
Norm-signalling punishment

Discussion Paper no. [2022-26](#)**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 punishment 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-regarding punishment reduces the social stigma of selfish behavior, while other-regarding punishment does not. As a result, self-serving punishment is less effective at encouraging compliance and is more likely to backfire compared to other-regarding punishment. Our findings have implications for the design of punishment mechanisms and highlight the importance of the punisher's motives in the norm-signalling function of punishment.

Keywords: punishment, norms, stigma, crowd out, experiment**JEL Classification:** D02

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

Norm-signalling punishment*

Daniele Nosenzo [†], Erte Xiao [‡], Nina Xue [§]

September 25, 2022

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 punishment 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-regarding punishment reduces the social stigma of selfish behavior, while other-regarding punishment does not. As a result, self-serving punishment is less effective at encouraging compliance and is more likely to backfire compared to other-regarding punishment. Our findings have implications for the design of punishment mechanisms and highlight the importance of the punisher's motives in the norm-signalling function of punishment.

JEL Classification: C91, C72, D02

Keywords: punishment, norms, stigma, crowd out, experiment

*We received helpful comments from 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 (DPDP16010274) and Aarhus University Research Foundation (AUFF Starting Grant 36835) and received approval from the Monash University Human Research Ethics Committee (project number 27176).

[†]Aarhus University, 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 an important function of punishment – which is crucial for its effectiveness – is to communicate information about society’s norms and values (e.g., Sunstein, 1996; Posner, 1997; Kahan, 1998; McAdams, 2000; Bénabou and Tirole, 2006; Bénabou and Tirole, 2011). This paper investigates whether the punisher’s *motivation* for imposing punishment can affect the message conveyed about underlying social norms, hence altering its effectiveness. 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”). Based on a simple theoretical framework, our hypothesis is that self-serving punishment transmits a weaker normative message compared to other-regarding punishment, and is hence more likely to trigger a crowding-out of prosocial behavior.

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. We focus on weak punishment that is not sufficient to change 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 Krupka and Weber (2013) and elicit personal norms, injunctive norms and descriptive norms about the agent’s behavior in the game.

Consistent with our hypothesis, we find that self-serving punishment sends a weaker normative message about the appropriateness of compliance relative to other-regarding pun-

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

ishment. 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.

Our findings have implications for how policymakers, enforcement agencies and institutions should design punishment mechanisms in order to avoid these detrimental crowding-out effects. Specifically, punishment sends a stronger normative signal 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 norm-transmitting 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 the transmission of social norms, 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

²Danilov 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.

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 the punishment is implemented out of a concern for the punisher or for others can affect the strength of the normative message conveyed and consequently decision-making.

2 Experimental design

We design 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 to two of the three players (Player B and another player, A or C, depending on treatment – see below) and offers £12 to each of the two “included” players while the “excluded” player receives a lower payoff (£6). Before Player B makes a choice, Player A (“the principal”) decides whether to impose a fixed fee to reduce the payoffs of each of the two players who are included in the EP by £2.

Our two treatments vary whether the player who is excluded from the EP is Player A or Player C (“the third party”). In the *Self* treatment, Player A is the excluded player and receives a higher payoff under CP than 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 the principal. In this sense, punishment is self-serving. In the *Other* treatment, Player C is the excluded player (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 (Player A), while in *Other* it can only benefit a passive third party (Player C). Moreover, punishment is weak in that the payoffs alone are not sufficient to incentivize Player B to change their behavior (Player B always earns more under EP than CP, regardless of whether Player A uses punishment). We elaborate in the next section on the role of punishment in changing the agent’s behavior by signaling the underlying norm of conduct. Thus, our treatments shed light on how self-serving motives underlying punishment may influence the perception of social norms and hence behavior.

In each treatment, Player A was asked to make a decision about whether to use punishment or not. We elicited Player B’s decisions using a strategy elicitation method, i.e.,

Figure 1: *Self* treatment

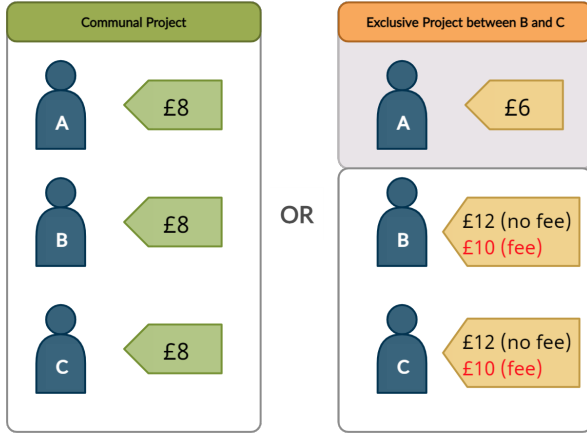
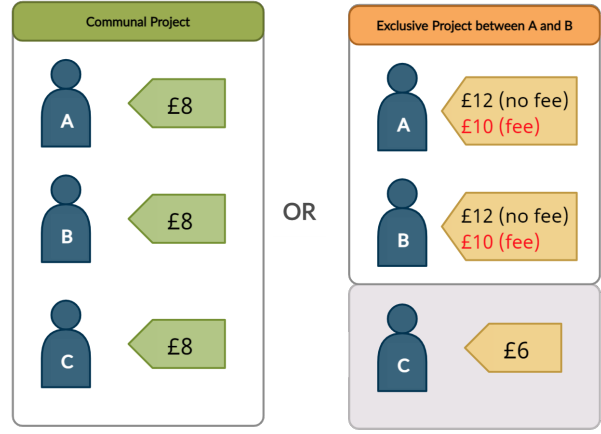


Figure 2: *Other* treatment



we asked Player B to make one choice in case A imposed a fee, and one choice in case A 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 whether A uses punishment; (ii) “Unconditional EP” if they choose the EP regardless of punishment; (iii) “Crowded-in” if they choose the CP when A uses punishment and the EP when A does not; and (iv) “Crowded-out” if they perversely choose the EP when A uses punishment and the CP when A does not. Our key question is whether the motive behind punishment affects the distribution of B’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 Player C, *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

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

⁴There 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 Players A and B 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)

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 footnote 5). We elicited a second-order belief in correspondence to each of the four first-order beliefs discussed above (appropriateness of choosing EP when A punishes; appropriateness of choosing CP when A punishes; appropriateness of choosing EP when A does not punish; appropriateness of choosing CP when A does not punish). 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 B’s behavior in the game.

Finally, we elicited subjects’ empirical beliefs by asking them to guess the percentage of Player B’s 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, a variation of the Becker-DeGroot-Marschak procedure (Becker et al., 1964).⁸ Descriptive norms can differ from injunctive norms and can be particularly useful in

⁵These 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 Player C’s were only asked unincentivized questions. We used the data from the pilot to incentivize Player C’s answers in the main experiment.

⁶See 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.

⁷Bašić and Verrina (2021) show that personal norms can differ from social norms (second-order beliefs) and are predictive of behavior.

⁸We chose the Karni mechanism due to its invariance to heterogeneous risk preferences. See Schwardmann

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 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 Theoretical framework and hypotheses

In this section, we present a theoretical framework to derive our hypotheses. Our main research question is whether the motive behind punishment affects (1) the normative message conveyed, and (2) the agent’s actual behavior in the game.

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

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).

⁹As 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.

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 Player B chooses the CP.¹⁰ 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 $f \in \{0, 2\}$, which is implemented only if the agent chooses $a_{Pun} = EP$.

Let $k > 0$ represent the agent's sensitivity towards norms and $S \geq 0$ 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:

$$\begin{cases} a_{NoPun}^* = CP, & \text{if } S_{NoPun} \geq 4/k \\ a_{NoPun}^* = EP, & \text{otherwise} \end{cases} \quad (1)$$

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:

$$\begin{cases} a_{NoPun}^* = CP, & \text{if } S_{Pun} \geq 2/k \\ a_{NoPun}^* = EP, & \text{otherwise} \end{cases} \quad (2)$$

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,

¹⁰Our 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).

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:

$$\{a_{NoPun}^*, a_{Pun}^*\} = \begin{cases} \{CP, CP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} \geq 2/k \\ \{EP, EP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} < 2/k \\ \{EP, CP\}, & \text{if } S_{NoPun} < 4/k, S_{Pun} \geq 2/k \\ \{CP, EP\}, & \text{if } S_{NoPun} \geq 4/k, S_{Pun} < 2/k \end{cases} \quad (3)$$

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.

Therefore, our first hypothesis concerns the effect that punishment has on the relative stigma against the EP. We conjecture that punishment that is devoid of self-serving motives sends a stronger normative message regarding what is considered appropriate behavior and therefore triggers a relatively stronger change in the stigma against the EP relative 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 hypothesis:

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

$$\Delta S^{Self} < \Delta S^{Other} \quad (4)$$

If our first hypothesis is confirmed, this can have direct implications for the distribution of agents' types we should observe across the two treatments. In particular, if $\Delta S > 0$, 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 Hypothesis 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 hypothesis:

Hypothesis 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 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 (Hypothesis 1). In Section 4.2 we examine agents' choices and the effectiveness of punishment (Hypothesis 2).

4.1 The normative message conveyed by punishment

We study Hypothesis 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 norm), second-order beliefs of social appropriateness (injunctive norm) and first-order beliefs of the frequency of agents' choices (descriptive norm). 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 .¹¹

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-

¹¹See Appendix B for the appropriateness ratings for personal and injunctive norms.

tailed Wilcoxon signed-rank test) and injunctive norms ($p < 0.01$).¹² 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 conjectures, $\Delta S^{Self} \leq 0$, while $\Delta S^{Other} \geq 0$.

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
<i>Self</i>	1.75 (1.77)	0.73 (1.75)	-1.02 (2.06)	1.76 (1.98)	0.62 (2.08)	-1.14 (2.87)	42.45 (30.50)	40.38 (28.70)	-2.07 (47.18)
<i>Other</i>	1.44 (1.76)	1.38 (1.59)	-0.06 (1.68)	1.11 (2.00)	1.41 (1.83)	0.30 (2.12)	29.66 (23.97)	51.26 (25.32)	21.60 (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.

We test Hypothesis 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, $p < 0.01$).¹³ 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, $p < 0.01$). These findings are corroborated by the regression analysis (which also controls for demographic variables), presented in Table 2. Thus, this analysis confirms our first hypothesis, as we summarize in the following result:

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

¹²Unless otherwise stated, we use two-tailed Wilcoxon signed-rank tests to compare changes in stigma due to punishment.

¹³Unless otherwise stated we use two-tailed Mann-Whitney tests to compare the change in stigma across *Self* and *Other* for each norm measure.

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

	Personal norm		Injunctive norm		Descriptive norm	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Self</i>	-0.96*** (0.22)	-0.93*** (0.23)	-1.45*** (0.29)	-1.52*** (0.32)	-23.67*** (4.94)	-25.15*** (5.26)
Constant	-0.06 (0.15)	0.05 (0.85)	0.30 (0.20)	1.73 (1.17)	21.60*** (3.43)	41.27** (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.

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 against choosing selfish behavior, while the latter strengthens it. We now assess whether these differences in the normative message transmitted by 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 (McNemar's test, $p = 1.00$). In *Other*, 38.5% choose the CP under no punishment, while 55.4% do so under punishment (McNemar's test, $p < 0.01$). Table 3 similarly shows that 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).¹⁴ Our findings suggest that punishment is effective at changing behavior, but only when it is motivated by a concern for others.

To test Hypothesis 2, we compare the distribution of agents' types between the two treatments. Across *Self* and *Other*, we find a similar 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

¹⁴We find no evidence of an order effect, see Appendix D.

Table 3: Likelihood of choosing the CP

	<i>Self</i>		<i>Other</i>	
	(1)	(2)	(3)	(4)
Pun	1.029 (0.177)	1.031 (0.194)	1.984*** (0.160)	2.155*** (0.178)
Constant	0.893 (0.168)	0.591 (1.273)	0.626*** (0.169)	1.037 (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.

punishment.¹⁵ 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 ($p = 0.03$).

This result is also supported by the 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. 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 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 2.83 times more likely to be a Crowded-out type than an Unconditional CP type ($p = 0.06$, column 3) and 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*

¹⁵Another 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.

($p = 0.04$, 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 Hypothesis 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: Likelihood 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)
<i>Self</i>	0.956 (0.294)	0.662 (0.377)	2.830* (0.548)	0.692 (0.366)	2.961** (0.541)	4.277** (0.586)
Constant	0.910 (1.019)	0.406 (1.632)	0.128 (2.102)	0.447 (1.610)	0.140 (2.059)	0.315 (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. Created using the Stargazer package (Hlavac, 2013) in R.

Result 2: Consistent with Hypothesis 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 all other types of agents.

5 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 the social stigma against self-interested choices, while self-serving punishment can have a detrimental effect by reducing this stigma. 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. (2021). 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. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.
- 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.
- 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.
- 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.
- Erkut, H., Nosenzo, D., and Sefton, M. (2015). Identifying social norms using coordination games: Spectators vs. stakeholders. *Economics Letters*, 130:28–31.

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

A Instructions

Figure A.1: The principal's choice

Task

Now you will participate in the "Choose-a-Project" task.








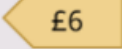

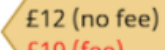

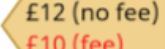
Recall, you are assigned to the role of Player A.

Please decide whether to impose a fee against Player B for choosing the Exclusive Project. If you impose the fee, the earnings of Player B and C will be **reduced by £2 each**.

I choose to impose a fee against Player B for choosing the Exclusive Project between Player B and C:

Yes

No

<p>If Player B chooses the Communal Project</p> <div style="border: 1px solid black; padding: 10px;"><p style="text-align: center; background-color: #8ebf4d; color: white; margin: 0;">Communal Project</p><div style="display: flex; justify-content: space-between; align-items: center; margin-bottom: 10px;"><div style="text-align: center;"><p>A</p></div><div style="text-align: center;"><p>£8</p></div></div><div style="display: flex; justify-content: space-between; align-items: center; margin-bottom: 10px;"><div style="text-align: center;"><p>B</p></div><div style="text-align: center;"><p>£8</p></div></div><div style="display: flex; justify-content: space-between; align-items: center;"><div style="text-align: center;"><p>C</p></div><div style="text-align: center;"><p>£8</p></div></div></div>	OR	<p>If Player B chooses the Exclusive Project</p> <div style="border: 1px solid black; padding: 10px;"><p style="text-align: center; background-color: #f4a460; color: white; margin: 0;">Exclusive Project between B and C</p><div style="display: flex; justify-content: space-between; align-items: center; margin-bottom: 10px;"><div style="text-align: center;"><p>A</p></div><div style="text-align: center;"><p>£6</p></div></div><div style="display: flex; justify-content: space-between; align-items: center; margin-bottom: 10px;"><div style="text-align: center;"><p>B</p></div><div style="text-align: center;"><p>£12 (no fee) £10 (fee)</p></div></div><div style="display: flex; justify-content: space-between; align-items: center;"><div style="text-align: center;"><p>C</p></div><div style="text-align: center;"><p>£12 (no fee) £10 (fee)</p></div></div></div>
--	----	--

Next

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

Task

Now you will participate in the "Choose-a-Project" task.

Recall, you are assigned to the role of **Player B**.

Before you find out whether Player A has actually imposed a fee, you will make **two choices**: one choice in case **Player A imposed the fee**, and one choice in case **Player A did NOT impose the fee**. Only one of these choices is implemented at the end of the study, depending on whether or not Player A actually imposed the fee.

Suppose Player A decided to impose a fee against you for choosing the Exclusive Project. Please select a project.

 ▼

Suppose Player A decided NOT to impose a fee against you for choosing the Exclusive Project. Please select a project.

 ▼

Next

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

Task

Now you will participate in the "Choose-a-Project" task.

Recall, you are assigned to the role of Player B.

Before you find out whether Player A has actually imposed a fee, you will make **two choices**: one choice in case **Player A did NOT impose the fee**, and one choice in case **Player A imposed the fee**. Only one of these choices is implemented at the end of the study, depending on whether or not Player A actually imposed the fee.

Suppose Player A decided NOT to impose a fee against you for choosing the Exclusive Project. Please select a project.

----- ▾

Suppose Player A decided to impose a fee against you for choosing the Exclusive Project. Please select a project.

----- ▾

Communal Project

← £8

← £8

← £8

OR

Exclusive Project between B and C

← £6

← £12 (no fee)
£10 (fee)

← £12 (no fee)
£10 (fee)

Next

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

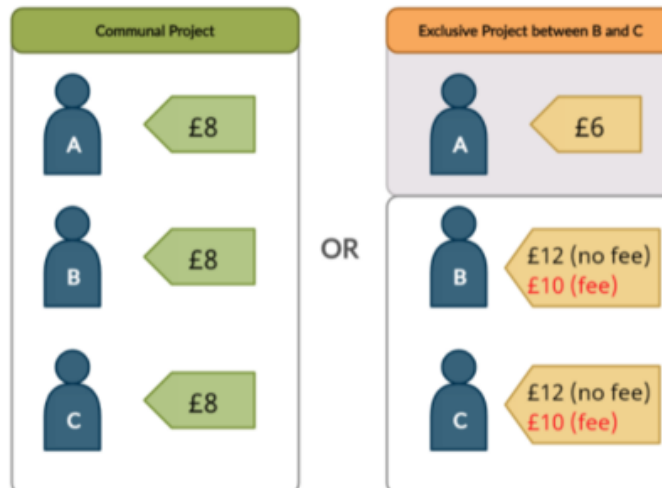


Figure A.5: Eliciting third-party beliefs about injunctive norms (Order 1)

Questions

2) We have surveyed the previous participants on what they personally believe is an appropriate choice by Players B in the “Choose-a-Project” task. We now ask you to guess, for each possible action by Player B, what the **most popular answer** was.

If your guess is correct, then you will receive an **additional £1** (for each response).

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

Questions

3) We would like to ask you to make a guess about the **choices made by the previous participants** in the "Choose-a-Project" task.

- If **Player A imposes a fee** against Player B for choosing an Exclusive Project, what share of Players B from the previous group would choose the Exclusive Project? (Recall that Player B is aware of Player A's decision when choosing a project.) To make a guess, think about out of every 100 Players B, how many would choose the Exclusive Project?
- If **Player A does NOT impose a fee** against Player B for choosing an Exclusive Project, what share of Players B from the previous group would choose the Exclusive Project between Player B and C?

You have the opportunity to earn an **additional £1** (for each response), depending on your guesses. Your chances of receiving the additional £1 are highest when you make your best guess.

If, before reporting your guess, you would like to learn the details of how reporting your best guess maximizes your chances of earning £1, click on "See more details".

[See more details](#)

[Report my guess](#)



Figure A.7: Payment mechanism

Questions

Payment mechanism

The payment mechanism works as follows. After you report your guess (a number between 0 and 100), the computer will randomly choose a number between 0 and 100 (let's call this number N), with each number being equally likely to be drawn.

- If N is higher or equal to your guess, then you will be paid according to a lottery where $N\%$ of the time you will earn £1, and $(100-N)\%$ of the time you will earn £0.
- If N is lower than your guess, then you will be paid according to a lottery where $X\%$ of the time you will earn £1 and $(100-X)\%$ of the time you will earn £0, where X is the actual share of Players B who chose the Exclusive Project.

Therefore, your chances of receiving the additional £1 are highest when you report your best guess of the actual share.

[Report my guess](#)

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

Task

Now you will participate in the "Choose-a-Project" task.

You are assigned to the role of Player C and will be randomly and uniquely matched with a Player A and a Player B. Your identity will remain anonymous, as will the identities of all other participants.

You have no choice to make. Your earnings from the task will depend on the choices of the Player A and Player B you are matched with.

If you are one of the 1 in 20 participants selected to receive a bonus payment, you will be notified on Prolific.

[Next](#)

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
<i>Self</i>	4.37 (0.91)	2.62 (1.19)	4.02 (1.13)	3.29 (1.03)
<i>Other</i>	4.36 (0.89)	2.92 (1.28)	4.24 (1.01)	2.86 (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

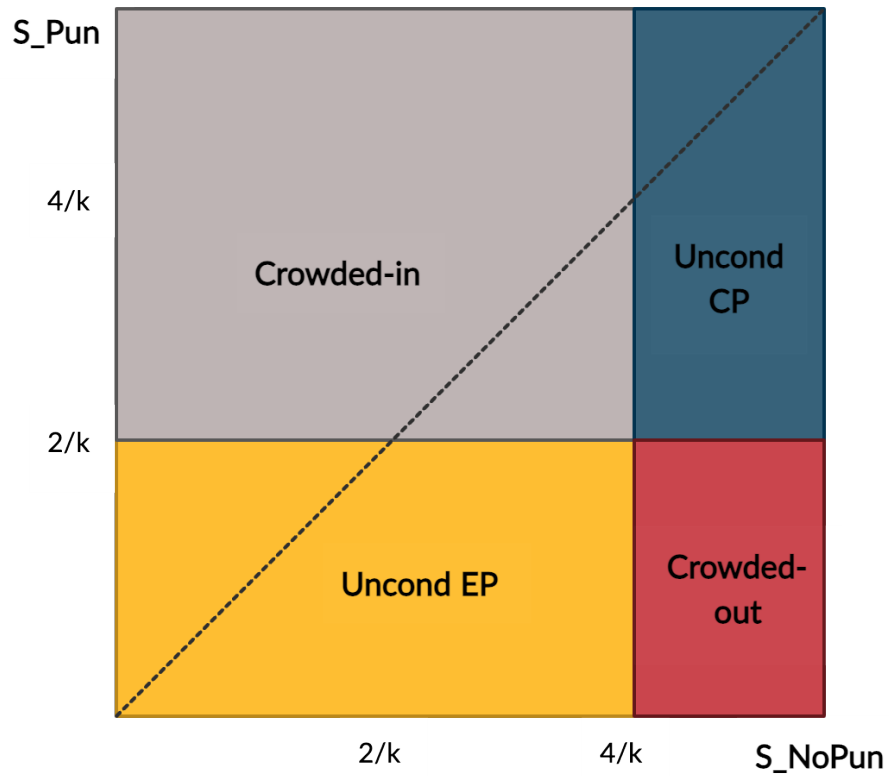
	NoPun		Pun	
	CP	EP	CP	EP
<i>Self</i>	4.35 (0.98)	2.59 (1.28)	3.96 (1.20)	3.34 (1.28)
<i>Other</i>	4.15 (1.07)	3.05 (1.34)	4.23 (1.01)	2.82 (1.24)

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: Likelihood of choosing the CP

	<i>Self</i>		<i>Other</i>	
	(1)	(2)	(3)	(4)
Pun	1.029 (0.177)	1.031 (0.194)	1.989*** (0.161)	2.155*** (0.178)
Order: Pun, NoPun	1.017 (0.291)	1.036 (0.332)	0.766 (0.294)	0.804 (0.333)
Constant	0.886 (0.213)	0.591 (1.273)	0.709 (0.224)	1.037 (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.