# Altruistic punishment as norm-signaling:
# A model and experimental evidence

January 29, 2024

Kiryl Khalmetski and Axel Ockenfels*

## Abstract

In social dilemmas, altruistic punishment is crucial for promoting cooperation. It is often posited that individuals use punishment to engage in reciprocal retaliation against non-cooperative counterparts, based on the assumption that people differ in their willingness or inclination to (conditionally) cooperate. We complement this literature by proposing an additional, purely informational motivation for punishment, which operates even among players with identical (social) preferences: If people care about being seen as no less norm-compliant than others, then punishment naturally emerges as an equilibrium strategy for credibly communicating one's understanding of the prevailing norm. When players face idiosyncratic uncertainty about the strength of the norm, our model allows for the emergence of both punishment and antisocial punishment. The main predictions of the model are confirmed in a public goods experiment where asymmetric information about the social norm is exogenously introduced.

**Keywords**: altruistic punishment, social norms, signaling model.

## 1. Introduction

Altruistic punishment is crucial for promoting cooperation in public goods games (Fehr and Gächter, 2000, 2002, Ostrom et al., 1992). People are willing to punish those who free ride on the cooperation of others, although the punishment is costly for them and yields no material

* Khalmetski (corresponding author): Department of Economics, University of Bath, Bath BA2 7AY, UK (email: kk917@bath.ac.uk, tel.: +44 1225 385593); Ockenfels: Department of Economics, University of Cologne, D-50932 Cologne, Germany and Max Planck Institute for Research on Collective Goods, D-53113 Bonn, Germany (email: ockenfels@uni-koeln.de).

gain. This in turn creates incentives for defectors to be cooperative in the first place. Yet, while punishment provides a solution to social dilemmas, the punishment of free riders constitutes a second-order public good, raising the question why people would punish even though it yields no material benefits for them.

The literature has suggested various explanations for the role of punishment in public good contexts. For instance, the (threat of) punishment can be used in a selfish and forward-looking way to encourage others to cooperate in repeated games (Fudenberg and Maskin, 1986; Ostrom et al., 1992), or it can be used in an altruistic and forward-looking way to make transgressors comply with social norms (Fehr and Fischbacher, 2004; Carpenter and Matthews, 2009, 2012; Reuben and Riedl, 2013). Similarly, backward-looking retaliatory punishment can be used to reciprocate unkind behavior (Rabin, 1993, Dufwenberg and Kirchsteiger, 2004). In all these cases, punishment is effective because it harms the punished player, which is indeed the intention of the punisher.

In this paper we analyse a different and potentially complementary function of punishment as a means of *communicating* with the opponent. We suggest that people who care about being seen as no less norm-compliant than others seek to change their opponents' beliefs about the appropriate social norm, and that punishment is a tool for doing so. As we will see, this leads them to punish as a way of 'teaching' others about what they perceive to be the prevailing norm if they are sufficiently sure of their own norm belief and that the opponent's belief could be changed. The motivation to change beliefs is different from the motivation to enforce norm-compliant behavior or punish deviant behavior examined in the papers cited above. For example, our norm-signaling motive of punishment would disappear once the social norm becomes common knowledge, as this precludes changing one's beliefs about the norm, whereas such information does not preclude punishments aimed at changing behaviour.

We study altruistic punishment as norm-signaling both theoretically and in a controlled laboratory setting, where both individual information about the norm and the transparency of this signal to others are exogenously varied. In our model, punishment emerges naturally as an equilibrium strategy for credibly communicating one's understanding of the prevailing norm. In contrast to standard models of altruistic punishment, our model allows for the emergence of both punishment of free-riders by cooperators and antisocial punishment of cooperators by free-riders.[1] The laboratory experiment confirms the main predictions of the model.

The model has two players in a one-shot public goods game with a subsequent costly punishment phase. Players have identical social preferences and care about monetary payoffs and about conforming to the social norm. The norm is assumed to be the same for all players and to be exogenous but uncertain. As is standard in the literature (Akerlof, 1980; Bernheim, 1994), the norm conformity element of utility creates an incentive for social behavior. However, we add a social comparison element to the literature: Players do not like to be seen as less norm-compliant than the opponent in a conflict. In our model, this kind of relative norm conformism concern implies that, in terms of psychological payoffs, it is worse to violate the social norm if the opponent is conforming. One interpretation of the motive is that whenever there is social conflict, i.e., whenever players differ in their cooperation level, there is scope for blaming and being blamed for inappropriate behavior. In this situation, changing others' beliefs about what would have been appropriate through punishment allows blame to be shifted to others, thereby increasing psychological utility.[2] Relative norm conformism is different from conditional cooperation in that players do not look at differences in behavior but at differences in conformism to the exogenous norm. This way, we extend the idea that social behavior is often comparative in nature. This is demonstrated, for example, by fairness models that extended pure altruism, which only considered 'absolute' payoffs, by emphasizing the importance of 'relative' payoff comparisons (Bolton and Ockenfels 1998, 2000; Fehr and Schmidt 1999).

The second key element of the model is the introduction of asymmetric information about the strength of the social norm. If there were no asymmetric information, there would be no reason to costly signal norm information in equilibrium. Each player gets a private signal about the strength of the norm. The signal can be 'precise' or 'imprecise'. The signal can also be 1 or 0, with the signal of 1 implying that the social norm of cooperation is rather strong, and the signal of 0 implying that the social norm of cooperation is rather weak. We show that there exists a

[1] Bénabou and Tirole (2011) provide a theoretical model predicting antisocial punishment in an abstract social dilemma setting where agents are driven by self-image concerns and have limited memory. Antisocial punishment would then take the form of social exclusion that prevents the punisher from observing the opponent's good behavior (and recalling it later), thus preserving the punisher's self-image. Thöni (2014) showed that inequity aversion may predict punishment of a cooperator by another cooperator in case if only the latter bears the costs of punishing free-riders.

[2] More generally, it has previously been observed that people are averse to others holding different beliefs, especially when others' beliefs are also perceived to be incorrect (Molnar and Loewenstein, 2020), and that being exposed to others who hold beliefs different from one's own can threaten one's identity (Golman et al., 2016).

symmetric equilibrium where only cooperators with a *precise* 1-signal about the norm punish defectors, thus separating themselves from players who receive an imprecise signal. Similarly, there exists a symmetric equilibrium where only defectors with a precise 0-signal about the norm punish cooperators. In both cases, costly punishment occurs in an attempt to convince the opponent that one's own behavior is consistent with the prevailing norm. There can be no punishment if both players chose the same action. That is, punishment endogenously acquires a norm-signaling value in equilibrium, which is beneficial to the punisher by shifting the punished player's belief about the norm so that it gets more aligned with the punisher's behavior, and less aligned with the opponent's behavior. The separating equilibrium is feasible because the expected informational benefit, in terms of influencing the punished player's belief about the appropriate norm, differs depending on whether the player received an imprecise or precise signal about the norm.[3]

As an example for an application of our model, note that the pattern of leaving negative feedback – a form of punishment – on eBay and other platforms in the sharing economy is well-understood in terms of altruistic punishment and (anti-social) counter-punishment (Bolton et al., 2013, 2018; Chen et al., 2021, and Ockenfels, 2023, provide surveys). However, there are indications that part of trader conflict is due to uncertainty about what to expect from each other: Buyers may have a different belief than sellers about what a used "Apple watch in good condition" on eBay or a "quiet neighbourhood" on Airbnb means, that is, about the appropriate social norm. In this setting, leaving negative feedback, including what looks like anti-social feedback, may serve not only as warning to future buyers, but also to teach the opponent about what one believes the prevailing social norm is. Two observations on such platforms appear to support this idea, and in fact motivated our research (other references below): Bolton et al. (2019) show that uncertainty about whether behavior conforms to the norm reduces punishment; and Bolton et al. (2020) show that people's willingness to punish depends on whether others share the same identity, which is in line with the idea that protecting one's social image by avoiding being seen by others as less norm-compliant is more valuable ingroup as compared to outgroup.

[3] As in our model, altruistic punishment is also a costly signaling strategy in Jordan et al. (2016) and Jordan and Rand (2017), yet to signal individual traits such as trustworthiness and willingness to cooperate. In contrast, our paper focuses on punishment as a signal of an objective social norm that is orthogonal to individual traits of a given player. Also in a different but related literature, costly signaling of own intrinsic type by conforming to a commonly known social norm was studied by Bernheim (1994) and Andreoni and Bernheim (2009). Te Velde (2022) analysed signaling of individual moral values (unlike common social norms).

We test our main hypotheses in an experiment. We entail players of a public goods game with noisy yet informative signals about the strength of the social norm by revealing to some of them the cooperation behavior of a subsample of previous participants. Since subsamples are drawn individually for each subject, players exogenously and randomly obtain heterogeneous private signals about the strength of the social norm (termed as 'Strong Norm' and 'Weak Norm' signals). Moreover, since only a fraction of subjects receive a signal, they also differ in the precision of obtained information in that subjects obtaining no experimental signal are assumed to have less precise information about the norm. We find strong evidence for our model predictions. Cooperators who received a Strong Norm signal punish defectors significantly more frequently than cooperators who obtained no or a Weak Norm signal. Moreover, this effect vanishes once we make subjects' private signals common knowledge between the players independently of the punishment. While this does not remove the motivation to punish *behavior*, as postulated by the models mentioned above, it deprives punishment of its capacity to signal what the punisher knows about the norm, and thus makes punishment useless in the context of our model.

To the best of our knowledge, no previous study provides an equilibrium analysis of the signaling mechanism, nor clean empirical evidence that signaling private information about the social norm is a significant part of individual motivation to punish others. However, much previous experimental evidence is consistent with our model. Punishment is more intense in situations where there is uncertainty about the social norm (Balafoutas and Nikiforakis, 2012), or potentially conflicting views on the norm (Reuben and Riedl, 2013). Also, people often choose informal (non-monetary) sanctions to punish others, if available, implying that pecuniary consequences of punishment may be not its primary purpose (Masclet et al., 2003, Xiao and Houser, 2005, Molnar et al., 2023). In such cases, however, our rational signaling model would suggest that costly punishment is more effective in shifting the opponents' beliefs, because informal sanctions are cheap talk. Brouwer et al. (2023) showed that parents accompanied by their children are more prone to punish norm violations of strangers than unaccompanied parents, suggesting that there is an educational value in punishment. Also, imposed sanctions are often interpreted by people as a signal about the normative, or expected behavior (Tyran and Feld, 2006, Drago et al., 2009, Galbiati and Vertova, 2008, Funk, 2007, Casoria et al. 2021, Lane et al., 2023). Indeed, Bicchieri et al. (2021) showed that punishment is perceived as more legitimate by the punished subjects if it is accompanied by information about the social norm that is consistent with the punishment. Xiao (2013) and Rai (2022) showed that punishment is less efficient in deterring norm violation if its implementation entails benefits for the punisher, which then obscure the potential informational function of punishment. In a similar vein, Chen et al. (2020) showed that punishment inflicted by a third party has a larger influence on the normative beliefs of external observers compared to that inflicted by the offended second party.

Molnar et al. (2023) independently run an experiment, which tested whether by punishing unfair behavior of others people are willing to affect the transgressors' beliefs about *why* they have been punished. They show that punishers were often willing to send an additional text message to a transgressor stating that she has been punished due to having been unfair towards the punisher, thus revealing punishers' preferences to explicitly affect beliefs of others about whether their behavior deserves punishment. One difference to our experiment is that we test whether punishers intend to signal an objective (yet unknown) social norm that exists independently of both the transgressor and the punisher, rather than a subjective punisher's perception of unfairness, and whether such behavior causally depends on exogenous information about the social norm. In another related and independent study, Dimant and Gesche (2023) showed that providing information about either descriptive or injunctive social norm to a third (unaffected) party increases the latter's propensity to punish the observed lying behavior of others. This observation is, in principle, consistent with our main supposition that punishment is intended to shift the beliefs of transgressors about the norm.[4] At the same time, the results of Dimant and Gesche (2023) also allow for an alternative explanation, primarily discussed by the authors, that individuals who perceive a social norm to be stronger after getting the signal are more willing to enforce this norm in the population by imposing norm-compliant behavior. This is different from the motive to shift others' *beliefs* about the norm which is at the centre of our analysis. Our experiment disentangles between these two possible motives behind punishment by varying the public observability of one's own signal about the norm, thus exogenously varying the potential of punishment as a costly signal while keeping the punisher's own beliefs about the social norm fixed.

The rest of the paper is organized as follows. Section 2 presents a theoretical model. Section 3 describes the experimental design and hypotheses. Section 4 discusses the experimental results, and Section 5 concludes.

[4] While, as shown in our model, a necessary prerequisite for the norm-signaling value of punishment to emerge are positive costs of implementing punishment, Dimant and Gesche (2023) employed costless punishment in their experiment. Indeed, studying punishment as a costly signal was not the primary purpose of their study.

## 2. Model

In order to organize thoughts, we introduce a model that elucidates how individuals' private beliefs regarding the appropriate social norm can serve as a motivating factor for both punishment of free-riders and antisocial punishment of cooperators, even when all players are identical in terms of social preferences and thus their inclination to cooperate. Building upon a substantial body of existing literature (see Bicchieri, 2016, and references therein), our starting point is the widely accepted premise that individuals do not like being perceived as transgressors of established social norms. In our model, we extend this framework of norm conformism to underscore the potential significance of relative norm violations: transgressions become more (less) tolerable as others engage in norm violations to an even greater (smaller) degree.

Specifically, we assume that there are two players, $A$ and $B$, playing a two-stage game. In Stage 1, both players simultaneously choose an investment action $a_i \in \{0,1\}$, $i = A, B$. For given player $i$, $a_i = 1$ is called *investment*, and $a_i = 0$ is called *free-riding*. In Stage 2, after observing each other's actions, both players simultaneously choose a punishment action $b_i \in \{0,1\}$, $i = A, B$. For given player $i$, $b_i = 1$ is referred to as *punishment*, and $b_i = 0$ as *no punishment*.

The payoff structure corresponds to a standard public goods game with a costly punishment option. That is, both players have an initial endowment normalized to 1. An investment choice $a_i = 1$ of player $i$ yields a payoff of $\gamma \in (0.5,1)$ to every player. If player $i$ decides to punish $j$ in Stage 2 by setting $b_i = 1$, it costs $c > 0$ to the punishing player $i$ while reducing $j'$s payoff by amount $f_{ij} \in (0, \bar{f}]$ chosen by $i$ (or being exogenously fixed). Thus, the payoff of player $i = A, B$ is given by

$$\pi_i(a_i, a_j, b_i, b_j) = 1 - a_i + \gamma \sum_{k=1,2} a_k - b_i c - b_j f_{ji}. \tag{1}$$

There are two ways to model uncertainty about the appropriate norm. One is to allow different norms to be possible, and the other one is to have one conceivable norm only, but to allow its perceived strength to differ. When the strength of a norm is perceived to be weaker, there is less shame associated with violating the norm. For simplicity, we adopt the latter approach, but note that the other approach (i.e., assuming that different degrees of cooperation can be possible behavioural norms) would yield qualitatively same results.

Specifically, we assume that cooperation, $a = 1$, is publicly regarded as the socially appropriate choice, but that the strength of the norm – the extent to which the norm is actually followed – may differ. We formalize this by a stochastic binary variable $N \in \{0,1\}$ that measures whether ($N = 1$) or not ($N = 0$) there are a sufficient number of people who choose $a = 1$, which would establish a strong descriptive norm.[5] Alternatively, $N$ can be interpreted as measuring whether there exists a sufficient majority who treat $a = 1$ as a morally appropriate choice, a strong prescriptive norm. We assume that $N$ is unknown to both players, as explained below.

People do not like to be seen as norm violators, in particular if the norm is strong. Thus, we define the norm violation of player $i$ as the difference between the normative action $a_i = 1$ and $i$'s actual action, weighted by the strength of the norm:

$$D_i(a_i) = N(1 - a_i). \tag{2}$$

We assume that each player cares about her *opponent's perception* of her relative position in terms of norm compliance. That is, if player $i$ has free-ridden while $j$ has invested, $i$ prefers that $j$ perceives the norm as weaker, thus reducing the scope for blaming $i$. On the other hand, if player $i$ has invested while $j$ has free-ridden, $i$ prefers that $j$ perceives the norm as stronger, so that it becomes clear that $i$'s behavior is norm-compliant and that $j$ is to blame for the social disagreement. In sum, $i$ tends to benefit if $j$ perceived her norm violation being larger while $i$'s norm violation is perceived to be less, thereby shifting blame and responsibility for social disagreement to the opponent. This is captured by the following utility function of player $i$:

$$u_i(a_i, a_j, b_i, b_j) = \pi_i(a_i, a_j, b_i, b_j) + \theta E_j\left[D_j(a_j) - D_i(a_i) | a_i, b_i\right], \tag{3}$$

where $E_j\left[D_j(a_j) - D_i(a_i) | a_i, b_i\right]$ is the relative norm violation as expected by the opponent after observing $i$'s investment and punishment choices, and $\theta$ is $i$'s sensitivity to $j$'s expected relative norm violation (the sensitivity is assumed to be the same for both players). Because the utility function directly depends on players' beliefs, we thus obtain a psychological game (Geanakoplos et al., 1989, Battigalli and Dufwenberg, 2022).

[5] Kölle and Quercia (2021) discuss the distinction between prescriptive (or injunctive) and descriptive social norms. The dependence of individual norm compliance on the share of population behaving close to the norm has been shown in d'Adda et al. (2020) and Dimant et al. (forthcoming), among many others. The analysis can be generalized by allowing $N$ to be a continuous variable taking values between 0 and 1. The binary version is retained for expositional simplicity.

We assume that $N$ is ex ante unknown to both players and that both players share a commonly known prior probability $q$ for $N = 1$. Players are potentially differently informed about $N$ through a privately observable and independently distributed signal about the norm $s_i$, prior to taking action $a_i$. To keep the model simple, the value of the signal can be 0 or 1. The ex ante probability of obtaining the correct signal for player $i$ (i.e., $s_i = N$) is $p_i$, which is called the precision of the signal. The precision can be either high ($p^H$) or low ($p^L$), with $1 > p^H > p^L > 0.5$. Each player knows the precision of her own signal, but not that of the opponent. In the following, we refer to a signal with high (low) precision as a precise (imprecise) signal, and to the player observing such signal as being of precise (imprecise) type. The ex ante probability that the signal of a given player is precise is $\kappa$. We say that signal $s'$ is higher than signal $s''$ if and only if the expected norm conditional on $s'$ is higher than conditional on $s''$; i.e., the signals are ordered as follows: 0 precise, 0 imprecise, 1 imprecise and 1 precise.

Below we show that both punishment and antisocial punishment can be informative about the norm and that both can occur in a perfect Bayesian equilibrium of our game. Because punishment is costly in terms of the material payoff, in order to be incentive compatible, punishment must have a positive effect on one's relative position $E_j[D_j - D_i]$ as perceived by the opponent. Consistent with empirical evidence (Herrmann et al. 2008, Nikiforakis, 2010), our model predicts that players do not punish those who have chosen the same action, because the term $D_j - D_i$ is always 0 then. Thus, only investors may punish free-riders and vice versa.[6]

The following proposition characterizes two relevant types of equilibria where punishment can be informative about the norm:[7]

**Proposition**.

*a) For certain parameter values, there exists an equilibrium with punishment of free-riders such that:*

*(i) In Stage 1, a given player* $i$ *invests if and only if* $s_i = 1$.

*(ii) In Stage 2, a given player* $i$ *punishes player* $j$ *by an arbitrary amount* $f \in (0, \bar{f}]$ *if and only if* $a_i = 1$, $a_j = 0$, *and* $i$ *holds a precise signal* $s_i = 1$.

[6] Also, because of the way we introduced norm uncertainty, an investment can never be a norm violation. A model that allows for competing norms to occur would work slightly differently.

[7] We are not interested here in trivial equilibria where all types of investors or cooperators pool on the same punishment action, in which case punishment does not reveal additional information about the player's signal on top of his observed investment choice. Such equilibria hinge on out-of-equilibrium beliefs that attribute no punishment to types with imprecise signals.

b) *For certain parameter values, there exists an equilibrium with (antisocial) punishment of cooperators such that:*

   (i) *In Stage 1, a given player $i$ invests if and only if $s_i = 1$.*

   (ii) *In Stage 2, a given player $i$ punishes player $j$ by an arbitrary amount $f \in (0, \bar{f}]$ if and only if $a_i = 0$, $a_j = 1$, and $i$ holds a precise signal $s_i = 0$.*

There do not exist pure-strategy equilibria where investment is increasing in the signal while at the same time players with imprecise signals are more likely to punish the opponent than players with precise signals. Thus, punishment is either totally uninformative (in the pooling equilibria where all types of investors or free-riders punish), or provides a favourable signal about the norm that makes the punisher's investment choice look more aligned with the norm.

To illustrate equilibrium punishment strategies given in part (*a*) of the Proposition, consider the case where player $i$ observes a precise signal $s_i = 1$, therefore invests, and eventually punishes the opponent $j$ who instead free-rode, thus, revealing $s_j = 0$. Given that all play is in equilibrium, punishment shifts the belief of $j$ about $N$ upwards by revealing that $i$ must have observed a precise signal of 1. This increases $j$'s perception of her own norm violation, while leaving her perception of $i$'s norm violation $D_i$ intact at 0, which eventually increases $i$'s utility function (3) through the relative norm violation term $E_j[D_j - D_i]$. If this psychological benefit overweighs the monetary cost of punishment $c$, punishment is optimal for $i$.

For this to be an equilibrium, we must have that a player $i$ observing an imprecise signal of 1 refrains from punishing, i.e., that her expected benefit from punishment is lower than that of the precise type. The latter follows from the fact that the precise and imprecise types differ in how they expect $j$ to update her beliefs in response to the punishment. Given the equilibrium strategy, $a_j = 0$ already reveals that the signal of $j$ must be 0. The event that $i$ and $j$ both obtain precise signals yet of different values is relatively unlikely. Consequently, the precise type of $i$ assigns a higher probability than the imprecise $i$'s type to the event that $j$'s signal is imprecise. Put differently, the precise $i$'s type expects $j$ to be less certain that the norm is weak. For certain priors, this implies that $j$'s belief responds stronger when additional information is revealed, which implies a stronger expected informational effect of punishment.[8]

[8] In particular, if $j$ has obtained a precise signal of 0, the additional precise signal of 1 revealed to $j$ through punishment would exactly offset her own signal and move her posterior belief back to the prior. If this prior is sufficiently low (below 0.5), this informational effect of punishment would be lower than if $j$ instead had privately obtained an *imprecise* signal of 0 before being punished. Thus, if $i$ puts a higher probability on the event that $j$'s

As a result, players observing precise and imprecise signals of 1 expect different benefits from punishing $j$. This enables a separating equilibrium where the monetary cost of punishment $c$ is large enough to deter the imprecise type from punishment, but low enough to still keep the punishment incentive compatible for the precise type.[9]

By the same logic, there exists an equilibrium where players who have obtained a precise signal of 0 and free-rode in Stage 1 punish investors. In this case, punishment shifts the investor's beliefs about the norm downwards, thus benefiting the free-rider by improving her image in the investor's eyes, who has now less reason to blame the norm violation of the free-rider.

Regarding players' incentives in the first stage, where the investment decision is made, we note that the expected disutility of player $i$ from free-riding, i.e., from potentially incurring the psychological cost $E_j\left[D_j(1) - D_i(0)\right]$ to oneself, is increasing in the expected norm strength $N$ (in particular, since then the opponent is expected to have a higher signal). This gives rise to the possibility, under certain parameter values, that the expected psychological cost from free-riding overweighs its expected monetary benefit, net of expected monetary punishment, if and only if one's own signal is equal to 1, consistent with the equilibrium strategies in the Proposition.

## 3. Experimental design and hypotheses

Our experiment tests whether altruistic punishment is (partly) explained by a motivation to teach others about the norm as captured by our model and as opposed to a retaliatory or simple norm-enforcement motives. Thus, the experimental game corresponds to the public goods game described in the model, except that we have increased the numbers of players to three. We decided to increase the group size to make the experiment more comparable to those in the literature, which typically involve three or more subjects per group (see Chaudhuri, 2011, for a review). Our proposition easily extends to our three-players game.

There are two stages. In Stage 1, each player decides whether or not to invest her entire endowment of 8 Euros in the public good or not. Each individual investment results in an additional payoff of 4 Euros for each of the three players, corresponding to $\gamma = 0.5$ in the

signal is imprecise, she would have larger incentives to punish $j$. If $i$ and $j$ differ in their prior beliefs, it is sufficient that only $q_j$ is below 0.5.

[9] The exact amount of punishment (conditional on being non-zero) does not matter for equilibrium incentives, i.e., any $f \in (0, \bar{f}]$ can be played in equilibrium. The only restriction is that parameter $\bar{f}$ should not be too large as otherwise free-riding in the first stage would never be incentive compatible given the equilibrium strategies.

model. In Stage 2, after observing the investments of the other group members, each player decides whether to punish one or two of the opponents by up to 4 Euro each. Punishment entails a total fixed cost of 1 Euro for the punishing player, regardless of whether she punishes one or both other players. At the end of the experiment, each player observes who has punished her and by how much. At the same time, the punishment decision of one randomly chosen player in the group is implemented.[10] Implementing the punishment of only one player eliminates some potentially confounding motives for (antisocial) punishment, such as coordination problems when two cooperators want to punish a single free-rider, or negative reciprocity on the punishment stage (Bolton et al., 2013), thus simplifying the interpretation of our data.

To study the causal effect of information asymmetries, we generate noisy signals about the social norm and privately disclose them to (some) subjects prior to their investment decisions. Our idea is to use the actual investment behavior of a random subsample of past participants as a noisy signal about the (descriptive or prescriptive) social norm. Thus, with 50% probability, subjects are informed about the behavior of 7 randomly selected participants from another baseline session in which subjects played the public goods game without being exogenously provided information about the norm. Informed subjects are told whether or not the majority of these 7 participants have chosen to invest into the public good: "*Of 7 randomly selected participants from another earlier experiment, the majority of the participants invested [did not invest] their initial endowment in the public account.*".[11] We drew random subsamples of 7 previous participants independently for each subject in the information conditions, ensuring heterogeneity of individual signals, and subjects were informed of this in the instructions (Appendix C). In this way, we randomly divided the subjects into three possible information conditions:

- *No Information*: Subjects were not shown information about the behavior of previous participants.
- *Strong Norm signal*: Subjects were informed that the majority of 7 random participants from the previous experiment chose to invest.

[10] If the punishment decision of a player is eventually *not* implemented, she does not have to pay the punishment cost of 1 Euro.

[11] We chose to disclose only the modal behavior rather than the exact number of investors to avoid informing subjects about the variance of behavior. Such more detailed distributional information could have potentially complicated our interpretation of behavior, because it might change subjects' (second-order) beliefs about the distribution of prior beliefs about the norm in the population, which is instead assumed to be uncorrelated with the signal in the model. For example, subjects learning that all 7 participants in the random subsample invested could think that all subjects, including free-riders, are likely to be quite certain about the (strong) social norm so that there is less scope to affect the opponent's beliefs about the norm by costly signaling.

- *Weak Norm signal*: Subjects were informed that the majority of 7 random participants from the previous experiment chose to not invest.

In terms of our model, a Strong Norm signal corresponds to a precise signal of 1, and a Weak Norm signal corresponds to a precise signal of 0. No Information corresponds to an imprecise signal, which must rely on the subjects' own cues about the social norm after they learn the decision situation.[12]

After the investment decision in Stage 1, we elicited subjects' first-order beliefs about the share of the other subjects in the current session who decided to invest, as well as their second-order beliefs about the first-order beliefs of both of their two group members after observing their investment decisions. Correct beliefs were rewarded with 2 Euro. This served to control for subjects' prior (first- and second-order) beliefs about the norm, and for the strength of belief updating caused by our exogenously varied norm signals.

We predict that investors who have received a precise signal that the social norm is strong (Strong Norm signal) and free-riders who have received a precise signal that the social norm is weak (Weak Norm signal) will be more likely than others to punish, in an attempt to teach – or inform – the opponents about what is known about the norm. However, there are competing motivations to punish that are unrelated to punisher's wish to change the opponents' beliefs. For example, players who received a Strong Norm signal may believe that the norm tends to be strong in the population. As a result, they may feel more obligated to 'protect' the norm by imposing monetary penalty on the norm-violator (Fehr and Fischbacher, 2004), or experience more negative emotions when the observed behavior of the opponent deviates more from the (expected) average behavior (Fehr and Gächter, 2002). Alternatively, a Strong Norm signal may lead to higher payoff expectations from the game, due to higher anticipated investment rate, and thus to a stronger retaliation when disappointed (Abeler et al., 2011). To disentangle our model from the class of potential motives that are not related to 'pure' norm-signaling, we implemented an additional treatment variation:

- In the *Private* treatment, the information condition of each subject (Weak Norm signal, Strong Norm signal, or No Information) was private knowledge, as in the model and explained above.

[12] We did not induce priors for the subjects and rather tested whether our predictions were robust to naturally occurring priors. We will return to this point in the concluding section.

- In the *Public* treatment, the subject's group members were informed about her information condition at the end of the experiment.

The advantage of the *Public* treatment is that it removes all pure teaching motives, because all information is directly observable anyway – but importantly, it leaves the various reciprocal and norm-enforcement motives intact. We randomly assigned each subject to either the Private or the Public treatment before the game started, and informed subjects about their assigned treatment prior to their punishment decision. In particular, there could be subjects in both Private and Public treatments within one group, and subjects did not know the treatment of the other group members until the end of the experiment.

Our predictions are summarized as follows.

**Hypothesis 1a:** *In the Private treatment, investors with Strong Norm signal are more likely to punish free-riders than investors with No Information or Weak Norm signal.*

**Hypothesis 1b:** *In the Private treatment, free-riders with Weak Norm signal are more likely to punish investors than free-riders with No Information or Strong Norm signal.*

**Hypothesis 2**: *In the Public treatment, the effect of information on punishment is mitigated.*

The game was played one-shot, and the experiment was conducted online, due to COVID-19, using the subject pool of the Cologne Laboratory for Economic Research. The software was programmed and run with z-Tree Unleashed (Fischbacher, 2007; Duch et al., 2020), and the participants were recruited with ORSEE (Greiner, 2015). A total of 731 complete individual observations were collected in 25 sessions, each including between 24 and 33 subjects (depending on the show-up rate).[13] For the first ten sessions, the signals were generated from the data of a baseline session, in which no information about the norm was provided to subjects and where 20 out of 30 subjects (66.7%) invested. The only purpose of the baseline session was to get the information session started, so we do not report further data here (see Table B.1 in Appendix B for the key statistics). For the last 15 sessions, signals were generated from the more representative data of the previous 10 sessions, where 177 out of 299 subjects (59.2%) invested. Because generating the Strong Norm signals was more likely than generating the Weak Norm signals, 74 (10.1%) subjects got a Weak Norm signal and 289 (39.5%) subjects

[13] Five additional subjects completed the investment stage but did not provide a punishment decision (with four of them being unable to reach the punishment stage as at least one of their group members did not provide an investment decision). Including these subjects in the analysis of investment choices (Fig. 1) does not change the results.

got a Strong Norm signal. As planned, roughly the other half of subjects, 50.3% (368), received no information. The assignment to Private vs. Public treatment proceeded individually for each subject with ex ante probability of 50%. Eventually, 50.5% of the subjects (369) were assigned to the Private treatment, and 49.5% (362) were assigned to the Public treatment. The average earning was 13.06 Euro (including the show-up fee of 2.50 Euro), and the average duration of experimental sessions was 45 minutes.

## 4. Results

### 4.1 *Cooperation and punishment*

Across all information conditions, 61.0% of our subjects invested in the first stage of the game. Our model predicts that a higher signal causes more cooperation, because it increases the expected strength of the norm, which leads to higher expected psychological disutility from (relative) norm violation. This is indeed what we find. Fig. 1 shows that a Strong Norm signal increases the cooperation rate by 22.5% compared to a Weak Norm signal ($p < 0.01$).[14] The cooperation rate of subjects having no information is in between.

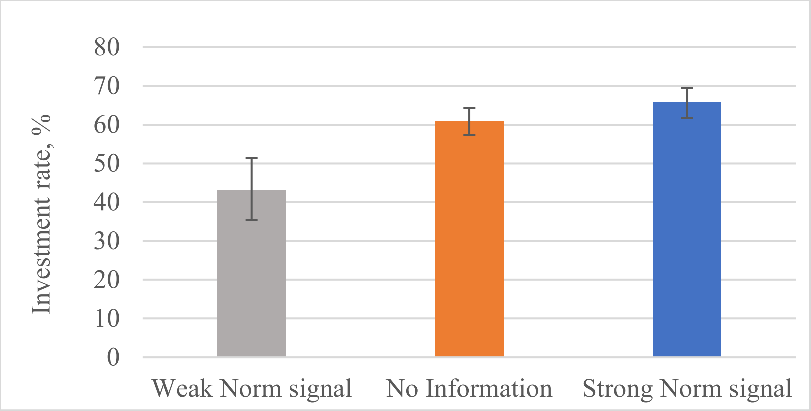

**Figure 1.** Investment rate conditional on information condition.

How does norm information affect punishment? Fig. 2 shows the punishment rate by investors facing free-riders, for which our model predicts higher rate of punishment for a Strong Norm signal than for No Information or a Weak Norm signal as long as the signal is private knowledge (Hypothesis 1a). The punishment rate measures the share of subjects who decide to

[14] Here and below the reported $p$-values for distribution comparisons are from the Mann-Whitney U test based on statistically independent observations.

punish the opponent thus incurring the fixed cost of punishment of 1 Euro.[15] Consistent with our prediction for the Private treatment, the punishment rate is the highest (66.7%) after a Strong Norm signal, and much lower in No Information and in Weak Norm signal conditions (42.9% and 20.0%, respectively). The differences in the punishment rate between Strong Norm signal and No Information and between Strong Norm signal and Weak Norm signal are both statistically significant ($p = 0.007$ and $p = 0.020$, respectively). Importantly, again as predicted, the exogenous signal about the norm in the Public treatment is much less relevant and has, in fact, no significant effect on the punishment rates. The effect goes, if at all, even in the opposite direction, and the $p$-values for the pairwise comparisons to Strong Norm signal are 0.138 for No information and 0.370 for Weak Norm signal. In particular, the punishment rate after a Strong Norm signal is significantly larger in the Private treatment than in the Public treatment ($p = 0.014$).[16] Thus, both Hypothesis 1a and Hypothesis 2 are strongly confirmed by our data.

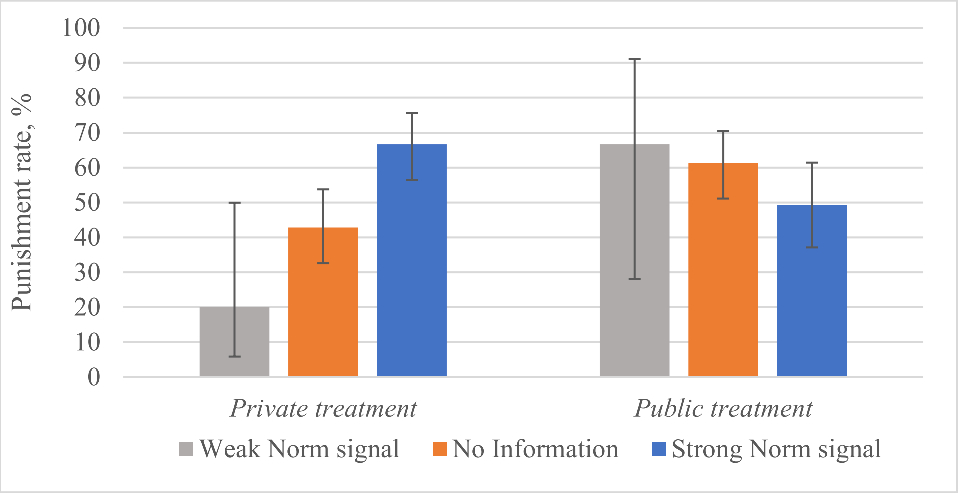

**Figure 2.** Punishment rate of investors facing free-riders conditional on information condition.

[15] The variation in the punishment amounts is discussed in Section 5. Each player makes punishment decisions separately for each of the other two group members so that we have two observations per subject. For nonparametric tests, the average punishment rate of a given subject is treated as an independent observation.

[16] For the other two signal conditions, Weak Norm signal and No Information, the punishment rate tends to be higher in the Public treatment than in the Private treatment, although the differences are not quite statistically significant ($p = 0.177$ and $p = 0.094$, respectively). We note that a higher punishment rate in these cases is not predicted by our simple model, yet a straightforward extension would capture this, as well as why investors punish even after receiving a Weak Norm signal or No Information, and it would be consistent with the observed sensitivity of punishments to the Public vs. Private treatment. All of this requires only that subjects use additional private signals beyond those induced in the experiment, e.g., from their social interaction experiences outside of the laboratory. However, since the effects not organized by our simple model are not statistically significant, it appears that such idiosyncratic signals are weak. Thus, we decided not to develop the extended model further in this study.

While the information effect for investors punishing free riders is economically and statistically strong, there is not much of an information effect in the other possible constellations. Specifically, there is hardly any antisocial punishment in our data suggested by Hypothesis 1b: Free-riders punish investors in only 4.1% of the possible cases, with insignificant variation between information conditions (see Fig. B.1 in Appendix B). We note here that, while punishment of defectors is a rather robust finding in the literature, punishment of cooperators is a less stable finding across experiments – we will return to this observation in our concluding section.[17]

The probit analysis in Table 1 corroborates our main result (the table notes include variable descriptions). Columns (1)-(2) for the Private treatment and (4)-(5) for the Public treatment show that punishment is significantly more likely if the opponent free-rode, and/or if the subject herself invested, and a Strong Norm signal significantly increases the probability of punishment, yet only in the Private treatment, as predicted. Columns (3) and (6) refer to the subsample of observations where investors face free-riders at the punishment stage (in the Private and Public treatments, respectively). The significant coefficient on Strong Norm signal in column (3) implies that that punishment of free-riders is significantly more likely if an investor has obtained a Strong Norm signal in the Private treatment. Again, this effect disappears in the Public treatment as is apparent from column (6).

[17] We also find that three investors punished investors. Also, some free riders punished other free riders, yet without any hint that the signal or that making the signal public makes any difference (see Fig. B.2 in Appendix B).

**Table 1.** The effect of information on the rate of punishment (marginal effects, probit).

| *Dep. variable:*<br>*Punishment* | | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treatment/subsample | Private | Private | Private, investors facing free-riders | Public | Public | Public, investors facing free-riders |
| Own investment | 0.114*** | 0.119*** | | 0.117*** | 0.107*** | |
| | (0.023) | (0.023) | | (0.019) | (0.018) | |
| Opponent free-rode | 0.244*** | 0.239*** | | 0.252*** | 0.240*** | |
| | (0.030) | (0.029) | | (0.034) | (0.037) | |
| Strong Norm signal | 0.064*** | 0.059** | 0.260*** | -0.022 | -0.014 | -0.101 |
| | (0.024) | (0.024) | (0.087) | (0.024) | (0.022) | (0.091) |
| Age | | -0.002 | -0.004 | | 0.006** | 0.024*** |
| | | (0.003) | (0.009) | | (0.002) | (0.009) |
| Female | | -0.032 | -0.057 | | -0.045** | -0.226** |
| | | (0.026) | (0.099) | | (0.021) | (0.090) |
| Observations | 738 | 732 | 188 | 724 | 724 | 172 |

*Notes*: Robust standard errors clustered by subjects in parenthesis. *** $p < 0.01$, ** $p < 0.05$. Columns (1)-(3) include data from the Private treatment, and columns (4)-(6) include data from the Public treatment. 'Own investment' is equal to 1(0) if the subject invested (free-rode), 'Opponent free-rode' is equal to 1(0) if the player to whom the punishment decision pertains free-rode (invested), 'Strong Norm signal' is equal to 1 if the subject observed a Strong Norm signal and to 0 otherwise, 'Age' is the subject's reported age, 'Female' is equal to 1 if the subject reported female gender and to 0 otherwise. The number of observations varies between columns (1)-(2) since 3 subjects did not report demographic variables at the end. Because each subject had to take two separate punishment decisions for each of the other group members, we have two observations per subject.

### 4.2 *Beliefs*

Our model predicts that actions are moderated through beliefs. Indeed, we find that the signals causally affected beliefs about the social norm, consistent with our theoretical mechanism. Fig. 3 shows that first-order beliefs about the investment rate are higher in Strong Norm signal than in either No Information or Weak Norm signal conditions, with both pairwise differences significant at the 1% level. The differences are also statistically significant for second-order beliefs ($p = 0.013$ and $p < 0.01$, respectively). Thus, subjects in Strong Norm signal condition expect both a higher social norm and higher beliefs of others about the norm. The latter fact is an important prerequisite for the separating equilibrium in our model as explained in the theory section. Moreover, Fig. 4 shows in line with our predictions that, for a given received signal, investors have significantly higher expectations of the social norm than free-riders ($p < 0.01$). In particular, conditional on no information from the experimenter, subjects are still characterized by significant heterogeneity of their beliefs which are, in turn, aligned with their investment decisions. This justifies our assumption that No Information condition in the experiment may correspond to an imprecise signal equal to either 0 or 1 in the model. Finally, the right side of Fig. 4 shows that subjects correctly anticipate this belief pattern if asked to provide their second-order beliefs about the first-order beliefs of their particular group members, who have either invested or not.

We conclude that not only behavior, but also the pattern of beliefs is consistent with the causal mechanisms predicted by our model.

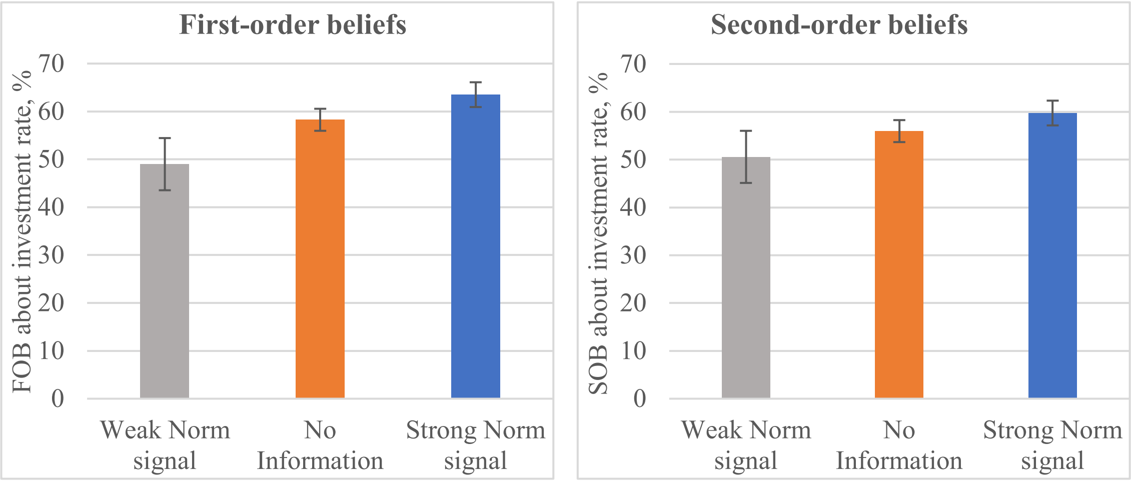

**Figure 3.** First- and second-order beliefs conditional on information condition.

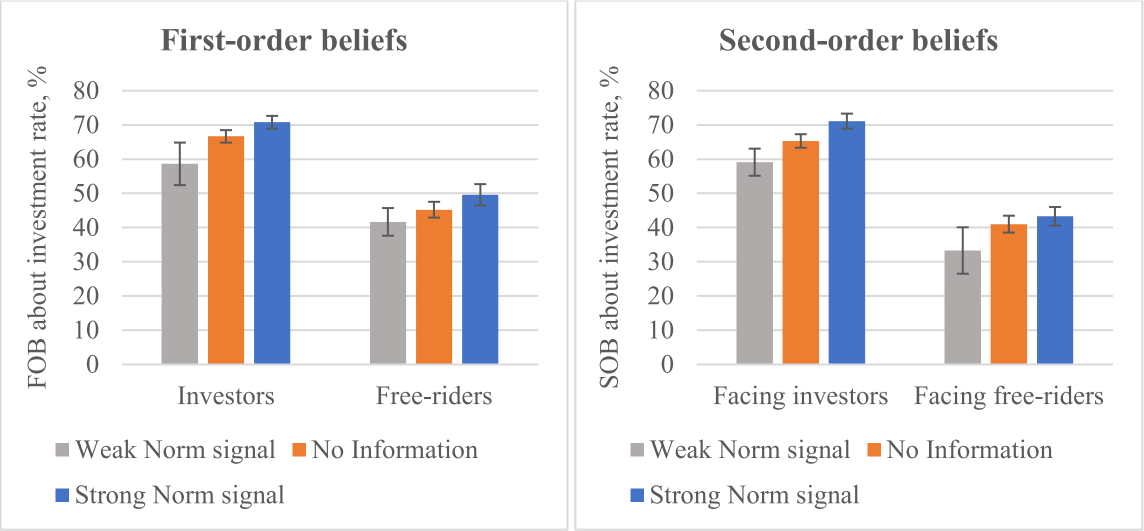

**Figure 4.** First-(second-)order beliefs conditional on own (opponent's) investment behavior and information condition.

## 5. Discussion and conclusion

Altruistic punishment is a robust phenomenon, perhaps because there is more than one reason to engage in it, even if it is not in one's own material self-interest. We investigate whether altruistic punishment in the public goods game can be (partly) driven by a desire to signal to transgressors what is individually known about the social norm. Our model is based on an idea of comparative norm conformism: People are particularly reluctant to be perceived as norm violators if others are not perceived as such. Similarly, if people follow the norm while others do not, they prefer to make the latter acknowledge this difference in norm compliance. In this case, altruistic punishment emerges straightforwardly and endogenously in equilibrium as a credible and costly signaling strategy. In particular, cooperators may punish a free rider only if they are sufficiently convinced that his behaviour deviates from a strong social norm. As a result, the punishment received has informational value and shifts the free rider's belief about the norm in a direction that shifts blame towards him, as desired by the punisher.

The results of our experiment confirm that cooperators are indeed more likely to punish free-riders after receiving a signal suggesting a strong social norm, and the first- and second-order belief patterns fully support our 'punishment as norm signaling' hypothesis. Importantly, the effect disappears when the private signal of the cooperator is publicly revealed to everybody independently of punishment. The positive effect of the normative signal on the rate of punishment thus needs an explanation that includes a desire of the punisher to influence the beliefs about the norm of the punished player.

While punishment in our model serves the desire to avoid being seen as a relatively strong norm violator or to be seen as a relatively strong norm complier, it is consistent with and complementary to other approaches in that it is also “reciprocal” and “altruistic”, although in different ways. Reciprocity comes in because teaching free-riders increases the free-riders’ psychological disutility from norm violations, which is believed to be more severe after punishment. This kind of reciprocity can complement any monetary retaliation motive that may or may not exist. Indeed, our experiment allowed different degrees of financial retaliation, and most punishers in fact chose the maximal monetary level of punishment for free-riders across all information conditions (4 Euros; see Fig. B.3 in Appendix B). Because all that matters for credible norm-signaling in our model is the cost of punishment, which is independent of the size of the punishment, this indicates that punishers like to teach free-riders about what they know *and* to financially harm the punished. It is conceivable that the desire to teach and the desire to harm are complementary, in that teaching is more effective when it is accompanied by more financial harm, or that knowing that the transgressor understands the underlying reason for the financial punishment and thus feels more responsible for the social conflict is more satisfying for the punisher (Molnar et al., 2023). The investigation of such potential complementarities is left to future research.

Regarding an altruistic element of punishment as norm-signaling, observe that teaching others about the social norm improves social cohesion of the group by helping aligning dispersed norm beliefs. To illustrate, suppose there is a larger social community, company or economic platform where punishment is possible, and people have uncertainty about the prevailing social norm, leading to different beliefs about what the norm is or should be. Applying our model and motivation to a setting where group members repeatedly interact with different group members, people’s beliefs would converge to a common belief about the social norm once punishment conveys private signals about the actual norm (see Aumann, 1976, and Geanakoplos and Polemarchakis, 1982, for belief convergence results in related contexts). That is, one’s rather selfish individual motive to protect one’s social image of not being a norm violator can be collectively beneficial if it leads to the emergence of consensus about the social norm, even if people initially differ in what they believe (see also Yuan et al., forthcoming). In this sense, because our punishment-as-signaling mechanism benefits the group by increasing social cohesion, it complements traditional views of the role of altruism in altruistic punishment. Future research could explore how these mechanisms reinforce each other.

One advantage of our model is that, unlike other models of punishment, norm-signaling naturally includes the possibility of antisocial punishment. Herrmann et al. (2008) document the widespread existence of antisocial punishment. They also find very strong cross-societal variation. In particular, they show that the inclination for antisocial punishment in Western countries such as Germany is low, as our data confirm. This heterogeneity of results is in so far consistent with our model that the prevalence of particular types of punishment depends on parameters and priors regarding the appropriate norm, which might well differ across cultures and societies – and which in turn leads to new hypotheses about the determinants of antisocial punishment and the underlying reason for the previously observed heterogeneity. For instance, according to the proof of our proposition, antisocial punishment arises only when the prior belief about the strength of the norm is sufficiently large (Appendix A). In our setup, the average first- and second-order beliefs about the investment rate in No Information condition are quite close to 50% (see Figure 3). This seems to reflect subjects' strong prior uncertainty about the behavior of others, suggesting that they tend not to believe in a strong social norm, which would correspond to a low prior belief $q$ in terms of our model. In this case, there would be less scope for downward revision of the opponent's beliefs about the norm, which prevents antisocial punishment in our setting. Disentangling these and other explanations (see our Introduction) for the (non-)occurrence of antisocial punishment, and testing our model in an environment more conducive to antisocial punishment, is another interesting avenue for further research.

That said, our study shows that altruistic punishment is more robust than previously thought, in the sense that it can occur both in rational equilibrium and in the laboratory even among similar individuals who dislike being seen as norm violators and thus care about communicating one's understanding of the appropriate social norm.

## Acknowledgements

We thank Eugen Dimant, Martin Dufwenberg, Ben Greiner, George Loewenstein, Sergey Popov, and participants of various conferences and workshops for their helpful comments and suggestions. Approval of the Ethics Committee of the University of Cologne from 1 September 2018.

## Funding

Ockenfels gratefully acknowledges support by the German Science Foundation through Germany's Excellence Strategy (EXC 2126/1 390838866) and through the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 741409).

## References

Abeler, J., Falk, A., Goette, L. and Huffman, D., 2011. Reference points and effort provision. *American Economic Review*, *101*(2), pp.470-492.

Akerlof, G.A., 1980. A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics, 94*(4), pp.749-775.

Andreoni, J. and Bernheim, B.D., 2009. Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), pp.1607-1636.

Aumann, R.J., 1976. Agreeing to disagree. *Annals of Statistics*, *4*(6), pp. 1236–1239.

Balafoutas, L. and Nikiforakis, N., 2012. Norm enforcement in the city: A natural field experiment. *European Economic Review*, *56*(8), pp.1773-1785.

Battigalli, P. and Dufwenberg, M., 2022. Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, *60*(3), pp.833-882.

Bénabou, R. and Tirole, J., 2011. Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics*, *126*(2), pp.805-855.

Bernheim, B.D., 1994. A theory of conformity. *Journal of Political Economy*, *102*(5), pp.841-877.

Bicchieri, C., 2016. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

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, pp.209-235.

Bolton, G., Greiner, B. and Ockenfels, A., 2013. Engineering trust: Reciprocity in the production of reputation information. *Management Science*, 59(2), pp.265-285.

Bolton, G.E., Greiner, B., and Ockenfels, A., 2018. Dispute resolution or escalation? The strategic gaming of feedback withdrawal options in online markets. *Management Science*, 64(9), pp.4009–4031.

Bolton, G.E., Kusterer, D.J. and Mans, J., 2019. Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, *65*(*11*), pp. 5371-5391.

Bolton, G.E., Mans, J., and Ockenfels, A., 2020. Norm enforcement in markets: Group identity and the volunteering of feedback. *Economic Journal*, 130, pp.1248–1261.

Bolton, G.E. and Ockenfels, A., 1998. Strategy and equity: An ERC-analysis of the Güth–van Damme game. *Journal of Mathematical Psychology*, *42*(2-3), pp.215-226.

Bolton, G.E. and Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *91*(1), pp.166-193.

Brouwer, T., Galeotti, F. and Villeval, M.C., 2023. Teaching Norms: Direct Evidence of Parental Transmission. *The Economic Journal*, 133(650), pp.872-887.

Carpenter, J. and Matthews, P.H., 2009. What norms trigger punishment? *Experimental Economics*, 12, pp.272-288.

Carpenter, J.P. and Matthews, P.H., 2012. Norm enforcement: anger, indignation, or reciprocity? *Journal of the European Economic Association*, *10*(3), pp.555-572.

Casoria, F., Galeotti, F. and Villeval, M.C., 2021. Perceived social norm and behavior quickly adjusted to legal changes during the COVID-19 pandemic. *Journal of Economic Behavior & Organization*, 190, pp.54-65.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14, pp.47-83.

Chen, Y., Cramton, P., List, J., Ockenfels, A., 2021. Market design, human behavior and management. *Management Science*, 67(9), pp.5317-5348.

Chen, H., Zeng, Z. and Ma, J., 2020. The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *Plos One*, 15(3), p.e0229510.

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

Dimant, E. and Gesche, T., 2023. Nudging enforcers: How norm perceptions and motives for lying shape sanctions. *PNAS Nexus*, 2(7), p.pgad224.

Dimant, E., Gelfand, M.J., Hochleitner, A. and Sonderegger, S., forthcoming. Strategic behavior with tight, loose, and polarized norms. *Management Science*.

Drago, F., Galbiati, R. and Vertova, P., 2009. The deterrent effects of prison: Evidence from a natural experiment. *Journal of Political Economy*, *117*(2), pp.257-280.

Duch, M.L., Grossmann, M.R. and Lauer, T., 2020. z-Tree unleashed: A novel client-integrating architecture for conducting z-Tree experiments over the Internet. *Journal of Behavioral and Experimental Finance*, *28*, p.100400.

Dufwenberg, M. and Kirchsteiger, G., 2004. A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), pp.268-298.

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

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

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

Fehr, E. and Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), pp.817-868.

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

Fudenberg, D., and Maskin, E., 1986. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica 54*(*3*), pp. 533-54.

Funk, P., 2007. Is there an expressive function of law? An empirical analysis of voting laws with symbolic fines. *American Law and Economics Review*, *9*(1), pp.135-159.

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

Geanakoplos, J.D. and Polemarchakis, H.M., 1982. We can't disagree forever. *Journal of Economic Theory*, *28*(1), pp.192-200.

Geanakoplos, J., Pearce, D. and Stacchetti, E., 1989. Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), pp.60-79.

Golman, R., Loewenstein, G., Moene, K.O. and Zarri, L., 2016. The preference for belief consonance. *Journal of Economic Perspectives*, *30*(3), pp.165-188.

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

Herrmann, B., Thoni, C. and Gachter, S., 2008. Antisocial punishment across societies. *Science*, *319*(5868), pp.1362-1367.

Jordan, J.J., Hoffman, M., Bloom, P. and Rand, D.G., 2016. Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), pp.473-476.

Jordan, J.J. and Rand, D.G., 2017. Third-party punishment as a costly signal of high continuation probabilities in repeated games. *Journal of Theoretical Biology*, *421*, pp.189-202.

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

Lane, T., Nosenzo, D. and Sonderegger, S., 2023. Law and norms: Empirical evidence. *American Economic Review*, *113*(5), pp.1255-1293.

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), pp.366-380.

Molnar, A., Chaudhry, S.J. and Loewenstein, G., 2023. "It's not about the money. It's about sending a message!" Avengers want offenders to understand the reason for revenge. *Organizational Behavior and Human Decision Processes*, *174*, p.104207.

Molnar, A. and Loewenstein, G., 2020. The False and the Furious: People are more disturbed by others' false beliefs than by differences in beliefs. *Working paper*.

Nikiforakis, N., 2010. Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, 68(2), pp.689-702.

Ockenfels, A., 2023. Behavioral market design. *Behavioral and Brain Sciences,* 46, e171.

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

Rabin, M., 1993. Incorporating fairness into game theory and economics. *American Economic Review*, pp.1281-1302.

Rai, T.S., 2022. Material benefits crowd out moralistic punishment. *Psychological Science*, *33*(5), pp.789-797.

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

te Velde, V.L., 2022. Heterogeneous norms: Social image and social pressure when people disagree. *Journal of Economic Behavior & Organization*, *194*, pp.319-340.

Thöni, C., 2014. Inequality aversion and antisocial punishment. *Theory and Decision, 76*, pp.529-545.

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

Xiao, E. and Houser, D., 2005. Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*(20), pp.7398-7401.

Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), pp.321-344.

Yuan, Y., Liu, T.X., Tan, C., Chen, Q., Pentland, A.S. and Tang, J., forthcoming. Gift contagion in online groups: Evidence from virtual red packets. *Management Science*.

# Appendix A: Omitted proofs

**Lemma 1.** *Assume that in equilibrium player* $i$ *invests in the first stage if and only if* $s_i = 1$*. Then,* $i$*'s expected probability that* $j$ *has observed a precise signal conditional on observing* $a_j \neq a_i$ *(and before observing* $b_j$*) is decreasing in the precision of* $i$*'s own signal. Formally,*

$$Pr_i[p_j = p^H | p_i = p^H, a_j \neq a_i] < Pr_i[p_j = p^H | p_i = p^L, a_j \neq a_i]. \tag{4}$$

*Proof:* Assume that players invest if and only if their signal is equal to 1. Consider the case where player *i* has observed a signal of 1 and the opponent's action of 0 in Stage 1. We need to show that

$$\begin{aligned} &Pr_i[p_j = p^H \big| p_i = p^H, s_i = 1, a_j = 0\big] - Pr_i[p_j = p^H \big| p_i = p^L, s_i = 1, a_j = 0\big] < 0 \\ &\Leftrightarrow Pr_i[p_j = p^H \big| p_i = p^H, s_i = 1, s_j = 0\big] - Pr_i[p_j = p^H \big| p_i = p^L, s_i = 1, s_j = 0\big] < 0, \end{aligned} \tag{5}$$

where the equivalence is due to players following their signals in the first stage. By the law of total probability

$$\begin{aligned} &Pr_i[p_j = p^H | p_i, s_i, s_j] = Pr_i[p_j = p^H | N = 1, p_i, s_i, s_j]\, Pr_i[N = 1 | p_i, s_i, s_j] \\ &+ Pr_i[p_j = p^H | N = 0, p_i, s_i, s_j]\, Pr_i[N = 0 | p_i, s_i, s_j] \\ &= Pr_i[N = 1 | p_i, s_i, s_j](Pr_i[p_j = p^H | N = 1, s_j] - Pr_i[p_j = p^H | N = 0, s_j]) \\ &+ Pr_i[p_j = p^H | N = 0, s_j] \end{aligned} \tag{6}$$

where for the last equality we used the fact that $Pr_i[p_j = p^H | N = 1, p_i, s_i, s_j] = Pr_i[p_j = p^H | N = 1, s_j]$ since the signals are distributed independently from each other conditional on a given state. Substituting this into (6) we obtain

$$\begin{aligned} &(Pr_i[N = 1 | p_i = p^H, s_i = 1, s_j = 0] - Pr_i[N = 1 | p_i = p^L, s_i = 1, s_j = 0]) \\ &\times (Pr_i[p_j = p^H | N = 1, s_j = 0] - Pr_i[p_j = p^H | N = 0, s_j = 0]) < 0. \end{aligned} \tag{7}$$

It is easily verifiable (by Bayes rule) that the first term in brackets is positive (the precise signal of 1 leads to higher beliefs about the norm than the imprecise signal of 1). Consider the second term. By Bayes rule (and given that $Pr_i[p_j = p^H | N] = Pr_i[p_j = p^H]$)

$$\begin{aligned}&Pr_i[p_j = p^H|N = 1, s_j = 0]\\&= \frac{Pr_i[s_j = 0|N = 1, p_j = p^H]\,Pr_i[p_j = p^H]}{Pr_i[s_j = 0|N = 1, p_j = p^H]\,Pr_i[p_j = p^H] + Pr_i[s_j = 0|N = 1, p_j = p^L]\,Pr_i[p_j = p^L]} \quad (8)\\&= \frac{(1 - p^H)\kappa}{(1 - p^H)\kappa + (1 - p^L)(1 - \kappa)}\end{aligned}$$

and, similarly,

$$\begin{aligned}&Pr_i[p_j = p^H|N = 0, s_j = 0]\\&= \frac{Pr_i[s_j = 0|N = 0, p_j = p^H]\,Pr_i[p_j = p^H]}{Pr_i[s_j = 0|N = 0, p_j = p^H]\,Pr_i[p_j = p^H] + Pr_i[s_j = 0|N = 0, p_j = p^L]\,Pr_i[p_j = p^L]} \quad (9)\\&= \frac{p^H\kappa}{p^H\kappa + p^L(1 - \kappa)}.\end{aligned}$$

It follows,

$$\begin{aligned}&Pr_i[p_j = p^H\big|N = 1, s_j = 0\big] - Pr_i[p_j = p^H\big|N = 0, s_j = 0\big]\\&= \frac{(1 - p^H)\kappa}{(1 - p^H)\kappa + (1 - p^L)(1 - \kappa)} - \frac{p^H\kappa}{p^H\kappa + p^L(1 - \kappa)} \quad (10)\\&< \frac{(1 - p^H)\kappa}{(1 - p^H)\kappa + (1 - p^H)(1 - \kappa)} - \frac{p^H\kappa}{p^H\kappa + p^H(1 - \kappa)} = 0.\end{aligned}$$

Consequently, (8) and hence (6) hold.

The proof for the opposite case where $a_i = 0$ and $a_j = 1$ is fully symmetric and hence omitted.

■

**Proof of Proposition 1.**

**Claim 1.** *Assume that in equilibrium player i invests in Stage 1 if and only if* $s_i = 1$*, while* $q \leq 0.5$*. Then, there exists a non-empty range of* $c/\theta$ *such that investors do not have incentives to deviate from the following punishment strategy in Stage 2 if it is played in equilibrium: investor i punishes player j by an arbitrary amount* $f \in (0, \bar{f}]$ *if and only if* $a_j = 0$ *and* $p_i = p^H$*.*

*Proof*: Assume that players invest in Stage 1 if and only if their signal is equal to 1. Let us show that no informational types of investor have incentives to deviate from the prescribed equilibrium strategy in Stage 2.

First, no investor has incentives to punish another investor, because the relative norm deviation $D_j - D_i$ is then always 0 independently of punishment, while punishment itself is costly. Thus, there are no incentives to deviate from the equilibrium strategy in Stage 2 in this case.

Consider the remaining case when $a_i = 1$ and $a_j = 0$. This is the case if and only if $s_i = 1$ and $s_j = 0$ by assumption. Then, by (3) $i$'s expected gain from punishing $j$ is (given that $D_i(1) = 0$)

$$\begin{aligned}
&E_i\big[u_i\big(a_i = 1, a_j = 0, b_i = 1\big)|p_i, s_i, a_j\big] - E_i\big[u_i\big(a_i = 1, a_j = 0, b_i = 0\big)|p_i, s_i, a_j\big] \\
&\quad = \theta\big(E_i\big[E_j\big[D_j(0)|a_i, b_i = 1\big]|p_i, s_i, a_j\big] - E_i\big[E_j\big[D_j(0)|a_i, b_i = 0\big]|p_i, s_i, a_j\big]\big) - c \\
&\quad = \theta\big(E_i\big[N_j^E(a_i, b_i = 1)|p_i, s_i, a_j\big] - E_i\big[N_j^E(a_i, b_i = 0)|p_i, s_i, a_j\big]\big) - c,
\end{aligned} \tag{11}$$

where $N_j^E(a_i, b_i)$ is the expectation of $j$ about the norm strength conditional on $a_i$ and $b_i$. In the considered equilibrium, it must hold that player $i$ has at least a weak preference for punishment if her signal precision is high, and has at least weak preference for no punishment if her signal precision is low. This is equivalent to the following set of conditions:

$$\begin{aligned}
&\theta\big(E_i\big[N_j^E(a_i, b_i = 1)|p_i = p^H, s_i, a_j\big] - E_i\big[N_j^E(a_i, b_i = 0)|p_i = p^H, s_i, a_j\big]\big) - c \geq 0, \\
&\theta\big(E_i\big[N_j^E(a_i, b_i = 1)|p_i = p^L, s_i, a_j\big] - E_i\big[N_j^E(a_i, b_i = 0)|p_i = p^L, s_i, a_j\big]\big) - c \leq 0,
\end{aligned} \tag{12}$$

or

$$c/\theta \in [\tau_1, \tau_2], \tag{13}$$

where

$$\begin{aligned}
\tau_1 &= E_i\big[N_j^E(a_i, b_i = 1)|p_i = p^L, s_i = 1, a_j\big] - E_i\big[N_j^E(a_i, b_i = 0)|p_i = p^L, s_i = 1, a_j\big], \\
\tau_2 &= E_i\big[N_j^E(a_i, b_i = 1)|p_i = p^H, s_i = 1, a_j\big] - E_i\big[N_j^E(a_i, b_i = 0)|p_i = p^H, s_i = 1, a_j\big].
\end{aligned} \tag{14}$$

Hence, for any given $\theta$ there exists a non-empty set of $c$ so that the incentive constraints are satisfied if and only if

$$\tau_2 - \tau_1 \geq 0. \tag{15}$$

Let us show under which parameter conditions this is the case. By the law of total probability

$$\begin{aligned}&E_i[N_j^E(a_i,b_i)|p_i,s_i,a_j] = N_j^E(a_i,b_i,s_j,p_j=p^H)\,Pr_i[p_j=p^H|p_i,s_i,a_j]\\&+N_j^E(a_i,b_i,s_j,p_j=p^L)(1-Pr_i[p_j=p^H|p_i,s_i,a_j])\\&=Pr_i[p_j=p^H|p_i,s_i,a_j](N_j^E(a_i,b_i,s_j,p_j=p^H)-N_j^E(a_i,b_i,s_j,p_j=p^L))\\&+N_j^E(a_i,b_i,s_j,p_j=p^L)\end{aligned} \tag{16}$$

Substituting this into (15) we obtain,

$$\begin{aligned}\tau_1 = Pr[p_j=p^H|p_i=p^L,s_i,a_j]\phi + N_j^E(a_i,b_i=1,s_j=0,p_j=p^L)\\ - N_j^E(a_i,b_i=0,s_j=0,p_j=p^L),\end{aligned} \tag{17}$$

$$\begin{aligned}\tau_2 = Pr_i[p_j=p^H|p_i=p^H,s_i,a_j]\phi + N_j^E(a_i,b_i=1,s_j=0,p_j=p^L)\\ - N_j^E(a_i,b_i=0,s_j=0,p_j=p^L),\end{aligned} \tag{18}$$

where

$$\begin{aligned}\phi = N_j^E(a_i,b_i=1,s_j=0,p_j=p^H) - N_j^E(a_i,b_i=1,s_j=0,p_j=p^L)\\ -(N_j^E(a_i,b_i=0,s_j=0,p_j=p^H) - N_j^E(a_i,b_i=0,s_j=0,p_j=p^L))\end{aligned} \tag{19}$$

Hence,

$$\tau_2-\tau_1 = (Pr_i[p_j=p^H|p_i=p^H,s_i,a_j] - Pr_i[p_j=p^H|p_i=p^L,s_i,a_j])\phi. \tag{20}$$

Here, the term in brackets is strictly negative by Lemma 1. Consider $\phi$. For given signal $s_j$

$$\begin{aligned}&N_j^E(a_i=1,b_i=1,s_j,p_j) = Pr_j[N=1|a_i=1,b_i=1,s_j,p_j]\\&=Pr_j[N=1|s_i=1,p_i=p^H,s_j,p_j]\\&=\frac{Pr_j[s_i=1,p_i=p^H,s_j,p_j|N=1]q}{Pr_j[s_i=1,p_i=p^H,s_j,p_j]},\end{aligned} \tag{21}$$

where the second inequality is due to the fact that punishment signals high precision of the punisher’s 1-signal in equilibrium, and the last equality is by Bayes rule. At the same time, since $s_i$ and $s_j$ are independently distributed, by the probability chain rule

$$\begin{aligned}&Pr_j[s_i,p_i,s_j,p_j|N] = Pr[s_i|p_i,N]\,Pr[p_i|N]\cdot Pr[s_j|p_j,N]\,Pr[p_j|N]\\&=Pr[s_i|p_i,N]\,Pr[p_i]\cdot Pr[s_j|p_j,N]\,Pr[p_j].\end{aligned} \tag{22}$$

Substituting this into (22) we obtain

$$N_j^E(a_i = 1, b_i = 1, s_j = 0, p_j) = \frac{Pr_j[s_i = 1, p_i = p^H, s_j = 0, p_j | N = 1]q}{Pr_j[s_i = 1, p_i = p^H, s_j = 0, p_j]}$$
$$= \frac{p^H(1 - p_j)q}{p^H(1 - p_j)q + (1 - p^H)p_j(1 - q)}. \tag{23}$$

Similarly,

$$N_j^E(a_i = 1, b_i = 0, s_j = 0, p_j) = \frac{Pr_j[s_i = 1, p_i = p^L, s_j = 0, p_j | N = 1]q}{Pr_j[s_i = 1, p_i = p^L, s_j = 0, p_j]}$$
$$= \frac{p^L(1 - p_j)q}{p^L(1 - p_j)q + (1 - p^L)p_j(1 - q)}. \tag{24}$$

Substituting (24) and (25) into (20) we obtain:

$$\phi = \frac{p^H(1 - p^H)q}{p^H(1 - p^H)q + (1 - p^H)p^H(1 - q)} - \frac{p^H(1 - p^L)q}{p^H(1 - p^L)q + (1 - p^H)p^L(1 - q)}$$
$$- \left(\frac{p^L(1 - p^H)q}{p^L(1 - p^H)q + (1 - p^L)p^H(1 - q)} - \frac{p^L(1 - p^L)q}{p^L(1 - p^L)q + (1 - p^L)p^L(1 - q)}\right) \tag{25}$$
$$= \frac{(p^H - p^L)^2(1 - q)q}{\big(q(p^H - p^L) + p^L(1 - p^H)\big)\big(p^H(1 - q) - p^L(p^H - q)\big)}(2q - 1).$$

The fraction term is clearly positive. Hence, $\phi \leq 0$ if and only if $2q - 1 \leq 0 \Leftrightarrow q \leq 0.5$. This together with (21) and Lemma 1 implies

$$\tau_2 - \tau_1 \geq 0 \Leftrightarrow q \leq 0.5. \tag{26}$$

Consequently, if $q \leq 0.5$, then for $c/\theta \in [\tau_1, \tau_2]$ investors do not have incentives to deviate from the prescribed equilibrium strategy in Stage 2. ■

**Claim 2.** *Assume that in equilibrium player $i$ invests in Stage 1 if and only if $s_i = 1$, while $q \geq 0.5$. Then, there exists a non-empty range of $c/\theta$ such that free-riders do not have incentives to deviate from the following punishment strategy in Stage 2 if it is played in equilibrium: free-rider $i$ punishes player $j$ by an arbitrary amount $f \in (0, \bar{f}]$ if and only if $a_j = 1$ and $p_i = p^H$.*

*Proof*: Assume that players invest in Stage 1 if and only if their signal is equal to 1. We need to show that no informational types of free-riders have incentives to deviate from the prescribed equilibrium strategy in Stage 2.

First, no free-rider has incentives to punish another free-rider, because the relative norm deviation $D_j - D_i$ is then always 0 independently of punishment, while punishment itself is costly. Thus, there are no incentives to deviate from the equilibrium strategy in this case.

Consider the remaining case when $a_i = 0$ and $a_j = 1$. This is the case if and only if $s_i = 0$ and $s_j = 1$ by assumption. Then, by (3) $i$'s expected gain from punishing $j$ is (given that $D_j(1) = 0$)

$$\begin{aligned} &E_i\left[u_i\left(a_i = 0, a_j = 1, b_i = 1\right)|p_i, s_i, a_j\right] - E_i\left[u_i\left(a_i = 0, a_j = 1, b_i = 0\right)|p_i, s_i, a_j\right] \\ &= \theta\left(-E_i\left[E_j[D_i(0)|a_i, b_i = 1]|p_i, s_i, a_j\right] + E_i\left[E_j[D_i(0)|a_i, b_i = 0]|p_i, s_i, a_j\right]\right) - c \\ &= \theta\left(E_i\left[N_j^E(a_i, b_i = 0)|p_i, s_i, a_j\right] - E_i\left[N_j^E(a_i, b_i = 1)|p_i, s_i, a_j\right]\right) - c, \end{aligned} \tag{27}$$

where, as before, $N_j^E(a_i, b_i)$ is the expectation of $j$ about the norm strength conditional on $a_i$ and $b_i$. In the considered equilibrium, it must hold that player *i* has at least a weak preference for punishment if her signal precision is high, and has at least weak preference for no punishment if her signal precision is low. This is equivalent to the following set of conditions:

$$\begin{aligned} &\theta\left(E_i\left[N_j^E(a_i, b_i = 0)|p_i = p^H, s_i, a_j\right] - E_i\left[N_j^E(a_i, b_i = 1)|p_i = p^H, s_i, a_j\right]\right) - c \geq 0, \\ &\theta\left(E_i\left[N_j^E(a_i, b_i = 0)|p_i = p^L, s_i, a_j\right] - E_i\left[N_j^E(a_i, b_i = 1)|p_i = p^L, s_i, a_j\right]\right) - c \leq 0, \end{aligned} \tag{28}$$

or

$$c/\theta \in [\widehat{\tau_1}, \widehat{\tau_2}], \tag{29}$$

where

$$\begin{aligned} \widehat{\tau_1} &= E_i\left[N_j^E(a_i, b_i = 0)|p_i = p^L, s_i = 0, a_j\right] - E_i\left[N_j^E(a_i, b_i = 1)|p_i = p^L, s_i = 0, a_j\right], \\ \widehat{\tau_2} &= E_i\left[N_j^E(a_i, b_i = 0)|p_i = p^H, s_i = 0, a_j\right] - E_i\left[N_j^E(a_i, b_i = 1)|p_i = p^H, s_i = 0, a_j\right]. \end{aligned} \tag{30}$$

Hence, for any given $\theta$ there exists a non-empty set of *c* so that the incentive constraints are satisfied if and only if

$$\widehat{\tau_2} - \widehat{\tau_1} \geq 0. \tag{31}$$

The proof that this is the case under $a_i = 0, a_j = 1, s_i = 0$ and $q \geq 0.5$ follows by analogous derivations as in (17)-(26) and is omitted. ■

**Claim 3.** *Assume $q \leq 0.5$, $c/\theta \in [\tau_1, \tau_2]$ and in equilibrium:*

- *Player $i$ invests in Stage 1 if and only if $s_i = 1$.*
- *Investor $i$ punishes player $j$ by an arbitrary amount $f \in (0, \bar{f}]$ if and only if $a_j = 0$ and $p_i = p^H$.*
- *Free-riders never punish. Out-of-equilibrium belief of player $i$ after being punished by a free-rider $j$ is that $p_j = p^L$.*

*Then, no player with a signal $s_i = 1$ has incentives to punish the other player in Stage 2 after deviating to $a_i = 0$ in Stage 1. Analogously, no player with a signal $s_i = 0$ has incentives to punish the other player in Stage 2 after deviating to $a_i = 1$ in Stage 1.*

*Proof:* Consider the incentives of player $i$ with signal $s_i = 1$ after deviating from the equilibrium play to $a_i = 0$ in Stage 1. This player does not have incentive to punish player $j$ in Stage 2 if $a_j = 0$, since then the relative norm violation expected by $j$, $E_j[D_j - D_i]$, is always 0 independently of punishment. Thus, the claim is satisfied in this case. Consider $a_j = 1$. In this case, by (2) the expected relative norm violation term is

$$E_i\big[E_j\big[D_j(1) - D_i(0)\big|a_i, b_i\big]\big|p_i, s_i\big] = -E_i[N_j^E(a_i, b_i)|p_i, s_i], \tag{32}$$

so that $i$ has incentives to reduce the opponent's belief about the norm. At the same time, punishment conditional on $a_i = 0$ would increase $N_j^E(b_i, a_i)$ due to the prescribed out-of-equilibrium beliefs, and hence will be suboptimal (as a punishing free-rider is believed to have obtained an imprecise signal of 0, while a non-punishing free-rider is believed to have obtained either precise or imprecise signal of 0). Thus, we have shown that no player $i$ with signal $s_i = 1$ has incentives to punish another player conditional on deviating to $a_i = 0$ in Stage 1.

Finally, consider the punishing incentives of player $i$ observing $s_i = 0$ and deviating from the equilibrium play to $a_i = 1$ in Stage 1. Recall that no player has incentives to punish another player who chose the same action in Stage 1. Hence, it remains to show that $i$ does not have incentives to punish $j$ if $a_j = 0$. The corresponding incentive constraint is (analogously to (13)):

$$E_i\big[N_j^E(a_i = 1, b_i = 1)|p_i, s_i = 0, a_j = 0\big] - E_i\big[N_j^E(a_i = 1, b_i = 0,)|p_i, s_i = 0, a_j = 0\big] \\ \leq c/\theta. \tag{33}$$

For the left-hand side, by (18) and (19) we have

$$\begin{aligned} &E_i\left[N_j^E(a_i=1,b_i=1)|p_i,s_i,a_j\right]-E_i\left[N_j^E(a_i=1,b_i=0)|p_i,s_i,a_j\right] \\ &= Pr\left[p_j=p^H\middle|p_i,s_i,a_j\right]\phi+N_j^E\left(a_i=1,b_i=1,s_j,p_j=p^L\right) \\ &\qquad -N_j^E\left(a_i=1,b_i=0,s_j,p_j=p^L\right), \end{aligned} \tag{34}$$

where

$$\begin{aligned} \phi &= N_j^E(a_i=1,b_i=1,s_j,p_j=p^H)-N_j^E(a_i=1,b_i=1,s_j,p_j=p^L) \\ &-\left(N_j^E\left(a_i=1,b_i=0,s_j,p_j=p^H\right)-N_j^E\left(a_i=1,b_i=0,s_j,p_j=p^L\right)\right). \end{aligned} \tag{35}$$

The only term on the right-hand side of (35) that depends on $i$'s signal is $Pr\left[p_j=p^H\middle|p_i,s_i,a_j\right]$. By (7) and the fact that $a_j=0$ if and only if $s_j=0$ by the prescribed equilibrium strategies, we have

$$\begin{aligned} &\Pr\left[p_j=p^H\middle|p_i,s_i,a_j=0\right]=\Pr\left[p_j=p^H\middle|p_i,s_i,s_j=0\right] \\ &=\Pr[N=1|p_i,s_i,s_j=0](Pr_i[p_j=p^H|N=1,s_j=0]-\Pr[p_j=p^H|N=0,s_j=0]) \\ &+\Pr\left[p_j=p^H\middle|N=0,s_j=0\right]. \end{aligned} \tag{36}$$

The term is brackets on the right-hand side is negative by (11), while $\Pr[N=1|p_i,s_i,s_j]$ is clearly increasing in $i$'s signal. Then, (37) implies that $\Pr\left[p_j=p^H\middle|p_i,s_i,a_j=0\right]$ is decreasing in $i$'s signal. Consequently, the left-hand side, and hence the right-hand side of (35) are increasing in $i$'s signal given also that $\phi\le 0$ by (26) and assumption of $q\le 0.5$. Finally, this implies that if the incentive constraint for non-punishment (34) is satisfied for a given signal $s_i$, it will also be satisfied for any lower signal. In particular, since (34) is satisfied for the type observing an imprecise signal of 1 (since $c/\theta\in[\tau_1,\tau_2]$ by assumption, where $\tau_1$ is given by by (15)), it will also be satisfied conditional on both precise and imprecise signals of 0.

Thus, player $i$ observing any $s_i=0$ would not have incentives to punish another player conditional on deviating to $a_i=1$ in Stage 1.

**Claim 4.** *Assume* $q\ge 0.5$, $c/\theta\in[\widehat{\tau_1},\widehat{\tau_2}]$ *and in equilibrium:*

- *Player* $i$ *invests in Stage 1 if and only if* $s_i=1$.
- *In Stage 2, player* $i$ *punishes player* $j$ *by an arbitrary amount* $f\in(0,\bar{f}]$ *if and only if* $a_i=0$, $a_j=1$, *while* $i$ *holds a precise signal* $s_i=0$.

- *Investors never punish. Out-of-equilibrium belief of player $i$ after being punished by an investor $j$ is that $p_j = p^L$.*

*Then, no player with a signal $s_i = 1$ has incentives to punish the other player in Stage 2 after deviating to $a_i = 0$ in Stage 1. Analogously, no player with a signal $s_i = 0$ has incentives to punish the other player in Stage 2 after deviating to $a_i = 1$ in Stage 1.*

*Proof*: The proof proceeds analogously to Claim 3 and is omitted.

**Claim 5.** *Assume $c/\theta \in [\tau_1, \tau_2]$, $\bar{f}$ is sufficiently small and $q \leq 0.5$ is sufficiently close to 0.5. Then, there exists an equilibrium such that:*

*(i)* *In Stage 1, player $i$ invests if and only if $s_i = 1$.*

*(ii)* *In Stage 2, player $i$ punishes player $j$ by an arbitrary amount $f \in (0, \bar{f}]$ if and only if $a_i = 1$, $a_j = 0$, while $i$ holds a precise signal $s_i = 1$.*

*(iii)* *Free-riders never punish. Out-of-equilibrium belief of player $i$ after being punished by a free-rider $j$ is that $p_j = p^L$.*

*Proof*: Let us show that no player type has an incentive to deviate from the equilibrium strategies given the assumed parameter restrictions.

*Step 1.* Consider Stage 2. The fact that no investor type has an inventive to deviate in Stage 2 (given the prescribed investment strategies in Stage 1) follows by Claim 1. In turn, a free-rider $i$ does not have an incentive to punish investor $j$ in Stage 2. Indeed, given the prescribed out-of-equilibrium beliefs, the free-rider is then believed to have obtained an imprecise signal of 0. In this case, an investor punished by a free-rider would update her beliefs about the norm strength upwards (relative to the case of observing a non-punishing free-rider). This would reduce the term $E_j[D_j - D_i]$ in the free-rider's utility function, thus making punishment suboptimal for the latter. Finally, no free-rider has incentives to incur monetary costs to punish another free-rider since then the term $E_j[D_j - D_i]$ is always 0.

*Step 2*. Consider the incentives to deviate in Stage 1. Let us show that there exists a non-empty set of $\theta$ such that no player has incentives to deviate in Stage 1, given the prescribed equilibrium strategies and beliefs, *conditional on deviating to $b_i = 0$ in Stage 2* (if $b_i = 0$ is prescribed by the equilibrium strategy of $i$, then she is assumed to just keep playing this strategy).

Note that $D_i(a_i = 1) = 0$ for any player $i$ so that, from the ex ante perspective (before taking action in Stage 1), the expected psychological utility from playing $a_i = 1$ is

$$\begin{aligned} &E_i\big[E_j\big[D_j\big(a_j\big) - D_i(a_i)\big|a_i = 1, b_i\big]|p_i, s_i\big] \\ &\qquad = E_i\big[N_j^E(a_i = 1, b_i)\big|p_i, s_i, a_j = 0\big]\Pr\big[a_j = 0\big|p_i, s_i\big] \\ &\qquad = E_i\big[N_j^E(a_i = 1, b_i)\big|p_i, s_i, s_j = 0\big]\Pr\big[s_j = 0\big|p_i, s_i\big], \end{aligned} \tag{37}$$

where the last equality is by equilibrium beliefs regarding $j$'s strategy in Stage 1. Analogously, the expected psychological utility from playing $a_i = 0$ is

$$\begin{aligned} &E_i\big[E_j\big[D_j\big(a_j\big) - D_i(a_i)\big|a_i = 0, b_i\big]|p_i, s_i\big] \\ &= -E_i[N_j^E(a_i = 0, b_i)|p_i, s_i, s_j = 1]\Pr\,[s_j = 1|p_i, s_i]. \end{aligned} \tag{38}$$

This together with (3) and the assumed equilibrium beliefs yields the following ex ante expected difference in utilities between free-riding and investment (given that $b_i = 0$ by initial assumption):

$$\begin{aligned} &E_i[u_i(a_i = 0, b_i = 0)|p_i, s_i] - E_i[u_i(a_i = 1, b_i = 0)|p_i, s_i] \\ &= \Delta_i - \theta\big(E_i\big[N_j^E(a_i = 0, b_i = 0)\big|p_i, s_i, s_j = 1\big]\Pr\big[s_j = 1\big|p_i, s_i\big] + \\ &E_i\big[N_j^E(a_i = 1, b_i = 0)\big|p_i, s_i, s_j = 0\big]\Pr\,[s_j = 0|p_i, s_i]\big), \end{aligned} \tag{39}$$

where $\Delta_i \equiv E_i\big[\pi_i\big(a_i = 0, a_j, b_i, b_j\big)|p_i, s_i\big] - E_i\big[\pi_i\big(a_i = 1, a_j, b_i, b_j\big)|p_i, s_i\big]$ denotes $i$'s expected monetary payoff difference between free-riding and investment. Since $i$ is initially assumed to deviate to $b_i = 0$ in Stage 2, by (1) this expected monetary difference is equal to $1 - \gamma - E\big[b_j|a_i = 0\big]f$. Then, by (40), player $i$ finds it optimal to invest in Stage 1 if and only if

$$1 - \gamma - E\big[b_j|a_i = 0\big]f \leq \theta\mu, \tag{40}$$

where

$$\begin{aligned} \mu \equiv\; &E_i\big[N_j^E(a_i = 0, b_i = 0)\big|p_i, s_i, s_j = 1\big]\Pr\big[s_j = 1\big|p_i, s_i\big] \\ &\qquad + E_i\big[N_j^E(a_i = 1, b_i = 0)\big|p_i, s_i, s_j = 0\big]\Pr\,[s_j = 0|p_i, s_i] \\ &= \lambda\Pr\big[s_j = 1\big|p_i, s_i\big] + E_i\big[N_j^E(a_i = 1, b_i = 0)\big|p_i, s_i, s_j = 0\big], \end{aligned} \tag{41}$$

where

$$\lambda \equiv E_i\left[N_j^E(a_i = 0, b_i = 0)\middle|p_i, s_i, s_j = 1\right] - E_i\left[N_j^E(a_i = 1, b_i = 0)\middle|p_i, s_i, s_j = 0\right]. \tag{42}$$

In the next step, we show that $\mu$ is increasing in $i'$s signal if $q$ is sufficiently close to 0.5.

*Step 3*. Let us first show that $\lambda$ is positive for $q$ sufficiently close to 0.5. Denote by $\hat{p}_i$ the probability of observing $i$'s signal conditional on $N = 1$ so that $\hat{p}_i = p_i$ if $s_i = 1$ and $\hat{p}_i = 1 - p_i$ if $s_i = 0$. Define function

$$\eta(a_i, s_j, \hat{p}_i) = E_i\left[\widehat{N_j^E}(a_i)\middle|\hat{p}_i, s_j\right],$$

where $\widehat{N_j^E}(a_i)$ is $j$'s expected norm after observing $a_i$ but before observing $b_i$, i.e., without knowing the precision of $i$'s signal. By the law of total probability,

$$\begin{aligned}
\eta\left(a_i, s_j, \hat{p}_i\right) &= E_i\left[\widehat{N_j^E}(a_i)\middle|\hat{p}_i, s_j\right] \\
&= N_j^E\left(a_i, s_j, p_j = p^H\right) Pr_i\left[p_j = p^H\middle|\hat{p}_i, s_j\right] \\
&\quad + N_j^E(a_i, s_j, p_j = p^L)(1 - Pr_i[p_j = p^H|\hat{p}_i, s_j]) \\
&= Pr_i\left[p_j = p^H\middle|\hat{p}_i, s_j\right]\left(N_j^E\left(a_i, s_j, p_j = p^H\right) - N_j^E\left(a_i, s_j, p_j = p^L\right)\right) \\
&\quad + N_j^E(a_i, s_j, p_j = p^L)
\end{aligned} \tag{43}$$

In turn,

$$\begin{aligned}
&Pr_i\left[p_j = p^H\middle|\hat{p}_i, s_j\right] \\
&= Pr_i\left[p_j = p^H\middle|N = 1, s_j\right] Pr_i\left[N = 1\middle|\hat{p}_i, s_j\right] + Pr_i[p_j = p^H|N = 0, s_j]\, Pr_i[N = 0|\hat{p}_i, s_j] \\
&= Pr_i[N = 1|\hat{p}_i, s_j](Pr_i[p_j = p^H|N = 1, s_j] - Pr_i[p_j = p^H|N = 0, s_j]) \\
&\quad + Pr_i\left[p_j = p^H\middle|N = 0, s_j\right],
\end{aligned} \tag{44}$$

where by Bayes rule

$$Pr_i\left[N = 1\middle|\hat{p}_i, s_j\right] = \frac{\hat{p}_i \Pr\,[s_j|N = 1]q}{\hat{p}_i \Pr\left[s_j\middle|N = 1\right] q + (1 - \hat{p}_i)\Pr\,[s_j|N = 0](1 - q)}. \tag{45}$$

Taking the second derivative of the right-hand side and simplifying, we obtain

$$\frac{\partial^2 Pr_i[N=1|\hat{p}_i,s_j]}{\partial \hat{p}_i^2} = \frac{2\Pr[s_j|N=1]\Pr[s_j|N=0](1-q)q(\Pr[s_j|N=0](1-q)-\Pr[s_j|N=1]q)}{\left(\hat{p}_i\Pr[s_j|N=1]q+(1-\hat{p}_i)\Pr[s_j|N=0](1-q)\right)^3}. \tag{46}$$

Since $q$ is sufficiently close to 0.5 by assumption, the sign of the right-hand side coincides with the sign of $\Pr[s_j|N=0]-\Pr[s_j|N=1]$. Consequently,

$$\frac{\partial^2 Pr_i[N=1|\hat{p}_i,s_j]}{\partial \hat{p}_i^2} < (>)0 \; if \; and \; only \; if \; s_j = 1(0) \tag{47}$$

Given that the term in brackets on the right-hand side of (45) is positive if and only if $s_j = 1$ (see (11)), (45) together with (48) imply

$$\frac{\partial^2 Pr_i[p_j=p^H|\hat{p}_i,s_j]}{\partial \hat{p}_i^2} < 0. \tag{48}$$

Finally, given that the term in brackets on the right-hand side of (44) is positive if and only if $s_j = 1$, (44) together with (49) imply

$$\frac{\partial^2 \eta(a_i,s_j,\hat{p}_i)}{\partial \hat{p}_i^2} < (>)0 \; if \; and \; only \; if \; s_j = 1(0). \tag{49}$$

Next, one can verify that for $q = 0.5$ it holds[18]

$$\eta(a_i=1,s_j=0,\hat{p}_i=0)_{q=0.5} = \eta(a_i=0,s_j=1,\hat{p}_i=0)_{q=0.5}, \tag{50}$$

$$\eta(a_i=1,s_j=0,\hat{p}_i=1)_{q=0.5} = \eta(a_i=0,s_j=1,\hat{p}_i=1)_{q=0.5}. \tag{51}$$

Define function

$$\xi(\hat{p}_i) = \eta(a_i=0,s_j=1,\hat{p}_i) - \eta(a_i=1,s_j=0,\hat{p}_i) \tag{52}$$

[18] Full derivations are available upon request.

This function is continuous, and concave by (50). Moreover, by (51) and (52) it crosses 0 at two points, namely at $\hat{p}_i = 0$ and $\hat{p}_i = 1$. Consequently, $\xi(\hat{p}_i) > 0$ between these points, or equivalently,

$$\eta\big(a_i = 0, s_j = 1, \hat{p}_i\big)_{q=0.5} > \eta\big(a_i = 1, s_j = 0, \hat{p}_i\big)_{q=0.5}\, for\ any\ \hat{p}_i \in (0{,}1). \tag{53}$$

Next, note that

$$E_i\big[N_j^E(a_i = 1, b_i = 0)\big|\hat{p}_i, s_j = 0\big] < E_i\big[\widehat{N_j^E}(a_i = 1)\big|\hat{p}_i, s_j = 0\big]. \tag{54}$$

Indeed, according to assumed equilibrium beliefs, no punishment by investor $i$ signals an imprecise investor type to $j$. Hence, after observing $a_i = 1$, $j$ updates beliefs about the norm upwards to a lesser extent if she observes $b_i = 0$ ($j$'s belief on the left-hand side of (55)) relative to the case where $b_i$ is unobserved ($j$'s belief on the right-hand side). Then, we obtain

$$\begin{gathered}\eta\big(a_i = 0, s_j = 1, \hat{p}_i\big) - \eta\big(a_i = 1, s_j = 0, \hat{p}_i\big)\\ = E_i\big[\widehat{N_j^E}(a_i = 0)\big|\hat{p}_i, s_j = 1\big] - E_i\big[\widehat{N_j^E}(a_i = 1)\big|\hat{p}_i, s_j = 0\big]\\ < E_i\big[N_j^E(a_i = 0, b_i = 0)\big|\hat{p}_i, s_j = 1\big] - E_i\big[N_j^E(a_i = 1, b_i = 0)\big|\hat{p}_i, s_j = 0\big],\end{gathered} \tag{55}$$

where the inequality is by (55). This together with (54) implies that $\lambda$ defined in (43) is strictly positive for $\hat{p}_i \in \{p_L, p_H, 1 - p_L, 1 - p_H\}$ and $q$ sufficiently close to 0.5. This implies that in the latter case, $\mu$ defined in (42) is strictly increasing in $i$'s signal, given that $\Pr\big[s_j = 1\big|p_i, s_i\big]$ and $E_i\big[N_j^E(a_i = 0, b_i = 0)\big|p_i, s_i, s_j = 1\big]$ are clearly increasing in $i$'s signal. Then, given also that $f$ can be set arbitrarily small, there exists a range of $\theta$ such that incentive constraint (41) is satisfied (i.e., $i$ prefers to invest in Stage 1) if and only if $s_i = 1$. Herewith, note that condition $c/\theta \in [\tau_1, \tau_2]$, required for incentive compatibility in Stage 2, does not restrict the absolute value of $\theta$ as far as the value of $c$ is also unrestricted.

*Step 4*. Let us take parameter values such that any given player has no incentive to deviate from the prescribed equilibrium strategy in Stage 1 given the equilibrium beliefs and playing $b_i = 0$ in Stage 2 (such parameter values exist as shown in Steps 2 and 3). Let us show that under such parameter values no player type has incentives to deviate in Stage 1 given the prescribed equilibrium strategies in both stages (and corresponding beliefs) as defined in Claim 5.

By the initial assumption,

$$E_i[u_i(a_i^*(s_i), b_i = 0)|p_i, s_i] \geq E_i[u_i(1 - a_i^*(s_i), b_i = 0)|p_i, s_i], \tag{56}$$

where $a_i^*(s_i)$ is the equilibrium strategy in Stage 1 conditional on $s_i$. By Step 1, no type has incentives to deviate from the prescribed equilibrium strategy $b_i^*(p_i, s_i)$ in Stage 2 (conditional on the prescribed equilibrium strategy in Stage 1), i.e.

$$E_i[u_i(a_i^*(s_i), b_i^*(p_i, s_i)|p_i, s_i] \geq [u_i(a_i^*(s_i), b_i = 0)|p_i, s_i]. \tag{57}$$

This together with (57) implies

$$E_i[u_i(a_i^*(s_i), b_i^*(p_i, s_i)|p_i, s_i] \geq E_i[u_i(1 - a_i^*(s_i), b_i = 0)|p_i, s_i]. \tag{58}$$

The left-hand side corresponds to $i$'s utility from the equilibrium strategy, and the right-hand side corresponds to $i$'s utility from the most profitable deviation (given that, by Claim 3, $i$ finds it optimal to play $b_i = 0$ in Stage 2 conditional to deviating to $1 - a_i^*(s_i)$ in Stage 1). It follows that no player has incentives to deviate from the prescribed equilibrium strategy also in Stage 1 (recall that the claim for Stage 2 was shown in Step 1). This completes the proof of Claim 5, and hence of Proposition 1(a). ■

**Claim 6.** *Assume $c/\theta \in [\widehat{\tau_1}, \widehat{\tau_2}]$, $\bar{f}$ is sufficiently small and $q \geq 0.5$ is sufficiently close to 0.5. Then, there exists an equilibrium such that:*

- *In Stage 1, player $i$ invests if and only if $s_i = 1$.*
- *In Stage 2, player $i$ punishes player $j$ by an arbitrary amount $f \in (0, \bar{f}]$ if and only if $a_i = 0$, $a_j = 1$, while $i$ holds a precise signal $s_i = 0$.*
- *Investors never punish. Out-of-equilibrium belief of player $i$ after being punished by an investor $j$ is that $p_j = p^L$.*

*Proof*: The proof proceeds analogously to Claim 5 (based on Claims 2 and 4) and is omitted. ■

# Appendix B: Additional experimental results

**Table B.1.** Share of positive punishment decisions depending on own and opponent's investment behavior in the baseline session with no information about the norm.

| | Facing investor | Facing free-rider |
|---|---|---|
| Investor | 0/24 | 11/16 |
| Free-rider | 2/16 | 0/4 |

*Note*: The left number in each cell is the number of decisions with a positive amount of punishment, the right number is the total number of observations for given constellation of investment decisions.

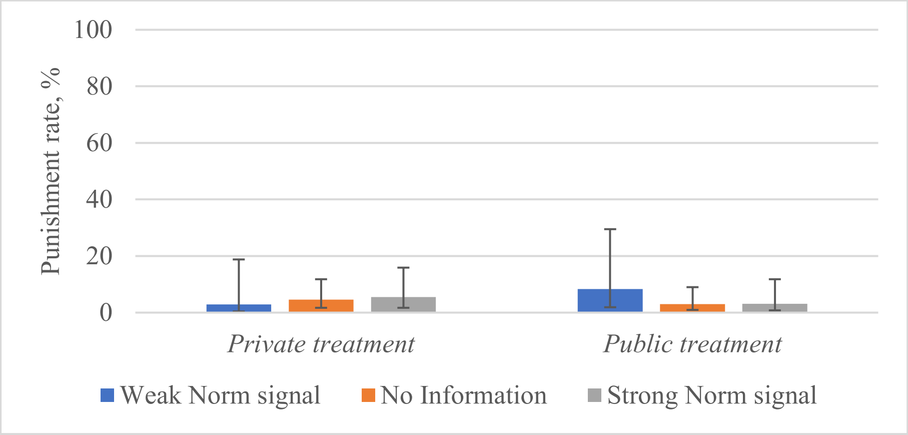

**Figure B.1.** Punishment rate of free-riders facing investors (antisocial punishment) conditional on information condition.

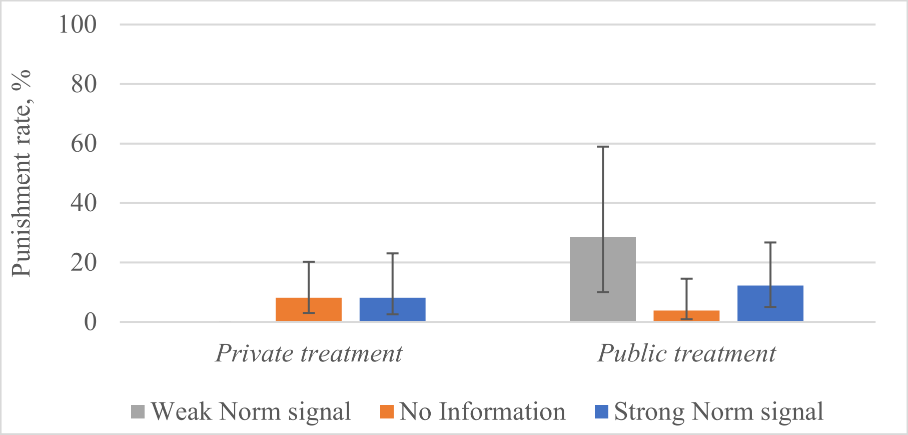

**Figure B.2.** Punishment rate of free-riders facing free-riders conditional on information condition.

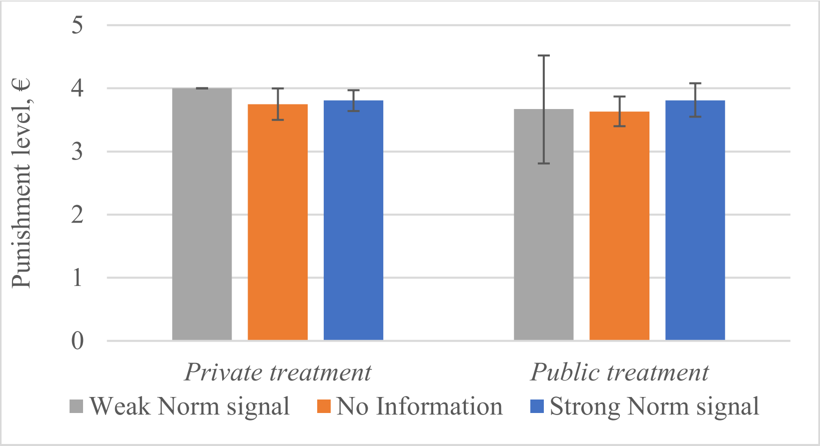

**Figure B.3.** Monetary value of punishment (€) for free-riders being punished by investors conditional on information condition.

## Appendix C: Experimental instructions

*English translation:*

### General information

Welcome to our experiment! The goal of the experiment is to study the behavior of individuals in certain situations. If you have a question, please send a message to the experimenter by clicking the button “Message to the experimenter” on your screen. **Other communication during the experiment is not permitted!**

In this experiment, you can earn money. The amount depends on your decisions and the decisions of other participants. You find further information on that below.

Your payment and your decisions are treated strictly confidentially. None of the participants find out during or after the experiment who they interacted with. Your decisions are therefore **anonymous**.

These instructions are the same for all participants.

### Information to the experiment

**Your decisions**

The experiment consists of **two** consecutive stages.

*Stage 1:* All participants are randomly matched in **groups of three**, i.e., every participant has two randomly selected partners. Each member of the group obtains **8 Euro**. You decide then whether to invest these 8 Euro in a public account or keep it for yourself. Each investment in the public account yields a **payoff of 4 Euro for each of the three group members**.

For example, if all group members have invested in the public account, then every group member obtains in total $3\times 4 = 12$ Euro from the public account. If nobody has invested in the public account, then every group member obtains in total only their own 8 Euro.

The following table shows the total payoffs in Stage 1 depending on the decisions of the group members:

| **The number of group members who invested in the public account** | The payoff in € of a group member who invested | The payoff in € of a group member who did *not* invest |
|---|---|---|
| 0 | | 8 |
| 1 | 4 | 12 |
| 2 | 8 | 16 |
| 3 | 12 | |

*Stage 2:* After all group members have made their investment decisions, you obtain information about the investment decisions of the other group members in Stage 1. Based on this, you can then decide whether you would like to **punish** other members of your group, i.e., by how much you would like to reduce the payoffs of the other group members (from 0 to 4 Euro). Herewith, you may select different amounts of punishment for the two other group members.

After all group members have made their punishment decisions, the computer randomly selects a single group member. Then, the punishment decision **of this group member** is implemented, while the punishment decisions of the other group members do not affect payoffs. The group member whose punishment decision is implemented has then to pay **a fixed fee of 1 Euro** *if* he has decided to punish at least one other group member. In case if no one was punished, no fees should be paid.

Regardless of which punishment decision is actually implemented, all group members are **informed** after Stage 2 whether and by how much the other group members **intended to punish them**.

### Information about behavior[19]

Before your investment decision, *with 50% probability* you obtain a signal about the investment behavior of other 7 randomly selected participants from a previous experiment with identical decision alternatives. The signal reveals whether the majority of these 7 participants have invested in the public account. The computer selects 7 random participants independently for each group member, so that different group members may obtain different signals.

[19] This section of the instructions was omitted in the very first session that was used to generate signals about the norm in subsequent sessions (see Section 3).

Each group member sees his own signal before making the investment decision (if he has received one), but no group member sees the signal of the other group members before making the investment decision. However, there is a 50% chance that the other two group members will find out **at the end of the experiment** based on which information you have made your investment decision. We will inform you before your punishment decision whether your information state will be revealed to the other two group members.

**Estimation questions**

During the experiment, you are asked about your expectations regarding the behavior of **all other** participants in this experimental session. In particular, you are asked to estimate the number of participants who decided to invest their endowment in the public account in Stage 1. Besides, you are asked to estimate the answer to this question of your group members. For any estimation which corresponds to the true value you receive additional 2 Euro.

**Payoffs**

Your payoff from the experiment will be paid to you via PayPal. You will receive detailed instructions for this at the end of the experiment.

This is the end of the instructions for this experiment. If you have questions, please send a message to the experimenter by clicking the button “Message to the experimenter”. If you have completely understood the instructions and do not have any more questions, please enter the code **381** into the input field on your screen and then press the button “Ready”.

*Original version (in German):*

## Allgemeine Informationen

Herzlich willkommen zu unserem Experiment! Das Ziel des Experiments ist, das Verhalten der Teilnehmer in bestimmten Situationen zu untersuchen. Falls Sie eine Frage haben, senden Sie bitte eine Nachricht an den Experimentleiter, indem Sie auf den Knopf „Nachricht an den Experimentleiter“ auf Ihrem Bildschirm klicken. **Eine andere Kommunikation ist während des Experiments nicht gestattet!**

In diesem Experiment können Sie Geld verdienen. Wie viel, hängt von Ihren Entscheidungen und den Entscheidungen anderer Teilnehmer ab. Genauere Informationen dazu finden Sie weiter unten.

Ihre Auszahlung und Ihre Entscheidungen werden streng vertraulich behandelt. Keiner der Teilnehmer erfährt während oder nach dem Experiment, mit wem er interagiert hat. Ihre Entscheidungen sind also **anonym**.

Diese Instruktionen sind für alle Teilnehmer gleich.

## Informationen zum Experiment

### Ihre Entscheidungen

Das Experiment besteht aus **zwei** aufeinander folgenden Stufen.

*Stufe 1:* Alle Teilnehmer werden zufällig **Dreiergruppen** zugeordnet, d.h. jeder Teilnehmer hat zwei zufällig ausgewählte Partner. Jedes Gruppenmitglied erhält **8 Euro**. Sie entscheiden dann, ob Sie die 8 Euro in ein öffentliches Konto investieren oder für sich selbst behalten. Jede Investition in das öffentliche Konto führt zu einer **Zahlung von 4 Euro für jeweils jeden der drei Gruppenmitglieder.**

Wenn zum Beispiel alle Gruppenmitglieder in das öffentliche Konto investieren, erhält jedes Mitglied insgesamt 3× 4 = 12 Euro vom öffentlichen Konto. Wenn niemand in das öffentliche Konto investiert, erhält jedes Mitglied insgesamt nur seine eigenen 8 Euro.

Die folgende Tabelle zeigt die Gesamtauszahlungen in Stufe 1 in Abhängigkeit von den Entscheidungen der Gruppenmitglieder:

| **Anzahl der Gruppenmitglieder, die in das öffentliche Konto investieren** | Auszahlung in € eines Gruppenmitglieds, das investiert | Auszahlung in € eines Gruppenmitglieds, das *nicht* investiert |
|---|---|---|
| 0 | | 8 |
| 1 | 4 | 12 |
| 2 | 8 | 16 |
| 3 | 12 | |

*Stufe 2:* Nachdem alle Gruppenmitglieder ihre Investitionsentscheidung getroffen haben, erhalten Sie Informationen über die Investitionsentscheidungen der anderen Gruppenmitglieder in Stufe 1. Sie können dann auf dieser Basis entscheiden, ob Sie andere Mitglieder Ihrer Gruppe **bestrafen** möchten, d.h. um wie viel Euro Sie die Auszahlung der

anderen Gruppenmitglieder reduzieren möchten (von 0 bis 4 Euro). Dabei können Sie unterschiedliche Strafen für die beiden anderen Gruppenmitglieder wählen.

Nachdem alle Gruppenmitglieder die Bestrafungsentscheidungen getroffen haben, wählt der Computer zufällig ein Gruppenmitglied aus. Dann wird die Bestrafungsentscheidung **dieses Mitglieds** umgesetzt, während die Strafentscheidungen anderer Mitglieder die Auszahlungen nicht beeinflussen. Das Mitglied, dessen Strafentscheidung umgesetzt wird, muss **eine feste Gebühr in Höhe von 1 Euro** zahlen, *falls* es sich entschieden hat, mindestens ein anderes Gruppenmitglied zu bestrafen. Falls niemand bestraft wird, muss auch keine Gebühr gezahlt werden.

Unabhängig davon, welche Bestrafungsentscheidung tatsächlich ausgeführt wird, werden alle Gruppenmitglieder nach Stufe 2 **darüber informiert**, ob und in welcher Höhe andere Gruppenmitglieder sie **bestrafen wollten**.

### Informationen über das Verhalten[20]

Sie erhalten vor Ihrer Investitionsentscheidung *mit 50% Wahrscheinlichkeit* ein Signal über das Investitionsverhalten von 7 zufällig ausgewählten Versuchspersonen aus einem früheren Experiment mit identischen Entscheidungsalternativen. Das Signal informiert darüber, ob die Mehrheit dieser 7 Teilnehmer in das öffentliche Konto investiert haben. Die 7 Versuchspersonen werden für die Gruppenmitglieder unabhängig voneinander ausgewählt, so dass die Gruppenmitglieder unterschiedliche Signale erhalten können.

Jedes Gruppenmitglied erfährt sein eigenes Signal vor seiner Investitionsentscheidung (falls es eins erhalten hat), doch kein Gruppenmitglied erfährt vor seiner Investitionsentscheidung von dem Signal anderer Gruppenmitglieder. Allerdings können die beiden anderen Gruppenmitglieder mit 50% Wahrscheinlichkeit **am Ende des Experiments** erfahren, auf Basis welcher Signal-Information Sie Ihre Investitionsentscheidung getroffen haben. Wir werden Sie vor Ihrer Bestrafungsentscheidung darüber informieren, ob die beiden anderen Gruppenmitglieder Ihren Informationsstand erfahren werden.

### Schätzfragen

Während des Experiments werden Sie nach Ihren Erwartungen bezüglich des Verhaltens **aller anderen** Teilnehmer dieser Experimentsitzung gefragt. Insbesondere werden Sie gebeten, die

[20] This section of the instructions was omitted in the very first session that was used to generate signals about the norm in subsequent sessions (see Section 3).

Anzahl der Teilnehmer zu schätzen, die sich dafür entscheiden, ihre Gelder in das öffentliche Konto in Stufe 1 zu investieren. Außerdem werden Sie gebeten, die Antwort Ihrer anderen Gruppenmitglieder auf diese Frage zu schätzen. Für jede Schätzung, die dem tatsächlichen Wert entspricht, erhalten Sie zusätzlich 2 Euro.

**Auszahlung**

Die Auszahlung bei diesem Experiment erfolgt über PayPal. Genaue Instruktionen dazu erhalten Sie am Ende des Experiments.

Dies ist das Ende der Instruktionen für das Experiment. Wenn Sie noch Fragen haben, senden Sie bitte eine Nachricht an den Experimentleiter, indem Sie auf den Knopf „Nachricht an den Experimentleiter“ klicken. Wenn Sie die Instruktionen vollständig gelesen haben und keine weiteren Fragen mehr haben, geben Sie bitte den Code **381** in das Eingabefeld auf Ihrem Bildschirm ein und drücken Sie anschließend den Knopf "Fertig".