MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

The motive matters: Experimental evidence on the expressive function of punishment

Discussion Paper no. 2024-09

Daniele Nosenzo, Erte Xiao and Nina Xue

Abstract:

The literature on punishment and prosocial behavior has presented conflicting findings. In some settings, punishment crowds out prosocial behavior and backfires; in others, however, it promotes prosociality. We examine whether the punisher's motives can help reconcile these results through a novel experiment in which the agent's outcomes are identical in two environments, but in one the pre-emptive punishment scheme is self-serving (i.e., potentially benefits the punisher), while in the other it is other-regarding (i.e., potentially benefits a third party). We find that self-serving punishment reduces the social stigma of selfish behavior, while other-regarding punishment does not. Self-serving punishment is thus less effective at encouraging compliance and is more likely to backfire. We further show that the normative message is somewhat weaker when punishment is less costly for the punisher. Our findings have implications for the design of punishment mechanisms and highlight the importance of the punisher's motives in expressing norms.

Keywords: punishment, norms, stigma, crowd out, expressive function of punishment

JEL Classification: C91, C72, D02

Daniele Nosenzo: Aarhus Univeristy, Denmark (email: Daniele.Nosenzo@econ.au.dk); Erte Xiao: Monash University, Australia (email: Erte.Xiao@monash.edu); Nina Xue: Monash University, Australia (email: Nina.Xue@monash.edu).

© The authors listed. All rights reserved. No part of this paper may be reproduced in any form, or stored in a retrieval system, without the prior written permission of the author.

The motive matters: Experimental evidence on the expressive function of punishment*

Daniele Nosenzo †, Erte Xiao ‡, Nina Xue §

29 March, 2024

Abstract

The literature on punishment and prosocial behavior has presented conflicting findings. In some settings, punishment crowds out prosocial behavior and backfires; in others, however, it promotes prosociality. We examine whether the punisher's motives can help reconcile these results through a novel experiment in which the agent's outcomes are identical in two environments, but in one the pre-emptive punishment scheme is self-serving (i.e., potentially benefits the punisher), while in the other it is other-regarding (i.e., potentially benefits a third party). We find that self-serving punishment reduces the social stigma of selfish behavior, while other-regarding punishment does not. Self-serving punishment is thus less effective at encouraging compliance and is more likely to backfire. We further show that the normative message is somewhat weaker when punishment is less costly for the punisher. Our findings have implications for the design of punishment mechanisms and highlight the importance of the punisher's motives in expressing norms.

JEL Classification: C91, C72, D02

Keywords: punishment, norms, stigma, crowd out, expressive function of punishment, experiment

*We received helpful comments from two anonymous referees, Eugen Dimant, Nick Feltovich, Xiaojian Zhao and audiences at the Norms and Behavioral Change Talk and Monash University. This work was supported by the Australian Research Council (DP16010274) and Aarhus University Research Foundation (AUFF Starting Grant 36835) and received approval from the Monash University Human Research Ethics Committee (project number 27176).

†Aarhus Univeristy, Denmark, Daniele.Nosenzo@econ.au.dk

‡Monash University, Australia, Erte.Xiao@monash.edu

§Monash University, Australia, Nina.Xue@monash.edu

1 Introduction

Evidence on the effectiveness of punishment in disciplining individual self-interest is mixed. In some settings, punishment appears to effectively restrain self-interest and promote prosocial behavior (e.g., Fehr and Gächter, 2002; Andreoni et al., 2003; Villatoro et al., 2014). However, another line of research shows that punishment can sometimes backfire and crowd out prosocial behavior (e.g., Gneezy and Rustichini, 2000; Fehr and Rockenbach, 2003; Galbiati et al., 2013).[1] The conflicting findings raise the question of why punishment crowds in prosocial behavior in some cases, but crowds out prosociality in others.

Scholars in law and economics have argued that punishment has an important "expressive function", in that it can convey information about society's norms and values (e.g., Sunstein, 1996; Posner, 1997; Kahan, 1998; McAdams, 2000; Sliwka, 2007; Bénabou and Tirole, 2011; van der Weele, 2012; Galbiati et al., 2013). This paper investigates whether the punisher's *motivation* for imposing punishment can affect the message conveyed about underlying social norms and shed light on previous findings. In particular, we compare two forms of punishment: (1) punishment that is designed to nudge an agent towards compliance for the punisher's own gain ("self-serving punishment"), and (2) punishment that encourages compliance for the benefit of a third party ("other-regarding punishment").

In our first study (Study 1), we design a novel principal-agent experimental paradigm to examine whether the same punishment mechanism (from the agent's perspective) can both crowd in and crowd out prosocial behavior, depending on whether punishment is motivated by self-interest, or by a concern for others. A key feature of our design is that our treatments hold the agent's payoffs constant, and only differ in whether punishment can be used to persuade the agent to take an action that increases the principal's payoff or the payoff of a passive third party. In our setting, the principal decides whether to impose punishment ex-ante (before the agent makes their choice). We focus on weak punishment in that the imposed penalty is no higher than the cost of compliance. We do so for two reasons. First, it allows us to focus on the expressive function of punishment through norms, rather than by changing equilibrium behavior. Second, in many real-world situations, punishment is weak due to the high costs of monitoring. To investigate social norms, we follow Bicchieri and Xiao (2009) and elicit personal norms, injunctive norms and descriptive norms about the agent's behavior in the game.

We conjecture that, in the context of our experiment, other-regarding punishment triggers a strong stigma against selfish behavior, while self-serving punishment may reduce the stigma. We construct a simple theoretical framework to show that, under this conjecture,

[1] For a review of the experimental literature on punishment, see Xiao (2018).

self-serving punishment is more likely to crowd out prosocial behavior than other-regarding punishment. Consistent with our framework, we find that self-serving punishment sends a weaker normative message about the appropriateness of compliance relative to other-regarding punishment. In fact, self-serving punishment actually reduces the social stigma of making the self-interested choice (compared to a scenario in which no punishment is used). In line with these effects on norms, self-serving punishment increases the prevalence of crowding-out, whereby agents who would behave prosocially in the absence of punishment, choose the self-interested action when the principal imposes punishment. This backfiring of punishment is significantly less likely in response to other-regarding punishment.

In a subsequent, additional study (Study 2), we design and run two additional treatments to further investigate the mechanisms underlying the findings from Study 1 and rule out potential alternative explanations. In a first treatment, we explore the extent to which the effectiveness of other-regarding punishment depends on the cost of punishment to the principal. Our findings suggest that punishment sends a somewhat weaker normative message when it is less costly for the punisher, although this has only small effects on agents' behavior. In a second treatment, we probe whether negative reciprocity can be a potential alternative explanation for the results of Study 1. The idea is that, when being threatened with punishment, the agent wishes to retaliate against the principal by reducing their payoff. This requires behavior akin to crowding-out when punishment is self-serving, but not when it is other-regarding. In our additional treatment, we find no evidence that negative reciprocity is a strong behavioral force in our setting.

Our findings have implications for how policymakers, enforcement agencies and institutions should design punishment mechanisms in order to avoid detrimental crowding-out effects. Specifically, punishment sends a stronger normative message when agents perceive it to be benefiting others, rather than simply the institution itself. Moreover, we contribute to the existing literature on how punishment affects prosocial behavior and the interplay between punishment mechanisms and social norms, which are increasingly recognized as an important driver of behavior (e.g., Bicchieri and Xiao, 2009; Krupka and Weber, 2013; Gächter et al., 2013; Kimbrough and Vostroknutov, 2016). Our findings shed light on a number of puzzling results from previous studies. For example, punishment has been shown to backfire in the trust game (e.g., Fehr and Rockenbach, 2003), but is often successful at raising contributions in public goods games (e.g., Fehr and Gächter, 2000). Although there are a number of differences between the two games, one key difference is that in trust games punishment only benefits the punisher, while in public goods games punishment can potentially benefit multiple members of the group. Thus, punishment can be perceived as "self-interested" in trust games and as more "other-regarding" in public goods games, which, as our paper

shows, has profound implications for the normative message transmitted by punishment.

Recent work has recognized the importance of the expressive role of punishment and emphasized the benefits of combining punishment with the provision of normative information (e.g., Kölle et al., 2020; Bicchieri et al., 2021).[2] Less is known, however, about which features of punishment can affect its expressive power, and how best to design punishment mechanisms to send a strong normative message. Bowles and Polania-Reyes (2012) emphasise the role of the contextual and institutional details of punishment mechanisms for their effectiveness. For example, punishment can be more effective when it is endogenously chosen by the group (Tyran and Feld, 2006), or implemented in public (Xiao and Houser, 2011). In a related study, Xiao (2013) shows that when punishment results in profits for the punisher, it is less effective in signaling to a third-party whether the punishee has lied or told the truth. Our paper differs from Xiao (2013) in that we design and study a context in which the punishee's choice is transparent, but the norm regulating his/her behavior is ambiguous. The design allows us to provide direct evidence on how the punishment motive affects beliefs about norms and the social stigma associated with certain actions. We show that whether punishment is implemented out of a concern for the punisher or for others can affect the strength of the normative message conveyed and subsequent decision making.

2 Study 1: Experimental design

In our initial study, we use a simple sequential principal-agent game with three players (Players A, B, and C). Player B ("the agent") chooses between a Communal Project (henceforth, CP) and an Exclusive Project (henceforth, EP). The CP provides the same payoff (£8) to each player. The EP offers a larger benefit (£12) to two of the three players (the agent and another player, A or C, depending on treatment – see below) while the "excluded" player, or the victim, receives a lower payoff (£6). Before the agent makes a choice, Player A ("the principal") decides whether to impose a fixed fee to reduce the payoffs of each of the two beneficiaries of the EP by £2. Player C ("the third party") has no choice to make in the game.

Our two treatments vary whether the player who is excluded from the EP is the principal or the third party. In the *Self* treatment, the principal is the excluded player and receives a higher payoff under the CP than the EP (see Figure 1). Thus, by imposing the fee, the principal can punish the agent if the agent takes an action (i.e. choosing the EP) that harms

[2]Danilov and Sliwka (2017) study the ability of positive incentives to signal norms and show that the choice of a fixed wage (over a performance-based wage) increases overall effort by changing agents' empirical expectations. See also van der Weele (2012) on this point.

the principal. In this sense, punishment is self-serving. In the *Other* treatment, the third party is the excluded player while the principal is a potential beneficiary of the EP (see Figure 2). By imposing the fee, not only does the principal punish the agent for choosing EP, but also reduces his/her own payoff. In this case, punishment cannot be self-serving and can only benefit the third party.

Note that an important feature of our design is that the two treatments are identical in all aspects (including the agent's incentives), except that in *Self*, punishment can be used to benefit the punisher, while in *Other* it can only benefit a passive third party. Moreover, punishment is weak in that the payoffs alone are not sufficient to incentivize the agent to change their behavior (the agent always earns more under the EP than the CP, regardless of whether punishment is imposed). We elaborate in the next section on the role of punishment in changing the agent's behavior by expressing the underlying norm of conduct. Thus, our treatments shed light on how the motives underlying punishment may influence the perception of social norms and hence behavior.

Figure 1: *Self* treatment

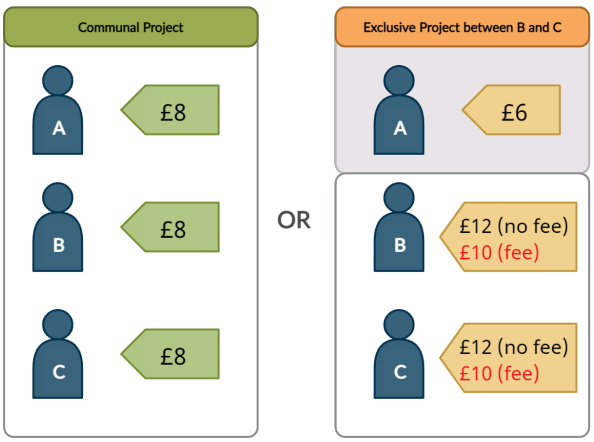
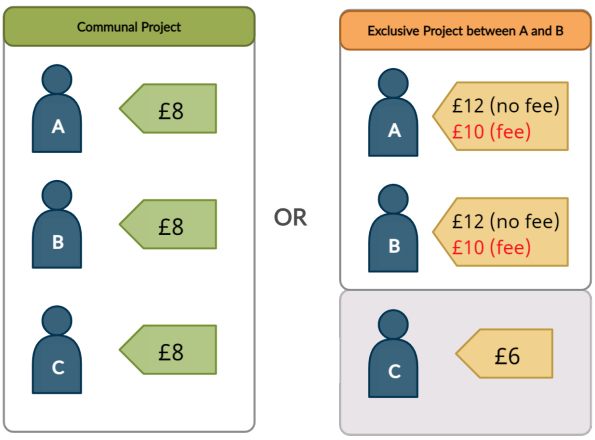


Figure 2: *Other* treatment



In each treatment, the principal was asked to make a decision about whether to use punishment or not. We elicited the agent's decisions using a strategy elicitation method, i.e., we asked the agent to make one choice in case the principal imposed a fee, and one choice in case the principal did not impose a fee.[3] Our analysis, which we pre-registered together with the experimental design on AsPredicted.org (pre-registration #64211), will focus on how the agent's strategies change based on the principal's punishment decision. In particular, depending on the agent's choices, we classify them as one of four possible types: (i) "Unconditional CP" if they choose the CP regardless of the punishment decision; (ii)

[3]We randomised the order in which we elicited these two choices to control for possible order effects.

"Unconditional EP" if they choose the EP regardless of punishment; (iii) "Crowded-in" if they choose the CP under punishment and the EP under no punishment; and (iv) "Crowded-out" if they perversely choose the EP under punishment and the CP under no punishment. Our key question is whether the motive behind punishment affects the distribution of the agent's types across the two treatments, and in particular, the share of subjects who are Crowded-out types.

The other key focus of the paper is on how punishment affects social norms across the two treatments. We elicited social norms from subjects assigned to the role of the third party, *before* we actually revealed their role to them, so that their normative beliefs would not be biased by any player-specific considerations.[4] These subjects were asked to answer a few questions about the behavior of previous participants in the task before being informed of their role.[5] After answering the questions in the first part, subjects moved to the second part, where there were told they would participate in the game (either *Self* or *Other*, depending on the treatment) they had just evaluated, in the role of Player C.

The norm-elicitation questions are based on the Bicchieri and Xiao (2009) procedure to elicit social norms.[6] We first asked participants for their first-order beliefs about the appropriateness of choosing the EP and the CP, with and without punishment (four questions in total). Subjects indicated their judgment using a 5-point scale ranging from "Very appropriate" to "Very inappropriate", and were told that by "appropriate" we meant behavior that they "personally believe is the correct or ethical thing to do". These first-order beliefs were not incentivized and can be interpreted as how participants personally felt about the appropriateness of each choice, or their *personal norms*, which may or may not align with the perceived views of the majority.[7]

Second, we elicited subjects' second-order beliefs by asking them to guess the most common first-order beliefs of participants in a previous session (the pilot experiment mentioned in

[4]There is mixed evidence regarding whether player-specific considerations affect elicited norms. Erkut et al. (2015) find little evidence that this is the case in a dictator game, but Heinicke et al. (2022) find the opposite result in a series of mini-dictator games with moral wiggle room. We did not elicit norms from the principal and agent before informing them of their role because we were worried that merely asking them to think about social norms may have altered their subsequent game behavior. This is known as the "focusing effect" of norms whereby focusing a decision-maker's attention on norms can activate norm compliance (e.g., Krupka and Weber, 2009; d'Adda et al., 2016)

[5]These previous participants were subjects recruited to take part in a pilot (N=120) that we used to conduct a power analysis to calibrate the study's sample size. The pilot was identical to the main experiment, except that the third parties were only asked unincentivized questions. We used the data from the pilot to incentivize the third parties' answers in the main experiment.

[6]See also Krupka and Weber (2013) for a related norm-elicitation procedure and Görges and Nosenzo (2020) for a review of the experimental literature on the elicitation of norms.

[7]Bašić and Verrina (2023) show that personal norms can differ from social norms (second-order beliefs) and are predictive of behavior.

footnote 5). We elicited a second-order belief in correspondence to each of the four first-order beliefs discussed above (appropriateness of choosing EP under punishment; appropriateness of choosing CP under punishment; appropriateness of choosing EP when punishment is not imposed; appropriateness of choosing CP when punishment is not imposed). Again, subjects indicated their responses on a 5-point scale ranging from "Very appropriate" to "Very inappropriate". We incentivized these responses by paying participants an additional £1 if their guess was correct for one of the four questions, randomly chosen. Since these guesses measure subjects' beliefs of what others consider appropriate or inappropriate, they express subjects' perception of the *injunctive norm* that surrounds the agent's behavior in the game.

Finally, we elicited subjects' empirical beliefs by asking them to guess the percentage of agents in a previous session (the pilot experiment) who actually chose the EP (by construction, the remainder would have chosen the CP), under punishment and under no punishment (two questions in total). These questions measure subjects' perception of the *descriptive norm* of behavior in the game. We incentivized empirical beliefs using the Karni (2009) mechanism.[8] Descriptive norms can differ from injunctive norms and can be particularly useful in explaining behavior when an injunctive norm is not followed in practice (e.g., Bicchieri and Xiao, 2009).

The experiment was programmed in oTree (Chen et al., 2016) and was conducted on Prolific in April 2021 (see Appendix A for screenshots of players' decision screens). We randomly matched three participants to form a group and randomly assigned each participant to one of the three roles in the game (A, B or C). Subjects were randomly assigned to a treatment (either *Self* or *Other*). We report data from N=883 participants with N=425 in *Self* and N=458 in *Other*.[9] The sample size was determined based on a power analysis conducted after we ran a small pilot with 60 subjects per treatment. In the pilot we observed a treatment effect on the distribution of types of size 0.33 (Cohen's d). We chose a sample of 150 subjects per role per treatment to be able to detect at least 75% of the effect size observed in the pilot (i.e., Cohen's d = 0.24), with 95% power and $\alpha = 0.05$. To improve data quality

[8]The mechanism was first introduced by Ducharme and Donnell (1973) as a variation of the Becker-DeGroot-Marschak procedure (Becker et al., 1964) and is also known as the "bets mode", "probability matching", "reservation probabilities", and "stochastic Becker-DeGroot-Marschak method" (Schotter and Trevino, 2014; Schlag et al., 2015). We chose the Karni mechanism due to its invariance to heterogeneous risk preferences. See Schwardmann and van der Weele (2019) for a similar elicitation question, presented as a multiple price list. Following Danz et al. (2020) who find that belief accuracy is higher with less information about the payment mechanism, we informed participants that their chances of receiving an additional £1 were highest when they made their "best guess" and gave the option to separately see more details about the payment mechanism if they wished (58% chose to do so).

[9]As specified in our pre-registration, we exclude from our analysis 49 participants who did not correctly answer all of the control questions (after two attempts). Our main results remain unchanged with the inclusion of these participants.

and homogeneity, we restricted participation to individuals residing in the United Kingdom, with an approval rate higher than 80% on Prolific. Participants received a completion fee of £1.50 and we selected 1 in every 20 participants to receive their earnings from the game as a bonus payment, as well as payments based on their second-order normative beliefs and empirical beliefs (if applicable). Decisions were anonymous and participants earned an average of £2.60 for a median completion time of 7.5 minutes.

3 Study 1: Theoretical framework

In this section, we sketch a simple theoretical framework to outline a possible mechanism through which the motive behind punishment may affect behavior in the game. In particular, our framework illustrates how the effectiveness of punishment in shifting behavior may depend on the strength of its expressive function.[10]

As a first theoretical benchmark, note that, if the agent only cares about maximizing material payoffs, in both versions of the game they have a dominant strategy to choose the EP, regardless of the punishment decision of the principal. Anticipating this, the principal chooses not to punish in *Other*, and is indifferent between punishing or not in *Self*.

Literature in behavioral economics has documented that agents care about more than material payoffs. We adopt a norm-based utility function framework in which utility depends on material payoffs and norm compliance: agents experience a disutility when they violate a social norm, due to the social disapproval or stigma they receive for breaking the norm (e.g., Bicchieri, 2005; Krupka and Weber, 2013). We further assume that, in the context of the game studied here, the norm prescribes that the agent chooses the CP.[11] When a player chooses the EP, they experience a disutility equal to the (positive) difference in appropriateness between choosing the CP and choosing the EP. The larger this difference, the stronger the relative stigma for choosing the EP over the CP. Crucially, below we will assume that the strength of this stigma depends on whether choosing the EP incurs punishment.

In the game, the agent chooses one action under no punishment ($a_{NoPun} \in \{CP, EP\}$), and one action under punishment ($a_{Pun} \in \{CP, EP\}$). Without punishment, the agent receives $\pi(CP) = 8$ and $\pi(EP) = 12$. The principal decides whether to impose a fee

[10]Note that our framework does not provide an explanation for why other-regarding punishment may have a stronger expressive function than self-regarding punishment, as a full-fledged analysis of the conditions under which punishment may have strong or weak expressive functions goes beyond the scope of our paper. We sketch our framework only to fix ideas about our suggested mechanism and leave the development of a full theory of the expressive function of punishment to future work. In this regard, see Sliwka (2007), van der Weele (2012) and Galbiati et al. (2013) for early models of the signalling function of punishment.

[11]Our norms data indeed confirms this since in all elicitations the appropriateness of choosing the CP is greater than the appropriateness of choosing the EP (see Appendix B).

$f \in \{0, 2\}$, which is implemented only if the agent chooses $a_{Pun} = EP$.

Let $k > 0$ represent the agent's sensitivity towards norms and $S \geq 0$ represent the relative stigma for choosing the EP instead of the CP. The agent's *net* utility for choosing the EP instead of the CP is therefore given by: $4 - f - k \cdot S$. We now analyze the agent's best-response to the principal's punishment decision, as a function of k and S.

Case 1: If the principal does not punish ($f = 0$), the agent's best-response is:

$$\begin{cases} a^*_{NoPun} = CP, & \text{if } S_{NoPun} \geq 4/k \\ a^*_{NoPun} = EP, & \text{otherwise} \end{cases} \tag{1}$$

For a given norm sensitivity parameter k, the greater the relative stigma of choosing the EP instead of the CP, the more likely it is that the agent chooses the CP. Similarly, the higher is k, the more likely it is that the agent chooses the CP, ceteris paribus.

Case 2: If the principal does punish ($f = 2$), the agent's best-response is:

$$\begin{cases} a^*_{NoPun} = CP, & \text{if } S_{Pun} \geq 2/k \\ a^*_{NoPun} = EP, & \text{otherwise} \end{cases} \tag{2}$$

As before, the agent's choice depends on the size of the relative stigma against the EP and the agent's norm sensitivity parameter. However, because the principal has imposed a fee, which makes the EP less attractive in monetary terms for the agent, the threshold values of S_{Pun} and k are lower than under the case of no punishment.

Taken together, these conditions define the threshold values of S and k that determine the agent's best-response strategy. There are four cases:

$$\{a^*_{NoPun}, a^*_{Pun}\} = \begin{cases} \{CP, CP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} \geq 2/k \\ \{EP, EP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} < 2/k \\ \{EP, CP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} \geq 2/k \\ \{CP, EP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} < 2/k \end{cases} \tag{3}$$

These four cases correspond to the four agent types that we defined in Section 2 (Unconditional CP; Unconditional EP; Crowded-in; Crowded-out). The framework clarifies that the

relative frequency of each type depends on the distribution of the norm sensitivity parameter and the relative stigma against the EP, which we assume is affected by punishment.

In particular, we conjecture that punishment that is devoid of self-serving motives may send a stronger normative message regarding what is considered appropriate behavior and therefore trigger a relatively stronger change in the stigma against the EP compared to the case without punishment. In particular, let ΔS^{Self} be the difference between S_{Pun} and S_{NoPun} in the *Self* treatment, and ΔS^{Other} be the difference in the *Other* treatment. We conjecture that ΔS^{Other} is likely to be positive since choosing the EP is likely to trigger strong stigma especially when a principal is willing to reduce his/her own payoffs to impose a fee when the agent's choice harms a third party. On the other hand, the effect may be smaller in the *Self* treatment, where the normative message of punishment may be "diluted" by the fact that the principal has a direct interest at stake in the choice of the agent. In fact, if punishment is perceived as self-servingly coercive (after all, choosing the EP maximizes joint profits and makes the agent and the third party better off), ΔS^{Self} may even be negative, i.e. punishment may reduce the stigma against the EP if choosing the EP is seen as a legitimate form of retaliation against self-serving punishment. We summarize these considerations in the following pre-registered conjecture:

Conjecture 1: Other-regarding punishment increases the stigma against choosing the EP more than self-serving punishment.

$$\Delta S^{Self} < \Delta S^{Other} \tag{4}$$

Our conjecture has direct implications for the distribution of agents' types we should observe across the two treatments. In particular, under the assumption that $\Delta S > 0$, there cannot be Crowded-out agents, because this type only emerges when the stigma against the EP, for any given k, is relatively larger under no punishment than under punishment (i.e., when $\Delta S < 0$; see (3) above and also Figure C.1 in Appendix C). Thus, if Conjecture 1 is confirmed and $\Delta S^{Other} > 0 > \Delta S^{Self}$, then we expect self-serving punishment to be more likely to backfire than other-regarding punishment. We summarize these considerations in our second pre-registered conjecture:

Conjecture 2: Punishment is more likely to backfire (i.e., induce more Crowded-out types) when it is motivated by self-interest compared to when it is motivated by other-regarding concerns.

4 Study 1: Results

The focus of this section is to study how punishment affects the normative message of punishment and its effectiveness. Overall, principals use punishment more often in *Self* (48.6%) than in *Other* (24.5%) and this difference is significant according to a χ^2 test ($p < 0.01$). In Section 4.1 we investigate how punishment affects the stigma for choosing the EP (Conjecture 1). In Section 4.2 we examine agents' choices and the effectiveness of punishment (Conjecture 2).

4.1 The normative message conveyed by punishment

We study Conjecture 1 by inspecting how punishment affects the relative stigma against the EP. Note that we have collected social norms data using three different norm-elicitation questions, pertaining to first-order beliefs of appropriateness (personal norms), second-order beliefs of social appropriateness (injunctive norms) and first-order beliefs of the frequency of agents' choices (descriptive norms). We can thus construct three distinct measures of stigma, based on personal norms, injunctive norms and descriptive norms. Table 1 reports data from these norm elicitations. The table reports both the average absolute levels of S_{Pun} and S_{NoPun} across our treatments, as well as the resulting values of ΔS.[12]

Punishment in *Self* reduces the relative stigma against the EP across all three norm measures. The drop in stigma is statistically significant for personal norms ($p < 0.01$; two-tailed Wilcoxon signed-rank test) and injunctive norms ($p < 0.01$).[13] The drop is instead insignificant for descriptive norms ($p = 0.56$). In contrast, punishment does not significantly change personal norms in *Other* ($p = 0.87$), but does increase relative stigma for the injunctive norm ($p = 0.04$), as well as for the descriptive norm ($p < 0.01$). Thus, in line with our conjecture, $\Delta S^{Self} \leq 0$, while $\Delta S^{Other} \geq 0$.

As per our pre-registration, we test Conjecture 1 by comparing ΔS^{Self} and ΔS^{Other} for each of our norm measures. We find that other-regarding punishment increases the stigma against the EP more than self-serving punishment, both when we look at personal norms (-1.02 vs. -0.06, $p < 0.01$; two-tailed Mann-Whitney test) and injunctive norms (-1.14 vs. 0.30, $p < 0.01$).[14] Moreover, subjects expect a larger increase in CP choices in response to other-regarding punishment as compared to self-serving punishment (-2.07 vs. 21.60,

[12] See Appendix B for a breakdown of the measures of stigma using appropriateness ratings based on personal and injunctive norms.

[13] Unless otherwise stated, we use two-tailed Wilcoxon signed-rank tests to compare changes in stigma due to punishment.

[14] Unless otherwise stated, we use two-tailed Mann-Whitney tests to compare the change in stigma across *Self* and *Other* for each norm measure.

Table 1: Stigma of choosing the EP

	Personal norm			Injunctive norm			Descriptive norm		
	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS
Self	1.75	0.73	**-1.02**	1.76	0.62	**-1.14**	42.45	40.38	**-2.07**
	(1.77)	(1.75)	**(2.06)**	(1.98)	(2.08)	**(2.87)**	(30.50)	(28.70)	**(47.18)**
Other	1.44	1.38	**-0.06**	1.11	1.41	**0.30**	29.66	51.26	**21.60**
	(1.76)	(1.59)	**(1.68)**	(2.00)	(1.83)	**(2.12)**	(23.97)	(25.32)	**(36.90)**

Notes: For personal and injunctive norms, in line with our theoretical framework, S is calculated as: (appropriateness of choosing CP) - (appropriateness of choosing EP). For descriptive norms, our measurement of S is simply the expected percentage of CP choices (note that this is a departure from our definition of S in the theoretical framework; adapting the framework to the empirical measure is however straightforward). ΔS is calculated as: S_{Pun} - S_{NoPun}. A positive value means punishment increases the stigma of choosing EP, while a negative value means punishment reduces the stigma. Standard deviations in parentheses.

$p < 0.01$). These findings are corroborated by the regression analysis (which also controls for demographic variables), presented in Table 2. Thus, this analysis confirms our first conjecture, as we summarize in the following result:

Table 2: How punishment changes the stigma against the EP (ΔS)

	Personal norm		Injunctive norm		Descriptive norm	
	(1)	(2)	(3)	(4)	(5)	(6)
Self	-0.96^{***}	-0.93^{***}	-1.45^{***}	-1.52^{***}	-23.67^{***}	-25.15^{***}
	(0.22)	(0.23)	(0.29)	(0.32)	(4.94)	(5.26)
Constant	-0.06	0.05	0.30	1.73	21.60^{***}	41.27^{**}
	(0.15)	(0.85)	(0.20)	(1.17)	(3.43)	(19.38)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.06	0.16	0.08	0.12	0.07	0.15
Adj. R^2	0.06	0.08	0.07	0.04	0.07	0.06
Num. obs.	292	291	292	291	292	291

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: OLS regression with standard errors in parentheses. The dependent variable is ΔS, computed using first-order beliefs of personal norms (Columns 1 and 2), second-order beliefs of injunctive norms (Columns 3 and 4) and first order beliefs of descriptive norms (Columns 5 and 6). The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation.

Result 1: Consistent with Conjecture 1, other-regarding punishment increases the relative stigma against the EP more than self-serving punishment.

4.2 The effectiveness of punishment

The previous section showed that there is a fundamental difference between self-serving and other-regarding punishment. The former reduces the stigma of selfish behavior, while the latter strengthens it. We now assess whether these differences in the expressive power of punishment translate into actual behavioral differences.

We first examine agents' choices in the two treatments, based on whether the principal chose to punish or not. In *Self*, 47.2% of agents choose the CP in the absence of punishment, and 47.9% choose the CP with punishment ($p = 1.00$, McNemar's test). In *Other*, 38.5% choose the CP under no punishment, while 55.4% do so under punishment ($p < 0.01$, McNemar's test). Table 3 corroborates these findings after controlling for demographic variables. When punishment is imposed in *Other*, it is 2.16 times ($p < 0.01$, column 4) more likely that agents will choose the CP, while in *Self*, choices are not significantly different when punishment is used ($p = 0.90$, column 2).[15] Our findings suggest that punishment is effective at changing behavior, but only when it is motivated by a concern for others.

Table 3: Effect of punishment on the choice of CP

	Self		*Other*	
	(1)	(2)	(3)	(4)
Pun	1.029	1.031	1.984***	2.155***
	(0.177)	(0.194)	(0.160)	(0.178)
Constant	0.893	0.591	0.626***	1.037
	(0.168)	(1.273)	(0.169)	(1.090)
Controls	No	Yes	No	Yes
AIC	397.00	414.31	404.73	418.76
BIC	404.30	501.88	412.11	514.71
Log Likelihood	−196.50	−183.15	−200.36	−183.38
Deviance	393.00	366.31	400.73	366.76
Num. obs.	284	284	296	296

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

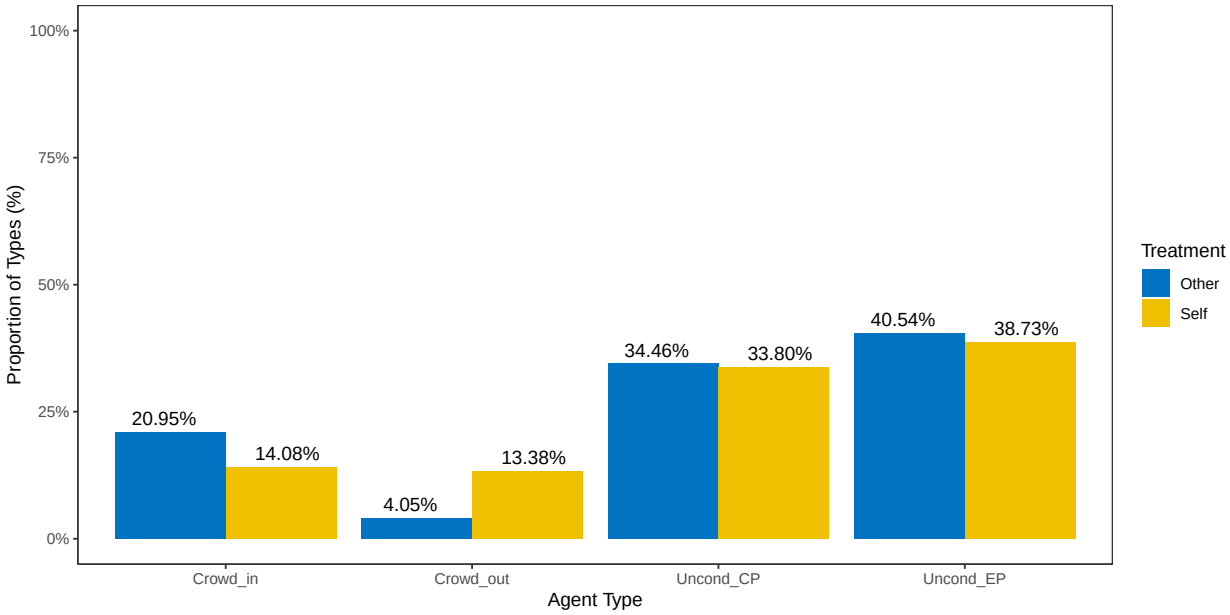
Notes: Odds ratio logistic regression with standard errors clustered at the individual level in parentheses. The dependent variable is the agent's choice (= 1 if they chose CP). The control variables are order in which the agent's choice was elicited, gender, age, education, religiosity, income and political orientation. Results are reported as factor changes in the odds ratios and hence an estimate below (above) 1 indicates a negative (positive) effect.

As per our pre-registration, we compare the distribution of agents' types between the two treatments (Figure 3) to test Conjecture 2. Across *Self* and *Other*, we find a similar

[15] We find no evidence of an order effect, see Appendix D.

share of Unconditional CP (33.8% vs. 34.5%) and Unconditional EP (38.7% vs. 40.5%) types. It is not surprising that these two types represent a majority of agents in our sample given that we examine a weak form of punishment.[16] We observe a smaller proportion of Crowded-in types (for whom punishment induced a switch from the EP under no punishment to the CP under punishment) in *Self* than in *Other* (14.1% vs. 20.9%). Conversely, we find a larger proportion of Crowded-out types (for whom punishment backfired) under self-serving punishment, compared to other-regarding punishment (13.4% vs. 4.1%). According to a χ^2 test, the distribution of types across *Self* and *Other* is significantly different ($p = 0.03$).

Figure 3: Agents' types in *Self* and *Other*



We next examine the robustness of these results after controlling for demographic variables in the (pre-registered) multinomial logistic regression analysis in Table 4, which compares the likelihood of observing each agent type against each of the other agent types under self-serving punishment, relative to other-regarding punishment. This analysis allows us to study how punishment affects the distribution of types, as well as the substitution of types across treatments. Columns 1-3 compare the likelihood of observing the Unconditional CP type against Unconditional EP, Crowded-in and Crowded-out types. In columns 4-5, we present the likelihood of the Unconditional EP type against Crowded-in and Crowded-out

[16] Another possibility is that the use of a strategy elicitation means we are more likely to observe consistency in agents' choices and might underestimate the number of Crowded-in and Crowded-out types. Our goal is not to draw conclusions about the levels of compliance or non-compliance, but rather to compare the relative effectiveness of punishment, given different underlying motivations.

types. Column 6 compares the likelihood of observing the Crowded-in type relative to the Crowded-out type. Relative to *Other*, agents in *Self* are 2.96 times more likely to be a Crowded-out type than an Unconditional EP type ($p = 0.01$, column 5). In *Self*, we are also 4.28 times more likely to observe a Crowded-out type than a Crowded-in type, relative to *Other* ($p = 0.04$, column 6). The relative shares of Unconditional CP, Unconditional EP and Crowded-in types against one another are instead unchanged across the two treatments. These results confirm Conjecture 2 and show that punishment is more likely to backfire when it is motivated by self-interest than by other-regarding motives, as we summarize in the following result.

Table 4: Odds of observing agents' types

	Uncond_CP			Uncond_EP		Crowd_in
	Uncond_EP	Crowd_in	Crowd_out	Crowd_in	Crowd_out	Crowd_out
	(1)	(2)	(3)	(4)	(5)	(6)
Self	0.956	0.662	2.830*	0.692	2.961**	4.277**
	(0.294)	(0.377)	(0.548)	(0.366)	(0.541)	(0.586)
Constant	0.910	0.406	0.128	0.447	0.140	0.315
	(1.019)	(1.632)	(2.102)	(1.610)	(2.059)	(2.431)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio multinomial logistic regression with standard errors in parentheses (N=290, AIC: 800.378). The dependent variable is agent's type based on their choices. The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation. Results are reported as factor changes in the odds ratios and hence an estimate below (above) 1 indicates a negative (positive) effect. Created using the Stargazer package (Hlavac, 2013) in R.

Result 2: Consistent with Conjecture 2, self-serving punishment is more likely to backfire and crowd out norm compliance compared to other-regarding punishment. Specifically, it increases the share of agents who react perversely to punishment (Crowded-out) compared to most other types of agents.

5 Study 2: Additional experiments

After completing the experiments described in the previous sections, we designed additional treatments to gain more insight into the mechanisms underlying our initial findings and rule

out alternative explanations. Specifically, these additional treatments serve two purposes. First, we wanted to examine whether the expressive function of punishment depends on the costs incurred by the principal for conveying the stigma against the EP. Second, we assessed the extent to which our previous results can be explained by an alternative mechanism: Reciprocity. Details of these further experiments are provided below. To preview our results, these additional treatments suggest that reciprocity is not a plausible explanation for our findings and that the normative meaning of punishment is somewhat weaker when the message is costless, although this only has small implications for behavior.

5.1 Does the cost of punishment matter for its expressive power?

In Study 1, we found that other-regarding punishment sends a stronger normative message than self-serving punishment, which leads to a higher effectiveness. Crucially, in our previous experiments, the use of other-regarding punishment required that the principal incurs a potential cost when choosing punishment (in the event that the agent chooses the EP, the principal would lose £2). That is, by choosing punishment in the *Other* treatment, the principal reveals a willingness to "self-harm" that may be essential to the credibility of the normative message that punishment conveys.[17] To what extent do our results depend on this element of self-harm?[18]

To answer this question, we designed an additional treatment, *Other-NoCost*, that only differs from our previous *Other* treatment in that the principal no longer incurs a cost of £2 when imposing punishment (see Figure 4). The principal now earns £12 in the EP regardless of the punishment decision, while the agent earns £12 in the EP if punishment is not chosen and earns a reduced £10 under punishment, as it was the case in our *Other* treatment (the payoffs are reproduced in Figure 5 for convenience). The payoffs of the third party are also the same as in *Other*. Thus, this additional treatment allows us to explore the importance of self-harm for conveying the normative message of punishment. If self-harm is necessary for punishment to convey social norms, we would expect punishment to be less effective in changing both the stigma of choosing the EP and agents' behavior in *Other-NoCost*, as compared to *Other*.

We ran this additional experiment in November 2023 on Prolific using the same procedures as in the initial experiment. Our experimental design and analysis were pre-registered on AsPredicted.org (pre-registration #146664). We recruited $N = 1022$ new subjects (who

[17] This idea echoes the mechanism proposed in Cukierman and Tommasi (1998) whereby the "credibility" of a policy reform – and hence its popular support – is increased when it is proposed by political parties that have the most to lose from the policy shift.

[18] We thank an anonymous referee for suggesting we explore this additional dimension of our initial results.

Figure 4: *Other-NoCost* treatment

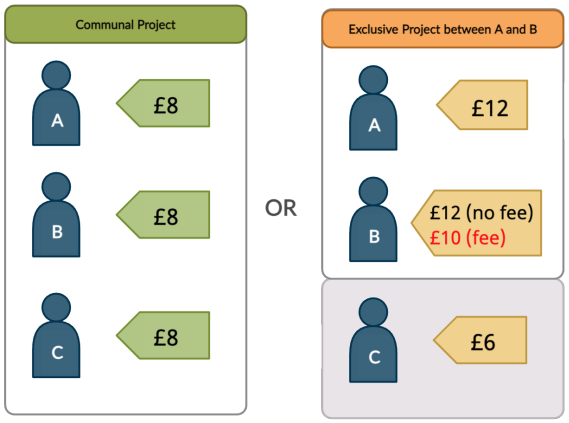
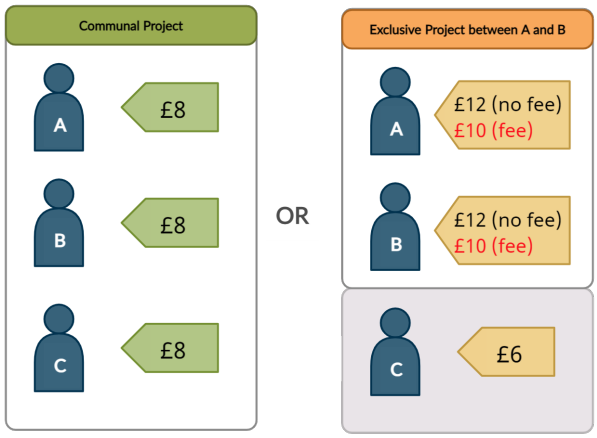


Figure 5: *Other* treatment



had not participated in our previous experiments) and randomly assigned them to the new treatment *Other-NoCost* ($N = 507$) or to the *Other* ($N = 515$) treatment, which we re-ran for comparability given that the original treatment was conducted in 2021.[19] Participants received a completion fee of £1.50 and we selected 1 in every 20 participants to receive their earnings from the game as a bonus payment. Participants earned an average of £2.90 for a median completion time of 5 minutes.

We start by examining differences in the stigma of choosing the EP across the two treatments (Table 5). First, note how our replication of the *Other* treatment produces very similar results as those already discussed in Table 1.[20] This is reassuring and lends further credibility to our initial findings. Second, Table 5 shows that there are some important differences in the stigma of choosing the EP across the two treatments. While we observe no significant difference in personal norms (-0.04 vs. 0.11, $p = 0.77$), the change in stigma is significantly greater in *Other* based on injunctive norms (0.42 vs. -0.14, $p = 0.01$) and descriptive norms (23.69 vs. 11.11, $p < 0.01$). OLS regressions which control for demographic variables corroborate these results (see Table E.1 in Appendix E). These results show that self-harm is an important component of the expressive power of punishment. When the principal incurs no cost for imposing punishment, the stigma against self-interested behavior is significantly reduced.

Next, we explore how agents' choices differ across the two treatments, by comparing

[19]The sample size was determined using a power analysis based on our initial results. We chose a sample size of 170 subjects per role per treatment to be able to detect at an effect size at least as large as the effect size from Study 1 of Cohen's d=0.179 with 80% power and $\alpha = 0.05$.

[20]The two treatments do not differ in ΔS based on personal norms ($p = 0.23$), injunctive norms ($p = 0.84$) and descriptive norms ($p = 0.69$).

Table 5: Stigma of choosing the EP

	Personal norm			Injunctive norm			Descriptive norm		
	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS
Other	1.51	1.47	**-0.04**	1.12	1.54	**0.42**	26.18	49.87	**23.69**
	(1.61)	(1.63)	**(1.73)**	(2.13)	(1.99)	**(2.09)**	(22.05)	(26.20)	**(34.49)**
Other-NoCost	1.35	1.46	**0.11**	1.45	1.31	**-0.14**	32.96	44.07	**11.11**
	(1.80)	(1.51)	**(1.82)**	(2.00)	(1.83)	**(1.98)**	(25.85)	(24.74)	**(37.73)**

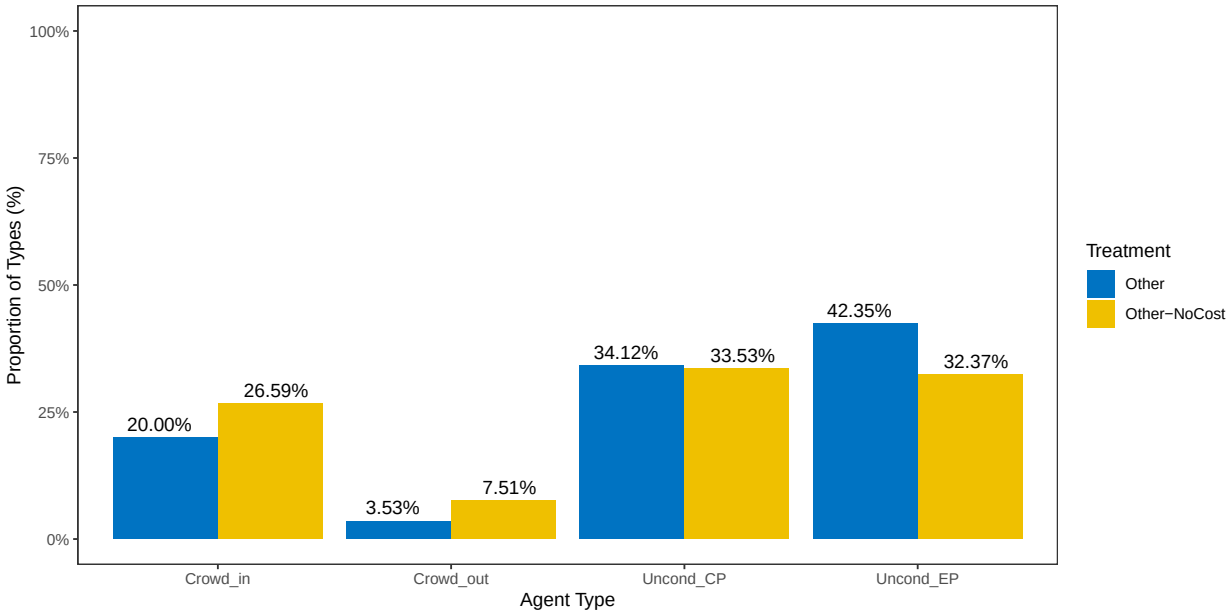
Notes: For personal and injunctive norms, in line with our theoretical framework, S is calculated as: (appropriateness of choosing CP) - (appropriateness of choosing EP). For descriptive norms, our measurement of S is simply the expected percentage of CP choices (note that this is a departure from our definition of S in the theoretical framework; adapting the framework to the empirical measure is however straightforward). ΔS is calculated as: S_{Pun} - S_{NoPun}. A positive value means punishment increases the stigma of choosing EP, while a negative value means punishment reduces the stigma. Standard deviations in parentheses.

the distribution of agents' types across *Other* and *Other-NoCost* (Figure 6). Again, note how the distribution of types in the replication of the *Other* treatment is very similar to our initial findings (see Figure 3).[21] Moreover, the distribution of types appears to be very similar across *Other* and *Other-NoCost* and we do not detect any significant difference on aggregate between the two treatments ($p = 0.10$, χ^2 test). The regression analysis (reported in Table E.2 in Appendix E), reveals some subtler shifts in the relative composition of types across treatments (agents in *Other-NoCost* are relatively less likely to be classified as Unconditional EP types than Crowded-out and Crowded-in types, compared to *Other*), although these post hoc results should be interpreted with caution given the lack of evidence of an overall difference in the distribution of types.

Overall, the results of this additional experiment reveal that the absence of costs for using other-regarding punishment reduces the strength of its normative message, but does not strongly affect the extent to which agents respond to it. One possible explanation for these mixed results comes from a closer inspection of Table 5. Although punishment in the absence of self-harm is less effective at increasing the stigma against the EP in *Other-NoCost* compared to *Other*, its impact is not completely eliminated, nor does it go in the opposite direction as observed in *Self*. Injunctive norms indicate that, without self-harm, punishment does not lower the perceived stigma against the EP (1.45 vs. 1.31, $p = 0.59$, two-tailed Wilcoxon signed-rank test). This differs from self-serving punishment, as observed in the *Self* treatment (see Table 1). Moreover, subjects expect a higher proportion of agents to choose the CP with than without punishment in *Other-NoCost*, unlike in *Self*, where

[21] According to a χ^2 test, the distributions of types are not significantly different between the original *Other* treatment and its replication ($p = 0.50$).

Figure 6: Agents' types in *Other* and *Other-NoCost*



self-serving punishment was expected to backfire. A two-tailed Wilcoxon signed-rank test indicates this effect on descriptive norms is significant (32.96 vs. 44.07, $p < 0.01$). Previous research suggests that descriptive norms play an important role in decision making and can even override injunctive norms when the two are in conflict (Bicchieri and Xiao, 2009). Thus, although the absence of self-harm somewhat reduces the strength of the normative message of punishment, it does not completely eliminate it - which may explain the small differences observed in agents' behavior across *Other* and *Other-NoCost*.

5.2 Can negative reciprocity explain the effectiveness of punishment in *Other*?

Other-regarding punishment in our *Other* (and *Other-NoCost*) treatment can be perceived as an unkind action by the principal towards the agent.[22] If the agent is motivated by (negative) reciprocity (e.g., Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006), they may respond in a similarly "unkind" manner by choosing the project that gives the principal the lowest payoff. The action that achieves this differs across our *Self* and *Other* treatments. In *Self*, the principal's payoff is lowest when the agent chooses the EP

[22]By imposing punishment in *Other*, the principal reduces the agent's payoff by £2 in the EP. Moreover, punishment may be perceived as "controlling" (e.g., Falk and Kosfeld, 2006; Kessler and Leider, 2016; Ziegelmeyer et al., 2012; Charness et al., 2012).

(£6 vs. £8) while in *Other*, the principal's payoff is lowest in the CP (£8 vs. £10 or £12). Thus, negative reciprocity also predicts that – after punishment – the agent is more likely to choose the EP in *Self* than in *Other*. This introduces a possible alternative explanation for the results presented earlier: The differences in the distribution of types in *Other* compared to *Self* may be driven by the reciprocal agents who react to punishment by choosing the EP in *Self* and the CP in *Other*. Can our results be explained by negative reciprocity? We believe this is not the case, for two reasons.[23]

First, we note that in *Self*, negative reciprocity would predict fewer CP choices under punishment than under no punishment since the agent can reduce the principal's payoff by choosing the EP to retaliate against punishment. Our data (also reported in Section 4.2), however, is not consistent with this prediction as the percentage of agents choosing the CP does not vary with the punishment decision: 47.2% choose the CP under punishment while 47.9% choose the CP under no punishment ($p = 1.00$, McNemar's test).

Second, we designed a new treatment to probe the explanatory power of negative reciprocity by directly testing whether retaliation plays a role in the choice of the CP under punishment in *Other*. The purpose of the new treatment is to shut down the ability of punishment to convey a clear normative message, while keeping the possibility of it triggering negative reciprocity. In our new treatment, *Other-R*, we keep the payoff structure the same as in *Other* for the principal and the agent but the payoff received by the passive third-party in the EP is £10 instead of £6 (as was the case in *Other*, see Figures 7 and 8). This means that the third party now earns more in the EP than in the CP. Thus, punishment in *Other-R* should not convey the message that choosing the CP is the right thing to do as it did in *Other*, since now by choosing the EP the agent maximizes the payoffs of all players.[24] On the other hand, the logic of reciprocity is unchanged across the *Other-R* and *Other* treatments, since the payoffs of the principal and agent are identical across these treatments.[25] Negative reciprocity would, therefore, predict the same response to punishment (choose the CP when there is punishment and the EP where there is no punishment) across the two treatments and hence the same proportion of Crowded-in types. Moreover, if the normative perceptions are informed by reciprocal considerations, the perceived stigma against the EP should be

[23]Moreover, if normative perceptions are based on reciprocal considerations, this mechanism can also explain why stigma against the EP is higher in *Other* than *Self*. We thank an anonymous referee to raising these points and suggesting we further explore this alternative explanation for our initial results.

[24]Punishment may not be viewed as totally meaningless, though, as it could be interpreted as a preference of the principal for equalizing payoffs in case the agent chooses the EP.

[25]The fact that the third party's payoff changes across these two treatments does not matter for individuals motivated by reciprocity, since the third party cannot affect the payoffs of the agent and hence the agent cannot perceive the third party as either kind or unkind (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006).

similar in the two treatments.

Figure 7: *Other-R* treatment

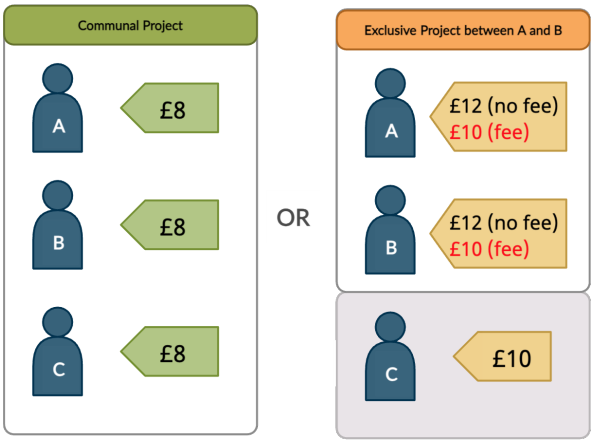
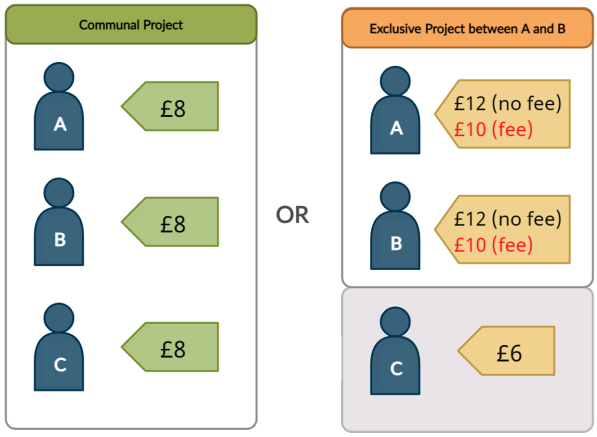


Figure 8: *Other* treatment



We ran the *Other-R* in November 2023 in the context of the new wave of experiments for Study 2 described in the previous subsection. The experiment was run on Prolific using the same procedures as in the original experiment with $N = 472$. Our experimental design and analysis were pre-registered on AsPredicted.org (pre-registration #146664).[26] Participants received a completion fee of £1.50 and we selected 1 in every 20 participants to receive their earnings from the game as a bonus payment. Participants earned an average of £2.90 for a median completion time of 5 minutes.

Figure 9 shows the distribution of agents' types across the new *Other-R* treatment and the *Other* treatment we re-ran in 2023 (we already displayed this data in Figure 6 and reproduce it here for convenience). The figure shows a strong and clear difference in the distribution of types between treatments ($p < 0.01$, χ^2 test). Our specific test of the reciprocity explanation relies on a comparison of the share of Crowded-in types in *Other* and *Other-R* (see our pre-registration and our reasoning above). According to a one-sided test of proportions, we are more likely to observe Crowded-in types in *Other* than *Other-R* (20% vs. 5.70%, $p < 0.01$). This result is corroborated by the regression analysis (Table E.3 in Appendix E). These results are in conflict with a rationalization of our data based on reciprocity: Although reciprocal considerations are held constant across our *Other* and *Other-R* treatments, behavior differs widely.

This conclusion extends to our analysis of the norms data, reported in Table 6. The

[26]The sample size of 170 subjects per role per treatment is consistent with the sample size for *Other* and *Other-NoCost* (see Section 5.1) and allows us to detect at least a 9.9 percentage point difference in behavior with 80% power and $\alpha = 0.05$.

Figure 9: Agents' types in *Other* and *Other-R*

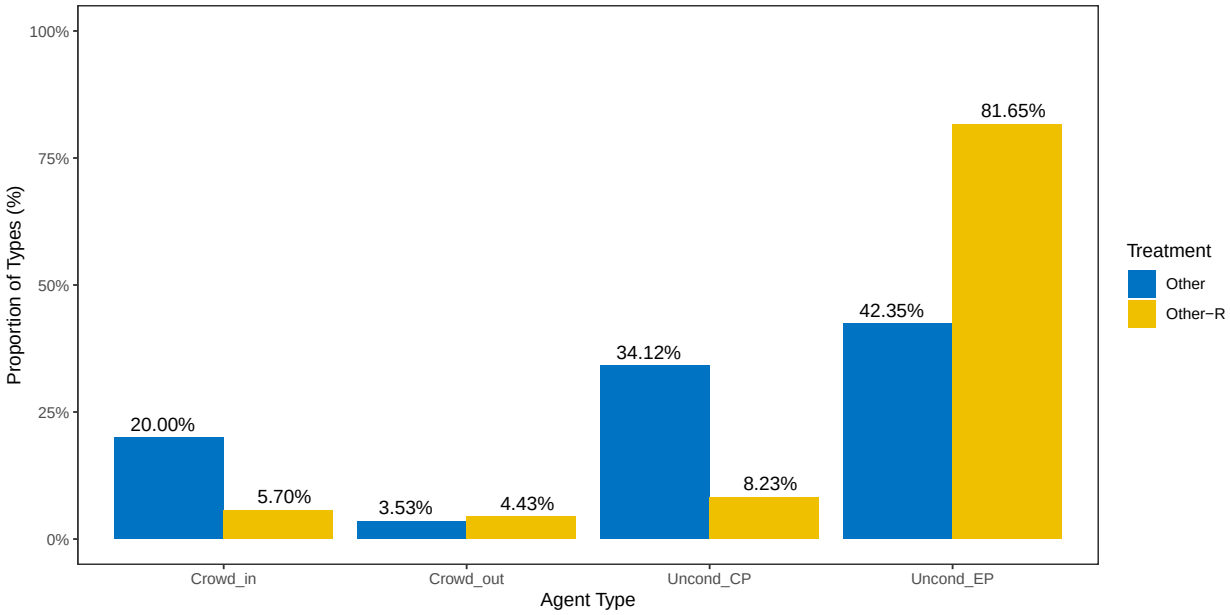


table shows personal, injunctive and descriptive norms in the new *Other-R* treatment as well as in *Other* (this data is the same as already reported in Table 5). The table shows that the effect of punishment on stigma is smaller in *Other-R*, especially for injunctive and descriptive norms – suggesting that the message conveyed by punishment differs across treatments. We detect significant differences in descriptive norms (23.69 vs. 9.57, $p < 0.01$), but not in personal (-0.04 vs. 0.05, $p = 0.89$) and injunctive norms (0.42 vs. 0.07, $p = 0.26$; this difference however becomes significant in our regression analysis reported in Table E.4 in Appendix E).

6 Conclusion

Punishment can be effective at encouraging prosocial behavior. However, the specific factors which lead to punishment crowding out or crowding in prosocial choices remain an open question. We investigate whether the perceived motive behind a punishment decision changes the normative message that is conveyed. We conjecture that punishment that is motivated by self-serving concerns is less effective at reigning in self-interest than punishment that is perceived to be motivated by other-regarding concerns.

Our key takeaways can be summarized as follows. First, by eliciting perceptions of norms (personal, injunctive and descriptive), we find that other-regarding punishment increases

Table 6: Stigma of choosing the EP

	Personal norm			Injunctive norm			Descriptive norm		
	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS	S_{NoPun}	S_{Pun}	ΔS
Other	1.51	1.47	**-0.04**	1.12	1.54	**0.42**	26.18	49.87	**23.69**
	(1.61)	(1.63)	**(1.73)**	(2.13)	(1.99)	**(2.09)**	(22.05)	(26.20)	**(34.49)**
Other-R	-0.16	-0.11	**0.05**	-0.34	-0.27	**0.07**	22.70	32.27	**9.57**
	(2.29)	(2.03)	**(1.99)**	(2.54)	(2.32)	**(2.38)**	(25.01)	(28.55)	**(35.58)**

Notes: For personal and injunctive norms, in line with our theoretical framework, S is calculated as: (appropriateness of choosing CP) - (appropriateness of choosing EP). For descriptive norms, our measurement of S is simply the expected percentage of CP choices (note that this is a departure from our definition of S in the theoretical framework; adapting the framework to the empirical measure is however straightforward). ΔS is calculated as: S_{Pun} - S_{NoPun}. A positive value means punishment increases the stigma of choosing EP, while a negative value means punishment reduces the stigma. Standard deviations in parentheses.

the social stigma against self-interested choices, while self-serving punishment can have a detrimental effect by reducing this stigma. In an additional treatment, we find evidence that the cost of punishment plays a role in changing this social stigma, with a higher cost sending a stronger message of what is appropriate and inappropriate behavior. An interesting topic for future work would be to formally model the mechanisms through which punishment can transmit social norms and examine the conditions under which it does so most effectively.

Second, consistent with these changes in social stigma and in line with a simple theoretical framework, when punishment is self-serving in nature, agents tend to respond in a perverse manner – by acting more prosocially when punishment is not used than when it is used. Punishment therefore backfires as agents respond to self-serving punishment by also pursuing their own self-interest. Conversely, punishment motivated by other-regarding concerns is effective at encouraging prosocial behavior.

Our results show that, in order for punishment mechanisms to be effective at constraining self-interest, punishment needs to communicate a strong normative message, and that the strength of this message crucially depends on the perceived motives behind punishment choices. Our findings have useful applications for the design of punishment mechanisms, and especially for mechanisms that are monetary in nature, such as fines and taxes. Our results caution that such mechanisms should be designed in a way that clearly communicates the benefits to the wider community (or a specific third party) and minimizes the chances that punishment is interpreted as a profit-making device, or used purely to benefit the enforcement agency.

This paper also sheds light on why punishment is generally effective at constraining self interest in public goods games when it can benefit multiple individuals, but tends to

backfire in trust games when it is used only to benefit the punisher. A promising avenue for future work is to examine other differences between the two punishment contexts which could affect the normative message that is conveyed by punishment, such as the number of potential benefactors of punishment and the nature of the punishment institution.

References

Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93(3):893–902.

Bašić, Z. and Verrina, E. (2023). Personal norms—and not only social norms—shape economic behavior. *MPI Collective Goods Discussion Paper*, (2020/25).

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral science*, 9(3):226–232.

Bénabou, R. and Tirole, J. (2011). Laws and norms.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge 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 and Organization*, 188:209–235.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.

Charness, G., Cobo-Reyes, R., Jiménez, N., Lacomba, J. A., and Lagos, F. (2012). The hidden advantage of delegation: Pareto improvements in a gift exchange game. *American Economic Review*, 102(5):2358–2379.

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

Cukierman, A. and Tommasi, M. (1998). When does it take a Nixon to go to China? *American Economic Review*, 88(1):180–197.

d'Adda, G., Drouvelis, M., and Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62:1–7.

Danilov, A. and Sliwka, D. (2017). Can contracts signal social norms? Experimental evidence. *Management Science*, 63(2):459–476.

Ducharme, W. M. and Donnell, M. L. (1973). Intrasubject comparison of four response modes for "subjective probability" assessment. *Organizational Behavior and Human Performance*, 10(1):108–117.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and economic behavior*, 47(2):268–298.

Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and economic behavior*, 54(2):293–315.

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

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

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

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

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences? *Journal of the European Economic Association*, 11(3):548–573.

Galbiati, R., Schlag, K. H., and van der Weele, J. J. (2013). Sanctions that signal: An experiment. *Journal of Economic Behavior and Organization*, 94:34–51.

Gneezy, U. and Rustichini, A. (2000). Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics*, 115(3):791–810.

Görges, L. and Nosenzo, D. (2020). Measuring social norms in economics: Why it is important and how it is done. *Analyse & Kritik*, 42(2):285–311.

Heinicke, F., König-Kersting, C., and Schmidt, R. (2022). Injunctive vs. descriptive social norms and reference group dependence. *Journal of Economic Behavior and Organization*, 195:199–218.

Hlavac, M. (2013). stargazer: Latex code and ascii text for well-formatted regression and summary statistics tables. *URL: http://CRAN. R-project. org/package= stargazer*.

Kahan, D. M. (1998). Social meaning and the economic analysis of crime. *The Journal of Legal Studies*, 27(S2):609–622.

Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, 77(2):603–606.

Kessler, J. B. and Leider, S. (2016). Procedural fairness and the cost of control. *The Journal of Law, Economics, and Organization*, 32(4):685–718.

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

Kölle, F., Lane, T., Nosenzo, D., and Starmer, C. (2020). Promoting voter registration: the effects of low-cost interventions on behaviour and norms. *Behavioural Public Policy*, 4(1):26–49.

Krupka, E. and Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3):307–320.

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

McAdams, R. H. (2000). A focal point theory of expressive law. *Virginia Law Review*, pages 1649–1729.

Posner, R. A. (1997). Social norms and the law: An economic approach. *The American Economic Review*, 87(2):365–369.

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

Schlag, K. H., Tremewan, J., and van der Weele, J. J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457–490.

Schotter, A. and Trevino, I. (2014). Belief Elicitation in the Laboratory. *Annual Review of Economics*, 6(1):103–128.

Schwardmann, P. and van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, 3(10):1055–1061.

Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, 97(3):999–1012.

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

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

van der Weele, J. (2012). The signaling power of sanctions in social dilemmas. *The Journal of Law, Economics, & Organization*, 28(1):103–126.

Villatoro, D., Andrighetto, G., Brandts, J., Nardin, L. G., Sabater-Mir, J., and Conte, R. (2014). The norm-signaling effects of group punishment: combining agent-based simulation and laboratory experiments. *Social Science Computer Review*, 32(3):334–353.

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

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

Xiao, E. and Houser, D. (2011). Punish in public. *Journal of Public Economics*, 95(7-8):1006–1017.

Ziegelmeyer, A., Schmelz, K., and Ploner, M. (2012). Hidden costs of control: Four repetitions and an extension. *Experimental Economics*, 15:323–340.

A Instructions

Figure A.1: The principal's choice

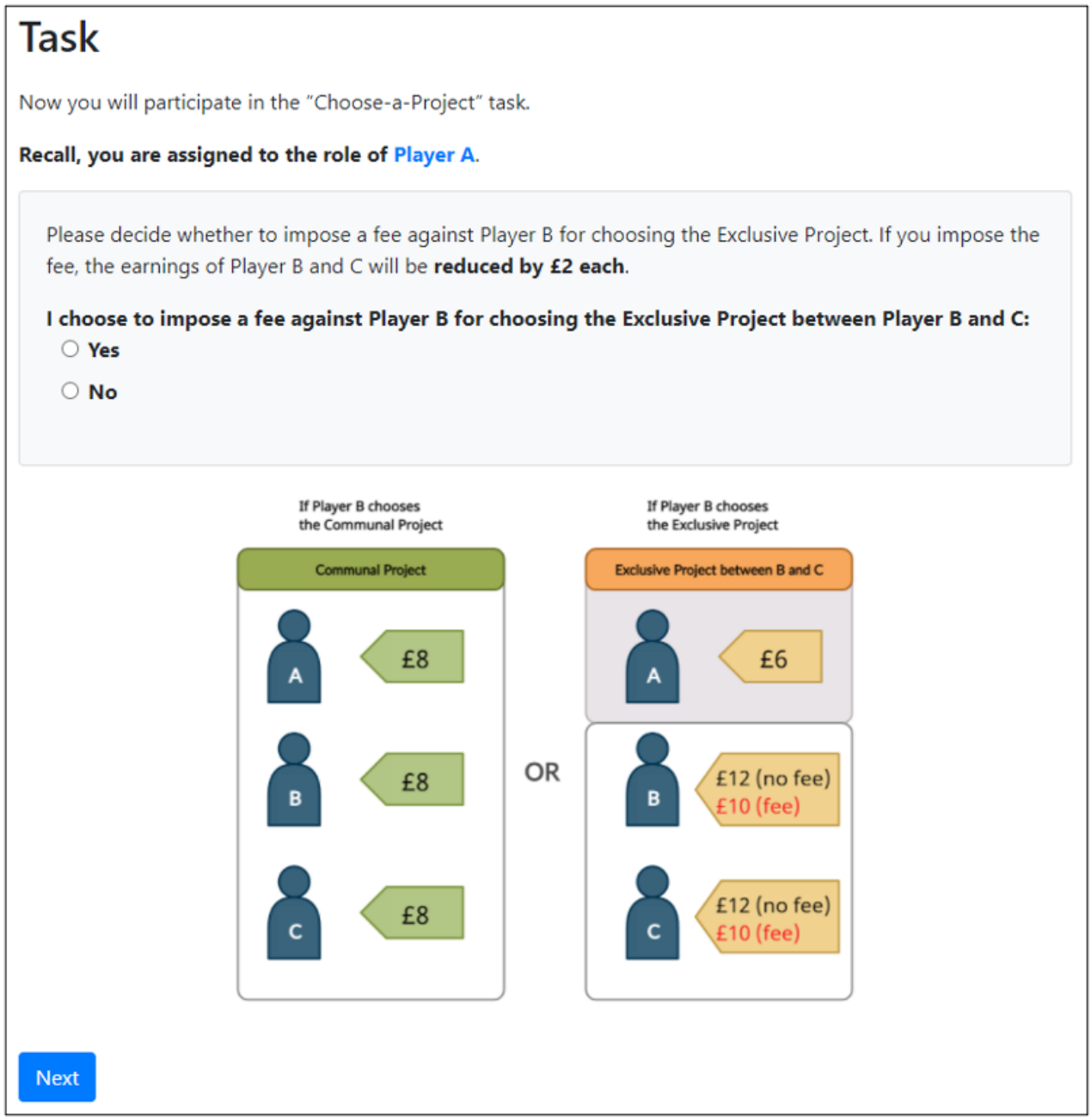


Figure A.2: The agent's choice (Order 1: Pun, NoPun)

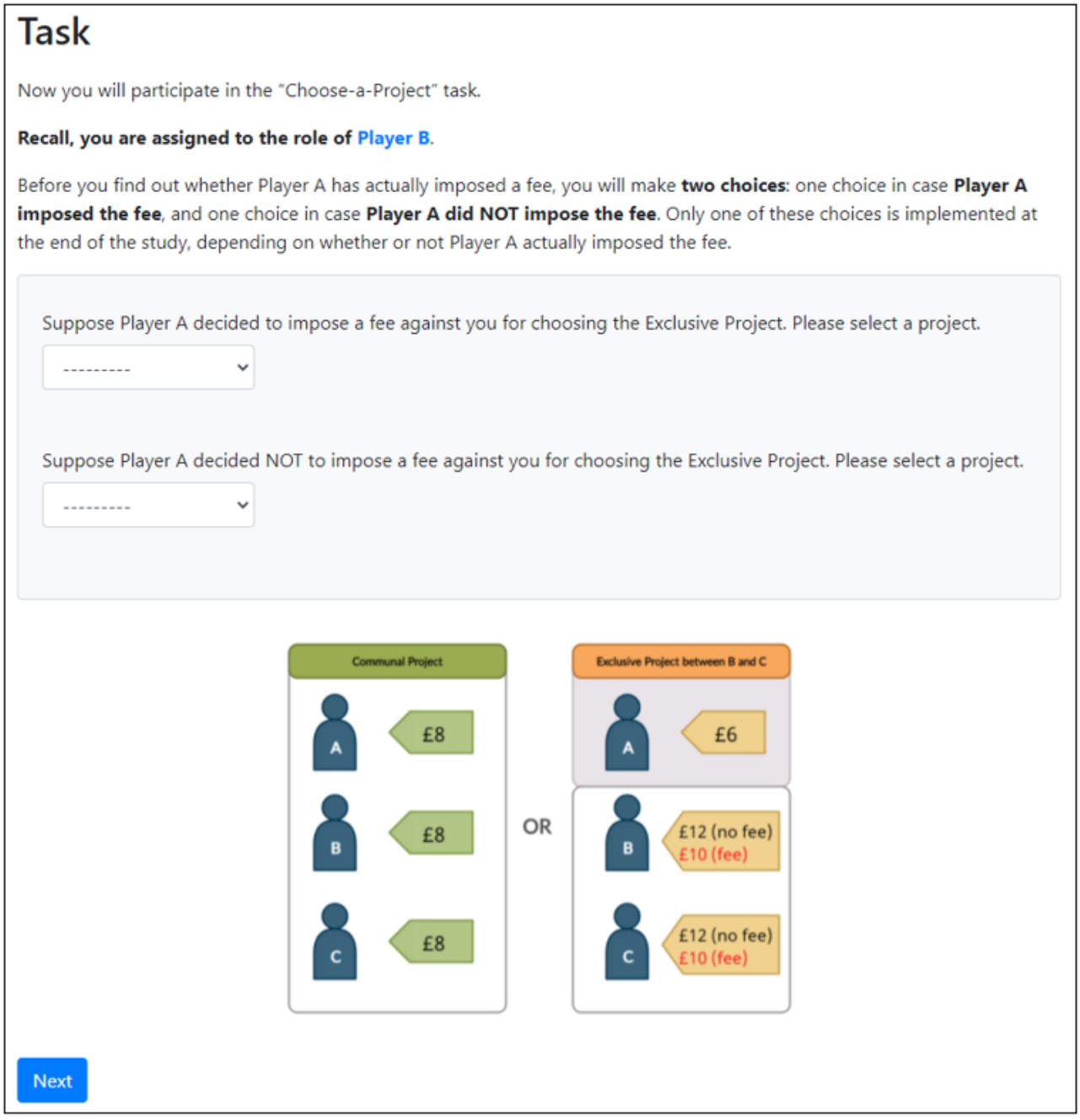


Figure A.3: The agent's choice (Order 2: NoPun, Pun)

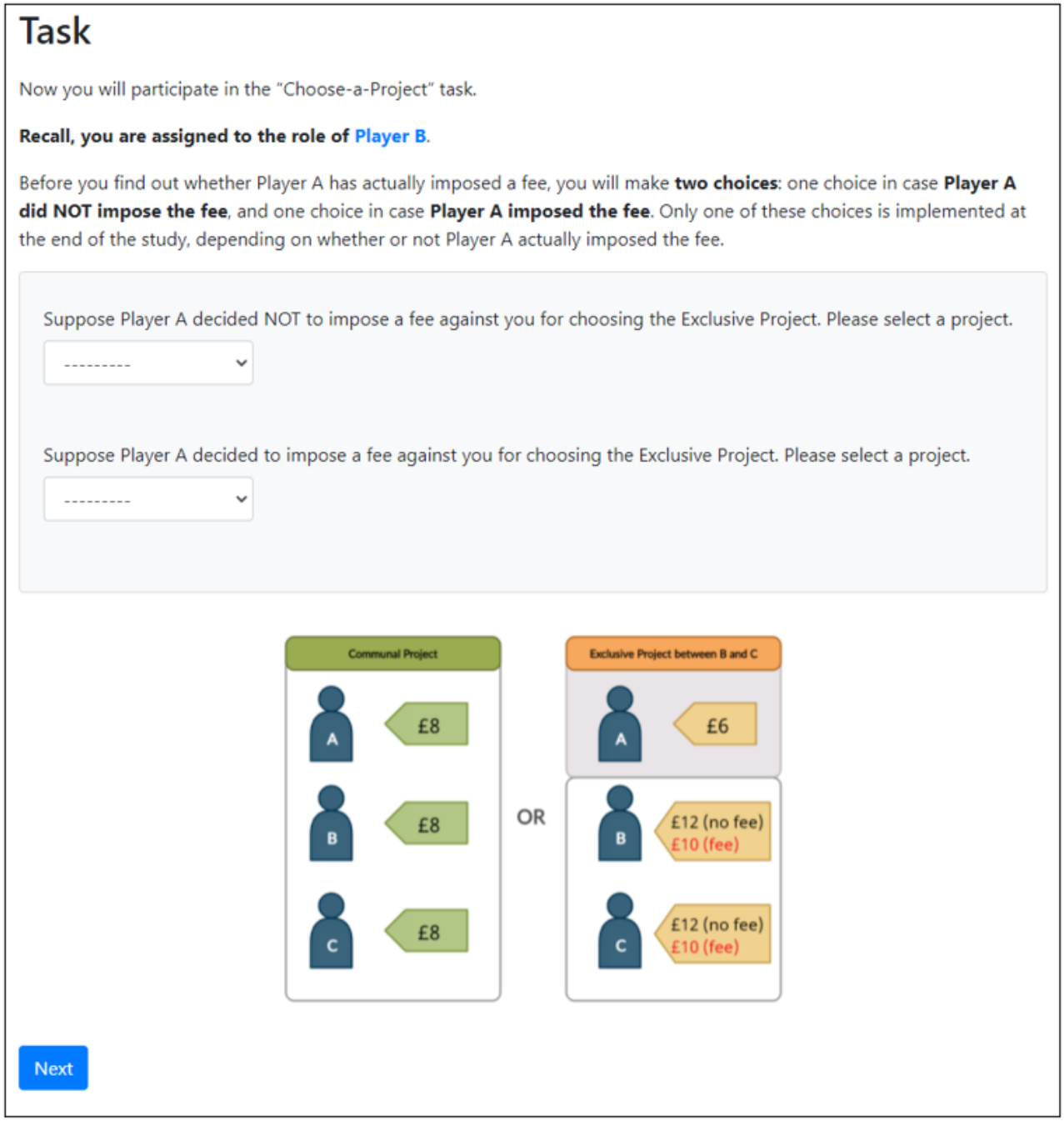


Figure A.4: Eliciting third-party personal norms (Order 1)

Questions

We would like to ask you a few questions about the "Choose-a-Project" task that was completed by a previous group of participants, recruited on Prolific. You may receive an additional payment depending on your answers to these questions.

1) For each possible action by Player B, please evaluate whether, in your opinion, the action is "appropriate" or "inappropriate". By appropriate, we mean behavior that you personally believe is the "correct" or "ethical" thing to do.

Suppose that Player A imposed a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

- Player B chooses the **Communal Project**

Suppose that Player A did NOT impose a fee against Player B for choosing the Exclusive Project and:

- Player B chooses the **Exclusive Project** between Player B and C.

- Player B chooses the **Communal Project**

Next

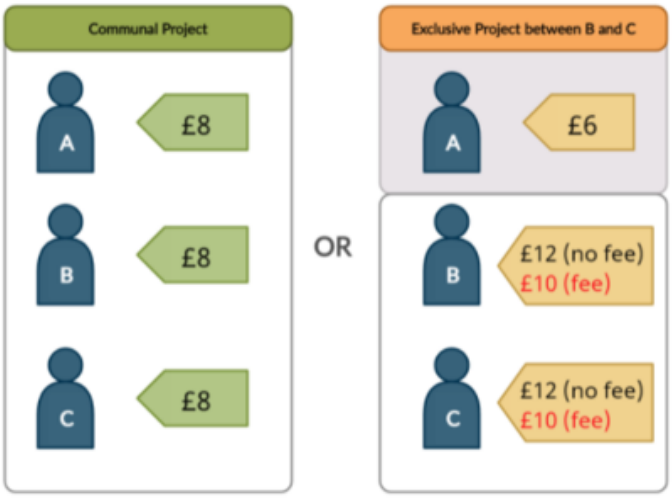


Figure A.5: Eliciting third-party beliefs about injunctive norms (Order 1)

Figure A.6: Eliciting third-party beliefs about descriptive norms (Order 1)

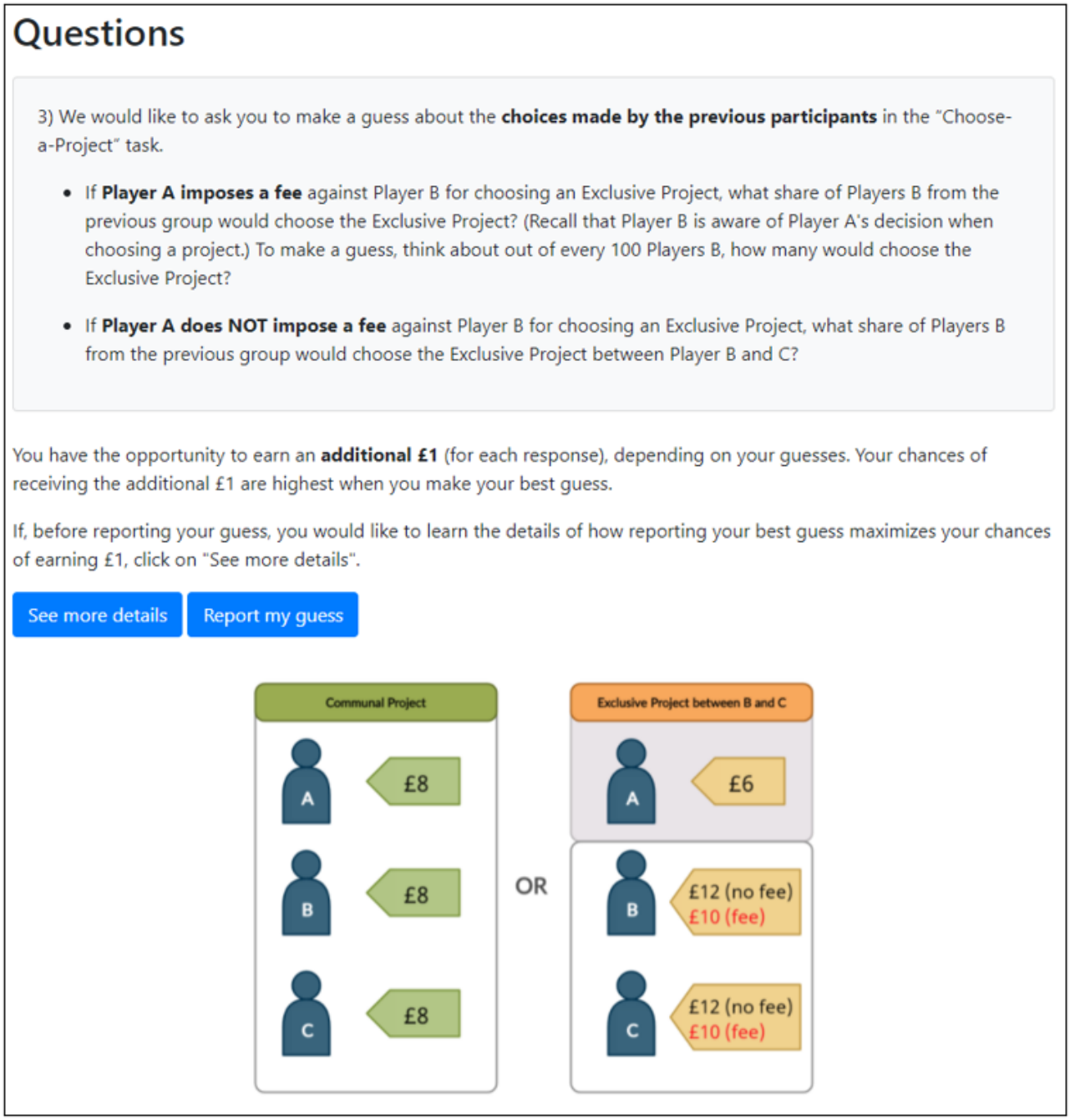


Figure A.7: Payment mechanism

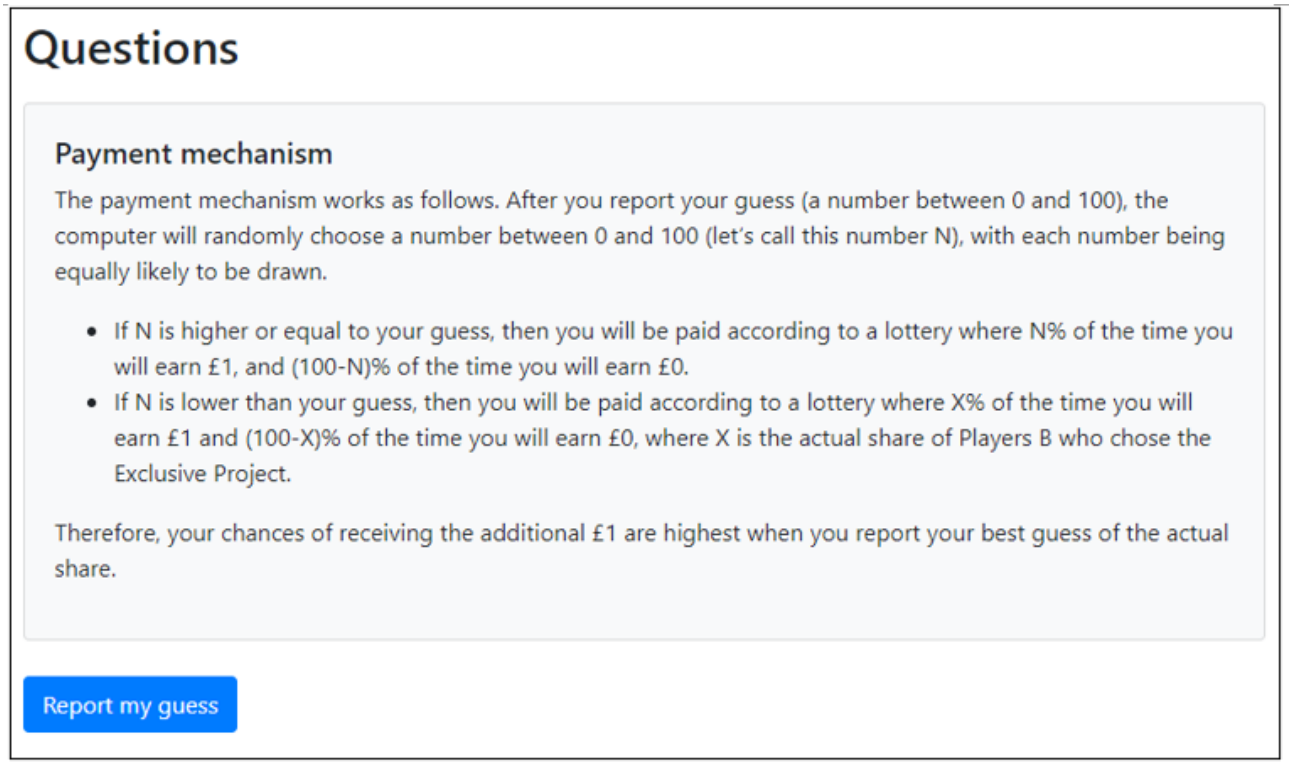
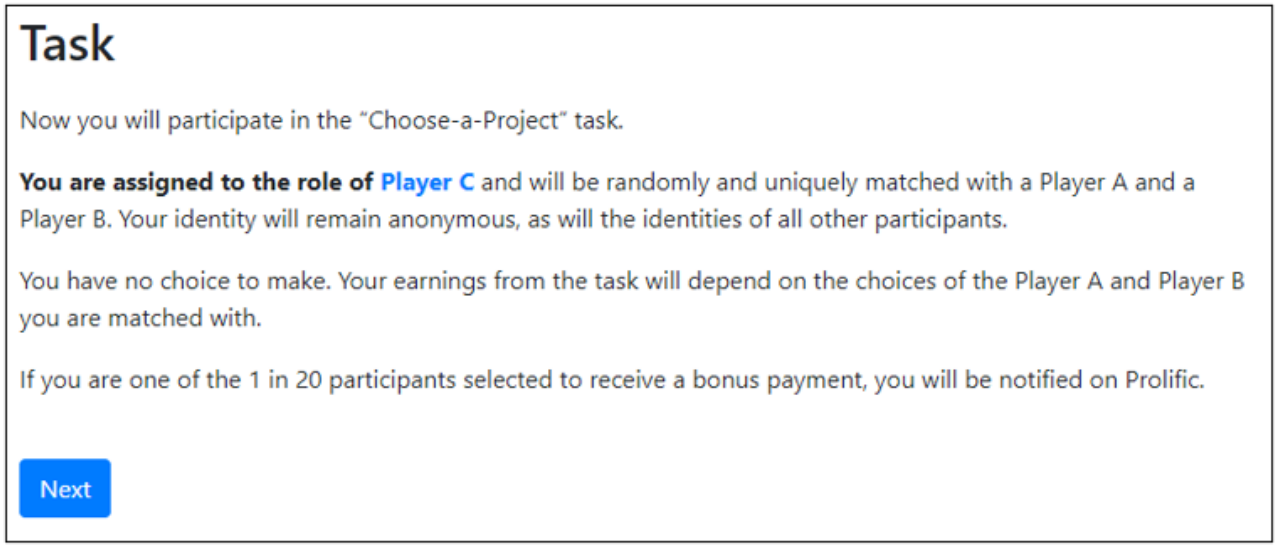


Figure A.8: The third party is informed of their role



B Normative beliefs

Table B.1 summarizes subjects' average personal norms (or first-order normative beliefs) while Table B.2 presents subjects' average injunctive norms (or second-order normative beliefs). In both *Self* and *Other*, across punishment and no punishment scenarios, choosing the CP is perceived to be more socially appropriate than choosing the EP ($p < 0.01$ in all comparisons, Wilcoxon signed-rank test).

Table B.1: Personal norms

	NoPun		Pun	
	CP	EP	CP	EP
Self	4.37	2.62	4.02	3.29
	(0.91)	(1.19)	(1.13)	(1.03)
Other	4.36	2.92	4.24	2.86
	(0.89)	(1.28)	(1.01)	(1.11)

Notes: Personal norms take a value from 1 to 5 with 1 = very inappropriate. Standard deviations in parentheses.

Table B.2: Injunctive norms

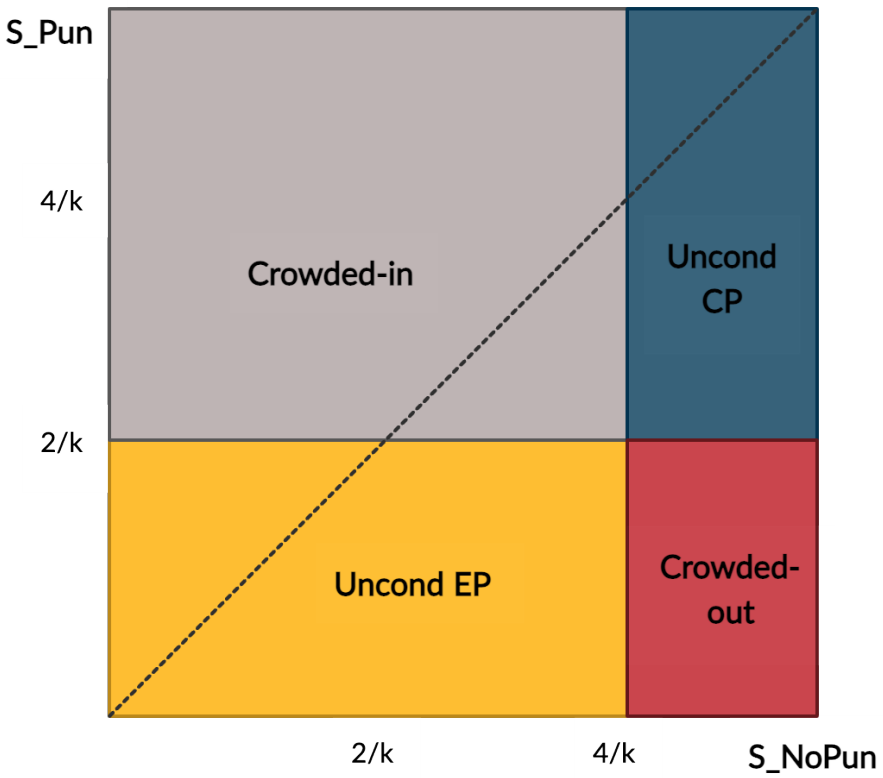
	NoPun		Pun	
	CP	EP	CP	EP
Self	4.35	2.59	3.96	3.34
	(0.98)	(1.28)	(1.20)	(1.28)
Other	4.15	3.05	4.23	2.82
	(1.07)	(1.34)	(1.01)	(1.24)

Notes: Injunctive norms take a value from 1 to 5 with 1 = very inappropriate. Standard deviations in parentheses.

C Agents' types

Figure C.1 presents the theoretical predictions of agents' types based on the stigma associated with choosing the EP under punishment (S_{Pun}) and no punishment (S_{NoPun}).

Figure C.1: Agents' types based on S_{Pun} and S_{NoPun}



Notes: The dotted line represents the cases in which $S_{Pun} = S_{NoPun}$, i.e. $\Delta S = 0$. The area below the line represents cases where $\Delta S < 0$, and area above the line cases where $\Delta S > 0$.

D Order effects

Table D.1 shows that the likelihood of the agent choosing the CP does not depend on the order in which the questions were asked (i.e., whether agents were first asked for their choice under punishment, or first asked for their choice under no punishment) in both *Self* ($p = 0.92$, column 2) and *Other* ($p = 0.51$, column 4).

Table D.1: Effect of punishment on the choice of CP

	Self		*Other*	
	(1)	(2)	(3)	(4)
Pun	1.029	1.031	1.989***	2.155***
	(0.177)	(0.194)	(0.161)	(0.178)
Order: Pun, NoPun	1.017	1.036	0.766	0.804
	(0.291)	(0.332)	(0.294)	(0.333)
Constant	0.886	0.591	0.709	1.037
	(0.213)	(1.273)	(0.224)	(1.090)
Controls	No	Yes	No	Yes
AIC	399.00	414.31	405.47	418.76
BIC	409.94	501.88	416.54	514.714
Log Likelihood	−196.50	−183.15	−199.73	−183.38
Deviance	393.00	366.31	399.47	366.764
Num. obs.	284	284	296	296

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio logistic regression with standard errors clustered at the individual level in parentheses. The dependent variable is the agent's choice (=1 if they chose CP). The baseline order is the choice without punishment, followed by the choice with punishment. The control variables are gender, age, education, religiosity, income and political orientation. Results are reported as factor changes in the odds ratios and hence an estimate below (above) 1 indicates a negative (positive) effect.

E Study 2 tables

Table E.1: How punishment changes the stigma against the EP (ΔS)

	Personal norm		Injunctive norm		Descriptive norm	
	(1)	(2)	(3)	(4)	(5)	(6)
Other-NoCost	0.14	0.08	-0.56^{**}	-0.62^{***}	-12.59^{***}	-11.98^{***}
	(0.19)	(0.20)	(0.22)	(0.23)	(3.92)	(4.04)
Constant	-0.04	0.46	0.42^{***}	1.65^{**}	23.69^{***}	17.72
	(0.14)	(0.69)	(0.16)	(0.79)	(2.77)	(14.08)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.00	0.09	0.02	0.10	0.03	0.11
Adj. R^2	-0.00	0.02	0.02	0.03	0.03	0.03
Num. obs.	341	341	341	341	341	341

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: OLS regression with standard errors in parentheses. The dependent variable is ΔS, computed using first-order beliefs of personal norms (Columns 1 and 2), second-order beliefs of injunctive norms (Columns 3 and 4) and first order beliefs of descriptive norms (Columns 5 and 6). The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation.

Table E.2: Odds of observing agents' types

	Uncond_CP			Uncond_EP		Crowd_in
	Uncond_EP	Crowd_in	Crowd_out	Crowd_in	Crowd_out	Crowd_out
	(1)	(2)	(3)	(4)	(5)	(6)
Other-NoCost	0.642	1.270	2.204	1.978^{**}	3.431^{**}	1.735
	(0.291)	(0.328)	(0.602)	(0.323)	(0.598)	(0.612)
Constant	1.560	0.631	0.041	0.405	0.026	0.065
	(0.995)	(1.164)	(2.236)	(1.151)	(2.226)	(2.293)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio multinomial logistic regression with standard errors in parentheses (N=343, AIC: 904.784). The dependent variable is agent's type based on their choices. The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation. Results are reported as factor changes in the odds ratios and hence an estimate below (above) 1 indicates a negative (positive) effect. Created using the Stargazer package (Hlavac, 2013) in R.

Table E.3: Odds of observing a "Crowded-in" type

	(1)	(2)
Other-R	0.242^{***}	0.208^{***}
	(0.393)	(0.430)
Constant	0.250^{***}	0.057^{*}
	(0.192)	(1.527)
Controls	No	Yes
Observations	328	328
Log Likelihood	−119.595	−108.611
Akaike Inf. Crit.	243.191	273.222

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Odds ratio logistic regression with standard errors in parentheses. The dependent variable is the agent's choice (= 1 if they are a "Crowded-in" type). The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation. Results are reported as factor changes in the odds ratios and hence an estimate below (above) 1 indicates a negative (positive) effect. Created using the Stargazer package (Hlavac, 2013) in R.

Table E.4: How punishment changes the stigma against the EP (ΔS)

	Personal norm		Injunctive norm		Descriptive norm	
	(1)	(2)	(3)	(4)	(5)	(6)
Other-R	0.09	0.01	−0.35	-0.58^{**}	-14.12^{***}	-15.96^{***}
	(0.21)	(0.22)	(0.25)	(0.25)	(3.88)	(4.03)
Constant	−0.04	1.03	0.42^{**}	1.85^{**}	23.69^{***}	42.95^{***}
	(0.14)	(0.76)	(0.17)	(0.87)	(2.69)	(14.12)
Controls	No	Yes	No	Yes	No	Yes
R^2	0.00	0.07	0.01	0.16	0.04	0.12
Adj. R^2	−0.00	−0.02	0.00	0.08	0.04	0.04
Num. obs.	326	326	326	326	326	326

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: OLS regression with standard errors in parentheses. The dependent variable is ΔS, computed using first-order beliefs of personal norms (Columns 1 and 2), second-order beliefs of injunctive norms (Columns 3 and 4) and first order beliefs of descriptive norms (Columns 5 and 6). The baseline treatment is *Other*. The control variables are the order in which agents' choices were elicited, gender, age, education, religiosity, income and political orientation.