Social influence on third-party punishment: An experiment

Social Influence on Third-Party Punishment: An Experiment[☆]

Marco Fabbri[b,c], Emanuela Carbonara[a,d]

[a] *Department of Economics, University of Bologna*
[b] *Institute of Private Law, Erasmus University Rotterdam*
[c] *Rotterdam Institute of Law and Economics, Erasmus University Rotterdam*
[d] *SAIS Europe, Johns Hopkins University*

Abstract

We study the effect of social influence on agents' decisions to engage in costly decentralized third-party punishment. In a laboratory experiment, participants play a modified Dictator game with third-party punishment and we elicit punishment decisions both in isolation and after providing information about peers' average punishment. Results show that social influence is a major driver of third-party punishment: after receiving information on peers' average punishment, participants revise initial punishment choices and seek conformity. Social influence effects are stronger when peers punished more than the individual punisher and conformity implies revising punishment upward. Adding to the information on peers' punishment the possibility for other participants to sanction or reward with an emoticon the choices of the individual punisher does not change results. Our findings contrast with the predictions of major theories of social preferences and are only explained by special cases of models incorporating aversion to norm-breaking.

Keywords: Conformism, Decentralized Enforcement, Dictator Game, Peer Effects, Social Norms.

[☆] Principal investigator and contact author: Marco Fabbri, fabbri@law.eur.nl. We thank Maria Bigoni, Stefania Bortolotti, Marco Casari, Robert Cooter, Andrea Geraci, Sven Hoeppner, Francesco Parisi, Matt Rabin, Matteo Rizzolli, Enrico Santarelli, Louis Visscher, Roberto Weber and seminar participants at Erasmus University Rotterdam, University of Bologna, University of Trento, the Annual conference of the European Association of Law and Economics 2014, the Annual conference of the Italian Society of Law and Economics 2014, the Annual conference of the American Law and Economics Association 2015 for helpful suggestions. Lorenzo Golinelli and Stefano Rizzo provided valuable research assistance. This paper is part of the SONIC project on behavioral and experimental economics at the University of Bologna, Miur Firb grant no. RBFR084L83. We are also grateful to the Alfred P. Sloan foundation for financial support. Marco Fabbri thanks the Behavioural Approach to Contract and Tort - BACT research program at Erasmus University Rotterdam for financial support during the completion of the project.

JEL Classification: D03, K04
PsycINFO Classification: 3020

1. Introduction

In this paper we study the impact of social influence (i.e. the influence of comparison peers' behavior when no material spillover between agents exists) on agents' willingness to engage in costly decentralized third party punishment. Third party punishment is a fundamental institution for the enforcement of social norms determining the cohesion and the functioning of human societies. Understanding how third party punishment is influenced by peer effects is important because governments can use social influence to shape social norms of behavior and to facilitate the enforcement of legal norms. This paper investigates the link between social influence and third party punishment by providing experimental evidence and by comparing the experimental results with the predictions of existing theories of third party punishment.

Third-party punishment is the sanctioning of a wrongdoer that involves the action of peers not directly affected by the consequences of the rule violation. It is defined in contrast to 'second-party punishment', where the sanctioning individual is directly harmed by the wrongdoing (Bendor and Mookherjee, 1990; Gintis, 2000). In this paper we focus on decentralized third party punishment occurring within a horizontal relationship among peers (generally all private citizens), as opposed to centralized third-party punishment, i.e., the public enforcement of legal rules via the legal system

that is characterized by a vertical relationship between the state and its private citizens.

Understanding the link between social influence and third party punishment is important for three reasons. First, third party punishment plays a key role in establishing and enforcing social norms in large organizations characterized by a predominance of one-shot and anonymous interactions (Balafoutas et al., 2014; Mathew and Boyd, 2011). In fact, scholars argue that second party punishment strategies are not evolutionarily stable in iterated pairwise interactions, contrary to strategies based on third party punishment that are stable (Bendor and Swistak, 2001). For this reason third party punishment is considered a fundamental ingredient of social cohesion (Fehr and Fischbacher, 2004). If third parties respond to peer effects in their decisions to enforce social norms, a central authority can effectively influence the process of norm production by strategically disclosing or hiding information regarding population aggregate behaviors (Benabou and Tirole, 2011).

Second, public enforcement is often burdened by limitations that might compromise its efficiency and efficacy. For instance, think about the problems for public authorities to contrast difficult-to-monitor criminal activities, such as terrorism, sexual violence, racial discrimination, or bullyism. Governments are thus recurring more and more to forms of 'shared enforcement', where the enforcement of legal norms is delegated or complemented by private citizens when transaction costs and better knowledge of situational factors make decentralized actions preferable to traditional legal interventions (Kaplow

and Shavell, 2007).[1] Social influence has been indicated as an effective and inexpensive tool that policymakers can use to increase citizens' engagement in decentralized enforcement (Dolan et al., 2012).

Third, evidence that social influence affects third party punishment has important implications for theories of social preferences. As discussed in a recent contribution by Thöni and Gächter (2015), who report evidence of peer effects on pro-social cooperation, several established theories of social preferences assume fixed preferences and cannot account for social influence effects. We deem it important to verify whether the existence of peer effects extends also to third-party punishment of distributional norms violation and to discuss the theoretical implications of these findings.

We measure social influence using the framework of a modified dictator game with third-party punishment. In a laboratory experiment individual punishment choices are elicited in isolation and after providing information regarding peers' punishment decisions. The paper explores and combines two possible channels through which social influence affects behavior: the so called *informational* social influence, consisting in the "need to be right",

[1] Recently there have been several examples of policies aimed at increasing third-parties' involvement. The most famous is probably the anti-terrorism campaign *'If you see something, say something'* promoted by the US Department of Homeland Security (https://www.dhs.gov/see-something-say-something). Another example is the campaign *'Bringing in the Bystander'* promoted in the UK by the National Sexual Violence Resource that employs advertising explicitly encouraging third-parties' intervention in situations of violence against women: 'Using a bystander intervention approach combined with a research component, this program assumes that everyone has a role to play in prevention [...] The Know Your Power campaign is the social marketing component of Bringing in the Bystander'(http://www.nsvrc.org/bystander-intervention-campaigns-and-programs.). Other examples of decentralized enforcement programs implemented by the US Government and by several universities against race and sex discrimination include the *Step Up!* (http://stepupprogram.org/) and *Green Dot* (http://www.livethegreendot.com/) campaigns.

and the *normative* social influence, that is the "need to be liked" by others (Deutsch and Gerard, 1955). In the 'informational' treatment we simply provide information regarding peers' punishment choices, making it clear that the decision to revise punishment will not be observed by peers. In the 'info+normative' we add to the informational treatment also a normative social influence components: after the punishment revision, peers will observe the individual decision of the third party punisher and they will send her a smiley or frowny emoticon.

While standard rationality predicts no third party punishment, several previous studies show that bystanders unaffected by the consequences of the rule violation engage in positive punishment even in one-shot interactions. Our experimental design rules out any material or strategic incentive for third parties to revise their initial punishment decisions after learning how their peers punished. Therefore, if participants revise punishment in order to conform to their peers we have evidence of social influence effects.

Results of our experiment show that social influence is a major driver of third-party punishment. After receiving information on peers' average punishment, third parties modify their punishment choices in order to seek conformity. Surprisingly, and in contrast with Thöni and Gächter (2015), social influence effects are the strongest when conformity implies a revision upward of the individual punishment and the consequent reduction of individual payoffs. In our sample, the further addition of a normative social influence component to informational social influence does not affect punishment decisions.

We then put our results in the context of the existing theoretical literature

on social preferences. While several approaches predict third-party punishment, peer effects are inconsistent with most existing theories. Models characterized by *distributional concerns* (inequity aversion, altruism) as well as models of (strong) reciprocity predict no peer effects. Theories of conformity, as well as theories of social norm abidance are best equipped to account for peer effects with third-party punishment. However, the most popular models assume that individuals react to either the number or the percentage of others conforming to the norm. In our experiment, we do not provide this particular information to participants, yet significant peer effects exist. We thus develop two novel models that explain peer effects in our specific setting. The first model introduces a variation in the utility function used by López-Pérez (2008). The second model assumes that players don't know the prevailing social norm regarding punishment and use the information about average punishment as a *focal point*. Both models result in theoretical predictions consistent with our experimental findings.

Before presenting the experimental design and procedure in section 3, we review the related literature in the next section. Section 4 reports the results, whereas section 5 compares our experimental findings with the predictions of theories of social preferences and proposes two novel extensions (technical details and proofs relative to this section are reported in Appendix A). Section 6 discusses the implications of our findings and suggests possible directions for future research.

2. Related Literature

Both social influence (Bond and Smith, 1996; Cialdini and Goldstein, 2004) and third party punishment (Balafoutas and Nikiforakis, 2012; Henrich et al., 2006) are well established empirical phenomena, supported by robust field and experimental evidence. Studies show that social influence affects decisions in a variety of settings.[2] Similarly, scholars have extensively studied the determinants of third-party punishment.[3] This paper contributes to the burgeoning literature that studies the link between peer effects and social preferences. Our study is the first to connect the research on social influence effects with the study of third-party punishment. Several papers study social influence effects on distributional preferences by focusing on the decision to share the endowment in a dictator game (Bicchieri and Xiao, 2009; Cason and Mui, 1998; Krupka and Weber, 2009) or ultimatum game (Ho and Su, 2009). Such contributions share with our experiment a framework that looks at distributional norms when reciprocity or payoff interdependence are absent. Our paper differs from these studies because our analysis is focused

[2]Researches have found significant social influence effects in settings such as teenage pregnancy (Akerlof et al., 1996), residential energy usage (Ayres et al., 2012), judicial voting patterns (Sunstein et al., 2006), labor productivity (Mas and Moretti, 2009), tax evasion (Fortin et al., 2007; Galbiati and Zanella, 2012) and other criminal activities (Glaeser et al., 1996; Falk and Fischbacher, 2002).

[3]For instance, Marlowe et al. (2008) suggest that societies characterized by complex organizations and subject to frequent market interactions engage in higher level of third-party punishment compared to less articulated ones. In a laboratory experiment, Kurzban et al. (2007) find that participants increase punishment when observers are present, arguing that third-party punishment is influenced by the so called 'audience effect', that is, the tendency to change one's behavior when in the presence of others. Subsequent works confirm that anonymity has a causal effect on third-party punishment (Piazza and Bering, 2008), suggesting that the third-party decisions to sanction wrongdoers is influenced by a cost-dependent reputation effect (Nelissen, 2008) and by emotions (Nelissen and Zeelenberg, 2009).

on the behavior of punishers rather than on the decisions of the dictator. The focus on third-party punishment and the setting of a dictator game distinguish our study from Falk et al. (2013), which looks at peer effects on voluntary cooperation in a public good game and on the ability to coordinate in a coordination game, from Gächter et al. (2012) and Thöni and Gächter (2015), which both study reciprocity in a three-person gift exchange game, and from Mittone and Ploner (2011), which studies reciprocity in a trust game.

A second contribution of this study consists in reviewing the predictions of established theories of social preferences to check whether they resonate with our empirical evidence. We perform this comparison looking at a range of theories that incorporate different motivations (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Cox et al., 2007; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Fehr and Schmidt, 1999; Levine, 1998; López-Pérez, 2008; Rabin, 1993).

3. The Experiment

The Game. Our experiment is based on a variation of the dictator game with third-party punishment. There are three possible roles: receiver (participant A), dictator (participant B) and third-party (participant C). The dictator has the opportunity to take part of or all the endowment from a passive receiver as in Cox et al. (2007) and List (2007). Third parties are then given the chance to impose a costly punishment to the dictator. Each participant receives an endowment of 30 tokens. Participant B has the opportunity to take between 0 and 30 tokens (in multiples of 5) from A. Therefore, there

are 7 possible actions available to B. Participants A do not take any action during the game. After participant B has made his decision, participant C has the opportunity to impose a costly punishment on B. Particularly, C could use up to 20 units of her initial endowment to reduce B's payoff. For each token used by C, B's payoff is reduced by 4 tokens.

Experimental Design. Participants do not know which role has been assigned to them until the end of the experiment. Each participant is required to make decisions both as a dictator and as a punisher (a receiver does not make decisions). Instructions stress that the only payoff-relevant decisions for a participant are those made in the role that will be randomly assigned at the end of the experiment. Decisions made in any other role are irrelevant for the computation of payoffs and will be discarded. To make sure that this feature of the design does not generate confusion, each participant had to answer a set of control questions correctly before starting the experiment. The elicitation of participants' punishment decisions under a 'veil of ignorance' regarding their actual role does not affect the validity of our results. As explained in the next section, our analysis focuses on the change in punishment between periods rather than on the absolute levels of third-party punishment. Even if uncertainty about the actual role somehow affects punishment choices (for instance, reducing punishment levels because of a lack of moral anger), this effect is bound to be the same across treatments, since participants are exposed to this identical design feature in all the three treatments. Moreover, the elicitation of the participants' choices both as a dictator and as a punisher has two advantages. First, it increases the

number of observations. Second, it makes it possible to combine the decisions taken in the two roles by each participant in order to classify different 'behavioral types' and so to verify whether different types respond to peer effects asymmetrically.[4]

[Figure 1 approx. here]

Figure 1 describes our experimental design. In the initial stage ('Beliefs' stage), we elicit the participants' beliefs about their peers' punishment choices by means of an incentivized coordination game.[5] Each participant is presented with a hypothetical situation, identical to the taking-dictator game with third-party punishment described above. The participant is then requested to guess the number of punishment tokens that the third-party punisher will use for each of the seven possible dictator's taking rates. After the participant states her guesses, one of the seven taking rates is randomly

[4]We identify four types of participants: (i) Altruist; (ii) Selfish; (iii) No Harm; (iv) Spiteful. An 'Altruist' is a participants who takes nothing when acting as a Dictator and punishes at least some of the Dictator's taking decisions when acting as Punisher. A 'Selfish' is a pure payoff-maximizer, always taking the maximum when acting as Dictator and punishing zero when acting as third-party punisher. A 'No-Harm' is a subject who takes zero as a Dictator, like an Altruist, but also does not punish when acting in the role of punisher. Finally, a 'Spiteful' type takes the entire endowment from the passive subject when acting as a Dictator (therefore showing little social concern) and sacrifices a positive fraction of his endowment to punish and reduce the Dictators' payoff (showing a positive utility from punishing others). We repeated the analysis presented in the next section excluding Selfish or No Harm types in our sample. The results remain qualitatively the same. We also verified that the effects of social influence on the punishment decisions of Altruist and Spiteful types in our sample do not differ significantly from those of the other subjects (results available upon request).

[5]In the main analysis these beliefs are included as a control, since our goal is to estimate peer effects on punishment decisions holding the initial individual expectations regarding peers' punishment choices constant. See section 4

selected. The participant earns 40 tokens if, for the selected taking rate, her choice matches the average answer provided by the other participants in the session, plus or minus one unit. In the second stage of the experiment ('Dictators choice 1' stage), each participant makes a choice as participant B. Therefore, the participant chooses to take one of the seven possible amounts of tokens from the participant C's endowment. In the third stage ('Punishment Period 1' stage), each participant makes seven choices as a participant C. We use the strategy method. For each of the seven dictator's taking rates, the participant C chooses how many tokens she wants to use to punish B. Therefore, while participant C makes seven punishment decisions, the only payoff-relevant choice is the one corresponding to the taking rate actually chosen by the participant B in her group.

Up to this point, the experiment is identical in all treatments. Starting from the 'Information stage' where participants receive some information that changes according to the treatment, we distinguish among three different treatments: the 'informational', the 'info+normative' and the 'control' treatment.

In the informational treatment, each participant learns the average number of tokens used by all the participants to the experimental session to punish dictators in the 'Punishment Period 1' stage. This information is provided separately for each of the seven taking rates. After receiving the information, each participant makes a choice as dictator ('Dictator choice 2' stage) and seven choices as third-party ('Punishment Period 2' stage) exactly as in the Punishment Period 1 stage.

In the info+normative treatment, we add a normative social influence com-

ponent to the informational treatment above. Specifically, as in the informational treatment, a participant receives information about the average number of tokens used by peers to punish for each of the seven taking rates. In addition, each participant is also told that the decisions she is going to take in the Punishment Period 2 stage will be revealed to five peers chosen randomly among the other participants in the experimental session. These five peers then vote to send her either a smiley or a frowny emoticon that will be displayed for one minute on the screen before the experiment ends. The instructions stress that the emoticon has no effect on earnings.[6] In the context of our info+normative treatment, participants C might expect that punishment above or below average could be disapproved by peers. Therefore, this anticipated sanction/reward ('anticipated' since an explicit threat of sanction or prospect of recognition is not even issued) might induce participant C to conform to average punishment levels (Crawford and Klotz, 1999; Xiao and Houser, 2009). After receiving information, as in the informational treatment, each participant makes again a choice as dictator (Dictator choice 2 stage) and seven choices as third-party (Punishment Period 2 stage). Finally, participants observe the punishment choices of some randomly selected peers, vote to send them the smiling or the frowny emoticon, and receive an emoticon appearing on their screen for one minute before the experiment ends.

In addition to these two treatments, we also have a control treatment in which no information about punishment choices is disclosed and participants

[6]Studies show that non-monetary sanctions can have a significant impact on individuals' behavior (Ellingsen and Johannesson, 2008; Masclet et al., 2003).

receive socially irrelevant information. Following Cason and Mui (1998), before the game starts we ask each participant to indicate her day of birth $\in$ [1; 31]. We then take the average and we communicate it to the participants C in the Punishment Period 2 stage of the control treatment. Since we do not disclose the year nor the month of birth, the participant C in the control treatment does not receive socially relevant information on her peers' behavior. In this treatment too, after receiving the information, each participant makes a choice as dictator (Dictator choice 2 stage) and seven choices as third-party (Punishment Period 2 stage) exactly as in the Punishment Period 1 stage.

Payoffs. In all treatments, the per-period payoffs are:

- Π_A = 30 - t
- Π_B = 30 + t - 4*p
- Π_C = 30 - p

where t are tokens taken by B from A and p are punishment tokens used by player C. In order to calculate individual earnings, participants are randomly divided in groups of size 3. Each group consists of one participant A, one B and one C. The final payment for each group follows this procedure:

- In the dictator game, after the second period ends, one of the two periods is randomly selected. This becomes the 'payment period". Earnings from that period only are paid.
- For each participant, earnings from the payment period are added to the earnings that were collected in the beliefs elicitation game.

- Tokens are converted into Euros at a rate of 5 tokens per Euro.
- A 5 euro participation fee is added to total payments.

Participants are informed that payoffs from the game can be negative, in which case the show-up fee is used to cover the losses. No subject in our sample experienced negative payoffs.

The Procedure.. The experiment was programmed using the software z-Tree (Fischbacher, 2007). All the sessions were conducted at the Bologna Laboratory for Experimental Social Sciences (BLESS) at the University of Bologna in November and December 2013 and in September 2016. All sessions were conducted by the same experimenter. The vast majority of participants were graduate and undergraduate students from the University of Bologna. Participants were recruited through the online system ORSEE (Greiner, 2015). In each session, participants were split into 5 groups of 3 participants. Overall, 14 sessions were run, which results in a total of 207 participants (55% female).[7]

At the beginning of each of the three experimental periods, instructions were distributed and read aloud.[8] Participants had additional images and tables summarizing the instructions on their computer screens. No communication

[7] In one session of the informational treatment, there were only 4 groups, for a total of 12 participants. Due to a technical problem, one session registered multiple interruptions and only part of the data were recorded. We excluded observations collected in that session from the analysis.

[8] Original instructions are in Italian and are available upon request. Instructions for the belief elicitation game and for the first period punishment are identical in all of the treatments. A copy of the instructions for the belief elicitation game and for the info+normative treatment, translated in English, are included in the Appendix.

among participants was allowed. Each participant earned on average approximately Euro 11 (including the participation fee). Sessions lasted about 40 minutes, including reading the instructions, answering control questions and completing a brief socio-demographic questionnaire. Each participant took part in one session only. Peers' identities remained unknown even after the end of the experiment. In order to guarantee anonymity, participants were individually and privately paid after the experiment finished.

4. Results

4.1. Initial Dictators' Taking Rate, Punishers' Beliefs and Choices

Table 1 reports mean, median, and standard deviation of the main variables we use in the analysis. Dictators' average taking rate is consistent with results from other comparable experiments (List, 2007; Krupka and Weber, 2013). Consistently with previous research, participants C are willing to use part of their endowment to punish the dictators (Fehr and Fischbacher, 2004). As expected, tables B.8 and B.9 in the appendix show that the mean amount of punishment is virtually 0 when dictators do not take money from the passive participant. Punishment then increases progressively with the share of the endowment taken by dictators, reaching an average of 6.26 tokens for the maximum taking rate. Participants on average overestimate the amount of third-party punishment that the dictators will receive. The variable *Beliefs* (i.e. a participant's initial belief about peers' average punishment decision) is on average statistically significantly larger than the variable *Punish1* that reports punishment choices in the Punishment Period 1 ($t(412)= 4.08$, $p< 1\%$). This difference suggests that participants expect

average punishment for the dictators to be 40% higher than what actually is. Controlling for initial beliefs is important because beliefs may affect punishment behavior. In the Punishment Period 1 stage, participants may be influenced in their choice of how much to punish by their beliefs about what their peers do. Moreover, participants might update their beliefs during the information stage, when actual peers' punishment is revealed and this might contribute to the possible change in their punishment decisions, in Punishment Period 2. Therefore, by controlling for *Beliefs*, we estimate how a participant modifies her initial punishment decisions due to social influence, holding her initial expectations regarding peers' punishment constant.

[TABLE 1 APPROX. HERE]

4.2. Main Analysis: Existence of Social Influence Effects on Third-party Punishment

We focus on how participants revise their initial punishment decisions after they are exposed to peer effects. If social influence has an effect on participants' punishment decisions through the desire to conform, we expect that those who receive socially meaningful information are more likely to converge toward average punishment compared to participants in the control treatment.

[FIGURE 2 APPROX. HERE]

Figure 2 provides a graphical representation of the punishment decisions in the three treatments. The x-axis repors the difference between peers' average punishment (the information received by subject in the treatment groups) and individual punishment in the first period (the initial punishment choice). On the y-axis the difference between individual punishment after the revision in the second and the initial punishment choice is reported.

When a participant seeks conformity with peers' average punishment, in the second period she revises the initial punishment choice in order to "close the gap" between peers' average punishment and initial individual punishment. If that is the case, observations plotted in figure 2 lay on the dotted 45-degrees diagonal line. A graphical inspection suggests that in both informational and info+normative treatments observations are distributed more closely to the dotted line compared to those in control.

We perform a series of parametric tests. First, we create the dummy variable *ConvergeDummy* that takes value 1 when a subject chooses a second-period punishment that reduces the gap between his or her first-period punishment and peers' average punishment.[9] We run a logistic regression to test whether the probability of reducing this gap is the same for participants in the control, informational and info+normative treatments. We control for individual beliefs regarding peers' average punishment and we include a set of socio-demographic controls.[10] Results are reported in table 2. We esti-

[9] *ConvergeDummy* =1 when ($diffAvgP1P1 > 0$ & $Punish2 - Punish1 > 0$) or ($diffAvgP1P1 < 0$ & $Punish2 - Punish1 < 0$).

[10] The socio-demographic controls that we used throughout the all data analysis include gender (variable *male*), field of study using dummy variables for *social* (social sciences and Medicine), *arts* (humanities) and *field_other* (not in *social* or *arts*), a measure of risk (*risk*), a measure of logical abilities (*logic*), a measure of impulsivity (*impulsivity*), and a measure

mate unstandardized regression coefficients and the table reports marginal effects at means.

[TABLE 2 APPROX. HERE]

In model 1 we include all the observations in our sample. The coefficients of both the dummies *informational* and *info+normative* are positive and statistically significant at the conventional level or better for any level of dictator taking rate considered. The marginal effect for informational suggests a 17.5 percentage point increase in the probability that participants modify punishment in order to converge toward the mean. The marginal effect for participants in the info+normative treatment indicates 16 point increase in the same measure.

In models 2 to 4 we also check whether social influence effects are stronger when dictators' behave more selfishly, since, as we discussed above (section 4.1), low taking rates are characterized by little or no punishment and so by limited variation in average punishment. In model 2, we include only punishment choices for dictators taking from the passive participant equal or more than the 25^{th} percentile of average taking in the sample (hence, taking 10 tokens or more from the passive player). In model 3, we restrict the attention to taking rates equal or higher than the 50^{th} percentile (taking 15 tokens or more). Finally, in model 4, the restricted sample includes

of the time used by the subject to answer the control questions (*Instruction*). Table B.7 in the appendix contains the definition of each variable employed in the regression and the explanation of how it was constructed.

only taking choices at least equal to the 75^{th} percentile of average taking (the entire endowment of 30 tokens). The coefficients of $informational$ and $info + normative$ remain statistically significant at the conventional level or better for each model specification, with marginal effects ranging between 15.2 and 23.4 percentage point increase according to the model specification. Therefore, we conclude that in our sample the likelihood to revise punishment choices and converge toward peers' average punishment is significantly higher for participants exposed to social influence.

As a second step, we estimate the magnitude of the convergence toward peers' average punishment. We regress the difference in punishment between periods 2 and 1 (variable *diffP2P1*) on the difference between peers' average first-period punishment and individual punishment (variable *diffAvgP1P1*), controlling for participants' beliefs regarding peers' average punishment and a set of socio-demographic characteristics. We use a Tobit model in order to take into account that punishment choices are censored by design. Standard errors are clustered by subject. We estimate unstandardized regression coefficients.

Models 1 in table 3 reports the results for the cases when we include all the observations. The coefficient of the interaction between the dummy *informational* and *diffAvgP1P1* is positive and statistically significant at the conventional level. The results suggest that an increase of one unit in the difference between average peers' punishment and individual punishment in the Punishment Period 1 stage induces the participant to modify her punishment in order to 'close the gap' by 0.21 units. The coefficient of the interaction between the dummy *info+normative* and *diffAvgP1P1* is also

positive, albeit not statistically different from zero at the conventional level.

[TABLE 3 APPROX. HERE]

In models 2 to 4, we include only the punishment for dictators taking respectively more than the 25^{th} percentile of average taking rate in our sample (hence, taking 10 tokens or more from the passive player), the 50^{th} percentile (taking 15 tokens or more) and the 75^{th} percentile (taking the entire endowment of 30 tokens). The coefficient of the dummy $informational$ is statistically significant at the conventional level or better for each model specification. When the punishment differential increases of one unit, the change in individual punishment aiming at 'closing the gap' ranges between .23 and .28 units depending on the model specification. For the variable $info + normative$, the estimated coefficient is always positive, albeit it is only statistically different from zero at the conventional level when the sample includes only taking rates equal or higher than the median.

We conclude that the amount of punishment tokens used for revising initial punishment decisions is higher for participants exposed to social influence than for participants in the control treatment. For subjects in the informational treatment, the result is statistically significant at the conventional level for each possible dictator taking rate. For participants in the info+normative treatment the result is statistically significant at the conventional level only for take-rate values equal or higher than the median.

Result 1. *Social influence is an important determinant of third-party punishment. Information on peers' punishment choices is effective in driving*

third-party punishment toward conformity with peers' average punishment.

4.3. The Effect of Adding a Normative Social Influence Component to Informational Social Influence

We check whether adding a normative social influence component (i.e. peers sending either a smiley or frowny emoticon) to informational social influence has an effect on third-party punishment. One possible result is that, if the desire to be liked is an important channel of transmission of social influence and the effects of informational and normative influence add up, participants in the info+normative treatment will display a greater convergence toward peers' average punishment. Conversely, if the main behavioral driver of social influence is the need to "perform the right action", or if adding the normative component described above to the informational channel does not boost social influence effects, then we should observe similar results in the informational and info+normative treatments.

Looking at descriptive statistics relative to punishment after the revision, we fail to reject the hypothesis that the distributions of second period punishment *Punish2* differs across informational and info+normative treatments (Wilcoxon rank-sum test, p-value $> .1$ for any dictator take level). Similarly, an Hotelling T-squared test for comparison of mean second period punishment reports no statistically significant differences across treatments (p-value $> .1$).

[TABLE 4 APPROX. HERE]

Table 4 reports the results of a logistic regression showing the likelihood to converge toward peers' average punishment for participants in informational and info+normative treatments. We regress *ConvergeDummy* on the dummy *info+normative*, controlling for participants' beliefs regarding peers' average punishment, subject's type and a set of socio-demographic characteristics. Standard errors are clustered by subject. We estimate unstandardized regression coefficients and the table reports marginal effects at means. The coefficient of the treatment dummy is never statistically different from zero, whether we include all the observations (model 1), only the punishment for dictators taking more than the 25^{th} percentile of average taking rate in our sample (taking 10 or more, model 2), the 50^{th} percentile (taking 15 or more, model 3) and the 75^{th} percentile (taking the entire endowment, model 4).

A Tobit regression confirms that in our sample participants in info+normative and informational fail to display statistically significant differences in their convergence toward peers' mean punishment. We regress the difference between punishment choices in the two periods (variable *diffP2P1*) on the interaction of the difference between average peers' first period punishment and individual punishment (*diffAvgP1P1*) and the treatment dummy *info+normative*, controlling for participants' beliefs regarding peers' average punishment, subject's type and a set of socio-demographic characteristics. We estimate unstandardized regression coefficients. Models (1) and (2) of table 5 show that the coefficient of the interaction between the variables *info+normative* and *diffAvgP1P1* is not statistically different from zero for any model specification.

[TABLE 5 APPROX. HERE]

Power Calculation. To further investigate the lack of a statistically significant difference between the informational and the info+normative treatments, a post hoc power analysis is conducted using the software Gpower3 (Faul et al., 2007). Considering 60 and 42 independent observations and 7 levels of *Punish2*, power analysis indicates that for a Hotelling T^2 mean vectors test there is a 81% chance of detecting a small effect size and a 99% chance of detecting a medium effect size (defined by Cohen, 1992, as .2 and .5 of a population standard deviation between the means respectively) between the two groups as significant at the 5% level (two tailed). For the difference of distributions, power analysis suggests that a Wilcoxon rank-sum test two-tailed has a 82 (87) % chance of detecting a small effect size if we assume a normal (logistic) distribution and a 99% chance of detecting a medium effect size. We perform the power analysis of the logistic model used for determining the likelihood of convergence toward peers' average punishment. The alpha level used for this analysis is $p < .05$. Considering an odd ratio equal to the one we have in our sample, 102 (714) observations, and a lognormal distribution, the resulting statistical power is 0.81 (0.99).

Result 2. *In our experiment the addition of a normative social influence component to the informational social influence treatment does not modify participants' third-party punishment decisions.*

4.4. Direction of Social Influence Effects

In a contribution investigating the interaction between peer effects and voluntary cooperation using a three-person gift exchange game, Thöni and Gächter (2015) found that participants exposed to peer effects significantly modify their initial effort choices only when conformity implies a downward revision of effort. This finding, that reminds to the 'moral wiggle room' found in other studies (Dana et al., 2007), suggests that social influence is more effective in convincing cooperative agents to break cooperation norms than to induce initially non-cooperative agents to adopt a pro-social behavior. If an asymmetric reaction to peer effects were confirmed also for individuals' decision to engage in decentralized third-party punishment, there would be important consequences for the possibility to implement policies based on social influence that selectively disclose information regarding peers' actions (Benabou and Tirole, 2011).

[FIGURE 3 APPROX. HERE]

We investigate whether, also in the context of third-party punishment, there is evidence of an asymmetric reaction to peer effects in cases where conformity implies reducing or increasing individual punishment. Figure 3 pools data from the informational and info+normative treatments and provides a scatter plot of the difference in average and initial individual punishment and the punishment revision after receiving information.[11] The bold line

[11] In a separate analysis, available upon request, we found no differences between informational and info+normative treatments for the results presented in this section.

provides a fit of the data, allowing for different slopes and origins when $diffAvgP1P1$, the difference between peers' average punishment and individual punishment, is positive or negative. The left panel of figure 3 plots the whole sample of data collected, while the right panel restricts the attention to those participants who punish a positive amount at least once.

A visual inspection of the left panel shows no obvious kink at zero on the horizontal axes, thus suggesting symmetric reactions to positive and negative initial punishment differentials. Results from a parametric regression reported in model 1 of table 6 confirm the visual impression. We regress the change in punishment after receiving information on the difference in average and initial individual punishment $diffAvgP1P1$, a dummy indicating that $diffAvgP1P1$ is positive, and the interaction between the two variables. We estimate unstandardized regression coefficients. The coefficient of the interaction term is not statistically different from zero.

[TABLE 6 APPROX. HERE]

In the right panel of figure 3 we exclude from the sample participants who never punish. The slope of the bold line produces a kink at zero, suggesting that a positive initial punishment differential (i.e. peers' average punishment is higher than initial individual punishment) triggers more conformism. The result from model 2 in table 6, identical to model 1 excepts for the exclusion of subjects who never punish, confirms this visual impression. The coefficient of the interaction term is positive and statistically significant at the 1% level. This finding indicates that, conditionally to engaging in third-party

punishment, participants respond significantly more to social influence when conformity implies *increasing* individual punishment. The coefficient of the variable $diffAvgP1P1$ is also positive and significant, showing that when the initial individual punishment is larger than average peers' punishment, so that conformity implies revising individual punishment downward, social influence effects are still affecting individual punishment decisions, but at a rate significantly lower.

This result is surprising since upward revisions of punishment imply a reduction of individual payoffs. This finding is also interesting since the asymmetry in social influence effects on voluntary cooperation found by Thöni and Gächter (2015) works in the opposite direction.

Result 3. *Social influence has significant effects on third-party punishment both when conformity implies reducing or increasing individual punishment. The effects are stronger when comparison peers punish more than the individual punisher, thus calling for a revision of punishment upward.*

5. The Theoretical Foundations of Peer Effects on Third-Party Punishment

In this section we examine whether theories of social preferences can explain social influence in third-party punishment. Among all the existing theories, we select representative models that can give account of third-party punishment and test whether peer effects can occur in those settings. We prove first that models with distributional preferences do not explain social influence in our game. We then revert to conformity and social norm theories. We

find that these theories are capable of accounting for peer effects and social influence in third-party punishment. Properly adapted to our experimental design, they are also fit to explain the peculiarities of our model. Particularly, they explain why third parties respond more when they learn that they are punishing below average. All the results presented in this section are proven in Appendix A.

5.1. Social Preferences and Third-Party Punishment

There are several theories of social preferences that can justify third-party punishment and its pattern. Social-preference theories assume that people's utility is affected not only by their own income, but also by the income of *relevant* others. In setting up a model accounting for third-party involvement, we can hypothesize that our third-party cares also for the payoffs of both the receiver A and the dictator B with whom he is matched. The three-person group consisting of A, B and C thus represents the 'reference group' for the third party, and her utility function accounting for social preferences takes the general form

$$u_C(\Pi_C, \Pi_A, \Pi_B) \tag{1}$$

where Π_i $(i = A, B, C)$ are the payoffs reported in Section 3.

Models of inequity aversion (Levine, 1998; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) assume utility functions that represent special versions of equation (1). They justify punishment as long as the dictator's actions determine inequality between the third-party's payoff and the payoffs of either the dictator or the receiver. These models can account for third-party punishment but they cannot explain peer effects. Even if we extended the

utility functions to include the payoffs of other dictators, receivers and third parties, the fact that C cannot exert any impact on the payoffs of the other subjects external to her reference group and that none of the actions taken by external subjects affects her reference group determines C's unwillingness to change her behavior when observing the choices of external punishers.

Likewise, some models of altruism can explain third-party involvement (Charness and Rabin, 2002; Cox et al., 2007, 2008) but are unable to account for peer effects. Moreover, to justify third-party involvement, often these models require extra assumptions. Consider, for instance, Cox et al. (2007). They assume a CES utility function that, adapted to our model, looks like

$$u_C(\Pi_C, \Pi_A, \Pi_B) = \frac{\Pi_C^\alpha + \theta_A \Pi_A^\alpha + \theta_B \Pi_B^\alpha}{\alpha} \tag{2}$$

where α determines the constant elasticity of substitution between C's payoff and the other players' payoffs ($\alpha \in (-\infty, 0) \cup (0, 1]$). The parameters θ_i ($i = A, B$) represent weights attached to the other players' payoffs (i.e., the emotions that C feels for A and B). This utility function implies that a positive punishment always occurs if $\theta_B < 0$, i.e., if the third party has negative feelings about the dictator. A negative θ_B, however, suggests that dictators might be punished even if they are not taking anything from the passive receivers.

This model can explain the positive relationship between takings and punishment observed in our data and in other TPP experiments and it can also explain why some subjects punish even when no taking occurred. However, it is not able to explain peer effects, since peers' actions have no impact on

any of the payoffs in the relevant group.

We then turn to the model in Charness and Rabin (2002). This is a very interesting contribution, since it allows us to combine altruism with strong reciprocity concerns. The paper presents two different utility functions. The first combines preferences for efficiency with distributional concerns, leaving out reciprocity. The format is

$$u_C(\Pi_C, \Pi_A, \Pi_B) = (1-\lambda)\Pi_C + \lambda[\delta\, min\{\Pi_A; \Pi_B; \Pi_C\} + (1-\delta)(\Pi_A + \Pi_B + \Pi_C)] \tag{3}$$

Thus, when making decisions, the third party considers both her own egoistic material payoff and social welfare. The weight attributed to social welfare is λ. The parameter δ measures concern for the worst-off person (maximin or Rawlsian criterion). It ranges from $\delta = 0$ (pure utilitarian concerns) to $\delta = 1$ (pure maximin).

In this simple form, the utility function in (3) cannot justify third-party punishment. All its arguments are in fact decreasing in punishment, so punishment is never positive in equilibrium. In order to account for third-party punishment, we need to introduce (strong) reciprocity. We follow again Charness and Rabin (2002), who, in the Appendix to their paper, introduce 'demerit' parameters d_i $(i = A, B)$ in their utility function (3), which becomes

$$\begin{aligned} u_C = {} & (1-\lambda)\Pi_C + \lambda[\delta\, min\{\Pi_C; \Pi_A + b\, d_A; \Pi_B + b\, d_B\} + \\ & + (1-\delta)(\Pi_C + \Pi_A(1 - k\, d_A) + \Pi_B(1 - k\, d_B)) - f(d_A\, \Pi_A + d_B\, \Pi_B)] \end{aligned} \tag{4}$$

where b, k are non-negative parameters and imply that, when C maximizes social welfare, the greater d_i, the less weight she places on the utility of individual i in her reference group ($i = A, B$). The parameter $f \geq 0$ implies that the third party is willing to reduce the payoff of those players she reckons are misbehaving. Since the receiver A is passive, he can never misbehave and we can safely set $d_A = 0$. The punishment behavior of the third party depends on the values of the parameters λ, δ and d_B and on the identity of the subject with the minimum payoff in the group.

Within this second specification, punishment occurs when the dictator's demerit d_B is high enough. Punishment is equal to $p = \frac{t}{2}$ for sufficiently high d_B and is equal to the maximum punishment $p = 20$ for very high d_B. This particular model thus explains the positive relationship between punishment p and taking t and also why a third party may apply the maximum punishment, depending on the strength of the demerit that the third party attaches to the dictator's behavior. Both functions in Charness and Rabin (2002), as the one by Cox et al. (2007), are not able to account for peer effects. As, before, peers' choices do not affect any of the components of the utility of the third party, since they do not affect neither Π_C nor Π_A and Π_B.[12]

[12] In principle, we could design an extension that introduces a theoretical justification for peer effects. Such an extension would imply a change in the definition of the society on which the third party bases her measure of social welfare. Particularly, she may consider not only her strict reference group (A, B and herself), but a broader society, including all dictators, receivers and third parties playing the same game. This however would require that she observed the payoffs of each single individual in that society. In our setting, she knows only the average punishment. So, even if we generated peer effects within these models with this extension, we would not be providing a theoretical justification for our specific game.

5.2. Conformity and Norm Compliance

To explain peer effects in our setting, we need to recur to models of conformity and norm compliance. A paper that accounts for a norm abiding behavior in a very simple and intuitive fashion is López-Pérez (2008) that presents a model in which individuals face a trade-off between maximizing their self interest and following a social norm. The social norm is exogenously given and is obtained maximizing a social welfare function that depends positively on the efficiency of income distribution and negatively on its inequality. Adapting López-Pérez (2008) to our notation, the social welfare function is

$$F_{\varepsilon\delta} = \varepsilon(\Pi_A + \Pi_B + \Pi_C) - \delta\left(max\{\Pi_A;\Pi_B;\Pi_C\} - min\{\Pi_A;\Pi_B;\Pi_C\}\right) \quad (5)$$

with ε, $\delta > 0$.

Given that we are concerned with the behavior of the third party, we look for the social norm regarding the level of punishment, which means that we look for the value of p that maximizes (5). In Appendix A, we prove that, according to the party holding the minimum payoff, the social norm can prescribe either zero punishment when the dictator has the minimum payoff, or a punishment equal to $p = \frac{t}{d}$ $(d = 2, 3)$ when the minimum payoff is held by the receiver.

Utility in this model is given by the function

$$u_C = \begin{cases} \Pi_C & \text{if } p = \hat{p} \\ \Pi_C - \gamma r & \text{otherwise} \end{cases} \quad (6)$$

where $\gamma > 0$ measures the taste for social-norm conformity and r represents the number of players who respected the norm up to a certain moment. A third party respects the norm if the cost of abidance is lower than the cost of norm deviation. Initially, a third party assumes that everybody respects the social norm. If the parameter γ is high enough, she too respects the norm. After the first period, she receives information about peers' behavior. Based on that information, the third party updates r and compares again the cost of abiding to the social norm to the loss from non conforming $\gamma\, r$. If the latter has decreased substantially because she observed many people deviating, she stops respecting the norm. Otherwise, she sticks to the norm. This model accounts for peer effects, explaining why people stops obeying to social norms, in response to lack of conformity by peers. However, it cannot explain changes from non-conformity to abidance. In other words, it explains why people might decrease punishment (the left hand side of figure 3) but not why they may increase it. Moreover, it does not explain why people might adjust more when they learn that others on average punished more (right panel of figure 3). In the next subsection we present two original models that are able to provide a theoretical foundation to our results.

5.3. Two Models of Norm Abidance and Conformity

The Trade Off Between Social Norms and Peer Effects and the Role of the Taste for Punishment.

In the first model, we keep the *fairmax* social norm and modify the utility function used by López-Pérez (2008), which now becomes

$$u_C = \begin{cases} \Pi_C(\hat{p}) + f\,\hat{p} & \text{if } p = \hat{p} \\ \Pi_C(p) - \gamma|\hat{p} - p| - \mu|\bar{p} - p| + f\,p & \text{otherwise} \end{cases} \tag{7}$$

where γ, $\mu > 0$ measure, respectively, the disutility the third party derives from not conforming to the social norm $\hat{p} = \frac{t}{d}$ $(d = 2, 3)$ and from following a punishment behavior different from the average $\bar{p}$ in the population. The parameter $f > 0$ represents the 'taste for punishment', i.e., the extra utility a third party receives from punishing (it might be the utility from enforcing a social norm or a simple pleasure he derives from punishing others).

The data from our experiment show that third parties tend to follow their peers, converging towards $\bar{p}$, no matter whether they initially punish more or less than average. In terms of the present model, this would imply high social influence μ relative to the weight γ attached to social norms. When we restrict the analysis to individuals punishing at least once (who are likely to be characterized by a higher taste for punishment parameter f than their counterparts who never punish), we observe that they tend to conform to their peers' behavior but they converge faster if their initial punishment is lower than average. This behavior too can be explained by a strong social influence parameter μ together with a high f. The role played by

f is important: it slows down convergence when subjects have to decrease punishment but it enhances convergence when an increase in punishment is required to conform to peers. Clearly, all these effects are magnified when social norms and peer effects drive punishment in the same direction, i.e., when they both call for an increases or for a decrease.

Focal Points and the Taste for Punishment

In this case individuals do not observe $\hat{p}$, the social norm. Thus, even if they are willing to conform, *a priori*, they do not know what is the socially prescribed punishment for every level of taking. The information they receive at the end of the first period might then represent a focal point, an indication of what a majority of people would deem admissible behavior in similar circumstances (Sugden, 1989).

If people like to conform and they take the information they receive as an indication of the social norm, their utility function becomes:

$$u_C = \begin{cases} \Pi_C(\bar{p} + f\,\bar{p} & \text{if } p = \bar{p} \\ \Pi_C(p) - \gamma|p - \bar{p}| + f\,p & \text{otherwise} \end{cases} \tag{8}$$

where, as before, $\bar{p}$ represents the average punishment disclosed to third parties at the end of the first period. Appendix A proves that, when $f > 1$, third parties are more likely to conform when they have to increase punishment after learning that they are punishing less than average. This implies that people with a large taste for punishment are more likely to conform when they have to increase punishment than when they have to decrease it. In a continuous model (where p varies continuously and the equilibrium is

an internal solution) this can be expressed in terms of the responsiveness of third parties to social information. Particularly, they might respond more to social information when they discover they were punishing less than average, thus getting closer to the social standard $\bar{p}$ than when they discover they were punishing more than average. This simple model therefore reproduces the results of our experiment, where individuals with a 'taste for punishment' (those who engage in positive punishment) increase their punishment more when they learn they are punishing below average than when they discover they are punishing above average.

6. Discussion

Societies and human organizations need enforcing mechanisms that prevent individuals to engage in opportunistic or anti-social behaviors. The recent increase in the use of policies based on decentralized interventions, which often complement centralized legal enforcement, as well as the possibility to influence the creation and enforcement of social norms, call for a better understanding of the nature and characteristics of third-party punishment. The results of this paper show that social influence is effective in regulating the likelihood that individuals engage in third-party intervention and it affects the intensity of third-party punishment. Hence, by leveraging social comparison effects, it is possible to promote norm-based campaigns that encourage third parties to act as social controllers when a centralized sanctioning authority has limited ability or resources (Benabou and Tirole, 2011; Cialdini and Trost, 1998; Kaplow and Shavell, 2007). For instance, social influence may improve the effectiveness of norm-based interventions when

the beliefs of the general population underestimate the tolerance and frequency of socially undesirable behaviors. Indeed, it is often the case that people's beliefs systematically overestimate the fraction of peers engaging in socially harmful behavior, such as in the case of perceived crime levels, benefit frauds or the percentage of non-voters.[13] In these and several other situations, policymakers can achieve welfare-improving results by means of inexpensive ad-hoc communication strategies aimed at de-biasing people's incorrect beliefs, generating stigma against the undesirable behaviors and inviting individuals to report unlawful activities.

Results show that informational social influence is a major driver through which social comparison affects third-party punishment, and that the addition of a normative social influence component is not necessary to influence behaviour. This finding has important consequences for the feasibility of policies based on social influence that target third-party punishment, since previous research found that informational social influence has stable and long-lasting effects on behavior. We also found that, in contrast with peer effects on pro-social cooperation, social influence effects are stronger when peers punished more and conformity calls for an upward revision of individual punishment.

A comparison of these findings with models of social preferences reveals that no one of the theories reviewed explains our empirical results. However, we

[13]For example, the Royal Statistical Society reports that 58% of the UK population estimates that crime is rising, while data suggest that the crime rate in the country is 19% lower than the previous year and 53% lower than 1995. For a discussion of other examples and additional details, see http://www.kcl.ac.uk/newsevents/news/newsrecords/2013/07-July/Perceptions-are-not-reality-the-top-10-we-get-wrong.aspx.

modify the utility function in López-Pérez (2008), showing that social-norm compliance theories can predict our results. We also show that theories of social conformism can explain our results. Together with the desire to conform, a key ingredient is how much individuals enjoy punishing violators of a social norm of fair distribution.

Finally, we acknowledge that our experimental design might create experimenter demand effects. Indeed, by eliciting punishment decisions right after providing information on peers' punishment, we might have steered participants toward conformity. On the other hand, our design includes also a feature that might have counteracted the previous problem, reducing the impact of social influence. Indeed, by eliciting beliefs regarding peers' punishment at the very beginning of the experiment, we made the social norm of third-party intervention salient to participants. It is possible that, absent beliefs elicitation, participants reacted even more strongly to information regarding peers' choices.

Future research should address whether and to what extent the results of our laboratory experiment can be replicated in real world situations. Field experimentation and observational data will play a crucial role in confirming the external validity of our findings. Moreover, the exposure to peer effects in our experiment is not repeated for additional periods. Future studies should investigate the evolution of peer effects in a dynamic setting.

References

Akerlof, G. A., Yellen, J. L., Katz, M. L., 1996. An analysis of out-of-wedlock childbearing in the united states. The Quarterly Journal of Economics

111 (2), 277–317.

Ayres, I., Raseman, S., Shih, A., 2012. Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage. Journal of Law, Economics, and Organization, ews020.

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

Balafoutas, L., Nikiforakis, N., Rockenbach, B., 2014. Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences 111 (45), 15924–15927.

Benabou, R., Tirole, J., 2011. Laws and norms. Tech. rep., National Bureau of Economic Research.

Bendor, J., Mookherjee, D., 1990. Norms, third-party sanctions, and cooperation. JL Econ & Org. 6, 33.

Bendor, J., Swistak, P., 2001. The evolution of norms1. American Journal of Sociology 106 (6), 1493–1545.

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

Bolton, G. E., Ockenfels, A., 2000. Erc: A theory of equity, reciprocity, and competition. American economic review, 166–193.

Bond, R., Smith, P. B., 1996. Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. Psychological bulletin 119 (1), 111.

Cason, T. N., Mui, V.-L., 1998. Social influence in the sequential dictator game. Journal of Mathematical Psychology 42 (2), 248–265.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The Quarterly Journal of Economics 117 (3), 817–869.

Cialdini, R. B., Goldstein, N. J., 2004. Social influence: Compliance and conformity. Annu. Rev. Psychol. 55, 591–621.

Cialdini, R. B., Trost, M. R., 1998. Social influence: Social norms, conformity and compliance. McGraw-Hill.

Cohen, J., 1992. A power primer. Psychological bulletin 112 (1), 155.

Cox, J. C., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59 (1), 17–45.

Cox, J. C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76 (1), 31–69.
URL http://dx.doi.org/10.1111/j.0012-9682.2008.00817.x

Crawford, N. C., Klotz, A., 1999. How sanctions work: a framework for analysis. In: How sanctions work. Springer, pp. 25–42.

Dana, J., Weber, R. A., Kuang, J. X., 2007. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory 33 (1), 67–80.

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.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., Vlaev, I., 2012. Influencing behaviour: The mindspace way. Journal of Economic Psychology 33 (1), 264–277.

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

Ellingsen, T., Johannesson, M., 2008. Anticipated verbal feedback induces altruistic behavior. Evolution and Human Behavior 29 (2), 100–105.

Falk, A., Fischbacher, U., 2002. "crime" in the lab-detecting social interaction. European Economic Review 46 (4), 859–869.

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

Falk, A., Fischbacher, U., Gächter, S., 2013. Living in two neighborhoods—social interaction effects in the laboratory. Economic Inquiry 51 (1), 563–578.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 39 (2), 175–191.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution and human behavior 25 (2), 63–87.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. The quarterly journal of economics 114 (3), 817–868.

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

Fortin, B., Lacroix, G., Villeval, M., 2007. Tax evasion and social interactions. Journal of Public Economics 91 (11), 2089–2112.

Gächter, S., Nosenzo, D., Sefton, M., 2012. The impact of social comparisons on reciprocity. The Scandinavian Journal of Economics 114 (4), 1346–1367.

Galbiati, R., Zanella, G., 2012. The tax evasion social multiplier: Evidence from italy. Journal of Public Economics 96 (5), 485–494.

Gintis, H., 2000. Beyond homo economicus: evidence from experimental economics. Ecological economics 35 (3), 311–322.

Glaeser, E. L., Sacerdote, B., Scheinkman, J. A., 1996. Crime and social interactions. The Quarterly Journal of Economics 111 (2), 507–548.

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

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J., Gurven, M., Gwako, E., Henrich, N., et al., 2006. Costly punishment across human societies. Science 312 (5781), 1767–1770.

Ho, T.-H., Su, X., 2009. Peer-induced fairness in games. The American Economic Review 99 (5), 2022.

Kaplow, L., Shavell, S., 2007. Moral rules, the moral sentiments, and behavior: toward a theory of an optimal moral system. Journal of Political Economy 115 (3), 494–514.

Krupka, E., 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., 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.

Kurzban, R., DeScioli, P., O'Brien, E., 2007. Audience effects on moralistic punishment. Evolution and Human behavior 28 (2), 75–84.

Levine, D. K., 1998. Modeling altruism and spitefulness in experiments. Review of economic dynamics 1 (3), 593–622.

List, J. A., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115 (3), 482–493.

López-Pérez, R., 2008. Aversion to norm-breaking: A model. Games and Economic behavior 64 (1), 237–267.

Marlowe, F., Berbesque, J., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., et al., 2008. More 'altruistic'punishment in larger societies. Proceedings of the Royal Society B: Biological Sciences 275 (1634), 587–592.

Mas, A., Moretti, E., 2009. Peers at work. American Economic Review 99 (1), 112–145.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.-C., 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. The American Economic Review 93 (1), 366–380.

Mathew, S., Boyd, R., 2011. Punishment sustains large-scale cooperation in prestate warfare. Proceedings of the National Academy of Sciences 108 (28), 11375–11380.

Mittone, L., Ploner, M., 2011. Peer pressure, social spillovers, and reciprocity: an experimental analysis. Experimental Economics 14 (2), 203–222.

Nelissen, R., 2008. The price you pay: cost-dependent reputation effects of altruistic punishment. Evolution and Human Behavior 29 (4), 242–248.

Nelissen, R., Zeelenberg, M., 2009. Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. Judgment and Decision Making 4 (7), 543–553.

Piazza, J., Bering, J., 2008. Concerns about reputation via gossip promote generous allocations in an economic game. Evolution and Human Behavior 29 (3), 172–178.

Rabin, M., 1993. Incorporating fairness into game theory and economics. The American economic review, 1281–1302.

Sugden, R., 1989. Spontaneous order. Journal of Economic Perspectives 3 (4), 85–97.

Sunstein, C. R., Schkade, D., Ellman, L. M., Sawick, A., 2006. Are judges political?: an empirical analysis of the federal judiciary. Brookings Institution Press.

Thöni, C., Gächter, S., 2015. Peer effects and social preferences in voluntary cooperation: A theoretical and experimental analysis. Journal of Economic Psychology 48, 72–88.

Xiao, E., Houser, D., 2009. Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. Journal of Economic Psychology 30 (3), 393–404.

Tables Main Text

Table 1: Summary Statistics

Treatment	Dictator take Period 1	Dictator take Period 2	Beliefs	Punish Period 1	Punish Period 2
Control					
(Mean)	16.57	16.38	5.63	3.87	4.02
(Median)	15	15	5	3	3
(SD)	10.54	11.01	4.77	4.35	4.21
Info+normative					
	18.33	17.00	5.27	4.14	3.67
	20	17.5	4	2	2
	12.08	12.08	4.08	5.14	4.47
Informational					
	16.90	16.67	5.89	3.81	3.26
	17.5	15	5	2	2
	11.46	11.91	5.04	4.53	3.73
Total					
	17.15	16.62	5.58	3.94	3.86
	15	15	5	2	3
	11.22	11.52	4.84	4.63	4.40

Table 2: Probability that Individual Second-Period Punishment Converges Toward Average Punishment

	(1)	(2)	(3)	(4)
Informational	0.175***	0.152**	0.155**	0.226***
	(0.05)	(0.06)	(0.06)	(0.08)
Info+normative	0.160***	0.181***	0.199***	0.234***
	(0.04)	(0.05)	(0.05)	(0.07)
Controls	Y	Y	Y	Y
N	1449	1035	828	207
pseudo R^2	0.160	0.159	0.164	0.185

Notes: Logistic regression: dep. var. $ConvergeDummy$ (takes value 1 when a punisher's second-period punishment reduces the gap between her first-period punishment and peers' average punishment, that is when ($diffAvgP1P1 > 0$ & $Punish2 - Punish1 > 0$) or ($diffAvgP1P1 < 0$ & $Punish2 - Punish1 < 0$)). Marginal effect at means, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 3: Convergence to Peers' Average Punishment

	(1)	(2)	(3)	(4)
diffAvgP1P1	0.323***	0.298***	0.247***	0.283***
	(0.08)	(0.08)	(0.07)	(0.09)
Informational	-0.962***	-1.081***	-1.122***	-1.392**
	(0.32)	(0.38)	(0.41)	(0.56)
Informational#	0.214**	0.225**	0.277***	0.256**
diffAvgP1P1	(0.10)	(0.10)	(0.10)	(0.11)
Info+normative	-0.860**	-0.914**	-0.985**	-0.775
	(0.34)	(0.38)	(0.42)	(0.64)
Info+normative#	0.184	0.164	0.208**	0.192
diffAvgP1P1	(0.12)	(0.10)	(0.10)	(0.12)
controls	Y	Y	Y	Y
N	1449	1035	828	207
pseudo R^2	0.063	0.058	0.058	0.062
Controls	Y	Y	Y	

Notes: Tobit regression: dep. var. $diffP2P1$, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 4: Prob. Indiv. Punish Second Period Converge Toward Average Punishment in Info and Info+normative Treatments

	(1)	(2)	(3)	(4)
Info+normative	-0.001	0.044	0.069	0.032
	(0.07)	(0.08)	(0.08)	(0.11)
N	714	510	408	102
pseudo R^2	0.219	0.236	0.259	0.325
Controls	Y	Y	Y	Y

Notes: Logistic regression: dep. var. $ConvergeDummy$ (takes value 1 when a punisher's second-period punishment reduces the gap between her first-period punishment and peers' average punishment, that is when ($diffAvgP1P1 > 0$ & $Punish2 - Punish1 > 0$) or ($diffAvgP1P1 < 0$ & $Punish2 - Punish1 < 0$)). Marginal effect at means, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 5: Convergence to Peers' Average Punishment in Informational and Info+normative treatments

	(1)	(2)	(3)	(4)
diffAvgP1P1	0.480***	0.440***	0.448***	0.492***
	(0.08)	(0.09)	(0.09)	(0.11)
Info+normative	-0.020	0.025	-0.013	0.441
	(0.34)	(0.42)	(0.47)	(0.72)
Info+normative#	-0.020	-0.034	-0.037	-0.034
diffAvgP1P1	(0.12)	(0.12)	(0.12)	(0.14)
Controls	Y	Y	Y	Y
N	714	510	408	102
pseudo R^2	0.092	0.082	0.082	0.085

Notes: Tobit regression, dep. var. $diffP2P1$, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Direction of Social Influence Effects

	(1)	(2)
diffAvgP1P1	0.550***	0.543***
	(0.12)	(0.11)
diffAvgP1P1 > 0	-0.210	0.690***
	(0.19)	(0.25)
diffAvgP1P1 > 0 (dummy)	-0.553	-1.723**
	(0.54)	(0.71)
_cons	-0.017	0.153
	(1.34)	(1.70)
Controls	Y	Y
N	714	497
R^2	0.3767	0.4386

* p<0.10, ** p<0.05, *** p<0.01

Notes: OLS regression, dep. var. $diffP2P1$, SE clustered by subject. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Figures Main Text

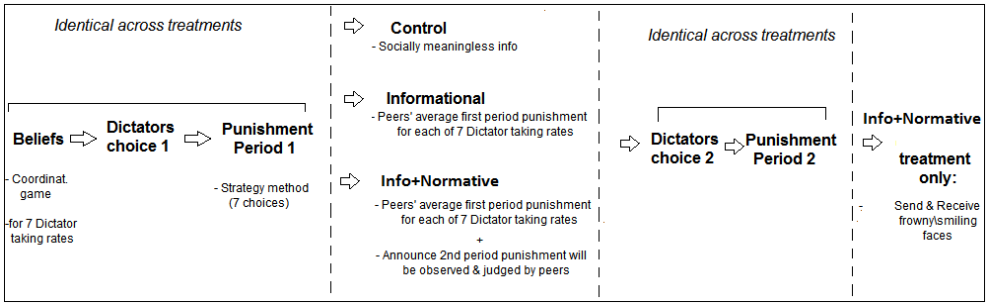


Figure 1: Experiment Design

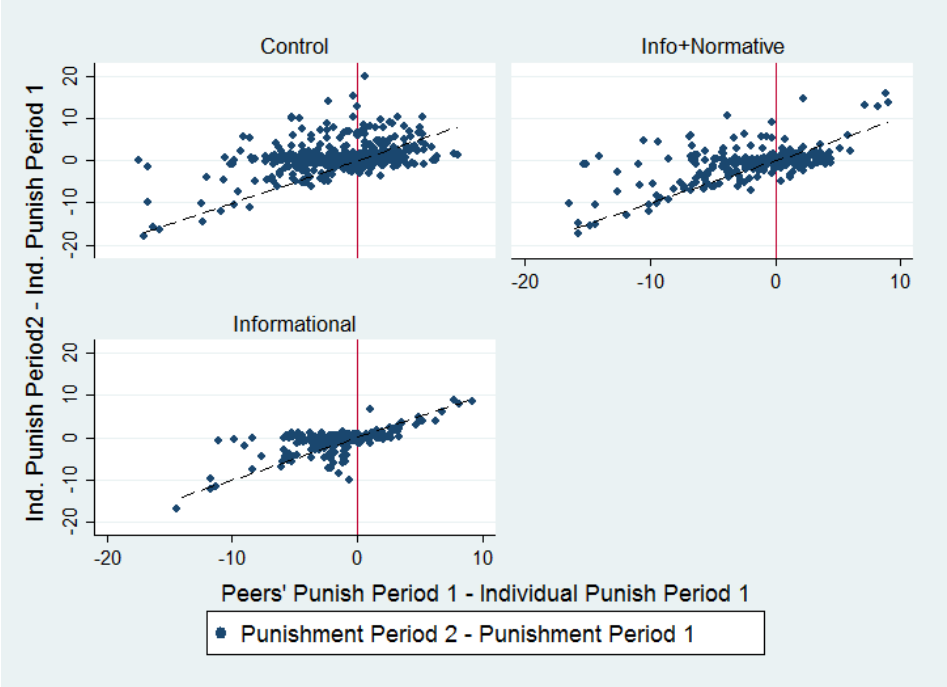


Figure 2: Change of punishment across periods. On the y-axis we report the difference between the second and the first period individual punishment. On the x-axis, we report the difference between peers' average first period punishment and individual first period punishment. An individual punisher seeking perfect conformity chooses a second period punishment such that the two measures are equal, hence the observation lies on the 45 degrees dotted line.

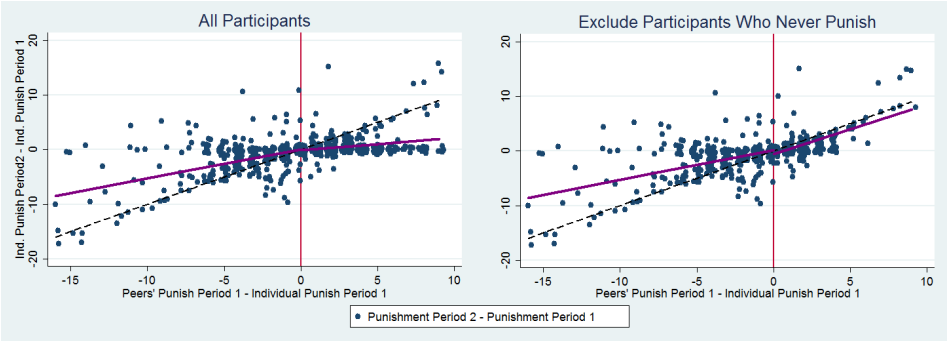


Figure 3: Direction of social influence effects (pooled data from informational and info+normative treatments). On the y-axis we report the difference between the second and the first period individual punishment. On the x-axis, we report the difference between peers' average first period punishment and individual first period punishment. An individual punisher seeking perfect conformity chooses a second period punishment such that the two measures are equal, hence the observation lies on the 45 degrees dotted line. The bold thick trend line fits the data. The intercepts and slopes are left unconstrained to differ between positive and negative punishment differentials. In the left panel all participants are included, in the right panel only participants punishing at least once a positive amount are included.

Appendix A. Theoretical Models

The Model by Cox et al. (2007)

Substituting the expressions for Π_i $(i = A, B, C)$ from Section 3 into the utility function (2) in the main text and differentiating with respect to the amount of punishment p, we obtain the first order condition for optimal p:

$$\frac{\partial u_c}{\partial p} = -(30 - p)^{\alpha-1} - 4\theta_B(30 + t - 4p)^{\alpha-1} = 0 \tag{A.1}$$

Since A's payoff does not depend on C's choice, θ_A drops out. Notice that an interior solution exists if and only if $\theta_B < 0$, i.e., if C feels negative emotions about the dictator. If $\theta_B \geq 0$, $\frac{\partial u_c}{\partial p} \leq 0$ always and, in equilibrium, $p^* = 0$. Solving A.1 with respect to p, we obtain:

$$p^* = \frac{\phi(30 + t) - 30}{4\phi - 1} \tag{A.2}$$

where $\phi = (-4\theta_B)^{\frac{1}{\alpha-1}} > 0$ given $\theta_B < 0$. Notice that p^* is increasing in t: the larger the taking by the dictator, the larger the punishment inflicted by the third party. However, if $t = 0$, $p^* > 0$ as long as $\phi < \frac{1}{4}$ or $\phi > 1$. Given that $\frac{\partial \phi}{\partial \theta_B} > 0$ and $\frac{\partial \phi}{\partial \alpha} < 0$, this means that dictators not taking anything from passive receivers might be punished anyway if either α is high (payoffs are highly substitutable in c's utility function) and θ_B low ($\phi < \frac{1}{4}$) or, viceversa, θ_B is high but α is low ($\phi > 1$).

The Model by Charness and Rabin (2002)

We consider the second utility function presented by Charness and Rabin (2002), augmented with strong reciprocity concerns and reported by function

(4) in the main text of our paper.

Following Thoni and Gachter (2015), we can write the derivative of equation (4) with respect to punishment p in a compact way as follows

$$\frac{\partial u_C}{\partial p} = -(1-\lambda) + \lambda\,\delta\,r - \lambda(1-\delta)[1+4(1-kd_B)] + 4\,\lambda\,f\,d_B \tag{A.3}$$

where

$$r = \begin{cases} -4, & \text{if } \Pi_B = min\{\Pi_A;\Pi_B;\Pi_C\} \\ -1, & \text{if } \Pi_C = min\{\Pi_A;\Pi_B;\Pi_C\} \\ \;\;0, & \text{if } \Pi_A = min\{\Pi_A;\Pi_B;\Pi_C\} \end{cases} \tag{A.4}$$

It can be seen that $\frac{\partial u_C}{\partial p} \geq 0$ if $\lambda[-4+\delta(5+r)+4d_B(f+k(1-\delta))] > 1$, that is if

$$d_B > \tilde{d}_B(r) = \frac{1+4\lambda-\lambda\delta(5+r)}{4(f+k(1-\delta))} \tag{A.5}$$

The punishment behavior of the third party depends on the values of the parameters λ and δ and on the different values of r, where r represents the marginal effect of a change in p on the minimum payoff in the group.

In this respect, it should be noted that Π_C can never be the lowest payoff, so that $r=-1$ is not an admissible value. To prove this, notice that $\Pi_C \geq \Pi_A$ if and only if $p \leq t$, whereas $\Pi_C \geq \Pi_B$ when $p \geq \frac{t}{3}$. Thus, if $p > t$, $\Pi_A > \Pi_C > \Pi_B$. If $p < \frac{t}{3}$, $\Pi_B > \Pi_C > \Pi_A$. If $\frac{t}{3} < p < t$, Π_C is the largest payoff, whereas $\Pi_A > \Pi_B$ if $p > \frac{t}{2}$. The relationship among payoffs in the $\{t,p\}$ space is depicted in figure A.4.

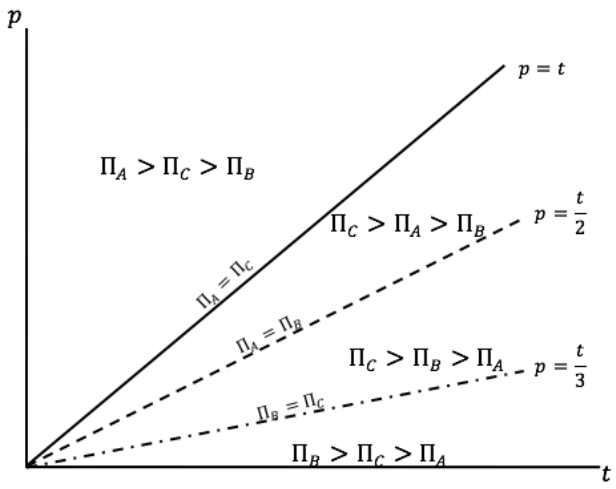


Figure A.4: Payoff rankings according to different combinations of t and p.

Thus, it is never the case that Π_C is the minimum payoff and we have to concentrate on the cases $r = -4$ and $r = 0$.

Substituting the values of $r = \{-4, 0\}$ in equation (A.5) yields

$$\tilde{d}_B(-4) > \tilde{d}_B(0).$$

We can then have the following cases:

1. $\tilde{d}_B(-4) > \tilde{d}_B(0) > d_B$. In this case, the demerit attached to B's hostile actions is too low and the third party never punishes.
2. $\tilde{d}_B(-4) > d_B > \tilde{d}_B(0)$. In this case, the inequality (A.5) is satisfied only for $r = 0$ and the third party is prepared to punish whenever the passive receiver's payoff is minimal. If $r = -4$ and $\Pi_B = min\{\Pi_A; \Pi_B; \Pi_C\}$ the third party does not punish. In fact, it takes a really high value of the demerit parameter to punish the subject

with minimum payoff when that imposes a cost $r = -4$ to the social welfare. We thus put ourselves in the case in which A has the minimum payoff. According to the minimax component of his utility, C should increase Π_A but he has no means to do that. At the same time, the positive $\frac{\partial u_C}{\partial p}$ implies that C is willing to punish up to the maximum allowed fine $p = 20$. However, such choice would not be feasible, since it would imply $\Pi_B < \Pi_A$, which violates our initial assumption that $\Pi_A = min\{\Pi_A; \Pi_B; \Pi_C\}$. Therefore, C sets $p = \frac{t}{2}$, equating $\Pi_A = \Pi_B$. In this case, there is a positive, monotonic relationship between t and p, as found in our experiment.

3. $d_B > \tilde{d}_B(-4) > \tilde{d}_B(0)$. In this case d_B is so high that $\frac{\partial u_C}{\partial p} > 0$ always and C applies the maximum punishment possible $p = 20$.

This function too is not able to account for peer effects.

The Model by López-Pérez (2008)

We start by computing the social norm concerning the level of punishment. We thus maximize equation (5) in the main text with respect to p.

Since Π_C is never the minimum payoff, as shown before, we have four different specifications for equation (5), each one corresponding to one area in figure A.4.

Starting from bottom to top:

1. $\Pi_B > \Pi_C > \Pi_A$**.** The social function is

$$F_{\varepsilon\delta} = \varepsilon(\Pi_A + \Pi_B + \Pi_C) - \delta\,(\Pi_B - \Pi_A)$$

The first order condition is

$$\frac{\partial F_{\varepsilon\delta}}{\partial p} = -5\varepsilon + 4\delta \tag{A.6}$$

which is linear in p. This implies that $\frac{\partial F_{\varepsilon\delta}}{\partial p} \geq 0$ if and only if $\delta > \frac{5}{4}\varepsilon$. Thus, the punishment will be the maximum possible compatibly with the payoff ordering and B will be punished until $\Pi_B = \Pi_C$, so that $p = \frac{t}{3}$.

2. $\Pi_C > \Pi_A > \Pi_B$**.** Here $\frac{\partial F_{\varepsilon\delta}}{\partial p} = -5\varepsilon - 3\delta p < 0$ always and punishment is zero.

3. $\Pi_C > \Pi_B > \Pi_A$**.** As in the first case, $\frac{\partial F_{\varepsilon\delta}}{\partial p} = -5\varepsilon + \delta \geq 0$ for $\delta > 5\varepsilon$. Again, the punishment will be the maximum possible compatibly with the payoff ordering. Here $p = \frac{t}{2}$ and $\Pi_B = \Pi_A$.

4. $\Pi_A > \Pi_C > \Pi_B$**.** Here, $\frac{\partial F_{\varepsilon\delta}}{\partial p} = -5\varepsilon - 4\delta < 0$ and no punishment occurs.

Thus, the exogenous social norm predicts $p = \hat{p} = \{\frac{t}{3}; \frac{t}{2}\}$ according to the initial distribution of income.

Given the social norm, a third party maximizes the utility function (6) in the main text.

In our game, a third party is not able to observe r, the number of other third parties who abided by the social norm, but is able to see whether the norm is respected on average or not. Thus, $r \in \{0, 1\}$, where $r = 1$ if $\bar{p} = \hat{p}$ ($\bar{p}$ defines the average punishment) and $r = 0$ if $\bar{p} \neq \hat{p}$.

A third party respects the norm if the cost $\hat{p}$ of respecting the norm is lower than the cost of norm deviation γr. The value of r is updated after the first

period, when third parties receive information about average punishment. In the first period, we start with $r = 1$ because no third party has had an opportunity to deviate so far.[14] Thus the third party punishes according to the norm if $\gamma < \frac{t}{d}$, where $d = \{2, 3\}$. If, after the revelation of information, the agent learns that the norm is not followed on average in the society, he too stops respecting it. Else, he keeps on sticking to it.

The Modified López-Pérez (2008) Model

Given the modified utility function (7) in the main text, we can distinguish the following cases.

1. $p < \hat{p}$ **and** $p < \bar{p}$**.** Utility from deviation is

$$u_{C1} = 30 - p - \gamma(\hat{p} - p) - \mu(\bar{p} - p) + f\,p$$

C sticks to the norm if and only if

$$\gamma > \tilde{\gamma}_1 = 1 - f - \frac{\mu(\bar{p} - p)}{\hat{p} - p}$$

Thus, C is more likely to conform (increasing punishment to $\hat{p}$) the lower $\tilde{\gamma}_1$, i.e., the higher his taste for punishment f, the greater social influence μ (that here pushes for a higher p) and the closer p to $\hat{p}$.

2. $p > \hat{p}$ **and** $p < \bar{p}$**.** In this case $\hat{p}$ is unambiguously lower than $\bar{p}$. C would

[14] Alternatively, we could set r according to the third party's belief about average punishment, so $r = 1$ if he believes that on average punishment will be $\hat{p}$ and $r = 0$ otherwise. This would not change the nature of the results.

conform if

$$\gamma > \tilde{\gamma}_2 = -1 + f - \frac{\mu(\bar{p} - p)}{p - \hat{p}}$$

that is, if f is small (here C has to lower his punishment towards the norm) and the smaller social influence μ is (again, pushes for a high p and an increase in p to match the norm would widen the difference $\bar{p} - p$). Finally, conformity is easier if p is close to $\hat{p}$.

3. $p < \hat{p}$ **and** $p > \bar{p}$**.** In this case, $\bar{p}$ is unambiguously higher than $\hat{p}$. C conforms to the norm (increasing p even if social influence pushes for a lower punishment) if and only if

$$\gamma > \tilde{\gamma}_3 = 1 - f - \frac{\mu(p - \bar{p})}{\hat{p} - p}$$

Conformity is more likely if C's taste for punishment f is high and social influence μ small.

4. $p > \hat{p}$ **and** $p > \bar{p}$**.** Here conformity requires

$$\gamma > \tilde{\gamma}_4 = -1 + f - \frac{\mu(p - \bar{p})}{p - \hat{p}}$$

i.e., a low taste for punishment and large social influence.

From the formal analysis above, it can be seen that f facilitates convergence to $\bar{p}$ when $p < \bar{p}$, while impeding it when $p > \bar{p}$.

For given f, cases characterized by $p < \hat{p}$ and $p < \bar{p}$, as well as cases with $p > \hat{p}$ and $p > \bar{p}$ entail faster convergence. Clearly, third parties with a high γ relative to μ end up closer to $\hat{p}$ than to $\bar{p}$ and vice-versa.

The Model with Focal Points and Taste for Punishment

Given the utility function (8) in the main text, in order to study people's decision to conform, we need to separate between the case in which they punish less than average in the first period from the case in which they punish more. As before $\bar{p}$ represents the average punishment disclosed to third parties at the end of the first period. We thus put ourselves at the beginning of the second period, when third parties take their punishment decisions after observing $\bar{p}$.

1.$p > \bar{p}$**.** Third parties conform, reducing their level of punishment if

$$30 - \bar{p} + f\,\bar{p} > 30 - p + fp - \gamma(p - \bar{p})$$

that is, if $\gamma > \gamma_1^F = f - 1$.

2.$p < \bar{p}$**.** Third parties conform, increasing their level of punishment if

$$30 - \bar{p} + f\,\bar{p} > 30 - p + fp - \gamma(\bar{p} - p)$$

that is, if $\gamma > \gamma_2^F = 1 - f$.

Notice that $\gamma_2^F < \gamma_1^F$ if $f > 1$, that is, if the third party receives a high utility from punishment. This means that people with a large taste for punishment are more likely to conform when they have to increase punishment than when they have to decrease it.

Appendix B. English Translation of the Instructions

Welcome! This is a study on individual decision-making. Participants' answers are completely anonymous. It will not be possible for the data analysts to link individual answers to the participants who provided them. You earned 5 Euros for showing up on time today. Additionally, you can collect other earnings. The amount of these earnings depends on your choices and on the choices other participants will make during this study. During the study you earn "tokens". For every 10 tokens you earn, 1 Euro will be paid to you. In the unlikely case you collect negative earnings, these losses will be subtracted from your participation fee. If you have questions at any time, please raise your hand and wait for a researcher that will answer your questions privately. Please switch off and remove from the table your mobile phone and any electronic devices. Do not talk or communicate with other participants during this study. This study comprises several parts. Earnings obtained in each part of the study are independent of those obtained in the other parts. Your final earnings consists of:

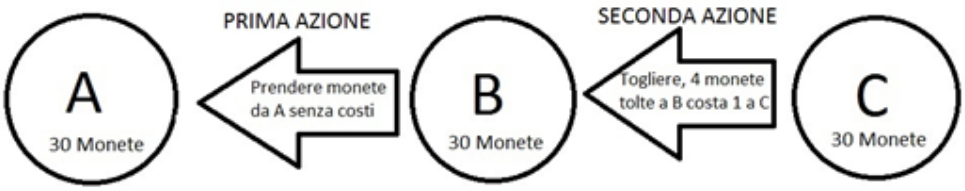
- 5 Euros as participation fee;
- Earnings collected in the first part of the study;
- Earnings collected in one part after the first one. At the end of the study the computer will randomly select the part from which earnings will be paid to you.

Final earnings will be paid privately and in cash at the end of the study.

Instructions for Part I: Description of the situation (Belief Elicitation Game. Instructions for this part are the same in the 3 treatments)
Consider a situation in which 3 people are present. A role is assigned to each person randomly: "participant A", "participant B" and "participant C". A, B and C can

make decisions and earn tokens.

- participant A receives 30 tokens and does not make decisions
- participant B receives 30 tokens. In addition, B could take some or all of A's tokens and add them to his/her own earnings without incurring costs. Precisely, B can take 0, 5, 10, 15, 20, 25 or 30 tokens from A.
- participant C receives 30 tokens, observes B's action and can eliminate some of B's tokens, incurring a cost. For each 4 tokens eliminated from B's earnings, C has to pay 1 token. Participant C could use up to 20 tokens to reduce B's earnings. C's decision does not affect A's earnings.

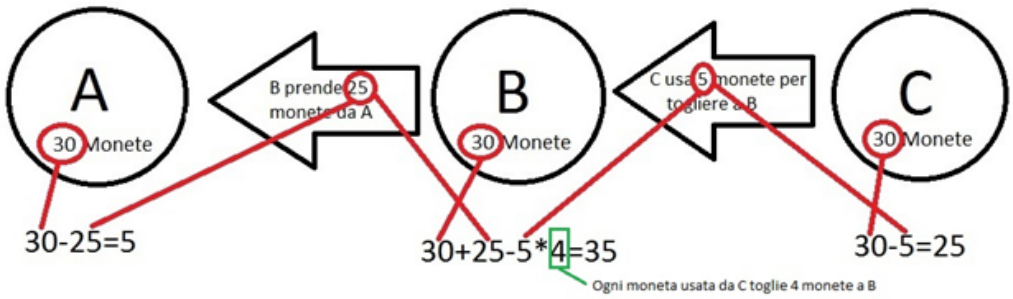


Therefore, A, B and C earnings are:

- participant A: (30 initial tokens) – (tokens taken by B)
- participant B: (30 initial tokens) + (tokens taken from A) – (4 × no. of tokens used by C)
- participant C: (30 initial tokens) – (tokens used to reduce B's earnings)

Example 1) (please look at your computer screen): B takes 25 tokens from A. After observing B's choice, C decides to use 5 tokens to reduce B's earnings. Therefore participants' final earnings are:

- participant A = 5 tokens (tokens B left)
- participant B = 35 tokens (30 initial tokens + 25 tokens taken from A – 5*4= 20 tokens coming from the 5 tokens used by C to reduce B's earnings)
- participant C = 25 tokens (30 initial tokens – 5 tokens used to reduce B earnings)

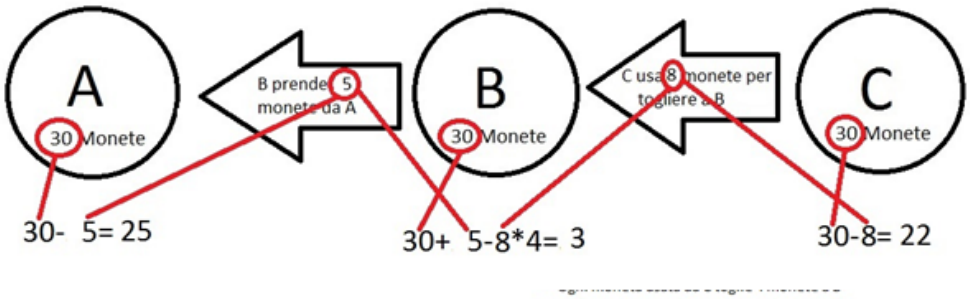


Example 2) (please look at your computer screen): B takes 5 tokens from A. After observing B's choice, C uses 8 tokens to reduce B's earnings. Therefore participants' final earnings are:

- participant A = 25 tokens (left by B)
- participant B = 3 tokens (30 initial tokens + 5 tokens taken from A – 8*4=32 tokens coming from the 8 tokens used by C to reduce B's earnings
- participant C = 22 tokens (30 initial tokens – 8 tokens used to reduce B's earnings)

Your actions and earnings

participant C observes how many tokens B takes from A. You and the other participants in the laboratory have to indicate the number of tokens (an integer between 0 and 20) that you believe C in will use. When everyone answered, we compute the average of the individual amounts indicated by you and by the other participants.

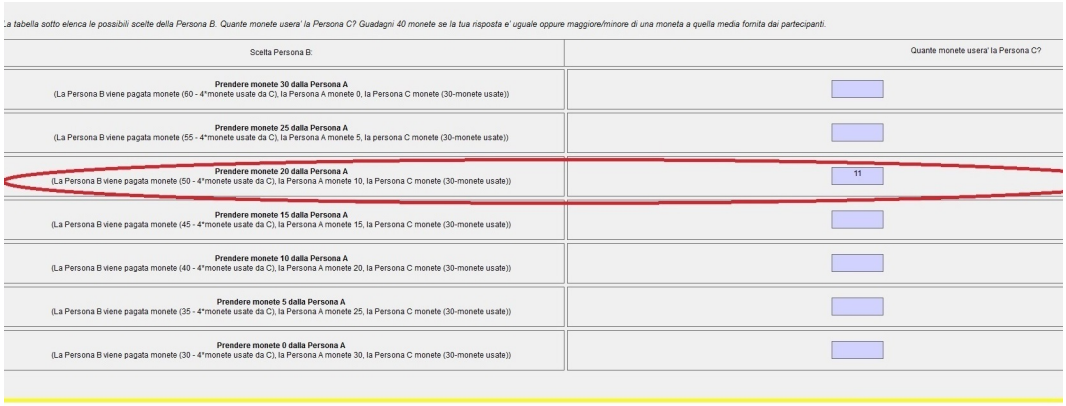


If the number you indicated is equal to the actual average or it either exceeds or falls short of the average by one token, you receive 40 tokens that will be added to your final earnings (if you indicate 1, you receive the forty tokens if the actual average is 0, 1 or 2; if you indicate 20, you receive the 40 tokens if the average is 20, 19 or 18). Conversely, you do not earn tokens in this part of the study if the number you indicate is bigger or smaller than the average by more than one unit.

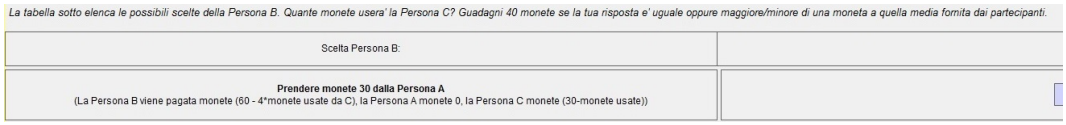
Example 1) (please look at your computer screen): Consider B's action "take 20 tokens from A, totalizing 50 tokens, leaving 10 tokens to A". You indicate that C uses 11 tokens. You receive 40 tokens if on average all the participants to the study indicated "11", "10" or "12" tokens. If the average of all individual responses is different, you will not earn tokens for this part of the study .

Example 2) (please look at you computer screen): Consider the action of B "take 0 tokens from A and collect 30 tokens, leaving 30 tokens to A". You indicate that C uses 3 tokens. You receive 40 tokens if on average all the participants to the study indicated "3", "2" or "4" tokens. If the average is different, you will not earn tokens for this part of the study.

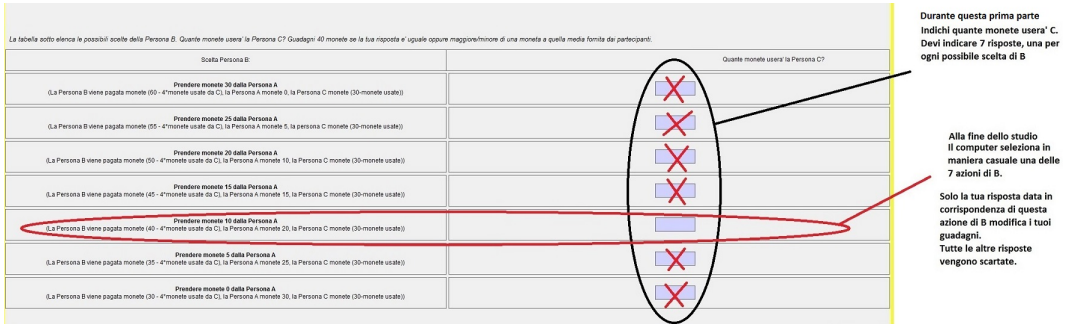
You are required to indicate how many tokens you believe a participant C uses for each of the possible 7 actions by B (B takes 30 tokens from A; B takes 25 tokens...; B takes 0 tokens from A). At the end of the study, the computer will randomly select one of the 7 actions by participant B. We will verify if you earned the 40



Guadagni le 40 monete se in media i partecipanti allo studio avranno indicato "11" oppure "10" oppure "12". Guadagni 0 se la media e' diversa da questi valori.



tokens specifically for the selected action. Your decisions and those of the other participants relative to other possible actions by B will be discarded and will not affect your final earnings.



Before the beginning of this first part of the study, we ask you to answer some *control* questions. Answers to these *control* questions will not affect your final

earnings.
(participants answer *control* questions on their computers. The Ztree file containing the *control* questions is available upon request from the authors).

Instructions for Part II: Description of the situation (Dictator Game: Instructions on this part are the same in all treatments)

Consider the same situation described in Part I, where there are 3 people present, A, B and C, who can make decisions and earn tokens. Exactly as in the first part:

- A receives 30 tokens and does not make decisions;
- B receives 30 tokens and could take some or all the tokens of A;
- C receives 30 tokens, observes B's action and can reduce B's earnings paying a cost (for every 4 tokens of reduction of B's earnings, C has to pay 1 token).

Your actions and earnings
In this second part you and the other participants have to make decisions first as "participant B" then as a "participant C". Therefore, you have to indicate:

- First, as "participant B", how many tokens you take from A.
- Afterwards, as a "participant C", how many tokens you use for reducing B's earnings for any possible action by B.

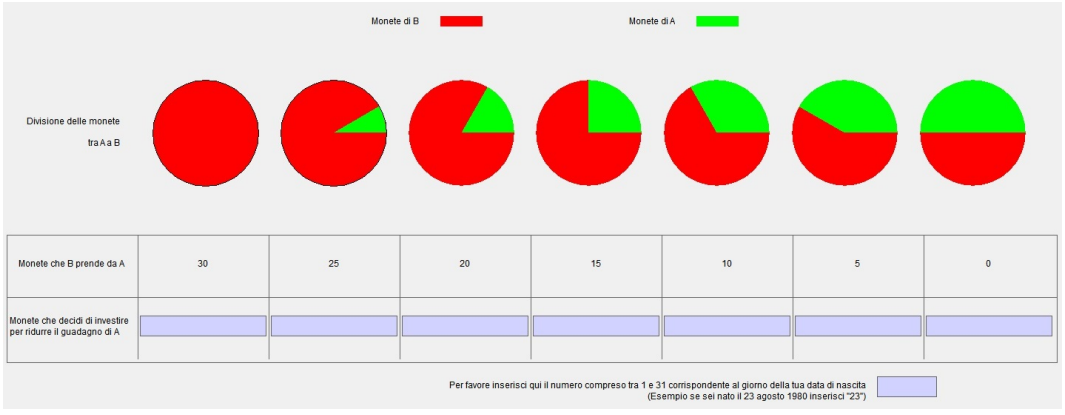
Why do you have to make decisions both as a "participant C" and as a "participant B"?
In calculating final earnings, each participant is associated to a unique role: either participant A or participant B or participant C. However, you and the other participants will not learn which role you have been assigned to until the end of today's study.

Seconda Parte

Prendi una decisione come Persona B.

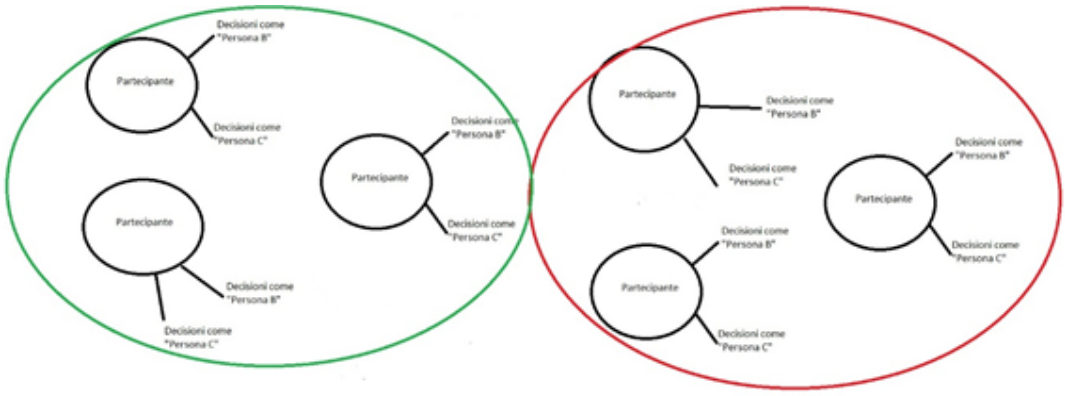
Quante monete vuoi prendere dalla Persona A?

- 0
- 5
- 10
- 15
- 20
- 25
- 30

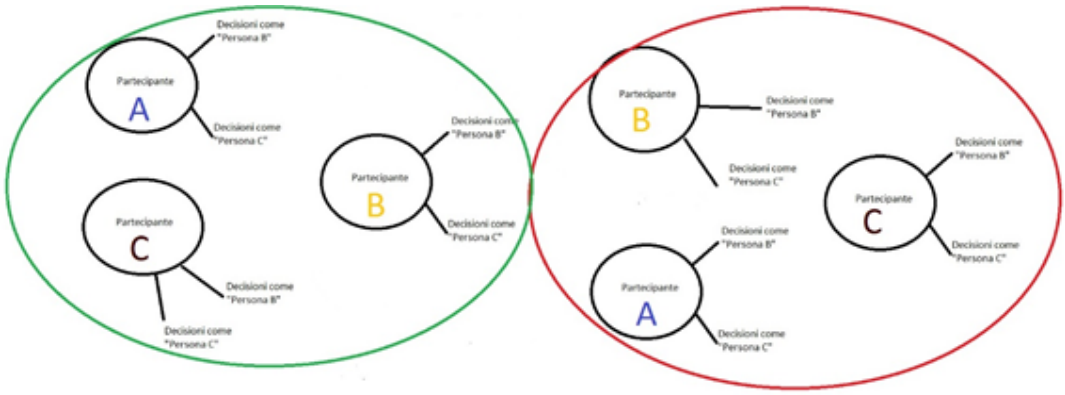


Monete che B prende da A	30	25	20	15	10	5	0
Monete che decidi di investire per ridurre il guadagno di A							

Per favore inserisci qui il numero compreso tra 1 e 31 corrispondente al giorno della tua data di nascita
(Esempio se sei nato il 23 agosto 1980 inserisci "23")

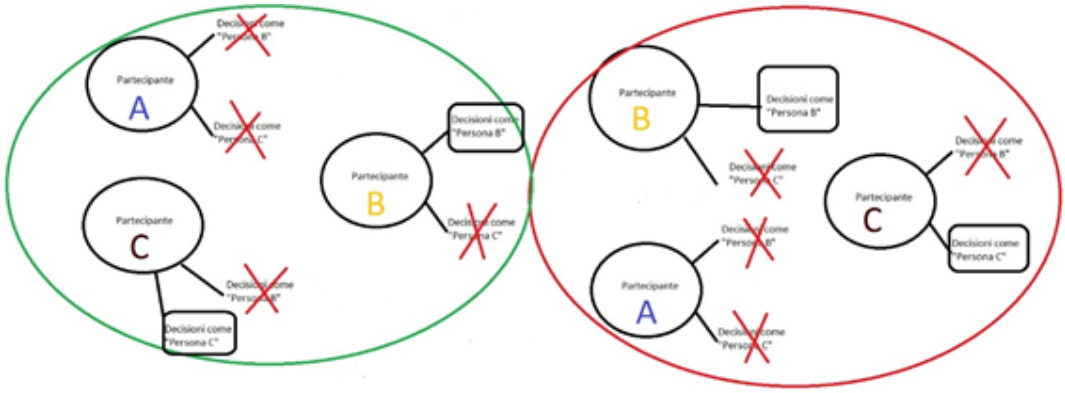


Indeed, you and the other participants will be randomly divided in groups of 3. Within the group, each one of the 3 participants is assigned either to role A, B or C.



Assignment to groups and assignment of roles is completely random and each participant has 1 possibility out of 3 of being assigned a specific role. Therefore, if you are assigned the role "participant A", your final earnings are determined by the tokens left to you by the participant B in your group. Other decisions you make as a participant B or C will be discarded and have no influence on your final earnings nor on the earnings of the other participants. Similarly, participants assigned to

the role "participant B" determine their final earnings and those of the other group components only by the decisions made as participant B. Decisions made as participant C have no effects on final earnings. Finally, also participants assigned to the role "participant C" influence final earnings only with the decisions they make as C.



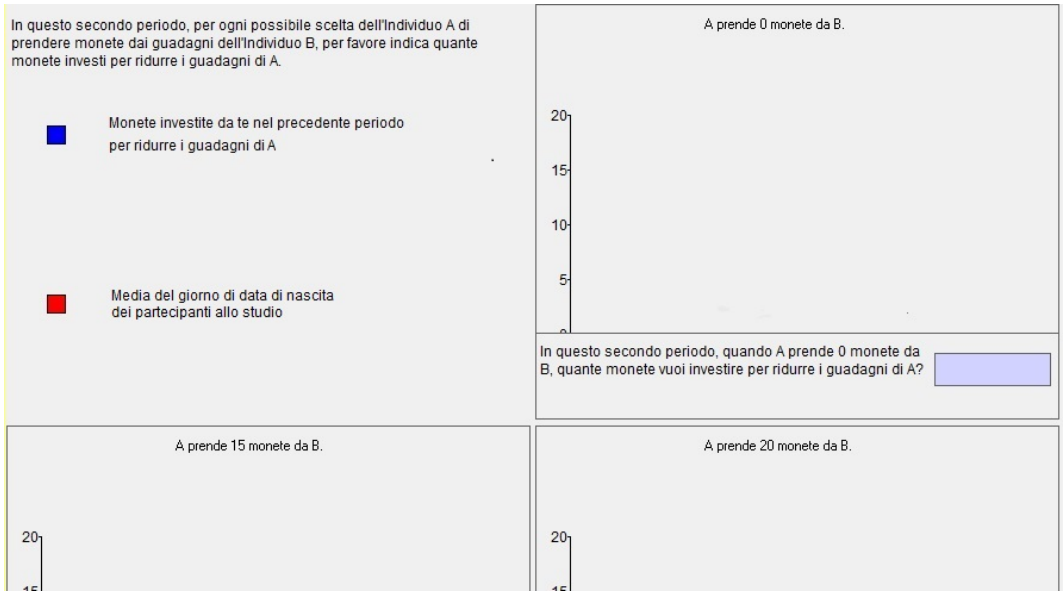
During this second part of the study we will also ask you to indicate the day of the month in which you where born (E.g. if you were born January 25th 1983 you should report "25").

Earnings for A, B and C in this second part are determined exactly as in the first part:

- participant A: (30 initial tokens) – (tokens taken by B)
- participant B: (30 initial tokens) + (tokens taken from A) – (4 × the no. of tokens used by C)
- participant C: (30 initial tokens) – (tokens used to reduce B's earnings)

Before starting this second part of the study, we ask you to answer some *control* questions. Answers to these *control* questions will not affect your final earnings.

Instructions Part III (*info+normative* Treatment; instructions for *control* and *informational* are available upon request)



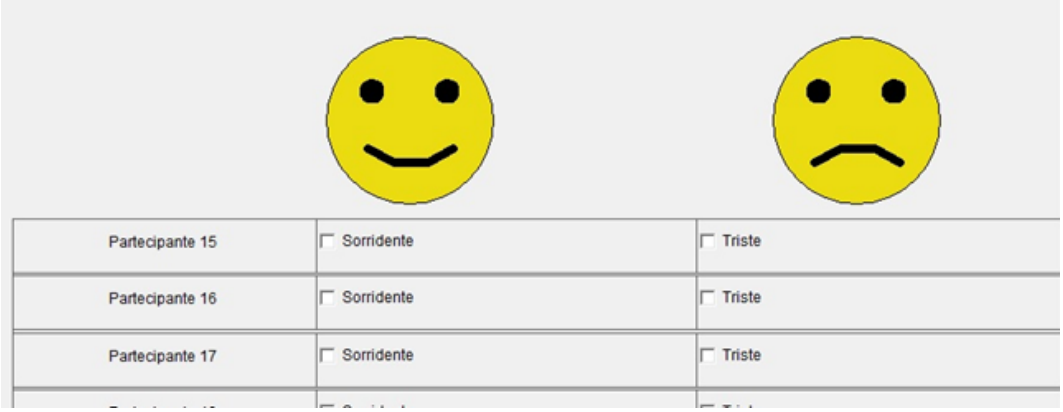
Now the third and last part of this study is about to begin. At the end of this part, we will ask you to fill in a brief questionnaire and then we will proceed with payments. Consider exactly the same situation you faced in the second part of this study, same roles of A, B and C, same possible decisions for B and C, same initial endowments and possible earnings. As in the second part, you have to make decisions first as a participant B then as a participant C. Moreover, in this third part, before making your decisions you will receive information regarding other participants. You will receive information on decisions made as participant C by the participants at today's study. You will be told how many tokens participants used on average in the second part of this study to reduce B's earnings. You will receive this information for any of the 7 possible B's choices.

Furthermore, before the end of the study, the decisions you are going to make as "participant C" in this third part will be revealed to 5 other randomly selected participants. Similarly, you will receive information about the individual choices made as participant C by 5 other participants.

Monete che B prende da A	**Media** monete riduzione guadagni B **parte 2**	Monete usate **Partecipante 1** parte 3	Monete usate **Partecipante 2** parte 3	Monete usate **Partecipante 3** parte 3	Monete usate **Partecipante 4** parte 3	Monete usate **Partecipante 5** parte 3
0	0	0	0	0	0	0
5	0	0	0	0	0	0
10	0	0	0	0	0	0
15	0	0	0	0	0	0

Each participant will be randomly assigned an ID number. The ID number assigned is independent of the number of the PC you are using. After you saw the individual

choices by the other 5 participants, you and the other participants will be able to vote to send a smiling or a sad emoticon



You receive a smiling emoticon if the majority of the five participants who saw your choices vote for "smiling". Otherwise you will receive a sad emoticon. The emoticon will remain on your screen for one minute, then it disappears automatically. After the minute passed, you will know your final earnings.

If you have questions, please raise your hand and we will answer privately. Otherwise press the "Continue" button and start sthe third part.

Tables Appendix

Description of the Variables Used in the Regressions

Table B.7: Variables

Variable	Description
degree	1 if subject completed 8th grade ('secondary school'), 2 if subject completed high school, 3 if subject has a bachelor degree or equivalent, 4 if subject has a master degree or equivalent, 5 if subject has a PhD or equivalent
job	dummy variable, 1 if worker, 0 otherwise
male	dummy variable, 1 if male, 0 if female
age	subject's age
social	dummy variable, 1 if subject is a student in social sciences and medicine, 0 otherwise
arts	dummy variable, 1 if subject is a student in arts or humanities, 0 otherwise
field_other	dummy variable, 1 if subject not in social or arts, 0 otherwise
DictatorTake1	tokes a subject take from the receiver period 1 when choosing as a dictator
DictatorTake2	tokes a subject take from the receiver period 2 when choosing as a dictator
risk	$\in$ [1, 10], 1 if the answer to the question 'In general, do you consider yourself ready to take risks?'is 'Not at all', 10 if the answer is 'Totally ready to take risks'

Variable	Description
logic	$\in$ [0, 2], 1 point for each correct answer.
impulsivity	$\in$ [0, 3], 1 point for each correct answer.
info+normative	dummy variable, 1 for participants in info+normative treatment
Treated	dummy variable, 1 for subject either in info+normative or in informational treatments, 0 otherwise
Punish1	punishment in the first period
Punish2	punishment in the second period
Beliefs	beliefs about peers' average punishment in the first period
diffBelAvgP1	Beliefs - Average peers' punishment period 1
ConvergeDummy	dummy, = 1 if (diffAvgP1P1>0 & Punish2-Punish1>0) or (diffAvgP1P1< 0 & Punish2-Punish1< 0)
Instructions	total time employed by participants to correctly answer control questions
diffAvgP1P1	Average peers' punishment period 1 - Punish1
diffP2P1	Punish2 - Punish1
Treat2016	dummy=1 sessions in 2016

Summary Statistics - First Period Punishment

Treatment	TakeRate_0	TakeRate_5	TakeRate_10	TakeRate_15	TakeRate_20	TakeRate_25	TakeRate_30
control							
(mean)	1.18	2.07	2.40	2.96	4.13	4.82	5.79
(median)	0	0	2	3	4	5	5
(SD)	3.31	4.09	2.74	2.97	3.94	4.68	5.72
info+normative							
	.36	1.73	2.58	3.58	4.29	4.87	5.73
	0	1	2	2	4	5	5
	1.05	2.86	3.13	4.36	4.44	5.03	6.01
informational							
	1.38	1.90	3.36	3.74	5.14	6.17	7.33
	0	1	3	4	5	6.5	8.5
	3.60	2.16	3.83	3.63	5.00	5.48	6.15
Total	.96	1.90	2.77	3.42	4.51	5.27	6.26
	0	1	2	3	5	5	6
	2.88	3.14	3.25	3.69	4.46	5.06	5.96

Table B.8: Average First Period Punishment by Levels of dictator Taking. StratP0 = dictator take 0 tokens from receiver

Summary Statistics - Second Period Punishment

Treatment	TakeRate_0	TakeRate_5	TakeRate_10	TakeRate_15	TakeRate_20	TakeRate_25	TakeRate_30
control							
(mean)	1.02	2.02	2.73	3.36	4.42	5.02	6.22
(median)	0	1	2	2	4	4	6
(SD)	3.18	3.45	3.49	3.58	4.38	5.24	6.26
info+normative							
	.62	1.69	2.47	3.18	3.87	4.2	5.18
	0	1	2	3	3	4	4
	1.99	2.86	3.27	3.63	4.31	4.83	6.02
informational							
	1.5	2.26	2.95	3.60	4.60	5.48	6.31
	0	1	3	4	5	6	8
	4.07	3.92	3.66	3.87	4.24	4.83	5.54
Total							
	1.04	1.98	2.71	3.37	4.29	4.89	5.89
	0	1	2	3	4.5	5	6
	3.16	3.41	3.45	3.67	4.29	4.96	5.93

Table B.9: Average Second Period Punishment by Levels of dictator Taking. Punish0 = dictator take 0 tokens from receiver