Title: Understanding Other-Regarding Mechanisms in Heterogeneous Populations

Authors: Teck-Hua Ho [a,b,c 1,2,3,4], Ming Hsu [d 1,2,3,4], Xing Zhang [e 2,3,4], Songfa Zhong [f 1,2]

a. Corresponding author
b. Haas School of Business, University of California at Berkeley, 545 Student Services #1900, Berkeley, CA 94720, USA. hoteck@berkeley.edu, +1 (510) 643-4272
c. NUS Business School, National University of Singapore, BIZ 1, 8-26, 15 Kent Ridge Drive, Singapore 119245.
d. Haas School of Business, University of California at Berkeley, 545 Student Services #1900, Berkeley, CA 94720, USA.
e. NUS Business School, National University of Singapore, BIZ 1, 8-26, 15 Kent Ridge Drive, Singapore 119245.
f. Department of Economics, Faculty of Arts and Social Sciences, National University of Singapore, AS2 #06-02, 1 Arts Link, Singapore 117570.

Author Contributions:
1. Design
2. Research
3. Data analysis
4. Writing

Keywords: social justice, type heterogeneity, social mechanism

Classification: Social Sciences; Economic Sciences, Political Sciences

Abstract (236/250)

Social injustice and altruism are daily occurrences in every society. When one individual treats another unfairly, a bystander may step in to correct the injustice by punishing the norm-violator or helping the victim. While the latter may be more efficient, experiments show that punishing the norm-violator is actually more effective in curbing undesirable behavior and maintaining social norms. We posit that a robust mechanism that enforces distributive norms must accommodate heterogeneous (selfish or other-regarding) types of individuals and their strategic interactions. To model such a mechanism, we combined laboratory games with latent class modeling to characterize the effects of the interactions between social justice mechanisms and a mixture of selfish and other-regarding types. Specifically, we investigated a three-person, repeated game in which a third-party bystander could monetarily help the victim or punish the norm-violator. We found that a model that allows for a mixture of types explains choice behavior significantly better than a representative agent model. Critically, we found that the superiority of the 'punish' condition in enforcing norm-compliance depended on the norm-violator and the third-party being other-regarding. In contrast, if either was selfish, norm-enforcement under the 'help' condition was equal or superior to that under the 'punish' condition. These results show that it is crucial to know the proportion of types of individuals in order to best understand and predict the effectiveness of a social mechanism so that it reinforces rather than impairs other-regarding behavior.

Significance Statement (116/120)

This paper shows the importance of knowing the composition of selfish and other-regarding individuals in a population before designing norm-enforcing mechanisms. Understanding this heterogeneity and the interactions between individuals is critical in determining whether it is more effective for a third-party to punish norm-violators or help victims. Intuitively, the worst results were observed when the norm-violator and the third-party were selfish, and the best when both were other-regarding. However, since societies are in reality heterogeneous, the results from such populations are the most interesting, showing that punishment and help can both be effective, albeit with different heterogeneities; punishment is better when the norm-violator and third-party are other regarding, while help is better when either is selfish.

\body

Introduction

A core question in every society is how to design social mechanisms so that they reinforce distributive norms and promote prosocial behavior (1). Numerous field observations have documented how self-organized institutions can solve collective action problems, reduce poverty, and promote justice, all of which are of central interest to social and biological scientists (2, 3).

Two major factors have been identified in theoretical and field studies as important determinants of a society's ability to enforce distributive norms: (i) the types of individuals interacting in the society, broadly categorized as selfish or other-regarding (4, 5, 6), and (ii) the types of mechanisms used to promote equity and other just outcomes, varying between those that punish individuals who act selfishly, and those that help individuals harmed by the selfish actions of others (7). A simple way to demonstrate the impact of such mechanisms on social justice outcomes is to study third-party games (8), which capture how an impartial third-party responds when one individual treats another unfairly. In a sanctioning mechanism like a consumer boycott campaign, the third-party punishes the norm-violator by reducing the violator's material payoffs, whereas in a subsidy mechanism like a charity organization, the third-party helps the victim by increasing the victim's well-being.

Evidence for the determinative nature of these factors has largely come from two sources. First, in field and laboratory studies, substantial evidence exists that variations in mechanism have a significant impact on the mean level of norm-compliance and the stability of compliance across different populations (1, 9). Second, even in controlled laboratory conditions, there is surprising variation in the level of norm-compliance observed across groups, ranging from extremely high to extremely low, even under relatively homogeneous conditions (5). This suggests the existence of unobserved heterogeneity in the composition of the participants.

Surprisingly, despite the importance ascribed to 'punish' and 'help' mechanisms, there is little direct evidence documenting conditions in which one mechanism outperforms the other, and how heterogeneity in the population affects this comparison. It is also unclear which types of individuals drive reductions in social injustice and how this reduction might vary across mechanisms. One widely discussed account on the superiority of 'punish' mechanisms shows that reducing the attractiveness of selfish behavior promotes cooperation. For example, in public goods games, allowing for the possibility of punishment produces a strong tendency towards cooperation (9, 10).

More broadly, existing approaches are limited by challenges related to (i) isolating the effects of social mechanisms on norm-enforcement, and (ii) characterizing types of individuals in terms of other-regarding concerns that are not directly observable. Although field evidence has been invaluable in demonstrating the importance of social mechanisms, isolating how the various types of individuals shape norm-enforcement in the field is extremely difficult, if not impossible. It is even more difficult to assess the effects of interactions between types of social mechanisms and types of individuals using field evidence because mechanisms tend to continue evolving over time (9, 11).

These are not issues in a controlled laboratory setting, where we can observe behavior at the individual level and statistically quantify types of individuals using a latent class approach (12, 13). Unlike approaches that use random effects or individual-level estimations where individual differences are distributed along a continuum, latent class models assume that the population is composed of discrete classes of individuals, mirroring the notion of 'types' frequently invoked in theoretical and agent-based simulations (14). Furthermore, once individual types are statistically calibrated using experimental data, simulations can then be performed, just like in agent-based simulations, to generate new insights (15, 16,

17). By combining laboratory games with latent class modeling, we can characterize the interactions between types of individuals, the social mechanisms used, and the consequences of this interaction on norm-compliance and social justice. In this paper, we focus on two social mechanisms that are pervasive in human societies in enforcing a distributive norm: (i) punishing norm-violators and (ii) helping poorer individuals (Fig. 1A).

The Third-Party Game

We studied the behavior of subjects in two types of three-person games – the third-party punishment game and the third-party help game – that capture the interactions between types of social mechanisms and types of individuals. Each game involves three players – a dictator (D), a recipient (R), and a third-party (TP). In both games, D is given an endowment of 100 monetary units (MU), which D can share with R in any proportion. After the units are shared, D's decision is revealed to TP, who is endowed with 120 MU. In the punishment game, TP must decide whether or not to spend 10 MU to punish D by 40 MU, whereas in the help game, TP must decide whether to spend 10 MU to help R by 40 MU (Fig. 1C).

One important feature of the games is that other-regarding actions are taken by D and TP on a *voluntary* basis rather than because of formal mechanisms. Standard game theory predicts that D won't distribute any proportion to R and that TP will neither punish D nor help R. Another feature of the game is that unlike the prisoner's dilemma game, other-regarding actions do not lead to Pareto improvement. In the punishment game, TP's other-regarding actions are driven by egalitarianism, whereas in the help game, other-regarding motives are driven by both egalitarianism and social welfare efficiency concerns, which potentially provide a stronger incentive for TP to take the other-regarding action (7, 18, 19). Note that, by design, TP's other regarding action always reduces the difference in payoffs between D and R by the same amount (i.e., 40 MU) independent of the social mechanism.

The reduction in inequity consists of two components – D's norm-compliance and TP's inequity reduction. The initial inequity is 100 MU, which is the difference in endowment between D and R. D's norm-compliance can reduce inequity at the outset; the inequity can then be further reduced by TP. Our focus is on final inequity, which is the outcome of the interactions between D and TP.

We designed the experiment with the following considerations: (i) TP's payoff is always higher than that of D and R. This is done to control social comparisons with the other players in the triplet. This means that TP's motivation is induced purely from a sense of social justice rather than jealousy or envy. (ii) Since we were interested in comparing the impact of different social mechanisms on social inequity, to make the mechanism exogenously determined, we did not offer TP the option to choose between punishing D and helping R. (iii) We adopted a repeated game with fixed matching protocol because we are studying a social setting with complete information where each party learns if the others are selfish or other-regarding after each play. Since moves are sequential and no additional motivation is provided for cooperation, the end-game effect is minimal.

Results

Aggregate Dictator and Third-Party Behavior

Fig. 2A shows the mean frequency of TPs' inequity reduction. On average, TPs chose more inequity reduction under the 'help' condition. Logistic regression analysis further confirms that TPs implemented inequity reduction depending on what D gave (Tab. S1). Not surprisingly, D was more norm-compliant and gave more under the 'punish' condition (Fig. 2B). The total inequity reduction was determined by D's norm-compliance and the likelihood of TP's inequity reduction. Fig. 2C shows that although TP was

more likely to reduce inequity under the 'help' condition, it was actually the 'punish' condition that led to a more equitable outcome. Put differently, the 'punish' mechanism is superior to the 'help' mechanism despite the fact that the latter is socially more efficient than the former.

Utility Model Characterization of the Heterogeneity of Types

These results emphasize the importance of the interaction between D and TP in different social mechanisms. However, the aggregate results do not account for the possibility that the outcome actually emerged based on the interaction between different types of individuals within a mechanism. It is also unknown how the outcome will change if the composition of individual types changes across societies and over time.

Arguably, in the 'help' mechanism, the greater the number of other-regarding TPs, the greater the reduction in social inequity. In such a scenario, even if D is other-regarding, he or she may wish to free-ride and reduce the amount given to R knowing that TP will act to reduce inequity. As a result, the superiority of the 'help' mechanism depends on the proportion of other-regarding Ds in the population. Similarly, in the 'punish' mechanism, the greater the proportion of other-regarding TPs, the greater the reduction in inequity. In this scenario, even if Ds are selfish, they may be forced to behave like other-regarding types in order to avoid being punished, decreasing *ex ante* inequity. Substantial field evidence suggests an alternative where the effectiveness of a social mechanism depends upon the relative composition of different types of individuals whose behavior depends upon the type of social mechanism used (1, 20).

To investigate the interaction between social mechanisms and the heterogeneity of types, we have to provide a benchmark to understand the degree of heterogeneity in the population in characterizing 'punish' and 'help' behavior. This requires a model that simultaneously accounts for (i) the interaction between the decisions made by D and TP, (ii) the effects of different social mechanisms, and (iii) the existence of different types of D and TP.

We simplified the population into a mixture of only two types of D and TP – selfish dictator (SFD) and other-regarding dictator (ORD), and selfish third-party (SFTP) and other-regarding third-party (ORTP) (Fig. 1B). One key insight from agent-based simulations is that the emergence of certain outcomes is sensitive to the fractions of types in the population and which types interacted, suggesting that a social mechanism must take heterogeneity of types into account in order to promote social justice. For example, consider SFD and ORD interacting with an other-regarding TP. SFD would presumably exhibit more norm-compliance in the 'punish' mechanism than the 'help' one. However, if TP is more likely to reduce inequity in the 'help' mechanism, determining which mechanism is more effective depends on whether D's norm-compliance or TP's inequity reduction causes a greater reduction in inequity.

We used an inequity-aversion model (19, 21) in which individuals make tradeoffs between their own monetary payoffs and inequity. In this model, selfish individuals only maximize their monetary payoff while other-regarding individuals also consider inequity. In addition to concerns of inequity between D and R, ORTPs also consider the efficiency of the social mechanism.

Using latent class modeling, we estimated the relative proportions of two types of individuals, assumed to be selfish or other-regarding, as well as the degree of the latter. Our results show that approximately 32% of individuals belonged to the other-regarding type (likelihood ratio test $p < 0.001$).

Conditions under Which 'Punish' Outperforms 'Help'

Using the calibrated model, we performed two sets of simulations. We first simulated four types of

interactions, illustrated in Fig. 1B, which serve as extreme cases where individual types are either selfish or other-regarding. Second, we changed the proportion of other-regarding individuals to examine the boundary conditions under which one social mechanism might be superior to another.

Fig. 3A presents the simulated results for norm-compliance. Comparing the top-left and bottom-left cells, ORD gives more to R and so is more norm-compliant than SFD. When examining the impact of social mechanism ('punish' or 'help') on D (top-left and top-right cells), we notice that SFD is non-norm-compliant independent of mechanism, giving the same low amount (4%) to R when interacting with both SFTP and ORTP. However, there is a striking difference in ORD's norm-compliance depending on the interaction between social mechanism and the type of TP (bottom-left and bottom-right cells). Under the 'punish' condition, ORD behaves significantly more norm-compliant in the presence of ORTP, giving 34%, which is twice as much as when SFTP is present (17%) under the same condition. Under the 'help' condition, when ORD is interacting with ORTP, we observe a 3.46% reduction in giving compared to when interacting with SFTP (17% to 13%), a result of ORD's freeriding on ORTP being other-regarding, an interesting phenomenon we call the *crowding out* effect.

TP's inequity reduction is shown in Fig. 3B. We see that when SFD and ORTP interacted under the 'help' condition (top-right cell), it led to greater inequity reduction (36%) than under the 'punish' condition (30%). In this scenario, the threat of punishment did not increase norm-compliance on the part of D, but it did increase the likelihood of TP reducing inequity. That is, when the impact of the efficiency concern on inequity reduction surpasses that of norm-compliance caused by punishment, the 'help' condition will lead to greater inequity reduction than the 'punish' condition. Interestingly, interaction between ORD and ORTP under the 'punish' condition (bottom-right cell) leads to greater inequity reduction by TP (79%) than under the 'help' condition (55%), which is primarily caused by substantially greater norm-compliance under the 'punish' condition.

We further simulated how changes in the proportion of other-regarding individuals in a society might influence the relative merits of the 'punish' and 'help' mechanisms. We first observed the proportion of ORDs and ORTPs that are equally effective in inequity reduction in both mechanisms (see Fig. 4). We plotted the difference in total inequity reduction between the 'punish' and 'help' mechanisms in Fig. 3D. As shown, the 'punish' mechanism is more effective than the 'help' one if the proportion of ORDs and ORTPs is sufficiently high. When the proportion of ORDs falls below a certain threshold, the 'help' mechanism is more successful at reducing social inequity.

Reinforcing and Crowding-out Effects of Heterogeneity and Mechanism

In Fig. 3C**,** we see how the interaction between different types of individuals affected outcomes, and how they differed by the type of mechanism used. In particular, we examined inequity reduction (i) by ORD if ORTP is not present (white bars, top), (ii) by ORTP if ORD is not present (light grey bars, bottom), and (iii) when ORD and ORTP are both present (dark grey bars).

Interestingly, we found that ORD and ORTP reduced inequity by similar amounts under the 'punish' and 'help' conditions. We also observed a mutually reinforcing effect in reducing inequity when ORD is paired with ORTP under the 'punish' condition. Together, they reduced inequity by 78.52%, which is more than the sum reduced by ORTP (30.4%) and ORD (33.39%) in isolation. In contrast, due to the crowding-out effect under the 'help' condition, ORTP and ORD together reduced inequity by less (54.94%) than the sum of the reduction by ORTP (36.03%) and ORD (33.39%) in isolation.

Discussion

A considerable number of observations from field studies have documented self-organized mechanisms being used to solve collective action problems, reduce poverty and promote justice (1, 20). By focusing on voluntary inequity reduction, we show that on average, the 'punish' mechanism leads to a more just outcome than the 'help' one, but when heterogeneous types are considered and the norm-violator is purely selfish, the 'help' mechanism is more effective in promoting justice. Our findings are of particular importance when considering how to design social mechanisms to reduce inequity in conditions where the effectiveness of the mechanism is sensitive to the heterogeneity of types in the population.

In this paper, we have demonstrated that when D is selfish, the 'help' mechanism reduces a greater amount of inequity than the 'punish' mechanism, but this should be interpreted with a few caveats. Apparently, when the cost of being punished is sufficiently large, SFD may exhibit more norm-compliant behavior. In our experiment, we set the punishment for SFD as 40 MU, which is 40% of the total endowment. We conducted more simulations to investigate how inequity reduction in the two mechanisms would be affected by inequity reduction (what D will lose and R will gain) at different costs to TP; see the Supporting Information for details. We varied experimental parameters such as consequence and cost while fixing the parameters for inequity-aversion, efficiency concern, and the proportion of OR-type individuals, estimated from data. We found that when the consequences of inequity reduction were sufficiently large, the 'punish' mechanism produced a greater reduction in inequity than the 'help' mechanism.

Aside from the 'punish' and 'help' mechanisms, we also occasionally observe other mechanisms used to reduce inequity. Besides reducing D's payoff, TP can transfer wealth from D to R, like Robin Hood, or reward a norm-compliant D. We have examined the effectiveness of different mechanisms when interacting with different types in the population, but how different mechanisms endogenously emerge, evolve, and shape the behavior of the individuals, remains to be answered. We aim to systematically examine these questions in future work.

Our combination of experimental data, a latent class model, and simulation offers a useful tool for understanding the interaction between social mechanisms and heterogeneous types of individuals. In our framework, all individuals are utility-maximizers, satisfying their own preferences in different social mechanisms. Our approach successfully uncovers preference in types of individuals and, more importantly, provides better estimates for preference parameters. Since this approach can investigate more complex heterogeneities in preference and behavior (e.g., expected vs. non-expected utility, Bayesian vs. reinforcement learning, myopic vs. forward-looking, etc.), we believe that it is worthwhile to investigate the design of social mechanisms in other issues where heterogeneity in risk attitude and dynamic behavior are key concerns.

Abbreviations

D	Dictator
R	Recipient
TP	Third-party
OR	Other-regarding
ORD	Other-regarding dictator
ORTP	Other-regarding third-party
SF	Selfish
SFD	Selfish dictator
SFTP	Selfish third-party

Glossary

Compliance	D's giving in the presence of ORTP, compared to SFTP
Crowding-out	Under the 'help' condition, ORD gives less when interacting with ORTP than with SFTP
Efficiency	The sum of D's and R's final payoff
Ex ante inequity	Payoff difference between D and R before TP's action
Ex post inequity	Payoff difference between D and R after TP's action
Mechanism	The experimental condition, either 'punish' or 'help'
Percentage of inequity reduction	We assume that initial inequity is 100. Any percentage in deduction is defined as (100−Deduction)/100

References

1. Ostrom E (1990) *Governing the commons: The evolution of institutions for collective action* (Cambridge University Press, Cambridge).
2. Boyd R, Richerson PJ (2002) Group Beneficial Norms Can Spread Rapidly in a Structured Population. *Journal of Theoretical Biology* 215(3):287-296.
3. Coakley S, Nowak M, Almenberg J (2013) *Evolution, Games, and God: the Principle of Cooperation* (Harvard University Press, Cambridge).
4. Camerer CF, Fehr E (2006) When does" economic man" dominate social behavior? *Science* 311(5757):47-52.
5. Fischbacher U, Gächter S, Fehr E (2001) Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71(3):397-404.
6. Kurzban R, Houser D (2005) Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences of the United States of America* 102(5):1803-1807.
7. Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O (2007) Egalitarian motives in humans. *Nature* 446(7137):794-796.
8. Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25(2):63-87.
9. Gürerk Ö, Irlenbusch B, Rockenbach B (2006) The competitive advantage of sanctioning institutions. *Science* 312(5770):108-111.
10. Gächter S, Renner E, Sefton M (2008) The long-run benefits of punishment. *Science* 322(5907):1510-1510.
11. Henrich J (2006) Cooperation, punishment, and the evolution of human institutions. *Science* 311(5769):60-61.
12. Bruhin A, Fehr-Duda H, Epper T (2010) Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica* 78(4):1375-1412.
13. Harrison GW, Lau MI, Williams MB (2002) Estimating individual discount rates in Denmark: A field experiment. *American Economic Review* 92(5):1606-1617.
14. Heckman J, Singer B (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society* 52(2):271-320.
15. Bowles S, Gintis H (2004) The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical population biology* 65(1):17-28.
16. Nowak MA, Sigmund K (1992) Tit for tat in heterogeneous populations. *Nature* 355(6357):250-253.
17. Janssen MA, Ostrom E (2006) Empirically based, agent-based models. *Ecology and Society* 11(2):37.
18. Engelmann D, Strobel M (2004) Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review* 94(4):857-869.
19. Charness G, Rabin M (2002) Understanding social preferences with simple tests. *Quarterly journal of Economics* 117(3):817-869.
20. Varughese G, Ostrom E (2001) The contested role of heterogeneity in collective action: some evidence from community forestry in Nepal. *World development* 29(5):747-765.
21. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Quarterly journal of Economics* 114(3):817-868.

Figure Legends

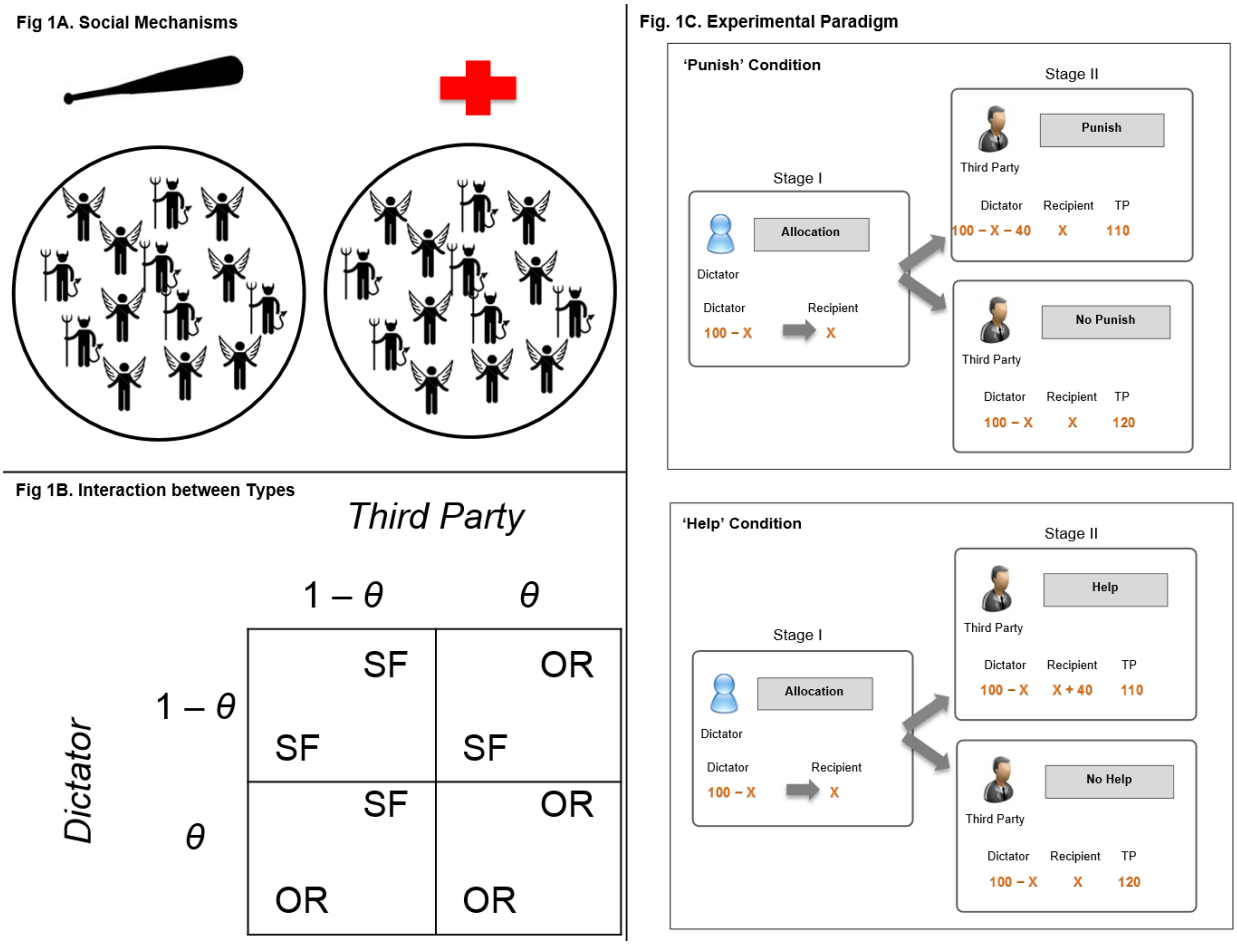

Figure 1 – The Social Mechanisms, Interaction between Types, and Experimental Paradigm

1A. Social Mechanisms – Two mechanisms ('punish' or 'help') are studied. Within each one, we consider social outcomes to be determined by interactions between heterogeneous types of individuals.

1B. Interaction between Types – We consider two types (SF and OR) in two roles (D and TP). The proportion of ORDs in the population is denoted by $\theta \in [0, 1]$, where the remaining $1 - \theta$ are SFD. Similarly, the proportion of ORTP is $\theta \in [0, 1]$. This results in four types of interactions: (SF, SF); (SF, OR); (OR, SF); and (OR, OR).

1C. Experimental Paradigm – D is endowed with 100 MU. In Stage 1, D must decide how much to give to R. This is observed by TP, who is endowed with 120 MU. In Stage 2, under the 'punish' condition, TP can punish D by 40 MU at the cost to TP of 10 MU; under the 'help' condition, TP can give 40 MU to R at the cost to TP of 10 MU. This is repeated for 20 rounds.

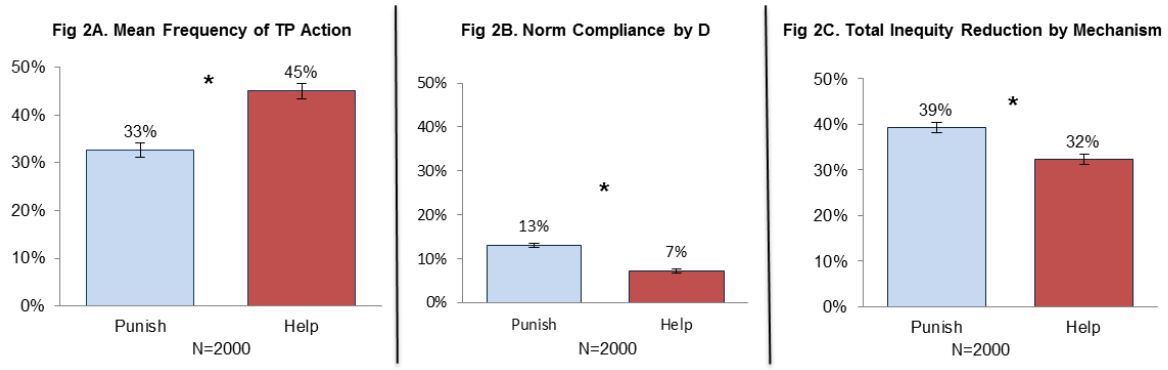

Figure 2 – Frequency of Third-Party Action, Norm Compliant Allocation, and Inequity Reduction

2A. Mean Frequency of TP Action (±SEM) – TP is more likely to reduce inequity under the 'help' condition, controlling for what D gives ($p < 0.001$).

2B. Average Norm-Compliance by D (±SEM) – D gives significantly more under the 'punish' condition than the 'help' one ($p < 0.001$).

2C. Total Inequity Reduction, by Mechanism (±SEM) – The 'punish' condition reduces more *ex post* inequity ($p < 0.001$).

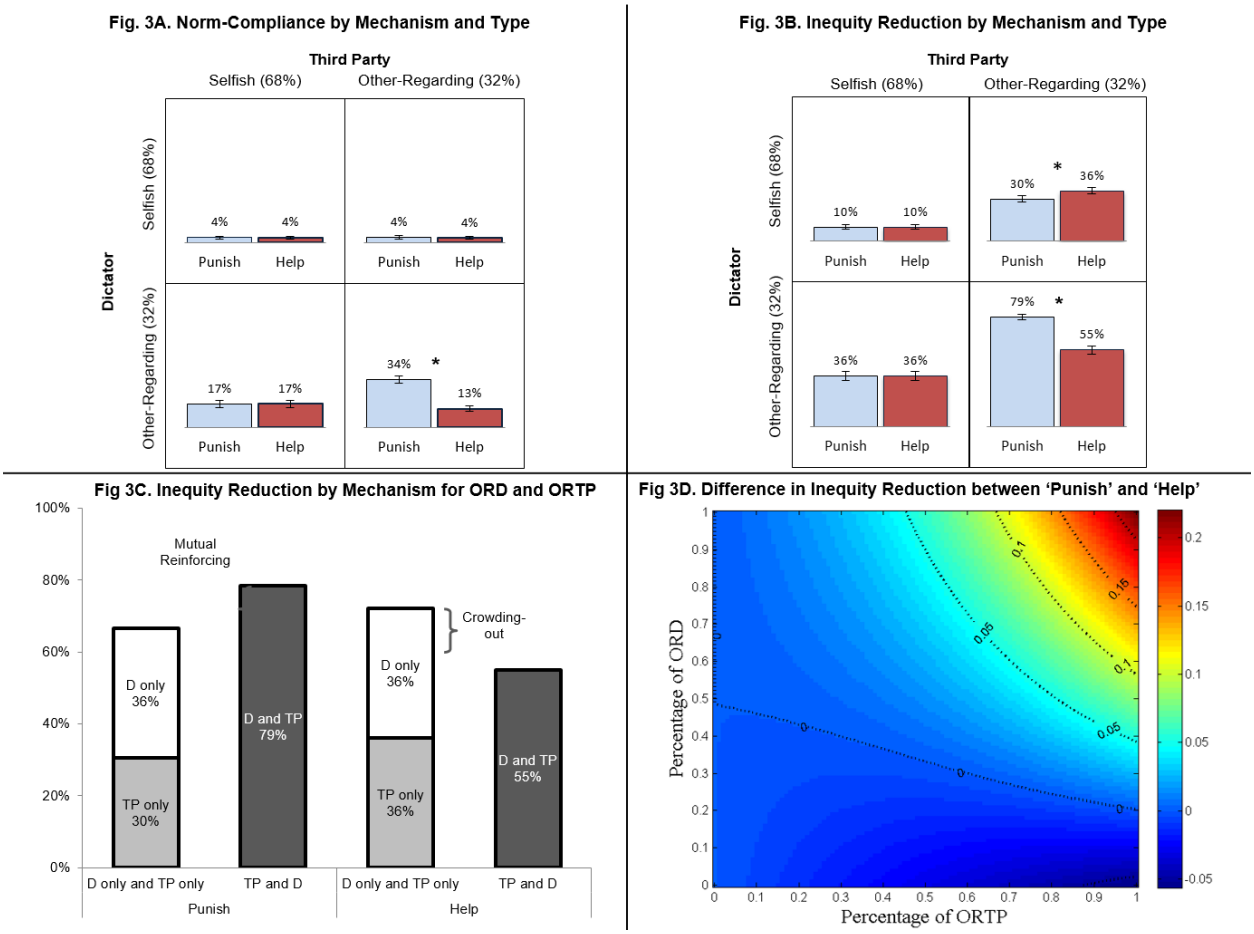

Figure 3 – Norm-Compliance and Inequity Reduction by Mechanism and Type

3A. Norm-Compliance by Mechanism and Type – When ORTP is present under the 'punish' condition, ORD is more compliant than SFD, giving 34% versus 4%. When SFTP is present in the same condition, SFD only gives 4%, but ORD gives 17.56%. When ORTP is present under the 'help' condition, the crowding out effect is observed and ORD reduces giving by 3.46% from 17% to 13%, compared to when SFTP is present. Bootstrapped SEMs (replication sample size = 1,000) are presented together with the mean.

3B. Inequity Reduction by Mechanism and Type – With SFD, ORTP reduces inequity more under the 'help' condition (36%) than the 'punish' one (30%). With ORD, ORTP reduces inequity more under the 'punish' condition (79%) than the 'help' one (55%). ORTP and ORD reduce a comparable amount of inequity when matched with a selfish counterpart under both conditions. Bootstrapped SEMs (replication sample size = 1,000) are presented together with the mean.

3C. Inequity Reduction by Mechanism for ORD and ORTP – There is a mutually reinforcing effect in reducing inequity when ORD is paired with ORTP under the 'punish' condition. Due to the crowding-out effect, under the 'help' condition, ORD and ORTP together reduce inequity by less than the sum reduced by ORTP and ORD in isolation.

3D. Difference in Inequity Reduction between 'Punish' and 'Help' – A positive value denotes that the 'punish' condition reduces inequity. If the proportion of other-regarding Ds and TPs is sufficiently high (the upper-right), the 'punish' mechanism is more effective than the 'help' one.

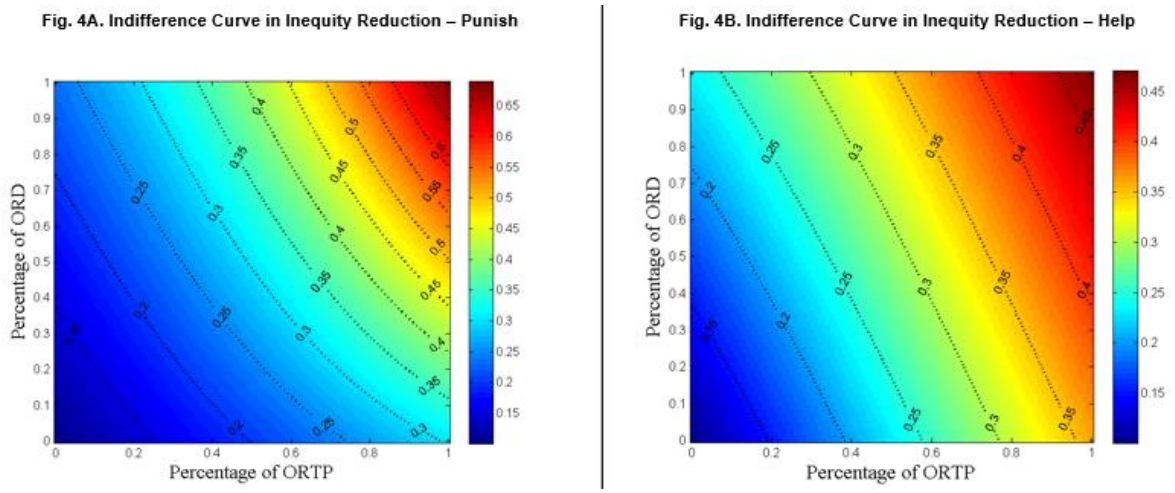

Figure 4 – Indifference Curve in Inequity Reduction under 'Punish' and 'Help'

Supporting Information

Experimental Method

Procedure

A total of 300 subjects participated in our study, which was undertaken at the laboratory of the Center for Behavioral Economics at the National University of Singapore (NUS). The participants were students from all fields of study at NUS. We posted recruitment advertisements on the university's online course management system. Participants received a show-up fee of $4[*] plus payments that were tied to decisions they made during the experiment. On average, subjects made approximately $15.70 in cash.

Upon arrival at the lab, each participant was randomly assigned a number that determined where he or she sat and his or her role (D, R, or TP). Subjects stayed in the same role throughout the experiment. The experiment was conducted online. Once everyone had successfully logged in to a specifically designed online system, the instructions appeared onscreen. The instructions (reproduced below in "Instructions") were read aloud to ensure that all participants knew of the procedures and payoffs. The participants could only communicate electronically and they were separated by dividers to ensure anonymity. Each experimental session lasted for twenty rounds. The same triplet of subjects made decisions for ten rounds, after which they were randomly matched with other subjects playing the other two roles for another ten rounds. We conducted sixteen experimental sessions with 300 participants in total ($Mean_{Age}$ = 21.6, SD = 2.3; 48% female). Each experimental session lasted about an hour and a half. Subjects were paid in cash before they left the laboratory.

Design

The experimental setup was based on Fehr and Fischbacher's third-party punishment game (1). We studied third-party responses to violations of a distributional norm under two experimental conditions in a between-subject design: 'punish' and 'help'. Participants were randomly assigned to one of three roles – D (red), R (blue), or TP (green) – forming a triplet with one of each role. Each condition comprised 150 participants in 50 triplets.

In the first stage, D decided how to split a stake of 100 MU with R. R must accept any amount allocated by D. TP was endowed with 120 MU and could either keep the endowment or take an action to reduce inequity at a cost of 10 MU. Under the 'punish' condition, if TP decided to reduce inequity, 40 MU was deducted from D's payoff. Under the 'help' condition, TP could reduce inequity by giving 40 MU to R.

We designed the experiment according to the fixed matching protocol. Each session consisted of 20 rounds in total. Participants underwent the decision making tasks repeatedly with the same triplet of players for ten rounds, after which they were randomly re-matched with players of the other two roles for another ten rounds.

Instructions

In the following textbox are the instructions that were given to the participants. Note that amounts in this section are in Singaporean dollars.

[*] Amounts are in USD, converted from Singaporean dollars at the average exchange rate of 1.0 USD = 1.25 SGD.

This is an experiment on decision making. The instructions are simple and if you follow them carefully and make good decisions, you could earn a considerable amount of money which will be paid to you in cash before you leave today. Different subjects may earn different amounts of cash. What you earn today depends partly on your decisions, partly on the decisions of others, and partly on chance.

There are [3X] subjects in this room. Subjects will be randomly assigned to one of the three colours: RED, BLUE, or GREEN. Each subject has an equal chance of playing the role of RED, BLUE and GREEN player. Each subject has a different role and will stay in the same role throughout the entire experiment. The experiment will consist of [20] decision making rounds. In the beginning of the 1st round, subjects will be randomly grouped into [X] triplets. Each triplet consists of one RED player, one BLUE player, and one GREEN player. You will undergo the decision making tasks repeatedly with the same triplet of players for [10] rounds. In the beginning of the 11th round, you will be rematched with completely different people of the other two colours based on the random assignment. You will play the same decision making tasks with the newly formed triplet for the remaining [10] rounds. The decision making task of each player will be explained below. The experiment is anonymous. Specifically, you do not know (and will not know) who the players are in your triplet. Similarly, the other players of your triplet do not know (and will not know) who you are.

It is important that you do not look at the decisions of others, and that you do not talk, laugh, or make noises during the experiment. You will be warned if you violate this rule. If you violate this rule twice, you will be asked to leave, and you will not be paid. That is, your earnings will be $0.

Experimental procedure

In each round, the decision making task occurs in 2 stages, namely, I and II. Each colored player undertakes its respective task as follows:

Stage I

In Stage I, RED will have a pot of 100 cents to divide between herself and BLUE player. RED can divide the pot of 100 cents in any way she pleases (giving BLUE player any amount ranging from 0 to 100 cents). BLUE player gets the division from RED player no matter what it is. For example, if RED decides to give BLUE 10 cents, then RED will earn 100-10 = 90 cents and BLUE will earn 10 cents. Similarly, if RED gives BLUE 90 cents, then RED will earn 100-90=10 cents and BLUE will earn 90 cents. Note that these numbers are chosen arbitrarily for illustrative purposes only. GREEN player will merely observe the RED player's division of the pot of 100 cents in Stage I.

Stage II

[PUNISH Condition]

GREEN player will be given a pot of 120 cents. GREEN player must decide whether or not to pay 10 cents to subtract 40 cents from RED player's earning in Stage I.

1. If GREEN player decides to pay 10 cents to subtract 40 cents from RED player's earning in Stage I, then GREEN player will earn 110 cents and RED player's earning in Stage I will be reduced by 40 cents.
2. If GREEN player decides not to pay 10 cents to subtract cash from RED player's earning in Stage I, GREEN will earn 120 cents and RED player's earning in Stage I will remain the same.
3. BLUE player's earning in Stage I stays the same.

[HELP Condition]

GREEN player will be given a pot of 120 cents. GREEN player must decide whether or not to pay 10 cents to add 40 cents to BLUE player's earning.

1. If GREEN player decides to pay 10 cents to add 40 cents to BLUE player's earning, then GREEN player will earn 110 cents and BLUE player's earning in Stage I earning will be increased by 40 cents.
2. If GREEN player decides not to pay 10 cents to add cash to BLUE player's earning, GREEN will earn 120 cents and BLUE player's earning in Stage I will remain the same.
3. RED player's earning in Stage I stays the same.

In each round, Players will be informed of their respective decision outcomes and cash earnings after Stage II. The above decision task is repeated for 20 rounds, during which [X] triplets will be formed twice. Each player will be matched with 2 other players of different colors in the beginning of the 1st and 11th round. Within the first 10 rounds and remaining 10 rounds, each player knows the decisions of 2 other players in previous rounds. Importantly, the information about decisions in round 1 to round 10 will not be revealed to the other matched players in round 11 to 20.

Payoffs

Your dollar earnings for the experiments are determined as follows. First, we will sum up your total dollar earnings from all 20 rounds. In addition, we will add a $5 show-up fee to this amount. You will be paid the total amount when you leave the experiment.

Model-Free Results

Fig. S1A and S1B show that there is a persistent difference in D's giving and TP's inequity reduction between the two conditions and throughout the rounds. On average, D gave 13.07 MU (SE = 0.54) under the 'punish' condition and 7.12 MU (SE = 0.45) under the 'help' condition. The difference is significant at the 0.05 level using a *t*-test ($t(1998) = 8.45$, $p < 0.0001$). TP reduced inequity at a rate of 33% (SE = 0.15) under the 'punish' condition and 45% (SE = 0.16) under the 'help' condition.

To determine whether the difference in TP's response in the two conditions could be attributed to the difference in D's giving, we ran a binary logistic regression analysis in STATA 12, which is shown in Table S1; the standard errors are in parentheses. From the regression analysis, we found that the likelihood of reducing inequity decreases with the amount given, meaning that the more D gave, the less likely that TP would reduce inequity. Controlling for giving, TP is more likely to reduce inequity under the 'help' condition. All the coefficients are significant at the 0.001 level.

Model, Estimation, and Simulation

Model Setup

Third Party Utility

TP's utility is modeled in the spirit of Charness and Rabin (2) and Fehr and Schmidt (3). In addition to inequity aversion, TP cares about efficiency. We add this efficiency concern to capture the fact that conditional on D giving x, TP is more likely to reduce inequity under the 'help' condition. We use two

parameters $(\alpha_{TP}, \kappa_{TP})$ to capture TP's degree of inequity aversion and efficiency concern. Hence, given D's giving x, TP's utility under the 'punish' condition is:

$$U(d|x, \alpha_{TP}, \kappa_{TP}) = \begin{cases} 120 - \alpha_{TP} \cdot max\{100 - 2x, 0\} + \kappa_{TP} \cdot 100, & if\ d = 0 \\ 110 - \alpha_{TP} \cdot max\{100 - 2x - 40,0\} + \kappa_{TP} \cdot 60, & if\ d = 1 \end{cases} \quad (1)$$

The utility under the 'help' condition is:

$$U(d|x, \alpha_{TP}, \kappa_{TP}) = \begin{cases} 120 - \alpha_{TP} \cdot max\{100 - 2x, 0\} + \kappa_{TP} \cdot 100, & if\ d = 0 \\ 110 - \alpha_{TP} \cdot max\{100 - 2x - 40,0\} + \kappa_{TP} \cdot 140, & if\ d = 1 \end{cases} \quad (2)$$

Dictator Utility

We denote $P(d = 1|x, \alpha_{TP}, \kappa_{TP})$ as the probability that TP will reduce inequity conditional on giving x and TP's other-regarding preference $(\alpha_{TP}, \kappa_{TP})$. Hence $P(d = 0|x, \alpha_{TP}, \kappa_{TP}) = 1 - P(d = 1|x, \alpha_{TP}, \kappa_{TP})$ is the probability that TP does not reduce inequity. $V(x, d)$ is D's *ex post* utility by giving x after TP's decision d. We assume that D only has inequity concerns α_D[†], and the degree is the same as TP's. Under the 'punish' condition, the *ex post* utility is defined as:

$$V(x|d, \alpha_D) = \begin{cases} 100 - x - \alpha_D \cdot max\{100 - 2x, 0\}, & if\ d = 0 \\ 100 - x - 40 - \alpha_D \cdot max\{100 - 2x - 40, 0\}, & if\ d = 1 \end{cases} \quad (3)$$

Under the help condition, the ex post utility is defined as:

$$V(x|d, \alpha_D) = \begin{cases} 100 - x - \alpha_D \cdot max\{100 - 2x, 0\}, & if\ d = 0 \\ 100 - x - \alpha_D \cdot max\{100 - 2x - 40, 0\}, & if\ d = 1 \end{cases} \quad (4)$$

Hence, D's expected utility for giving x is:

$$EU(x|\alpha_D) = P(d = 1|x, \alpha_D) \cdot V(x|d = 1, \alpha_D) + P(d = 0|x, \alpha_D) \cdot V(x|d = 0, \alpha_D) \quad (5)$$

We assume that D has rational expectation on the probability $P(d = 1|x, \alpha_D)$.

Discrete Choice Modeling

We add an independent and identically distributed extreme value error term, ε, to utilities in equation (1), (2), and (5), representing the components in the utility that are unobserved by the researcher. Then we have a logit specification of the choice probability that is consistent with utility maximization (4, 5). The choice probability of giving x for individual i is:

$$P_i(x) = \exp(\lambda \cdot EU_i(x|\alpha_D)) \,/\, \sum_{x'=0}^{100} \exp(\lambda \cdot EU_i(x'|\alpha_D)) \quad (6)$$

[†] D's role is to decide how much money to allocate to the recipient, so we believe D's consideration is less likely to be affected by social efficiency. More importantly, under the 'punish' condition, punishment reduces D's payoff, as well as inequity and efficiency. It is not distinguishable which one is the concern for D. As a result, we assume $\kappa_D = 0$ for D.

In this equation, the parameter $0 \leq \lambda \leq 1$ reflects the sensitivity of the choices to utility differences. When $\lambda = 0$, the individual is completely insensitive to the differences in utility and the model would predict equal probability of the individual choosing either alternative. When $\lambda \to \infty$, the probability of choosing the alternative with a higher utility approaches one (5). Similarly, the choice probability of decreasing inequity (or justice) j for individual i is:

$$P_i(j|x) = \exp\left(\lambda \cdot U(j|\alpha_{TP}, \kappa_{TP})\right) / \sum_{j'=0}^{1} \exp\left(\lambda \cdot U(j'|\alpha_{TP}, \kappa_{TP})\right) \qquad (7)$$

Mixture Modeling

We adopt the finite-mixture model developed by Heckman and Singer (6) to investigate heterogeneity in the population. We assume that there are two types of individuals ($s = s_1, s_2$) who have heterogeneous preferences over the fairness concern (α). Within each segment, subjects have homogeneous preferences. Specifically, one segment of Ds are fair-minded ($\alpha^{s_1} \geq 0$) and the remainder are purely selfish ($\alpha^{s_2} = 0$). Let $0 \leq \theta \leq 1$ be the relative size of the first segment, and $1 - \theta$ the size of the remaining segment. Conditional on D i being a member of segment s, we can write the probability of giving x as:

$$P_i^s(x) = \exp(\lambda \cdot EU_i(x|\alpha_D^s)) / \sum_{x'=0}^{100} \exp(\lambda \cdot EU_i(x'|\alpha_D^s)) \qquad (8)$$

We can write TP's choice probability in segment s as:

$$P_i^s(j|x) = \exp\left(\lambda \cdot U(j|\alpha_{TP}^s, \kappa_{TP}^s)\right) / \sum_{j'=0}^{1} \exp\left(\lambda \cdot U(j'|\alpha_{TP}^s, \kappa_{TP}^s)\right) \qquad (9)$$

The probability of giving x, unconditional on segment membership is:

$$P(x) = \theta \cdot P_i^{s_1}(x) + (1 - \theta) \cdot P_i^{s_2}(x) \quad (10)$$

TP's choice probability, unconditional on segment membership is:

$$P(j|x) = \theta \cdot P_i^{s_1}(j|x) + (1 - \theta) \cdot P_i^{s_2}(j|x) \qquad (11)$$

Defining D i's choice history as $H_i = x(t), t = 1, 2, \ldots T$, the likelihood of this subject's choice history can be computed as:

$$L(H_i) = \theta \cdot L(H_i|s = s_1) + (1 - \theta) \cdot L(H_i|s = s_2) \qquad (12)$$

$$L(H_i|s) = \prod_{t=1}^{T} \prod_{x=0}^{100} (P_{it}^s(x))^{I[x(t)=x]} \qquad (13)$$

$I[x(t) = x]$ is the indicator function, which is 1 if $x(t) = x$; otherwise, it is 0. Then we have the likelihood for the observed set of Ds:

$$L_D = \prod_{i=1}^{N} L(H_i) \qquad (14)$$

Using the same procedure, we can obtain the likelihood for TP's choice L_{TP}. In this estimation, we maximize the likelihood $L = LL_D \cdot LL_{TP}$ over the parameters $(\alpha_D, \alpha_{TP}, \kappa_{TP}, \lambda, \theta)$. The estimation was performed using the OPTIMUM package in GAUSS 13. The standard errors of the coefficients were calculated using the Delta method (8). To avoid the local maxima problem, we randomly chose 200 sets of initial values, drawn to ensure that the maximum is global.

Based on the estimated parameters and observed choice history for individual i, we can calculate the probability of the individual belonging to the other-regarding segment s_1. The probability is obtained by updating the base-rate θ in a Bayesian fashion:

$$P(i \in s_1|H_i) = \frac{L(H_i|s_1)\cdot\theta}{L(H_i|s_1)\cdot\theta+(H_i|s_2)\cdot(1-\theta)} \qquad (15)$$

We use these posterior probabilities to predict each individual i's choice in round t.

Estimation Results

We presented the estimation results from three model specifications: Model I, in which we assume that both D and TP are purely selfish ($\alpha_D = 0, \alpha_{TP} = 0, \kappa_{TP} = 0$); Model II, in which D and TP are homogeneous, with an other-regarding preference ($\alpha_D \geq 0, \alpha_{TP} \geq 0, \kappa_{TP} \geq 0$); and Model III, in which there are two types of players, selfish and other-regarding, with approximately 32% of players being other-regarding. From the estimation results shown in Table S2, the mixture model dramatically improves fitness relative to Model I, in which we assume purely selfish players, and Model II, in which we assume homogeneous, other-regarding players.

In Fig. S2, we also plotted the predicted average rate of giving by D and the predicted probability of TP's response. The mixture model outperforms other models in terms of predicting participants' behavior.

Simulation

Based on the estimated coefficients $\{\widehat{\alpha_D}, \widehat{\alpha_{TP}}, \widehat{\kappa_{TP}}, \hat{\lambda}\}$, we can predict TP's response probability $P(j|x, \alpha_{TP}, \kappa_{TP})$, where $j = 0, 1$; $x = 0, \ldots, 100$. More specifically, SFTP's response probability is defined as $P(j|x, \alpha_{TP} = 0, \kappa_{TP} = 0)$, and ORTP's response probability is $P(j|x, \alpha_{TP} = \widehat{\alpha_{TP}}, \kappa_{TP} = \widehat{\kappa_{TP}})$. We assume D's belief about TP's response is consistent with TP's predicted response probability. Then we can predict the probability of each giving level $P(x|\alpha_D, \alpha_{TP}, \kappa_{TP})$, which depends on D's type ($\alpha_D = 0$ or $\alpha_D = \widehat{\alpha_D}$), and TP's type.

The *ex post* inequity is:

$$\sum_{x=0}^{100} P(x|\alpha_D, \alpha_{TP}, \kappa_{TP}) \cdot [P(j = 0|x, \alpha_{TP}, \kappa_{TP})(100 - 2x) + P(j = 1|x, \alpha_{TP}, \kappa_{TP})(100 - 2x - 40)]$$

(16)

Bootstrapped Standard Errors

Given the large cross-sectional dimension (100 Ds and 100 TPs) and a smaller temporal dimension (T = 10), we resampled the data in a cross-sectional dimension by sampling subjects' identities. Cameron and Trivedi (7) discussed using this resampling scheme when the number of individuals is large but the time series is small and fixed. We carried out 1,000 bootstrapped samples and estimation exercises based on the replications. After we obtained the coefficients, we simulated the outcomes for different social interactions among heterogeneous players.

Simulation Results

In this section, we report the simulation results on interaction among different types of players. Fig. S3A presents giving by norm-compliant Ds across different mixtures of types under both conditions. With SFDs, under the 'punish' condition, we found that having ORTP only drives SFD to give 0.2% more than having SFTP. With ORD, a sizable drop in giving (3.46%) is observed in the presence of SFTP (17%) compared to ORTP (13%). In Fig. 3A, we present the marginal effect of ORTP on norm-compliance by taking the difference in giving when D is interacting with ORTP and with SFTP. By comparing SFD vs. ORTP and ORD vs. ORTP under the 'punish' condition, we noticed that the impact of ORTP is mainly on ORD.

Fig. S3B presents total inequity reduced across different mixtures of types under both conditions. With SFD, ORTP reduces inequity more under the 'help' condition (36%) than the 'punish' one (30%). When ORD and ORTP are together, ORTP reduces more inequity under the 'punish' condition (79%) than the 'help' one (55%).

Supporting References

1. Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evolution and Human Behavior* 25(2):63-87.
2. Charness G, Rabin M (2002) Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics* 117(3):817-869.
3. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114(3):817-868.
4. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, ed Zarembka P (Academic Press, New York), pp 105–142.
5. McFadden D (1981). Econometric Models of Probabilistic Choice. *Structural Analysis of Discrete Data with Econometric Applications*, eds Manski C, McFadden D (MIT Press, Cambridge), pp 198-272.
6. Heckman J, Singer B (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica* 52(2):271-320.
7. Cameron AC, Trivedi PK (2005). *Microeconometrics: Methods and Applications*, (Cambridge University Press, Cambridge).
8. Wooldridge JM (2010). *Econometric Analysis of Cross Section and Panel Data*, (MIT Press, Cambridge).

Supporting Figure Legends

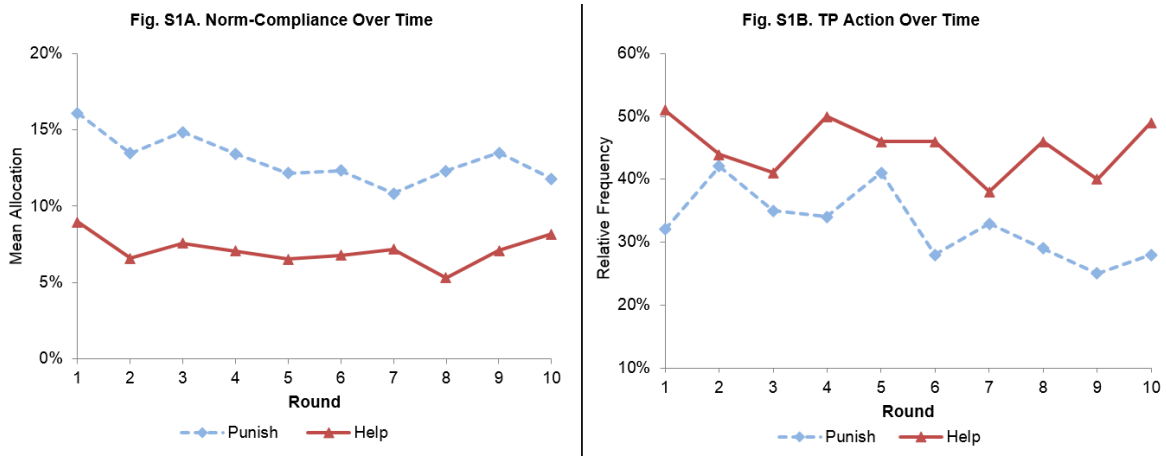

Figure S1 – Norm-Compliance and Third-Party Action over Time

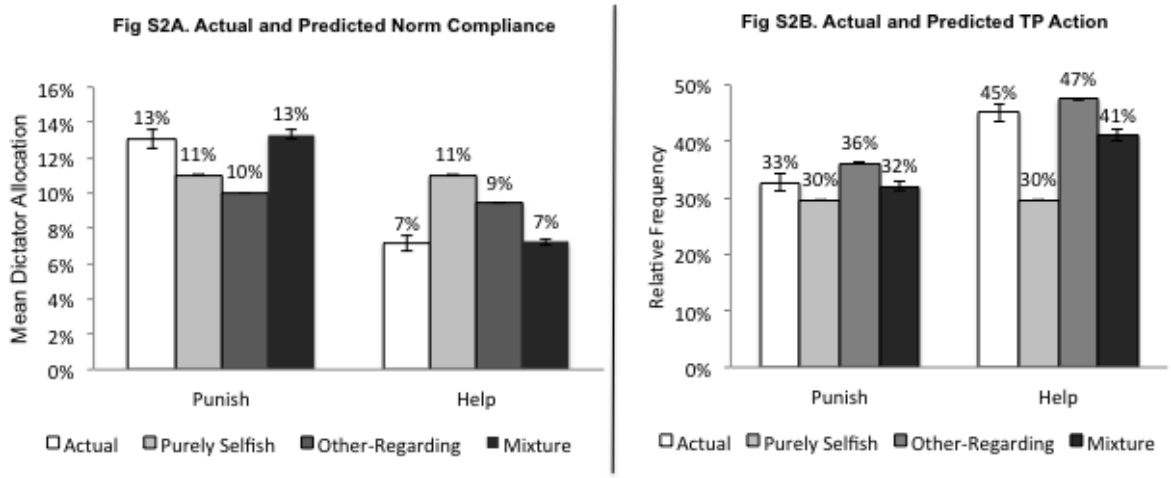

Figure S2 – Actual and Predicted Norm-Compliance by D and Justice Action by TP

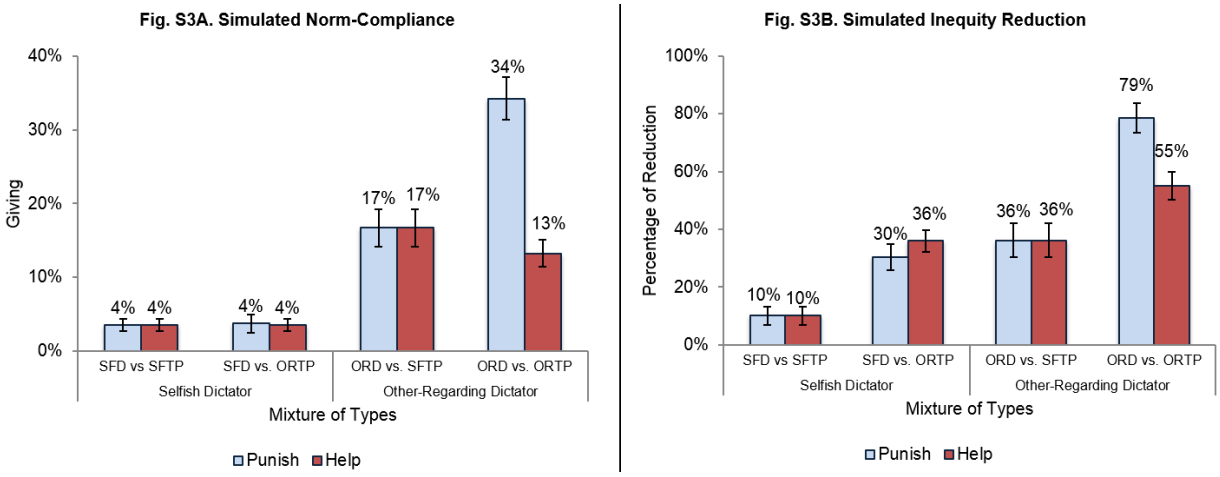

Figure S3 – Simulated Norm-Compliance and Inequity Reduction

Supporting Tables

Table S1 – Logistic Regression Analysis of TP's Response

Independent Variable	Value	SEM
D's Giving	−0.015	(0.003)
HELP Dummy	0.45	(0.094)
Constant	−0.9858	(0.159)
No. of Observations	2,000	
Log Likelihood	−1309.33	
χ^2	54.52	

Table S2 – Estimation Results

	Model I		Model II		Model III	
	Value	SEM	Value	SEM	Value	SEM
Dictator's Inequity Aversion	-	-	0	(0)	0.4129	(0.0052)
TP's Inequity Aversion	-	-	0.177	(0.0119)	0.3138	(0.0076)
TP's Efficiency Concern	-	-	0.0559	(0.0114)	0.033	(0.0068)
Payoff Sensitivity	8.6675	(0.1812)	10.0691	(0.2418)	25.0793	(0.9085)
Probability of Being Other-Regarding	-	-	-	-	0.3247	(0.012)
Log Likelihood	−8103.4906		−8005.5417		−7156.074	
χ^2	-		195.90		1894.83	
No. of Observations	4000		4000		4000	

Abbreviations

D	Dictator
R	Recipient
TP	Third-party
OR	Other-regarding
ORD	Other-regarding dictator
ORTP	Other-regarding third-party
SF	Selfish
SFD	Selfish dictator
SFTP	Selfish third-party

Glossary

Compliance	D's giving in the presence of ORTP, compared to SFTP
Crowding-out	Under the 'help' condition, ORD gives less when interacting with ORTP than with SFTP
Efficiency	The sum of D's and R's final payoff
Ex ante inequity	Payoff difference between D and R before TP's action
Ex post inequity	Payoff difference between D and R after TP's action
Mechanism	The experimental condition, either 'punish' or 'help'
Percentage of inequity reduction	We assume that initial inequity is 100. Any percentage in deduction is defined as (100−Deduction)/100