary payoff, its expected social appropriateness, and two weighting parameters, assuming both arguments to enter the Bernoulli utility function linearly and the subjects to be risk neutral.[24] To construct the expected utility, I use the average elicited belief about the percentage of dictators choosing the selfish option in the hidden information treatment (52.58%). Assuming a random utility model in which the errors follow a type I extreme value distribution, which leads to the

[24] For the reasons mentioned in Krupka and Weber (2013) I am not able to estimate the parameters with the data at hand and I thus use the parameters these authors estimated based on dictator game data from List (2007).

convenient choice probabilities of the logit model, I then predict the choice probability of each strategy. In the second step, I derive the predicted distribution from the empirical proportion of fair and selfish dictator choices.[25]

Table 2: Predicted and empirical distribution of third-party choices

	Predicted proportion	Empirical proportion
Reveal	67.31%	63.63%
Not punish selfish dictator	14.10%	18.18%
Punish selfish dictator	19.55%	14.77%
Not punish fair dictator	30.74%	29.55%
Punish fair dictator	2.92%	1.14%
Ignore	32.69%	36.36%
Not punish	31.41%	35.23%
Punish	1.28%	1.14%

Note: The table shows the percentage of third parties predicted to choose a choice combination in the hidden information treatment (column 1) and the percentage of third parties actually choosing the choice combination (column 2). Predictions are based on a weight of $\beta = 1.456$ ($\gamma = 1.941$) on expected monetary payoffs (expected mean appropriateness) of the strategy.

Table 2 contains for each of the six choice combinations the predicted and the actual empirical proportion of third parties taking the choice combination. Consider, for example, the proportion of third parties revealing a selfish dictator choice and choosing not to punish. The prediction is that this represents 14.10% of third parties, which comes close to the 18.18% who reveal a selfish dictator and do not punish him in the actual experiment. The same holds true for the other choice combinations. Hence, the predicted probabilities match the actual proportions of the choice combinations remarkably closely. Beyond that, predicting behavior based on norms and monetary payoffs also does a good job in predicting the proportion of third parties remaining ignorant (32.69% compared to 36.36% in the experimental data). These findings are especially striking given that the parameters I use for the predictions were obtained from dictator game choices. They imply that in the third-party punishment

[25] For instance, to arrive at the prediction that 30.74% percent of third parties reveal the dictator's choice, observe a fair dictator choice and choose not to punish, I sum up the percentage of third parties who are predicted to choose the strategy "reveal, do not punish a selfish dictator, and do not punish a fair dictator" (25.76%) and who are predicted to choose the strategy "reveal, punish a selfish dictator, and do not punish a fair dictator" (35.71%), and multiply this by the empirical percentage of fair dictator choices in the hidden information treatment (50.00%).

game with the possibility of remaining ignorant, based on monetary payoffs, social norms, and the empirical distribution of dictator choices, quite accurate predictions on aggregate behavior can be made.

Finally, note that the preceding analysis depends on the social norms being both exogenous and homogeneous. However, it is possible that what an individual perceives as appropriate is biased in a self-serving way and, more generally, that norm perceptions of individuals are heterogeneous. That is, people might be able to convince themselves that it is appropriate to remain ignorant to differing degrees. In turn, the third parties who do not reveal the information might be those that are able to find sufficient excuses for their choice to remain ignorant and not to punish ("mind your own business," etc.). This would imply that the norms that govern a subject's behavior are measured with (systematic) error and the predictive power of monetary incentives and social norms is even higher than presented here in Table 2.

5.2 Sorting of Types

The preceding analysis explained different choices by a random utility component. An alternative approach is to pin down the individual level differences that give rise to different choices and to analyze whether subjects who remain ignorant differ from those who reveal, or more broadly, whether there is a sorting of types into actions. A sorting of types is predicted by the model of Grossman and van der Weele (2017) who investigate willful ignorance in dictator games. In their model, individuals differ with respect to their degree of prosociality and self-image concerns. If their model also explains how people avoid information about the adverse welfare consequences of the self-interested decisions of others then there should be three types: Selfish types should always choose not to punish and high social types should always act prosocially by punishing a selfish dictator choice. Low social types should punish an egoistic dictator only if it cannot be avoided without substantially deteriorating their self-image.[26] Thus, the third parties who reveal the dictator behavior should mostly be the high social types such that, conditional on revealing the choice of the dictator, the fraction of altruistic choices should be higher than the fraction of altruistic choices in the game with full information, as this average is taken over all individuals in the population.[27]

An alternative approach is to directly measure the social type and the self-image type of subjects and analyze whether the average values of these measures follow the model's predictions. For the measure of social type, applying the model would imply that third

[26] This requires the assumption that self-image concerns are also present for altruistic punishment.

[27] As discussed before, an alternative approach for explaining differences in choices is that third parties differ in their ability to bias their perception regarding how appropriate it is to remain ignorant.

parties who do not reveal the choice of the dictator are more prosocial than those who reveal that the dictator chose selfishly and choose not to punish, but are less prosocial than those who reveal that the dictator chose selfishly and choose to punish. Regarding self-image concerns, third parties who reveal that the dictator chose selfishly and do not punish should care less about their self-image than those who reveal that the dictator chose selfishly and choose to punish or than those who remain ignorant.

I first analyze sorting of prosocial types with respect to observed choices. Sixty-one percent of the third parties in the main experiment choose to punish in the baseline treatment if the dictator behaved egoistically. Of the third parties who revealed the same information, only 45% choose to punish the dictator (FET, p=0.292). Hence, I find little evidence of sorting if I compare the punishment rates of selfish dictators between the baseline treatment and the hidden information treatment conditional on revealing the dictator's choice.[28] This finding can, for instance, be explained by more individuals of the selfish type revealing, since they are indifferent, but can learn about the action of another person, while more high social types remain ignorant.[29]

I also consider direct measures of the types of subjects that I obtained in the main experiment after subjects had played the game, starting with their social type by employing Murphy et al.'s (2011) measure of social value orientation.[30] The measure is obtained by letting the subjects make choices between different allocations of money between themselves and another individual. It is higher the higher the concern a subject has for others.

The average social value orientation of third parties who reveal a selfish dictator and punish her is 35.52 and higher than for those who do not reveal the choice of the dictator (diff: 5.56, MW-test, p=0.024). Equally, for third parties who reveal a selfish dictator and punish her, the index is higher than for those who reveal this information but refrain from making the costly punishment (diff: 10.69, MW-test, p=0.001). Comparing the social value orientation between thirds who remain ignorant and thirds who reveal but do not punish, I find that the former is marginally significantly higher (MW-test, p=0.075). This result offers some support of the model of Grossman and van der Weele (2017) and, more generally, indicates that more prosocial types are more likely to punish norm violations.

As a measure of the importance of self-image I again follow Grossman and van der Weele (2017) in using Aquino and Reed's (2002) measure of self-importance of moral identity.

[28] Note that there is mixed evidence of sorting in generosity decisions, as there is significant sorting into revealing in Grossman and van der Weele (2017) but not in Dana et al. (2007), Larson and Capra (2009) and Kajackaite (2015).

[29] If the selfish type reveals and the share of the selfish type is sufficiently large, prosocial behavior may actually be lower among those who reveal (see Grossman and van der Weele, 2017).

[30] I obtained this measure of social value orientation for 93 of the subjects.

The measure is based on asking individuals to indicate their agreement to six statements about the importance of certain moral characteristics for their sense of self. The attributes I consider match the attributes of someone who is willing to engage in altruistic punishment and are hence “compassionate,” “caring,” and “fair.” I create a linear index taking on values between 0 and 30 with higher values indicating a higher self-importance of moral identity.

The reported self-image concerns of third parties who reveal that the dictator opted for the selfish choice and sanction her are not different to those who reveal that the dictator opted for the selfish choice and do not sanction her (MW-test, p=0.930) or to those who remain ignorant (MW-test, p=0.734). Thus, there is no evidence that third parties who reveal the information and then act selfishly care less about their self-image than third parties who reveal the information and behave altruistically or than those who do not reveal.

In summary, I find only limited support for a sorting of types into different actions. While the findings with respect to the social value orientation are in line with the predictions of the model of Grossman and van der Weele (2017) and indicate that more prosocial individuals are more likely to punish norm violations, the findings regarding the self-image index and the actual choices do not offer support for the model and provide no evidence of sorting.

6 Informing about Willful Ignorance

Although significantly decreasing the proportion of altruistically punished norm violators and changing dictators’ payoff-maximizing choices, willful ignorance does not decrease the proportion of selfish dictator choices. This finding is puzzling, though it is in line with the evidence in Kriss et al. (2016) who also find that dictators do not behave differently between the treatment in which they have moral wiggle room and the treatment in which they do not (while dictators do behave differently between the second-party punishment and third-party punishment treatment). It seems that, at least in one-shot games in laboratory experiments, individuals have trouble anticipating that others (second-movers) will exploit the moral wiggle room they have. At the same time, it is crucial whether the exploitation of moral wiggle room leads to differences in norm compliance. If it does not, the same moral behavior can be sustained with and without the wiggle room, and we might not consider the exploitation of moral wiggle room an urgent issue, at least from a policy perspective.

I hence ran two additional treatments, “baseline informed” and “hidden information informed,” which replicate the baseline treatment and the hidden information treatment, with only one difference: Before making their respective decisions, all players were informed about the average behavior of the opposite players in previous sessions. That is, dictators were given information about the proportion of dictators that were punished conditional on their choice (selfish or fair) in the baseline and hidden information treatment, respectively, and the third parties learned about the proportion of dictators that chose the selfish option in each treatment. I hence moved from a set-up in which dictators have no information about how moral wiggle room affects their income conditional on their choice (other than describing the game in the instructions) to a set-up in which they have perfect information about its effects (in expectation). Thereby, I evoke correct beliefs in the dictators. I can exploit the fact that because third-party punishment renders antisocial behavior unprofitable in the baseline, but not under hidden information, a third party that maximizes the expected value of its income should choose the fair option in the baseline treatment, but the selfish option in the hidden information treatment.[31] I provided the information as part of the general instructions.[32] To investigate whether a potential difference in dictator choices across treatments is related to a difference in beliefs about the consequences of their choices, I elicit subjects' beliefs after they make their choices but before they are informed about the outcome of the game.[33]

I ran 12 sessions in June 2019 in Berlin. Two hundred and sixty-four subjects participated such that I observe 45 (43) dictator choices and 90 (86) third-party choices in the baseline informed treatment (hidden information informed treatment).[34]

When dictators are fully informed about the consequences of willful ignorance for altruistic punishment there is a sizable treatment difference in dictator behavior of 51%. The majority of dictators in the baseline select the fair option (62%) and the majority of dictators under hidden information choose the selfish option (70%) (FET, $p = 0.003$). In line with the notion of third parties holding accurate beliefs under full information (baseline treatment) but struggling to anticipate the effects of initially hiding the information about whether a norm violation has taken place (hidden information treatment), dictators in the baseline informed treatment do not significantly respond to the information provided (diff. to baseline treatment: 9pp, FET, $p = 0.481$), while dictators in the hidden information informed treatment adapt their choices substantially (diff. to hidden information treatment: 20 pp, FET, $p = 0.081$).

Dictators in both treatments hold the same beliefs about the proportion of fair dictators

[31] The results of these treatments are not obvious, because i) it is unclear what fraction of dictators who chose the fair option in the original sessions did so due to their social preferences, ii) the elasticity of dictators' beliefs with respect to the information is unknown, and iii) dictators' risk-preferences are unknown.

[32] Hence, the information was provided in a natural way. Assuming that the dictators do not leave money on the table to act in accordance with an experimenter demand when beliefs are elicited, I can check whether a difference in dictator behavior goes along with a difference in beliefs rather than being driven by an experimenter demand effect.

[33] I did not elicit the beliefs at the start, because eliciting beliefs itself may or may not affect behavior and I intended to measure the causal effect of the information on dictator behavior (see Nyarko and Schotter (2002) and Costa-Gomes and Weizsäcker (2008), but Kovářík (2007) and Gächter and Renner (2010)).

[34] These are about the same numbers as in the hidden information treatment, which I aimed for.

being punished (10.71% in the baseline informed treatment, 11.12% in the hidden information informed treatment, t-test, $p = 0.915$), but differ with respect to their beliefs about the proportion of selfish dictators being punished (63.40% and 37.30%, t-test, $p < 0.001$). As a consequence, the fraction of dictators that should choose the fair option if maximizing the expected value of their payments and responding optimally to their beliefs is 49% in the baseline and 2% in the hidden information (FET, $p < 0.001$). I also elicit beliefs about the proportion of punished dictators conditional on their choice from third parties.[35] For these, I do not find any difference in the beliefs about the proportion of selfish dictators being punished (t-test, $p = 0.514$). Hence, the third parties do not anticipate the treatment effect, which is in line with the original finding of dictators without information behaving in the same way across treatments.

Analyzing third-party behavior, I find overall smaller levels of third-party punishment. Twelve out of 34 third parties punish under full information, only 14 out of 60 do so under hidden information. This leads to a treatment difference of 34%. Due to the fact that the punishment level is lower this difference is not statistically significant on conventional levels (FET, $p = 0.237$).[36] As in the original sessions, ignorant third parties almost never punish.[37]

These additional treatments hence generate two takeaways. First, if dictators are perfectly informed about the likelihood of being punished – as they might be at least in some dynamic settings that allow for repeated interactions – moral wiggle room affects norm compliance. Second, whether moral wiggle room affects norm compliance in any given set-up is likely to depend on the salience of its effect. In particular, in set-ups in which the outcome of interest is only indirectly linked to an individual having the possibility to exploit the wiggle room, we might expect the effect of moral wiggle room to be muted.

[35] Remember that the third parties do not receive any information about previous punishment behavior. They are, however, informed about the proportion of dictators choosing selfishly in their treatment.

[36] This might be caused by third parties who engage in punishment believing that a high fraction of dictators are fair. Learning about approximately 50% of dictators being selfish might then discourage punishment. Accordingly, third parties who punish selfish dictators in the original sessions believe that the fraction of selfish dictators is lower than third parties who do not (diff.: 21 pp, t-test, $p < 0.001$). In addition, the beliefs about the fraction of selfish dictators vary slightly between the baseline informed and the hidden information informed treatment (diff.: 6 pp, t-test, $p = 0.027$), which might also contribute to the result.

[37] If I pool the data of the baseline treatment and the baseline informed treatment, and the hidden information treatment and the hidden information informed treatment, respectively, the treatment effect is 20 percentage points or 42% and highly statistically significant (FET, $p = 0.011$). Forty-eight percent of third parties remain ignorant. Across both treatments only three out of 82 third parties punish a fair dictator.

7 Conclusion

Obscuring information about choice-relevant behavior decreases altruistic punishment. More than a third of subjects remain ignorant about how a dictator chooses to allocate money between himself and a passive recipient. The ignorant third parties exploit the moral wiggle room provided by the information structure and act as if no norm violation has taken place. They comply with social norms that consider it malicious to punish a dictator without knowing how she behaved and that consider it *okay* to remain ignorant and not punish. As a consequence, if third parties can remain ignorant about the behavior of the dictator, less than one-third of norm violations are altruistically punished. This implies that in situations in which people can remain ignorant about potential norm violations, it is likely that norm violations remain unpunished.

The present study remains silent about what constitutes an optimal level of punishment. It might be that the punishment rate under full information is in fact too high in the sense that the desired proportion of moral behavior could be sustained with lower punishment rates. In this regard, even more important than whether willful ignorance impacts third-party punishment seems to be the question of whether it affects dictator behavior (i.e., the norm it is supposed to enforce). While dictator behavior is not affected by the potential willful ignorance of the third party when no information about its consequences for punishment rates is given, it does when information is provided. I analyze a rather extreme change in the information provided. Hence, in scenarios where the salience of its effect is low, the effect of willful ignorance on norm compliance might be small. At the same time, when the equivalent information is given to third parties, punishment levels and the effect of willful ignorance on punishment are lower, indicating that willful ignorance can only become effective in scenarios where there is substantial moral behavior to begin with.

From a policy perspective, providing people with the possibility to reveal choice-relevant information might be insufficient to induce desirable punishment behavior and, in turn, to sustain certain norms of distribution or cooperation or, more generally, moral behavior. If sustaining a certain behavior is of high importance, exogenously providing information in a way that enforces that people process the information *and* ensuring that the consequences of violating a norm are known is highly advisable.[38] At the same time, as observed choices seem to be in line with the elicited social norms, a public discussion on how to judge deliberate ignorance might be beneficial. In this regard, recent discussions about whistleblowing and ethical misconduct within companies can be considered a step in the right direction.

[38] More broadly, whether an exogenous information provision is beneficial also depends on other economic considerations, such as the monitoring costs or the costs of punishment relative to the benefits of sustaining a norm.

References

Andreoni, J., J. M. Rao, and H. Trachtman (2017): "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy*, 125, 625–653.

Aquino, K. and A. I. Reed (2002): "The self-importance of moral identity," *Journal of personality and social psychology*, 83, 1423.

Balafoutas, L., K. Grechenig, and N. Nikiforakis (2014): "Third-party punishment and counter-punishment in one-shot interactions," *Economics Letters*, 122, 308–310.

Balafoutas, L. and N. Nikiforakis (2012): "Norm enforcement in the city: a natural field experiment," *European Economic Review*, 56, 1773–1785.

Balafoutas, L., N. Nikiforakis, and B. Rockenbach (2016): "Altruistic punishment does not increase with the severity of norm violations in the field," *Nature communications*, 7, 13327.

Bartling, B., F. Engl, and R. A. Weber (2014): "Does willful ignorance deflect punishment?–An experimental study," *European Economic Review*, 70, 512–524.

Bartling, B. and U. Fischbacher (2012): "Shifting the Blame: On Delegation and Responsibility," *The Review of Economic Studies*, 79, 67–87.

Bartling, B., R. A. Weber, and L. Yao (2015): "Do Markets Erode Social Responsibility?" *The Quarterly Journal of Economics*, 130, 219–266.

Cain, D. and J. Dana (2012): "Paying people to look at the consequences of their actions," Working paper.

Cappelen, A. W., C. Cappelen, and B. Tungodden (2018): "Second-Best Fairness Under Limited Information: The Trade-Off between False Positives and False Negatives," NHH Department of Economics Discussion Paper No. 18/2018.

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

Carpenter, J. P. and P. H. Matthews (2012): "Norm Enforcement: Anger, Indignation, Or Reciprocity?" *Journal of the European Economic Association*, 10, 555–572.

Charness, G., R. Cobo-Reyes, and N. Jiménez (2008): "An investment game with third-party intervention," *Journal of Economic Behavior & Organization*, 68, 18–28.

Coffman, L. C. (2011): "Intermediation Reduces Punishment (and Reward)," *American Economic Journal: Microeconomics*, 3, 77–106.

Costa-Gomes, M. A. and G. Weizsäcker (2008): "Stated Beliefs and Play in Normal-Form Games," *The Review of Economic Studies*, 75, 729–762.

Cox, J. C., M. Servátka, and R. Vadovič (2017): "Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission," *Experimental Economics*, 20, 1–18.

Dana, J., D. M. Cain, and R. M. Dawes (2006): "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, 100, 193–201.

Dana, J., R. A. Weber, and J. X. Kuang (2007): "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 33, 67–80.

DellaVigna, S., J. A. List, and U. Malmendier (2012): "Testing for altruism and social pressure in charitable giving," *The quarterly journal of economics*, 127, 1–56.

Egas, M. and A. Riedl (2008): "The economics of altruistic punishment and the maintenance of cooperation," *Proc. R. Soc. B*, 275, 871–878.

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

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

Fehr, E. and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

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

Feiler, L. (2014): "Testing models of information avoidance with binary choice dictator games," *Journal of Economic Psychology*, 45, 253–267.

Felgendreher, S. (2018): "Do consumers choose to stay ignorant? The role of information in the purchase of ethically certified products," Working paper in economics 717, Department of Economics, University of Gothenburg.

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

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

Gächter, S., D. Nosenzo, and M. Sefton (2013): "Peer effects in pro-social behavior: Social norms or social preferences?" *Journal of the European Economic Association*, 11, 548–573.

Gächter, S. and E. Renner (2010): "The effects of (incentivized) belief elicitation in public goods experiments," *Experimental Economics*, 13, 364–377.

Goeschl, T. and J. Jarke (2016): "Second and third party punishment under costly monitoring," *Journal of Economic Psychology*, 54, 124–133.

Greiner, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Grossman, Z. (2014): "Strategic ignorance and the robustness of social preferences," *Management Science*, 60, 2659–2665.

Grossman, Z. and J. J. van der Weele (2017): "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 15, 173–217.

Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker (2006): "Costly Punishment Across Human Societies," *Science*, 312, 1767–1770.

Kajackaite, A. (2015): "If I close my eyes, nobody will get hurt: The effect of ignorance on performance in a real-effort experiment," *Journal of Economic Behavior & Organization*, 116, 518–524.

Kandul, S. (2016): "Ex-post blindness as excuse? The effect of information disclosure on giving," *Journal of Economic Psychology*, 52, 91–101.

Kovářík, J. (2007): "Belief Formation and Evolution in Public Good Games," LABSI Working Paper.

Kriss, P. H., R. A. Weber, and E. Xiao (2016): "Turning a blind eye, but not the other cheek: On the robustness of costly punishment," *Journal of Economic Behavior & Organization*, 128, 159–177.

Krupka, E. L. and R. A. Weber (2013): "Identifying social norms using coordination games: Why does dictator game sharing vary?" *Journal of the European Economic Association*, 11, 495–524.

Larson, T. and C. M. Capra (2009): "Exploiting moral wiggle room: Illusory preference for fairness? A comment," *Judgment and Decision Making*, 4, 467.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012): "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 4, 136–63.

Leibbrandt, A. and R. López-Pérez (2012): "An exploration of third and second party punishment in ten simple games," *Journal of Economic Behavior & Organization*, 84, 753–766.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of economic dynamics*, 1, 593–622.

Lewisch, P., S. Ottone, and F. Ponzano (2011): "Free-Riding on Altruistic Punishment? An Experimental Comparison of Third-Party Punishment in a Stand-Alone and in an In-Group Environment," *Review of Law & Economics*, 7, 161–190.

List, J. A. (2007): "On the interpretation of giving in dictator games," *Journal of Political economy*, 115, 482–493.

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

Matthey, A. and T. Regner (2015): "Do reciprocators exploit or resist moral wiggle room? An experimental analysis," *Jena Economic Research Papers*, 9.

Moradi, H. and A. Nesterov (2017): "Moral wiggle room reverted: Information avoidance is myopic," Working paper.

Murphy, R. O., K. A. Ackermann, and M. J. Handgraaf (2011): "Measuring social value orientation," *Judgment and Decision Making*, 6, 771–781.

Nikiforakis, N. (2008): "Punishment and counter-punishment in public good games: Can we really govern ourselves?" *Journal of Public Economics*, 92, 91–112.

Nikiforakis, N. and D. Engelmann (2011): "Altruistic punishment and the threat of feuds," *Journal of Economic Behavior & Organization*, 78, 319–332.

Nikiforakis, N. and H. Mitchell (2014): "Mixing the carrots with the sticks: third party punishment and reward," *Experimental Economics*, 17, 1–23.

Nyarko, Y. and A. Schotter (2002): "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70, 971–1005.

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

Regner, T. (2018): "Reciprocity under moral wiggle room: Is it a preference or a constraint?" *Experimental Economics*, 21, 779–792.

Trachtman, H., A. Steinkruger, M. Wood, A. Wooster, J. Andreoni, J. J. Murphy, and J. M. Rao (2015): "Fair weather avoidance: unpacking the costs and benefits of "Avoiding the Ask"," *Journal of the Economic Science Association*, 1, 8–14.

van der Weele, J. J. (2014): "Inconvenient truths: Determinants of strategic ignorance in moral dilemmas," SSRN Working Paper.

van der Weele, J. J., J. Kulisa, M. Kosfeld, and G. Friebel (2014): "Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior?" *American Economic Journal: Microeconomics*, 6, 256–264.

Discussion Papers of the Research Area Markets and Choice 2019

Research Unit: **Market Behavior**

Azar Abizada, Inácio Bó Hiring from a pool of workers	SP II 2019-201
Philipp Albert, Dorothea Kübler, Juliana Silva-Goncalves Peer effects of ambition	SP II 2019-202
Yves Breitmoser, Sebastian Schweighofer-Kodritsch Obviousness around the clock	SP II 2019-203
Tobias König, Sebastian Schweighofer-Kodritsch, Georg Weizsäcker Beliefs as a means of self-control? Evidence from a dynamic student survey	SP II 2019-204
Rustamdjan Hakimov, Dorothea Kübler Experiments on matching markets: A survey	SP II 2019-205
Puja Bhattacharya , Jeevant Rampal Contests within and between groups	SP II 2019-206
Kirby Nielsen, Puja Bhattacharya, John H. Kagel, Arjun Sengupta Teams promise but do not deliver	SP II 2019-207
Julien Grenet, Yinghua He, Dorothea Kübler Decentralizing centralized matching markets: Implications from early offers in university admissions	SP II 2019-208
Joerg Oechssler, Andreas Reischmann, Andis Sofianos The conditional contribution mechanism for repeated public goods – the general case	SP II 2019-209
Rustamdjan Hakimov, C.-Philipp Heller, Dorothea Kübler, Morimitsu Kurino How to avoid black markets for appointments with online booking systems	SP II 2019-210
Thibaud Pierrot Negotiation under the curse of knowledge	SP II 2019-211
Sabine Kröger, Thibaud Pierrot What point of a distribution summarizes point predictions?	SP II 2019-212
Sabine Kröger, Thibaud Pierrot Comparison of different question formats eliciting point predictions	SP II 2019-213
Hande Erkut, Shaul Shalvi Working until you drop: Image concerns or prosocial motives?	SP II 2019-214
Robert Stüber The benefit of the doubt: Willful ignorance and altruistic punishment	SP II 2019-215

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice

Research Unit: **Economics of Change**

Kai Barron, Steffen Huck, Philippe Jehiel SP II 2019-301
Everyday econometricians: Selection neglect and overoptimism when learning from others

Marta Serra-Garcia, Nora Szech SP II 2019-302
The (in)elasticity of moral ignorance

Kai Barron, Robert Stüber, Roel van Veldhuizen SP II 2019-303
Motivated motive selection in the lying-dictator game

Maja Adena, Steffen Huck SP II 2019-304
Can mass fundraising harm your core business? A field experiment on how fundraising affects ticket sales

Maja Adena, Rustamdjan Hakimov, Steffen Huck SP II 2019-305
Charitable giving by the poor: A field experiment on matching and distance to charitable output in Kyrgyzstan

Maja Adena, Steffen Huck SP II 2019-306
Personalized fundraising: a field experiment on threshold matching of donations

Kai Barron SP II 2019-307
Lying to appear honest

Research Unit: **Ethics and Behavioral Economics**

Daniel Parra, Manuel Muñoz-Herrera, Luis Palacio SP II 2019-401
The limits of transparency as a means of reducing corruption

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice