Spillover Effects of Third-Party Punishment on Cooperation: A Social Norms-Based Explanation

Authors: Chen Sijing, Xing Yilin, Weng Yijing, Li Chang, Li Chang

Date: 2021-03-04T00:00:00+00:00

Abstract

Third-party punishment may maintain cooperation through either economic functions or norm-signaling functions. Prior research has not distinguished between these two functions, thus failing to answer whether punishment can still promote cooperation when it is insufficient to affect the gains from violations. Experiment 1 (N = 252) found that even when third-party punishment could not reduce the gains from violations, it still inhibited selfish behavior. Experiment 2 (N = 179) found that punished violators exhibited higher cooperation levels in subsequent dictator games. Experiment 3 (N = 179), employing a 2 (whether observing punishment) × 2 (before/after observation) design, showed that participants' cooperation levels after observing punishment were significantly higher than before observation, and also higher than those of participants who did not observe punishment. In the latter two experiments, social norms mediated the relationship between punishment and cooperation. This further confirms that punishment' s promotion of cooperation is largely achieved through norm activation, and that two types of spillover effects exist: punishment inhibits selfish behavior in new game contexts among both former violators (longitudinal spillover effect) and observers (horizontal spillover effect). The discovery of these two spillover effects complements the dominant economic explanations in the literature and provides a new perspective for understanding long-term, large-scale cooperation in human society.

Full Text

Spillover Effects of Third-Party Punishment on Cooperation: A Norm-Based Explanation

CHEN Sijing[1], XING Yilin[1], WENG Yijing[1], LI Chang[2]
([1] School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou 310023, China)

([2] School of Business Administration, Zhejiang Gongshang University, Hangzhou 310018, China)

Abstract

Third-party punishment can sustain cooperation through either economic or norm-signaling functions. Previous research has not distinguished between these two functions, leaving unanswered whether punishment can promote cooperation when it is insufficient to affect the gains from defection. Experiment 1 (N = 252) found that third-party punishment suppressed selfish behavior even when it could not reduce defection payoffs. Experiment 2 (N = 179) revealed that previously punished defectors exhibited higher cooperation levels in a subsequent dictator game. Experiment 3 (N = 179), using a 2 (whether observing punishment) × 2 (before vs. after observation) design, showed that participants' cooperation levels after observing punishment were significantly higher than before observation and also higher than those who did not observe punishment. In the latter two experiments, social norms mediated the relationship between punishment and cooperation. These findings confirm that punishment promotes cooperation largely through norm activation and demonstrate two types of spillover effects: punishment suppresses selfish behavior in both former defectors (vertical spillover effect) and bystanders (horizontal spillover effect) in new game contexts. The discovery of these two spillover effects supplements the dominant economic explanations in the literature and provides a new perspective for understanding long-term, large-scale cooperation in human societies.

Keywords: third-party punishment, social norm, cooperation, focus theory, spillover effect
Classification Code: B849: C91

1. Introduction

In social sciences, cooperation refers to behavior where individuals incur costs to benefit others (Nowak, 2006; Rand, 2016). Extensive cooperation among non-kin individuals is crucial for the smooth functioning of human society (Fehr & Schurtenberger, 2018), which has led to the development of social norms of cooperation (de Kwaadsteniet et al., 2007). These norms are widely accepted behavioral guidelines regarding cooperation that differ from explicit regulations such as laws (Cialdini & Trost, 1998). Although cooperative norms exist universally across cultures, compliance with them is not automatic (de Kwaadsteniet et al., 2019). Third-party punishment—punishment administered by uninvolved parties against norm violators—is generally considered an important mechanism for reducing violations and sustaining cooperative norms (Balliet et al., 2011; Fehr & Gächter, 2002; Halevy & Halali, 2015).

Building on this foundation, scholars have examined the role of norms in how punishment influences cooperation. For instance, Bicchieri et al. (2018) found that punishment needs to be combined with certain social norms to exert pos-

itive effects. Similarly, Fehr and Williams (2018) noted that third-party punishment promotes cooperation only when group members share corresponding normative consensus; when such consensus is absent, punishment accelerates community collapse. Additionally, Lois and Wessa (2019) explored the moderating role of social norms on third-party punishment. However, a more fundamental question concerns why third-party punishment reduces violations (and promotes cooperation), and we observe that research from a normative perspective is relatively absent in addressing this question.

Currently, a widely accepted explanation is based on an economic perspective: third-party punishment changes the payoff structure for defectors. In the presence of third-party punishment, the cost of defection rises substantially to exceed the gains from violation (韦倩, 姜树广, 2013; Bicchieri et al., 2018; Carpenter & Matthews, 2004; Nelissen & Mulder, 2013; Rand et al., 2010). In this situation, the dominant strategy for rational individuals is to cooperate rather than defect.

However, this economic perspective may have several problems. First, numerous studies show that people do not always follow the homo economicus principle in decision-making (Alkan, 2020; Camerer & Fehr, 2006; Henrich et al., 2001). Therefore, unless we presuppose that defectors are always purely rational economic agents—a premise whose reasonableness remains debatable—an exclusively economic explanation is insufficient to account for how third-party punishment inhibits violations. Second, previous research has found that punishers' motivations significantly affect punishment' s effectiveness (谢东杰, 苏彦捷, 2019; Raihani & Bshary, 2015). For example, Rand et al. (2009) noted that whether punishment is perceived as legitimate can greatly influence the punished individual' s response. Fehr and Rockenbach (2003) also observed that when punishment is perceived as self-serving (e.g., intended to obtain more personal benefit), although punishment significantly reduces defection payoffs (by 40% of the initial amount), punished defectors do not show higher cooperation levels; rather, their cooperation decreases markedly. If punishment promotes cooperation primarily by reducing defection payoffs, these findings become difficult to explain. Third, if punishment' s inhibitory effect on defection mainly stems from increasing defection costs, we would expect that unless every violation is punished, past punishment experience would be insufficient to automatically improve behavior in new situations. However, as Shreedhar et al. (2018) point out, if a group must punish every violation, the cost would be extremely high, potentially exceeding the positive effects of punishment. In other words, ubiquitous punishment not only fails to maintain cooperation in large-scale communities but also causes such groups to lose competitive advantage.

For these reasons, we argue that a purely economic perspective is insufficient to explain how third-party punishment sustains cooperative norms. Chen et al. (2015) proposed, based on the Focus Theory of Normative Conduct (Cialdini et al., 1991), that third-party punishment itself is a process of social norm activation, providing an alternative theoretical starting point for understanding third-party punishment. Focus Theory suggests that people may violate norms

simply because they are unaware of them; therefore, making norms salient can significantly reduce violations. Indeed, researchers from this perspective have found that third-party punishment can activate social norms (陈思静等, 2015), and Chen et al. (2020) noted that third-party punishment significantly affects people' s norm perceptions. However, in previous research, third-party punishment typically altered defectors' payoff structures, meaning researchers could not strictly distinguish between two functions: enhancing cooperation by reducing defection payoffs (economic effect) versus enhancing cooperation by activating social norms (normative effect). This paper aims to provide a useful supplement to existing literature in this regard.

Specifically, Experiment 1 will test the norm activation function of third-party punishment while controlling for defector payoffs. If results show that punished defectors still exhibit higher cooperation despite punishment not substantially reducing their gains, we can conclude that the normative effect is a function independent of the economic effect, providing new empirical evidence for Focus Theory: activating people' s norms can change their behavior. Second, human cooperation exhibits long-term and large-scale characteristics (Bingham, 1999). If punishment' s role only manifests as activating cooperation norms for the punished individual in a specific scenario, we face a theoretical dilemma similar to that of economic explanations: if every individual must be reminded of norms through punishment in every situation, societal operation costs become extremely high, rendering third-party punishment meaningless (Shreedhar et al., 2018). Therefore, we hypothesize that third-party punishment' s norm-signaling effect not only suppresses defectors' immediate selfish behavior but also extends this norm activation effect to new contexts (vertical spillover effect or temporal spillover, Experiment 2) and to bystanders who witness rather than personally experience punishment (horizontal spillover effect or spatial spillover, Experiment 3), even when no potential punishers exist in these situations. If these hypotheses hold, we can partially explain why cooperation in human society proceeds orderly despite third-party punishment not occurring constantly in real life.

Finally, social norms, as widely accepted behavioral guidelines distinct from legal regulations (Cialdini & Trost, 1998; Forquesato, 2016), are typically distinguished in social science literature as descriptive norms and injunctive norms (Cialdini et al., 1991). The former refers to common behavioral patterns (e.g., the prevalence of cooperative behavior), while the latter refers to widely held attitudes of approval or disapproval toward a behavior (e.g., approval of others' cooperative behavior). Social norms significantly influence behavior by simplifying decision-making and providing behavioral guidance in complex, uncertain, or dangerous situations (McDonald & Crandall, 2015). However, researchers have noted differences in how these two norms affect behavior. Deutsch and Gerard (1955) suggest that people process descriptive norm information faster than injunctive norm information, making descriptive norms more likely to influence behavior. Petty and Cacioppo (1986), comparing the two norms from the perspective of personal involvement, note that injunctive norms have stronger

effects when personal involvement is high. For this paper, a worthwhile question is: when punishment influences cooperation through norm activation, does it activate one type of norm or both? If both are activated, do they have different mechanisms? We explore these questions in Experiments 2 and 3. Additionally, since Focus Theory primarily examines descriptive norms, if our results show that injunctive norms are also activated and significantly affect the process through which punishment influences cooperation via norm activation, this could supplement Focus Theory.

Based on this literature review, we propose the following research questions as the main objectives:

Research Question 1: When third-party punishment cannot reduce defector payoffs, can it still effectively reduce violations (and promote cooperation)? (Experiment 1)

Research Question 2: Can third-party punishment' s cooperation-enhancing effect through norm activation spill over to new contexts? (Experiment 2)

Research Question 3: Can third-party punishment' s cooperation-enhancing effect through norm activation spill over to bystanders? (Experiment 3)

Research Question 4: Do descriptive and injunctive norms have similar mechanisms in the process through which punishment influences cooperation via norm activation? (Experiments 2 and 3)

In summary, this paper explains the mechanism through which third-party punishment affects cooperation from a social norm perspective. We argue that norm activation is an independent function of third-party punishment: even when it cannot reduce defection payoffs, punishment can still inhibit violations (and promote cooperation) (Experiment 1). Moreover, this effect spills over to new contexts lacking punishment mechanisms (Experiment 2) and to bystanders who witness punishment (Experiment 3). We also examine the mechanisms of both types of norms in this process (Experiments 2 and 3) and discuss the theoretical and practical implications of these findings.

2. Experiment 1: Norm Activation Function of Third-Party Punishment

2.1 Participants

A power analysis using G*Power 3.1 with a medium effect size of f = 0.25 and significance level $\alpha = 0.05$ indicated that at least 252 participants were needed for a one-way ANOVA across three groups to achieve 95% statistical power (1-β). Given the "4+1" experimental design where 4 out of every 5 participants provided valid data for statistical analysis (see Section 2.2), we recruited 315 undergraduate students from various majors at Zhejiang Gongshang University. All participants read detailed instructions and signed informed consent forms before the experiment. They familiarized themselves with the experimental rules

through practice exercises (examples in Appendix). The 252 valid participants had a mean age of 21.42 ± 2.25 years, with 58.33% female. Their majors were distributed as follows: science and engineering (34.92%), social sciences (28.57%), humanities (25.40%), and arts and others (11.11%).

2.2 Design and Procedure

Experiment 1 used a 3-group between-subjects design (control group, high-benefit group, and low-benefit group). The paradigm was a public goods game conducted via computer using z-Tree software (Fischbacher, 2007). Participants sat in separate cubicles and could not communicate. Each group of 5 participants played together: 4 were players who participated in the public goods game, while the remaining 1 was an enforcer[1]. The enforcer did not participate in the game; in the control group, this role was a tax collector, while in the other two conditions, the enforcer was a punisher. To exclude potential effects of direct reciprocity (Trivers, 1971), indirect reciprocity (Nowak & Sigmund, 1998), and costly signaling (Gintis et al., 2001), the 4 players were randomly labeled A, B, C, and D in each round, while the enforcer was always labeled E. Group members were randomly assigned by computer each round, and player and enforcer roles were not interchangeable. After each round, participants were informed of each member' s contribution and earnings (including punishment information when applicable), but they did not know their group members' past performance in new rounds. Additionally, to avoid end-game effects, participants were not told the total number of rounds in advance.

Participants were randomly assigned to one of three conditions: control (n = 84), high-benefit (n = 84), or low-benefit (n = 84). In the control group, each participant (including players A/B/C/D and enforcer E) began with an initial endowment of 25 tokens (equivalent to 5 RMB). In each round, players decided whether to contribute 10 tokens from their initial endowment to a public account. Tokens contributed to the public account were doubled and evenly distributed among all group players. For players, the dominant strategy was to retain the initial endowment (defect) while encouraging others to contribute, rather than contributing 10 tokens (cooperate). However, if everyone did this, each person' s final earnings would decrease. Players were told that retaining 10 tokens required paying a 1-token income tax to member E, which did not enter the public account nor was returned to any member. After these steps, the next round began, for a total of 10 rounds. After 10 rounds, the experiment ended, and participants received feedback and payment. Earnings consisted of a 10 RMB show-up fee plus tokens from a randomly selected round (5 tokens = 1 RMB).

The procedures in the high-benefit and low-benefit groups were similar to the control group, with the main difference being that after the computer revealed players' choices, member E could punish defectors[2]: (1) In the high-benefit group, if the enforcer chose to punish a player, the enforcer paid 5 tokens while the punished player paid only 1 token as defection cost, resulting in high defec-

tion benefits; (2) In the low-benefit group, the enforcer paid 5 tokens while the punished player paid 10 tokens as defection cost, resulting in low defection benefits. In each round, the enforcer could punish multiple defectors simultaneously but could punish each defector only once. After participants completed their punishment decisions, the computer announced these decisions and each participant' s earnings for that round. Table 1 summarizes the payoffs for cooperation and defection in each experimental condition.

Table 1 Payoffs for Cooperation and Defection Under Different Experimental Conditions

Condition	Cooperation Payoff (Uc)	Defection Payoff (Ud)
Control	25 - 10 + 2(10xC+10)/4	25 + 2(10xC+10)/4 - 1
High-Benefit	25 - 10 + 2(10xC+10)/4	25 + 2(10xC+10)/4 - 10xD
Low-Benefit	25 - 10 + 2(10xC+10)/4	25 + 2(10xC+10)/4 - 10xD

Note: Uc represents payoff for cooperation; Ud represents payoff for defection; xC represents number of other cooperators (xC {0,1,2,3}); xD represents number of times punished (xD {0,1}).

2.3 Results and Discussion

Differences in cooperation levels across gender ($t = 0.83$, $p = 0.408$) and major ($F = 1.54$, $p = 0.204$) were not significant, and age was not significantly correlated with cooperation level ($r = -0.03$, $p = 0.597$). A one-way ANOVA comparing cooperation levels across the three groups revealed significant differences ($F = 15.24$, $p < 0.001$, $d = 0.65$, 95% CI = [0.38, 0.92]). Post-hoc comparisons (Tukey' s HSD) showed that cooperation in the high-benefit group ($M = 4.75$, $SD = 2.57$, $n = 84$) was significantly higher than in the control group ($M = 3.55$, $SD = 2.80$, $n = 84$) ($p = 0.012$, 95% CI = [0.22, 2.19]). Cooperation in the low-benefit group ($M = 5.86$, $SD = 2.76$, $n = 84$) was significantly higher than in both the control group ($p < 0.001$, 95% CI = [1.32, 3.30]) and the high-benefit group ($p = 0.023$, 95% CI = [0.12, 2.09]). Figure 1 [Figure 1: see original paper] visually displays these differences.

Figure 1 Cooperation Levels Across Three Groups

These results confirm the importance of reducing defection payoffs for increasing cooperation, meaning that economic cost-benefit considerations indeed play a significant role in how punishment inhibits violations. This is evident in the low-benefit group' s significantly higher cooperation compared to the high-benefit group, indicating that changing defectors' payoff structure through third-party punishment can incentivize reduced violations and enhance cooperation (Balliet et al., 2011; Gächter et al., 2008). However, economic factors alone cannot fully explain Experiment 1' s results. Comparing defection payoffs (Ud) between the high-benefit and control groups in Table 1 shows that the expected payoff for

defection in the high-benefit group was always greater than or equal to that in the control group. According to homo economicus logic, more (fewer) participants in the high-benefit group should choose defection (cooperation). In reality, however, the high-benefit group showed significantly higher cooperation than the control group, indicating that even when third-party punishment did not essentially reduce defection payoffs, it still effectively inhibited violations (and promoted cooperation). This suggests that the psychological mechanism through which punishment reduces violations involves more than changing defectors' payoff structures; other important factors must exist. In other words, the principle that people do not always follow the homo economicus assumption also applies to defectors, though this result may violate our intuitions.

By addressing Research Question 1—that third-party punishment' s promotion of cooperation does not entirely depend on reducing defection payoffs—Experiment 1 also supports Focus Theory's perspective: often people violate norms not purely for profit but simply because they are unaware of the norm' s existence (Cialdini et al., 1991). In Experiment 1, comparing the high-benefit and control conditions, the only difference was defection cost: xD in the high-benefit group versus 1 in the control group, where xD $\leq$ 1. Yet xD had a stronger inhibitory effect on defection, suggesting that the difference in violation inhibition stems mainly from qualitative rather than quantitative differences in the two costs. The defection cost manifested as punishment signals moral disapproval of violations, thereby activating social norms of cooperation (陈思静等, 2015), whereas the income tax in the control group was relatively neutral and lacked this function. Although previous studies (e.g., 陈思静等, 2015; Chen et al., 2020) have suggested that punishment has norm-signaling functions, because punishment typically affects violators' economic interests, they could not strictly answer whether third-party punishment can effectively promote cooperation when it cannot change defectors' payoff structures. Experiment 1 provides clear empirical evidence for third-party punishment' s norm-signaling function by controlling for economic effects through a randomized controlled design, demonstrating that punishment' s normative effect does not require its economic effect as a prerequisite—a significant supplement to existing research.

3. Experiment 2: Temporal Spillover Effect of Punishment

Experiment 1 provided evidence for the pure norm-signaling function of third-party punishment. Experiment 2 further examines whether punishment' s cooperation-enhancing function can spill over to new contexts without punishment mechanisms and compares the mechanisms of descriptive and injunctive norms to answer Research Questions 2 and 4.

3.1 Participants

Three hundred students from various majors participated in Experiment 2, reading detailed instructions and signing informed consent forms before the experiment. Experiment 2 first required screening for defectors. Based on Experiment

1' s effect size of d = 0.65, with $\alpha = 0.05$, G*Power 3.1 calculations indicated that at least 104 defectors were needed to achieve 95% statistical power (1-β). Through Stage 1 operations, we obtained 179 defectors. These 179 participants had a mean age of 21.30 $\pm$ 1.97 years, with 54.19% female. Their majors were distributed as: science and engineering (35.75%), social sciences (31.84%), humanities (24.02%), and arts and others (8.38%).

3.2 Design and Procedure

Experiment 2 used a 2-group between-subjects design (control vs. punishment). The paradigm was a dictator game with a third party. In Stage 1, participants were told they would complete 5 rounds of a dictator game with 2 other participants. In all 5 rounds, participants played as dictators, while the recipients and third parties were virtual participants—computer programs preset by the experimenters[3]. Participants were told that dictators, recipients, and third parties began each round with 10, 0, and 2 tokens respectively. Dictators could freely allocate their initial endowment between themselves and recipients, who could not reject offers, but third parties could punish unfair allocations by paying 2 tokens to reduce the dictator' s payoff by 6 tokens. Participants also learned that group members were randomly selected by computer each round with no outcome feedback. Based on previous literature (Csukly et al., 2011; Fehr & Fischbacher, 2003), we used the following criterion to identify violations: allocations to recipients below 30% of the initial endowment were considered violations, while others were cooperative. After completing 5 rounds, 179 participants who had violated at least once were randomly assigned to two groups: 90 were told they had been punished by third parties in the previous 5 rounds (punishment group), while the remaining 89 received no feedback (control group)[4].

3.2.2 Stage 2: Dictator Game and Public Goods Game After grouping, participants in both conditions completed the following tasks: (1) a one-round dictator game without third parties with another participant, where they continued as dictators but with different allocation rules: each participant had 20 tokens and could freely choose any integer amount between 0 and 10 to allocate to the recipient, with explicit instructions that no punishment would occur regardless of their allocation; (2) a one-round public goods game without third parties with 3 other participants, where they could freely contribute any integer amount between 0 and 20 from their 20-token endowment to a public account, with contributions doubled and evenly distributed among 4 members, and explicit instructions that no punishment would occur regardless of their choice. To avoid potential order effects, half the participants read instructions for the dictator game first, while the other half read the public goods game instructions first.

Participants then estimated: (1) the percentage of dictators allocating 0, 1, 2··· 10 tokens to recipients; (2) the percentage of participants approving of allocating 0, 1, 2···10 tokens to recipients; (3) an integer between 0–10 representing how

much they would allocate to recipients; and for the public goods game: (4) an integer between 0–20 representing how much they would contribute to the public account. After completing these steps, the experiment ended, participants received feedback and payment. Earnings consisted of a 10 RMB show-up fee plus tokens from a randomly selected round.

We measured participants' norm activation levels in two ways: First, following Chen et al. (2020), we used the weighted averages of items (1) and (2) as operational definitions of descriptive and injunctive norm activation levels. Second, following Bicchieri and Xiao (2009), Voisin et al. (2016), and Sood et al. (2020), we used participants' estimates of how common a behavior was or how widely it was approved to represent norm activation levels. Specifically, we used the estimated percentage of dictators allocating 7, 8, 9, and 10 tokens to recipients as operational definitions[5]. We primarily used the first definition to test research questions and the second for robustness checks to examine whether results differed qualitatively across definitions, thereby strengthening conclusions. Finally, following Huang and Zhang' s (2013) definition of cooperation as behavior that costs oneself to benefit others or public goods, we used items (3) and (4) to represent cooperation levels in the two games, with higher numbers indicating greater cooperation.

3.3 Results and Discussion

Using the first norm activation measure, we found no significant differences in descriptive norms, injunctive norms, or cooperation levels across gender ($t = 0.07$–1.26, $p = 0.209$–0.941) or major ($F = 0.18$–1.43, $p = 0.236$–0.911), and no significant correlations between age and these variables ($r = 0.03$–0.05, $p = 0.540$–0.736). As shown in Figure 2 [Figure 2: see original paper], punished participants showed significantly higher descriptive norm activation ($M = 3.80$, $SD = 2.45$, $n = 90$) than controls ($M = 2.83$, $SD = 1.85$, $n = 89$) ($t = 2.97$, $p = 0.003$, $d = 0.44$, 95% CI = [0.15, 0.74]). Punished participants also showed significantly higher injunctive norm activation ($M = 5.62$, $SD = 2.79$) than controls ($M = 4.10$, $SD = 2.56$) ($t = 3.82$, $p < 0.001$, $d = 0.57$, 95% CI = [0.27, 0.87]). Additionally, punished participants showed higher cooperation in the dictator game ($M = 3.55$, $SD = 2.83$) than controls ($M = 2.46$, $SD = 2.75$) ($t = 2.59$, $p = 0.009$, $d = 0.39$, 95% CI = [0.09, 0.68]). These results provide preliminary answers to Research Question 2, showing that third-party punishment significantly activated both social norms and increased defectors' cooperation in new contexts. In Stage 2' s dictator game, no third party could implement punishment, and the only difference between conditions was that the punishment group was reminded of their past punishment. Thus, a reasonable explanation for Experiment 2' s results is that third-party punishment' s norm-signaling function spilled over to a new context where punishment mechanisms were absent, yet activated social norms still enhanced cooperation.

Using the second norm activation measure yielded similar results: punished participants showed significantly higher descriptive norm activation ($t = 4.18$,

$p < 0.001$) and injunctive norm activation ($t = 4.80$, $p < 0.001$) than controls, indicating robust findings.

To answer Research Question 4 (Do the two norms have different mechanisms?), we examined the psychological mechanism through which punishment affects cooperation, using punishment experience as the independent variable, descriptive and injunctive norms as mediators, and cooperation level as the dependent variable. Following researchers'recommendations that bias-corrected nonparametric percentile bootstrap methods provide more accurate confidence intervals for indirect effects than Sobel tests (方杰, 张敏强, 2012; 温忠麟, 叶宝娟, 2014), we used the PROCESS 3.5 macro (Model 4) developed by Preacher and Hayes (2004) for mediation analysis.

Results in Table 2 show that punishment significantly affected both norms in Models 1 and 2. Compared to Model 3, Model 4' s R^2 increased by 0.24 after including both norms, indicating they explained 24% of variance in cooperation. Further analysis of indirect effects revealed that both descriptive norms (Effect = 0.57, BootSE = 0.22, BootLLCI = 0.18, BootULCI = 1.06) and injunctive norms (Effect = 0.26, BootSE = 0.13, BootLLCI = 0.04, BootULCI = 0.55) had significant indirect effects, as their confidence intervals excluded zero. However, the direct effect of punishment (Effect = 0.24, SE = 0.39, t = 0.64, p = 0.523, LLCI = -0.51, ULCI = 1.01) was nonsignificant, as its confidence interval included zero. Thus, punishment' s promotion of cooperation was largely achieved through activating both social norms, with indirect effects accounting for 77.20% of the total effect: descriptive norms contributed 53.08% and injunctive norms 24.12% (Figure 3 [Figure 3: see original paper]). The difference between the two indirect effects was not significant (BootSE = 0.09, BootLLCI = -0.01, BootULCI = 0.30), suggesting similar mediating roles in Experiment 2.

Using the second norm activation measure for robustness checks yielded similar results: descriptive norms showed significant indirect effects (Effect = 0.76, BootSE = 0.21, BootLLCI = 0.38, BootULCI = 1.21); injunctive norms showed significant indirect effects (Effect = 0.33, BootSE = 0.16, BootLLCI = 0.05, BootULCI = 0.68); and the direct effect was nonsignificant (Effect = -0.01, SE = 0.41, t = -0.01, p = 0.989).

Analyzing participants' cooperation in Stage 2' s public goods game deepens our understanding of punishment' s spillover effects. Results showed that the punishment group not only exhibited significantly higher cooperation than controls in the dictator game (similar to Stage 1) but also in the different public goods game context ($M = 5.24$, $SD = 5.70$, $n = 90$ vs. $M = 3.76$, $SD = 4.23$, $n = 89$) ($t = 1.97$, $p = 0.050$, $d = 0.30$, 95% CI = [0.001, 0.592]). This indicates that punishment' s spillover effect appears not only in contexts similar to the original but also in completely different situations. Comparing cooperation levels across the two games helps understand the spillover mechanism. Since dictator and public goods games are different contexts, we first standardized cooperation levels. Following Peysakhovich and Rand (2016) and Rand et al. (2014), we set

allocating 10 of 20 tokens to the recipient as the maximum (1) in the dictator game (highest cooperation) and 0 tokens as the minimum (0) (lowest cooperation). Similarly, we set contributing all 20 tokens to the public account as 1 (highest cooperation) and 0 tokens as 0 (lowest cooperation) in the public goods game. Analysis showed no significant difference in cooperation between games for the control group (dictator game: M = 0.19, SD = 0.21, n = 89; public goods game: M = 0.24, SD = 0.26, n = 89) (t = 1.53, p = 0.127), suggesting the games themselves did not affect cooperation. In contrast, the punishment group showed significantly higher cooperation in the dictator game (M = 0.36, SD = 0.28, n = 90) than in the public goods game (M = 0.26, SD = 0.28, n = 90) (t = 2.35, p = 0.020, d = 0.35, 95% CI = [0.06, 0.65]).

These results further confirm punishment' s spillover effect while also indicating that cooperation enhancement through norm activation, though transferable across contexts, is weaker in different contexts than in similar ones. This can be explained by Rand et al.' s (2014) social heuristics hypothesis: real-life interactions are often non-anonymous and repeated (Dreber et al., 2008; Rand et al., 2016). In the long run, cooperation is a more advantageous strategy, which people internalize as a cooperative heuristic and apply intuitively across contexts. However, contextual differences trigger conscious deliberation, through which people may realize cooperation is not optimal for self-interest in the new context (Peysakhovich & Rand, 2016). In other words, deliberation inhibits cooperation in new contexts. In Experiment 2, when transitioning from Stage 1' s dictator game to Stage 2's public goods game, participants needed to think to understand similarities and differences, and this deliberation reduced cooperation in the public goods game. In contrast, Stage 2' s dictator game was essentially the same as Stage 1, allowing intuitive responses without deliberation, resulting in higher cooperation.

4. Experiment 3: Spatial Spillover Effect of Punishment

Experiment 2 verified the temporal spillover effect of third-party punishment: it increased defectors' cooperation in subsequent new contexts through norm activation, even when punishment mechanisms were absent. Experiment 3 further explores whether punishment' s norm-signaling function can spill over to bystanders or potential defectors (spatial spillover effect) and compares the mechanisms of the two norms to answer Research Questions 3 and 4.

4.1 Participants

A power analysis using G*Power 3.1 with f = 0.25 and $\alpha = 0.05$ indicated that at least 158 participants were needed to achieve 95% statistical power (1-β). The actual sample consisted of 160 undergraduate students from various majors (mean age = 21.9 $\pm$ 1.93 years; 42.50% female). Their majors were distributed as: science and engineering (34.38%), social sciences (28.13%), humanities (26.25%), and arts and others (11.25%). Participants read written instructions and signed informed consent forms before the experiment.

4.2 Design and Procedure

Experiment 3 used a 2 (before vs. after observation) × 2 (defection group vs. norm group) mixed design. Participants were told they would watch one round of a dictator game with 3 members and needed to calculate each member' s earnings after the game. After learning the rules (dictators had 20 tokens and could allocate any integer 0–10 to recipients, who could not intervene, but third parties could pay 2 tokens to reduce unfair dictators' payoffs by 6 tokens), participants were randomly assigned to two conditions (80 in defection group, 80 in norm group). All participants estimated: (1) the percentage of dictators allocating 0, 1, 2···10 tokens to recipients; (2) the percentage approving of these allocations; (3) how much they would allocate if they were dictators (0–10), with explicit instructions that no punishment would occur. After these estimates, participants watched one round of the dictator game on their computer screens: the defection group saw a dictator allocate 20% to the recipient and receive punishment, while the norm group saw a 5:5 allocation with no punishment. Participants then calculated earnings and made the same three estimates again about the game they had just observed.

As in Experiment 2, we calculated norm activation levels in two ways: first, using weighted averages of items (1) and (2) for descriptive and injunctive norms; second, using estimates of the percentage of dictators allocating 7, 8, 9, and 10 tokens. We primarily used the first definition for hypothesis testing and the second for robustness checks. Cooperation level was measured by item (3). After these steps, the experiment ended, participants were debriefed and paid. Earnings consisted of a 10 RMB show-up fee plus tokens from a randomly selected allocation.

4.3 Results and Discussion

Using the first norm activation measure, we found no significant differences in descriptive norms, injunctive norms, or cooperation levels across gender ($t = 0.45$–1.51, $p = 0.133$–0.652) or major ($F = 0.08$–1.04, $p = 0.374$–0.972), and no significant correlations between age and these variables ($r = -0.05$ to -0.03, $p = 0.420$–0.602). A 2$×$2 mixed-design ANOVA with group (defection, norm) and round (before, after observation) revealed significant main effects for both factors and a significant interaction (Table 3). Post-hoc comparisons (Figure 4 [Figure 4: see original paper]) showed that the defection group' s cooperation after observing punishment ($M = 4.54$, $SD = 2.59$, $n = 80$) was significantly higher than before observation ($M = 2.30$, $SD = 2.37$, $n = 80$) ($SE = 0.42$, $p < 0.001$, 95% CI = [1.42, 3.06]) and significantly higher than the norm group' s post-observation level ($M = 2.87$, $SD = 2.73$, $n = 80$) ($SE = 0.42$, $p < 0.001$, 95% CI = [0.85, 2.49]). The norm group showed no significant difference between pre-observation ($M = 2.80$, $SD = 2.82$, $n = 80$) and post-observation ($SE = 0.42$, $p = 0.855$, 95% CI = [-0.90, 0.74]), and no significant difference existed between groups before observation ($SE = 0.42$, $p = 0.235$, 95% CI = [-0.32, 1.31]). These results indicate that observing punishment significantly increased bystanders' co-

operation, demonstrating that punishment' s cooperation-enhancing effect spills over to observers. This spillover effect was not due to repeated measurement, as the norm group' s cooperation did not change significantly across observations, providing an affirmative answer to Research Question 3.

Notably, the two groups observed essentially two faces of the same norm: compliance without punishment (norm group) versus violation with punishment (defection group). Yet these different presentations produced dramatically different effects, suggesting that displaying punished violations may make people more aware of social norms and effectively change behavior patterns than showing normative behavior alone. This finding echoes Cialdini et al. (1990): a small amount of litter on the ground activated people' s norm awareness and increased pro-environmental behavior more than a completely clean scene. This may be because violations remind people of the norm' s existence, while minimal violations (Cialdini et al., 1990) or punished violations (this study) signal widespread disapproval, thereby promoting cooperation. This has implications for policy practices aimed at enhancing cooperation.

Table 3 Results of Two-Way ANOVA

Source	F	p	^{2}
Group	11.23	0.001	0.034
Round	16.45	<0.001	0.050
Group × Round	9.87	0.002	0.030

Note: R^2 = 0.094 (adjusted R^2 = 0.086).

Figure 4 Multiple Comparisons of Cooperation Behavior

After observing punishment, the defection group showed significantly higher descriptive norms (M = 3.37, SD = 2.20) than the norm group (M = 2.98, SD = 1.89) (t = 2.30, p = 0.023, d = 0.36, 95% CI = [0.06, 0.73]), and significantly higher injunctive norms (M = 4.97, SD = 2.77) than the norm group (M = 4.18, SD = 2.51) (t = 3.32, p = 0.001, d = 0.52, 95% CI = [0.32, 1.27]). This suggests that increased cooperation may result from norm activation through observing punishment. We further tested whether norm activation mediated the relationship between observing punishment and cooperation, using group (whether seeing punishment) as the independent variable, descriptive and injunctive norms as mediators, and cooperation level as the dependent variable. Bootstrap analysis revealed partial mediation (Figure 5 [Figure 5: see original paper]). The direct effect of seeing punishment on cooperation was significant (Effect = 1.23, SE = 0.43, t = 2.89, p = 0.004, LLCI = 0.39, ULCI = 2.08). The indirect effect through descriptive norms was significant (Effect = 0.30, BootSE = 0.16, BootLLCI = 0.04, BootULCI = 0.85). However, the indirect effect through injunctive norms was not significant (Effect = 0.13, BootSE = 0.14,

BootLLCI = -0.16, BootULCI = 0.40), primarily because injunctive norm activation did not significantly change cooperation, despite punishment significantly affecting injunctive norm activation.

Figure 5 Mediating Role of Descriptive Norms

Robustness checks using the second norm activation measure yielded similar results: descriptive norms ($t = 3.96$, $p < 0.001$) and injunctive norms ($t = 4.89$, $p < 0.001$) were both significantly higher in the defection group after observation. Descriptive norms showed significant indirect effects (Effect = 0.41, BootSE = 0.19, BootLLCI = 0.10, BootULCI = 0.83), while injunctive norms did not (Effect = 0.30, BootSE = 0.20, BootLLCI = -0.06, BootULCI = 0.72), though the direct effect remained significant (Effect = 1.43, SE = 0.49, $t = 2.90$, $p = 0.004$). These results suggest that the two norms have different mediating mechanisms, contrasting sharply with Experiment 2.

Comparing the two norm pathways, the experimental manipulation activated both norms, as confirmed by tests of mean differences. The difference lies in activated descriptive norms increasing cooperation, while injunctive norms did not. One explanation is that people are generally more influenced by descriptive norms (陈思静等, 2015; Cialdini et al., 1991) because descriptive norms involve factual judgments ("What do people do?") while injunctive norms involve value judgments ("What do people think should be done?"), with factual information processed faster than value judgments (Deutsch & Gerard, 1955). Comparing Experiments 2 and 3 reveals a key difference: both norms showed significant mediation in Experiment 2 with no significant difference, while in Experiment 3, only descriptive norms mediated significantly. We speculate this may reflect differences in personal involvement: Experiment 2 participants personally experienced punishment, while Experiment 3 participants merely observed others being punished. Thus, personal involvement was likely higher in Experiment 2. Petty and Cacioppo (1986) argue that injunctive norms have stronger effects when personal involvement is high, which may explain the difference: higher involvement in Experiment 2 made injunctive norms' effect on cooperation more pronounced, while low involvement in Experiment 3 rendered injunctive norms' influence nonsignificant.

5. General Discussion

5.1 Theoretical Implications

Extensive literature has examined how third-party punishment inhibits violations and promotes cooperation (e.g., Fehr & Gächter, 2002; Grimalda et al., 2016; Halevy & Halali, 2015). However, less attention has been paid to how this effect occurs, with existing literature primarily adopting an economic perspective that changing defectors' payoff structures is the core mechanism (韦倩, 姜树广, 2013; Carpenter & Matthews, 2004; Nelissen & Mulder, 2013; Rand et al., 2010). This explanation contradicts a key finding in behavioral economics: individuals are not necessarily rational decision-makers with purely self-regarding prefer-

ences (Kahneman, 2011; Thaler, 2016), unless we presuppose that defectors are always rational economic agents.

Unlike the economic perspective, Chen et al. (2015) and Chen et al. (2020) view third-party punishment as a norm-signaling mechanism that enhances cooperation by activating internalized cooperative norms (Rand et al., 2014) without involving economic interests. However, to strictly conclude this, we must exclude punishment' s economic effects, as punishment typically reduces payoffs in mainstream research. From this logic, Experiment 1 first tested punishment' s norm-signaling function while controlling its economic effects. Results showed that even when punishment losses were smaller than defection gains, third-party punishment still significantly inhibited violations and enhanced cooperation. In other words, even defectors are not necessarily benefit maximizers. The ancient Greek philosopher Socrates famously argued that people do evil out of ignorance (汪子嵩等, 2004). This study partially confirms Socrates' wisdom: often people violate norms simply because they are unaware of them, and activating norm awareness can significantly reduce selfish behavior, with third-party punishment being an important means of activation. A practical implication is that economic punishment, which reduces violators' payoffs, may be inefficient because it incurs costs that may reduce collective net benefits (Dreber et al., 2008). In policy practice, we must identify whether violations stem from lack of norm awareness or pure self-interest; blanket punishment without distinguishing motivations may reduce social efficiency.

Second, Experiment 1' s results can explain several previous findings. Rand et al. (2009) and Fehr and Rockenbach (2003) found that punishment legitimacy greatly influences punished individuals' cooperation. Pure economic perspectives cannot fully explain this, but viewing punishment as norm-signaling resolves the issue: as a norm signal, punishment must itself conform to norms —that is, possess moral legitimacy. Punishment violating norms cannot have norm-signaling effects and thus loses its positive cooperation-promoting role. A corollary is that if punishment has no economic function, we can largely exclude illegitimate motivations (e.g., punishment to gain relative advantage). In such cases, following Experiment 1, we should observe cooperation-promoting effects. Indeed, researchers have noted that verbal condemnation of violations (also called social or moral punishment) can have similar effects (Noussair & Tucker, 2005) without causing monetary losses, sometimes proving more effective than economically motivated punishment (Wu et al., 2016). Experiment 1 explains this phenomenon: although verbal condemnation does not change payoffs, like third-party punishment, it signals norm existence, while largely excluding self-serving illegitimate motivations, thereby effectively reducing selfish behavior. Some researchers argue verbal condemnation can also be explained economically: van den Berg et al. (2012) suggest costs take various forms, and though verbal condemnation may not increase monetary defection costs, it may increase interpersonal costs, thus reducing defector payoffs. However, laboratory verbal condemnation is typically mild (e.g., "I think your allocation is unfair" (Nelissen & Mulder, 2013) or "So-and-so only cares about themselves"

(崔丽莹等, 2017)) and often occurs in anonymous settings (陈思静, 徐烨超, 2020), making it unlikely to cause substantial interpersonal damage. Therefore, we argue Experiment 1' s results better explain how verbal condemnation enhances cooperation.

Third and most importantly, this paper proposes a social norm perspective on how third-party punishment maintains long-term, large-scale human cooperation. Economic perspectives cannot explain this because rational individuals' past punishment experiences would not improve behavior in new contexts unless punishment mechanisms remained present. Yet as Shreedhar et al. (2018) note, ubiquitous punishment would impose enormous social costs. Experiments 2 and 3' s two spillover effects explain why third-party punishment can maintain extensive cooperation: Experiment 2 shows punishment' s norm activation not only inhibited violations in the current context but also enhanced cooperation in subsequent different contexts—what we call "vertical spillover effect." Experiment 3 shows this spillover effect occurs not only across contexts for the same individual but also among bystanders who merely observe punishment—what we call "horizontal spillover effect." These results mean maintaining large-scale cooperation does not require ubiquitous punishment, as a specific punishment' s effects extend temporally and spatially: punishment serves largely as a norm signal, so punished individuals or bystanders activate internal norms that inhibit potential violations and maintain cooperation at relatively high levels without constant external monitoring and punishment. Thus, this study' s spillover effects provide new theoretical insights into how punishment sustains extensive human cooperation. We note that our interpretation is not unique: Gintis and Fehr (2012) propose an alternative explanation that punishment' s cooperation-enhancing effect still relies on reducing defection payoffs, but punishment need not cause actual losses—merely the fear of potential losses can be effective. This view can economically explain why occasional punishment maintains large-scale cooperation. However, to exclude this competing hypothesis, Experiments 2 and 3 explicitly informed participants that no punishment would occur regardless of violations. This largely eliminates the possibility that fear of potential punishment increased cooperation. Thus, Experiments 2 and 3 support the norm-signaling explanation over the deterrence explanation. Of course, as one of the biggest puzzles in social sciences (Bear & Rand, 2016), cooperation may have no single explanation. Our norm-based explanation and Gintis and Fehr' s (2012) theory may not be mutually exclusive but complementary, providing a more complete answer to cooperation' s evolution.

5.2 Limitations and Future Directions

Despite meaningful results, this study has limitations. First, we used economic punishment as the mainstream research paradigm, where both punishment and defection costs are monetary. While this yields clear conclusions, real-life costs take various forms (Guala, 2012) with different effects (陈思静等, 2020). Whether our conclusions hold when costs are non-monetary (e.g., time, effort, interper-

sonal resources) warrants further exploration.

Second, although we clearly identified two spillover effects, the short interval between sessions (less than 1 hour) prevents us from determining whether these effects persist over longer periods (e.g., weeks or months). Conducting such experiments over longer time spans would strengthen our conclusions.

Third, in Experiment 3 we compared two norm presentation methods (compliance without punishment vs. violation with punishment). From a broader theoretical perspective, a more meaningful comparison might be between "compliance rewarded" and "violation punished." However, as our focus was on punishment' s role in inhibiting violations and promoting cooperation, we did not analyze this comparison, leaving it for future research.

Finally, we observed that punishment' s spillover effect diminished during transfer, suggesting limits to third-party punishment' s ability to maintain large-scale cooperation. This aligns with observations that punishment's effectiveness weakens as community size increases (Greif, 1993) and indicates that bottom-up third-party punishment alone may be insufficient to explain extensive human cooperation. Incorporating other mechanisms such as top-down pool punishment (Baldassarri & Grossman, 2011) or coordinated punishment (韦倩等, 2019) may help better understand human cooperation.

---

[1] To avoid emotional connotations that might affect participants, terms like "player," "enforcer," or "punishment" were replaced with "Role A/B/C/D," "Role E," and "deduction" in the experiment.

[2] In Experiment 1, participants could only punish defectors (those retaining 10 tokens) to avoid antisocial punishment—punishment of cooperators (Herrmann et al., 2008)—interfering with results.

[3] In instructions, dictator, recipient, and third party were labeled Role A, Role B, and Role C, respectively.

[4] The 121 participants who did not violate norms in Stage 1 did not continue to Stage 2. To ensure smooth experimentation, they were told they would complete a round testing "the effect of foreign language thinking on norm perception," involving reading an English passage about gift-exchange norms in an African tribe and answering questions.

[5] Previous literature shows stable cross-cultural perceptions of what constitutes violation/cooperation: allocating less than ~30% is considered a violation (Csukly et al., 2011; Fehr & Fischbacher, 2003), with some arguing this stability has biological foundations (Wallace et al., 2007). With a 20-token endowment, 6 tokens is the threshold, so allocations above 6 tokens are considered cooperative.

References

Alkan, H. I. (2020). A challenge to homo economicus: Behavioral economics. In I. Akansel (Ed.), *Examining the relationship between economics and philosophy* (pp. 176–195). Hershey, PA: IGI Global.

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences of the United States of America, 108*(27), 11023–11027.

Balliet, D., Mulder, L. B., & van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences of the United States of America, 113*(4), 936–941.

Bicchieri, C., Dimant, E., & Xiao, E. T. (2018). Deviant or wrong? The effects of norm information on the efficacy of punishment (PPE Working Papers 0016). Philadelphia, PA: Philosophy, Politics and Economics of University of Pennsylvania.

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

Bingham, P. M. (1999). Human uniqueness: A general theory. *The Quarterly Review of Biology, 74*(2), 133–169.

Camerer, C. F., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science, 311*(5757), 47–52.

Carpenter, J. P., & Matthews, P. H. (2004). Why punish? Social reciprocity and the enforcement of prosocial norms. *Journal of Evolutionary Economics, 14*(4), 407–429.

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

Chen, S. J., He, Q., & Ma, J. H. (2015). The influence of third-party punishment on cooperation: An explanation of social norm activation. *Acta Psychologica Sinica, 47*(3), 389–405.

Chen, S. J., Hu, H. M., & Yang, S. S. (2020). Payment vs. retaliation: Impact of cost form on third-party punishment. *Journal of Psychological Science, 43*(2), 416–422.

Chen, S. J., & Xu, Y. C. (2020). Warmth and competence: Impact of third-party punishment on punishers' reputation. *Acta Psychologica Sinica, 52*(12), 1436–1451.

Cialdini, B., Kallgren, A., & Reno, R. (1991). A focus theory of normative conduct. *Advances in Experimental Social Psychology, 24*, 201–234.

Cialdini, B., Reno, R., & Kallgren, A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Cialdini, B., & Trost, M. (1998). Social influence: Social norms, conformity, and compliance. In T. Gilbert, T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 2, pp. 151–192). Boston, MA: McGraw-Hill.

Csukly, G., Polgár, P., Tombor, L., Réthelyi, J., & Kéri, S. (2011). Are patients with schizophrenia rational maximizers? Evidence from an ultimatum game study. *Psychiatry Research, 187*(1-2), 11–17.

Cui, L. Y., He, X., Luo, J. L., Huang, X. J., Cao, W. J., & Chen, X. M. (2017). The effects of moral punishment and relationship punishment on junior middle school students' cooperation behaviors in public goods dilemma. *Acta Psychologica Sinica, 49*(10), 1322–1333.

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology, 84*, 103800.

de Kwaadsteniet, E. W., van Dijk, E., Wit, A., de Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin, 33*(12), 1648–1660.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51*(3), 629–636.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don' t punish. *Nature, 452*(7185), 348–351.

Fang, J., & Zhang, M. Q. (2012). Assessing point and interval estimation for the mediating effect: Distribution of the product, nonparametric Bootstrap and Markov Chain Monte Carlo methods. *Acta Psychologica Sinica, 44*(10), 1408–1420.

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

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

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

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

Fehr, E., & Williams, T. (2018). Social norms, endogenous sorting and the culture of cooperation. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.

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

Forquesato, P. (2016). Social norms of work ethic and incentives in organizations. *Journal of Economic Behavior & Organization, 128*, 231–250.

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science, 322*(5907), 1510.

Gintis, H., & Fehr, E. (2012). The social structure of cooperation and punishment. *Behavioral and Brain Sciences, 35*(1), 28–29.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology, 213*(1), 103–119.

Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *The American Economic Review*, 525–548.

Grimalda, G., Pondorfer, A., & Tracer, D. P. (2016). Social image concerns promote cooperation more than altruistic punishment. *Nature communications, 7*, 12288.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(1), 1–15.

Halevy, N., & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences of the United States of America, 112*(22), 6937–6942.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review, 91*(2), 73–78.

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

Huang, S. A., & Zhang, S. (2013). How did cooperative behavior evolve: A summary and review. *Social Sciences in China, 7*, 79–91.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Macmillan.

Lois, G., & Wessa, M. (2019). Creating sanctioning norms in the lab: The influence of descriptive norms in third-party punishment. *Social Influence, 14*(2), 50–63.

McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences, 3*, 147–151.

Nelissen, R. M., & Mulder, L. B. (2013). What makes a sanction "stick" ? The effects of financial and social sanctions on norm compliance. *Social Influence, 8*(1), 70–80.

Noussair, C., & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry, 3*(3), 649–660.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science, 314*(5805), 1560–1563.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology, 194*(4), 561–574.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York, NY: Springer.

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

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology and Evolution, 30*(2), 98–103.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science, 27*(9), 1192–1206.

Rand, D. G., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology, 265*(4), 624–632.

Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General, 145*(4), 389–396.

Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science, 325*(5945), 1272–1275.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*, 3677.

Shreedhar, G., Tavoni, A., & Marchiori, C. (2018). Monitoring and punishment networks in a common-pool resource dilemma: Experimental evidence (GRI Working Papers 292). London, England: Grantham Research Institute on Climate and the Environment.

Sood, S., Kostizak, K., Lapsansky, C., Cronin, C., Stevens, S., Jubero, M., ··· & Obregon, R. (2020). ACT: An evidence-based macro framework to examine how communication approaches can change social norms around Female Genital Mutilation. *Frontiers in Communication, 5*, 29.

Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review, 106*(7), 1577-1600.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35-57.

van den Berg, P., Molleman, L., & Weissing, F. J. (2012). The social costs of punishment. *Behavioral and Brain Sciences, 35*(1), 42-43.

Voisin, D., Girandola, F., David, M., & Aim, M. A. (2016). Self-affirmation and an incongruent drinking norm: Alcohol abuse prevention messages targeting young people. *Self and Identity, 15*(3), 262-282.

Wallace, B., Cesarini, D., Lichtenstein, P., & Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences of the United States of America, 104*(40), 15631-15634.

Wang, Z. S, Fan, M. S., Chen, C. F., & Yao, J. H. (2004). *A History of Greek Philosophy (Volume 2)*. Beijing: People' s Publishing House.

Wei, Q., & Jiang, S. G (2013). How is social cooperative order possible: Exploring mysteries. *Economic Research Journal*, (11), 140-151.

Wei, Q., Sun, R. Q., Jiang, S. G., & Ye, H. (2019). Coordinated punishment and the evolution of human cooperation. *Economic Research Journal*, (7), 174-187.

Wen, Z. L., & Ye, B. J. (2014). Different methods for testing moderated mediation models: Competitors or backups? *Acta Psychologica Sinica, 46*(5), 714-726.

Wu, J., Balliet, D., & van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports, 6*, 23919.

Xie, D. J., & Su, Y. J. (2019). The evolutionary and cognitive mechanisms of third-party punishment. *Journal of Psychological Science, 42*(1), 216-222.

Note: Figure translations are in progress. See original paper for figures.