The Perverse Costly Signaling Effect on Cooperation under the Shadow of the Future

Kamei, Kenju

Durham University

26 September 2020

The Perverse Costly Signaling Effect on Cooperation under the Shadow of the Future

Kenju Kamei

Department of Economics and Finance, Durham University
Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk

First version: May 2019

This version: September 2020

Abstract: A literature in the social sciences proposes that humans can promote cooperation with strangers by signaling their generosity through investment in unrelated pro-social activities. This paper studied this hypothesis by conducting a laboratory experiment with an infinitely repeated prisoner's dilemma game under random matching. A novel feature of the experiment is that each player first decided how much to donate to a charitable organization, the British Red Cross, and then this donation information was conveyed to the player's matched partner. Surprisingly, the donation activities significantly undermined cooperation. This negative effect of charitable-giving was consistently observed regardless of whether players had a post-interaction opportunity to punish the partners. A detailed analysis suggests that the negative effect (a) resulted from the transmission of the charitable-giving information, not from the fact that subjects engaged in the charitable-giving, and (b) was caused by mis-coordination between the two parties who can both costly signal their generosity. This suggests that letting players have an implicit costly signaling opportunity has damaging unintended consequences for their interactions among strangers. Possible ways to encourage players to use costly signaling for mutual cooperation, such as partner choice, are also discussed in the paper.

JEL classification: C92, C73, D91

Keywords: experiment; cooperation; prisoner's dilemma; charitable-giving; costly signaling

Acknowledgment: The author thanks John Hey for his hospitality in letting him conduct the experiments at the University of York, and Artem Nesterov for his assistant on performing structural estimation in the paper. The author also thanks Pedro Dal Bó, Christian Thöni, Matthew Embrey, Hong Il Yoo, Hajime Kobayashi, Matthew Walker, and the audience at GREDEG (Sophia Antipolis), University of Tokyo, Keio University, and Osaka University for helpful comments. This project was supported by a grant-in-aid from the Yoshida Hideo Memorial Foundation in Japan. The Kyoto University Institute of Economic Research (KIER) foundation provided additional funding.

1. Introduction

An influential literature in the social sciences and biology proposes that humans can achieve cooperation among non-kins (strangers) using costly signals of own quality, such as cooperativeness, through investing in unrelated pro-social activities (e.g., Zahavi, 1975 and 1977; Grafen, 1990; Boone, 1998; Smith and Bliege Bird, 2000; Bliege Bird and Smith, 2001; Gintis *et al.*, 2001; Lotem *et al.*, 2003; Leimar and Hammerstein, 2011; Fehrler and Przepiorka, 2013). This theme dates back to Darwin (1874), who discovered animals' development of characters, such as coloration (e.g., plumage) and songs of birds, which, while conferring no or even negative survival value, are nevertheless used to signal their quality to peers (Zahavi, 1975). Prior research has consistently proposed that this costly signaling proposition may extend to humans.

A key underlying mechanism for the costly signaling hypothesis is that revealing honest information benefits not only observers, but also signalers through the responses induced in the observers (see, e.g., Smith and Bliege Bird [2005] for a survey). Since signalers' interests are aligned with observers' interests, they do not commit cheating (e.g., Zahavi, 1975; Grafen, 1990). This condition could be met under repeated interactions, because theoretically a cooperative equilibrium can emerge even with random matching, if humans are sufficiently patient (e.g., Kandori, 1992; Ellison, 1994). Ethnographic and historical examples of humans' seemingly wasteful or costly displays and behavior for signaling purposes abound: food sharing (e.g., turtle hunting by islanders to serve at a public feast in funerary rituals among the Meriam of Torres Strait, Australia), participation in ritual activities, or recreational community activities (e.g., religious ritual; dance and festivals in groups), holding redistributive feasts, contribution to conflict by attending group raiding and defense, and artistic elaboration (see, e.g., Smith and Bliege Bird, 2000; Hawkes and Bliege Bird, 2002; Sosis and Alcorta, 2003; Sosis and Bressler, 2003; Hagen and Bryant, 2003; Bliege Bird and Smith, 2005). But how does information on pro-social acts per se affect peers' decisions to cooperate in the community?

Costly activities often discussed in the literature are some social interactions, or contributions to provision of public goods in a community, which directly benefits the community members. It is also stressed that people can form alliances with high types based on available information. Community compositions are therefore endogenously determined. Thus, according to the author's view, there are at least three channels considered by past studies to play important roles in the positive impact of costly signaling. The first channel is the impact of information itself. People can send information on their own inclinations to cooperate with fellow community members. These signals may facilitate cooperation in the community at large. Second, some endogenous group formation process may not only encourage people to invest in costly pro-social activities outside the on-going dilemma interactions for a signaling purpose, but

may also make the signaling more credible. For example, those whose behavior deviates from their signals may be excluded from the group. Third, if costly activities take a form of social interactions (e.g., feasts), the activities can help community members share group identity or build social cohesion, and accordingly foster cooperation norms. Alternatively, if a costly pro-social activity helps their community as in the example of food sharing, the community members may want to reciprocate the signaler's pro-social action. This reciprocation helps people sustain more cooperative relationships within the community. It remains unclear which aspect plays the most important part in improving cooperation and what the value of the pro-social information in itself is. This paper studies how the information regarding people's pro-social behaviors in one dimension affects their dilemma interactions in another dimension. The attempt here is to identify the signaling value of costly pro-social activities in a controlled manner, without the confounding aspects such as direct reciprocation and endogenous partner selection.

The method of a laboratory decision-making experiment is used to collect clean data. Recruited subjects play an infinitely repeated prisoner's dilemma game under random matching (supergame, hereafter) multiple times. The setup of infinite repetition captures a wide range of people's interactions in real lives (e.g., Mailath and Samuelson 2006). The multiple-supergame design allows us to also study people's possible learning behaviors. The experiment uses a perfect stranger matching protocol across the supergames: they play each supergame with a different set of players. A random matching protocol is used for matching within supergames. That is, in each round in any given supergame, subjects are randomly matched with a member in their community and interact with each other once. Experimental parameters are set so that mutual cooperation, in addition to mutual defection, holds as an equilibrium outcome. A novel feature of the experiment is to let subjects decide whether to donate to a real charitable organization at the beginning of each round. The donation decisions are then informed to the interaction partners before their prisoner's dilemma interactions begin. The signaling effect can be studied by examining how the charitable-giving information affects subjects' action choices. As people often have peer-to-peer punishment opportunities in reality (e.g., Fehr and Gächter, 2000 and 2002), in half of the treatments subjects are given a post-interaction punishment stage at the end of each round. This enables us to check the robustness of the impact of charitable-giving institution to the presence of peer-to-peer punishment.

The experiment data surprisingly showed that charitable-giving information significantly undermines cooperation under random matching. While two kinds of charitable-giving formats were used in the experiment, this perverse effect was robust to the format. In addition, the perverse effect was commonly observed, independent of whether subjects had a post-interaction opportunity to punish their peers or not. Nevertheless, consistent with the costly signaling argument, the size of

a subject's donation decision was still a reliable signal for her likelihood to cooperate in the experiment.

Using an additional experiment, this paper shows that charitable-giving decision per se does not cause this negative impact; rather, the transmission of the information is the likely cause. Several additional experiments were further conducted in an attempt to better explain subjects' behaviors, as the main findings obtained were surprising ones. These experiments revealed that the perverse effects of charitable-giving are not alleviated even if positive framing is added by emphasizing the signaling value of charitable-giving activities or the charitable-giving process is simplified. This paper thus concludes that letting players have an indirect costly signaling opportunity through pro-social activities can backfire for the interactions among strangers. This also suggests that a stronger coordination device might be required to induce players to use charitable-giving to achieve mutual cooperation and that the costly signaling proposition discussed in prior research may be driven by such coordination devices. As an illustration, this paper experimentally demonstrates that the charitable-giving activities do have a strong positive effect on cooperation if the information is provided along with partner choice (an option to select an interaction partner based on charitable-giving activities).

2. Experimental Design

The experiment consists of two sets of treatments – implemented with a 3×2 factorial design (Table 1). Each treatment is built on the framework of an indefinitely repeated prisoner's dilemma game with random matching. The two sets (the second dimension of the 3×2 design) are identical except that a post-interaction punishment stage is included in each round in one set. There are three conditions in the first dimension of the design. Two treatments are designed in each set so that subjects have a stage in which they make donation decisions before prisoner's dilemma interactions in a given round. Once they make donation decisions, subjects are informed of their matched counterparts' donation decisions *before* deciding on their action choice. There is a control setup in which subjects have no donation opportunities. The three treatments without punishment are named the "No Donation, No Punishment" (N-N) treatment, the "Donate in advance, No Punishment" (Donate-N) treatment, and the "Commit, No Punishment" (Commit-N) treatment. The three treatments with punishment are named the "No Donation, Punishment" (N-P) treatment, the "Donate in advance, Punishment" (Donate-P) treatment, the "Commit, Punishment" (Commit-P) treatment.

This section will first explain the common features of the six treatments (Section 2.1). It will then explain each design piece (Sections 2.2 and 2.3). The experimental procedure will be described in Section 2.4.

2.1. The common structure of the experiment

The experiment design is built on the framework of Camera and Casari (2009).[1] Subjects play five supergames (infinitely repeated prisoner's dilemma games) in sequence. At the onset of a given supergame, subjects will be randomly assigned to a group of four. Subjects do not interact with anyone outside their own groups. In each supergame, subjects are given an initial endowment of 20 points and can accumulate earnings through interactions with peers.[2] Each round has the same structure. At the onset of a given round, each subject is randomly matched with another subject in their group. They then simultaneously select C (cooperate) or D (defect). The payoff matrix is shown in Figure 1.

The infinite repetition is modeled by use of a random termination rule. An integer, ranging from 1 to 100, is randomly drawn at the end of each round (each integer is drawn with a probability of 1%). The supergame will be over if the integer is greater than 90 and continues if it

Table 1: *Summary of Treatments (including Additional Treatments)*

Punish:	Charitable-giving opportunities: No [control treatments]	Yes: Commit to donate	Yes: Donate before PD interaction
	Six main treatments		**Additional treatments**
No	**N-N** (No donation, No Punishment)	**Commit-N** (Commit, No Punishment)	**Donate-N** (Donate in advance, No Punishment) / **Donate-N (Not Informed)**
Yes	**N-P** (No donation, Punishment)	**Commit-P** (Commit, Punishment)	**Donate-P** (Donate in advance, Punishment) / **Donate-P (Not Informed)**
			Stag hunt (Section 4.2.5)
			Donation decisions are kept private (Section 5)
			Donate-P (Framing): Players are given information about correlations between subjects' decisions to donate and cooperate in the Donate-P treatment (Section 6).
			Donate-P (1 point): Subjects' donation decisions are binary (Section 6).
			Partner Choice / **Random Match Comparison**: Players select partners based on their past donation behaviors (Section 6).

Note: In addition to the additional treatments listed in the table, two other treatments, called the Donate-N (One-way) and Donate-P (One-way) treatments, were also conducted – see Appendix B for the detail.

[1] See Kamei (2017) also.

[2] Initially endowed with 20 points, subjects in the Donate-N and Donate-P treatments can donate in advance even in round 1. In order to make the other four treatments parallel to the Donate-N and Donate-P treatments, subjects in all the treatments including the two control treatments are assigned the initial endowment in each supergame.

Figure 1: *Payoff Matrix*

		Player 2	
		C (cooperate)	D (defect)
Player 1	C	25, 25	5, 30
	D	30, 5	10, 10

is less than or equal to 90. Once a given supergame ends, subjects move on to the next supergame and are randomly assigned to a new group with three different subjects with perfect stranger matching. The structure of each supergame is identical.[3] It should be worth noting that the threshold value of the continuation probability above which mutual cooperation holds as an equilibrium outcome ($\bar{\delta}$) is 44.3% in this framework, much less than 90.0% (see Camera and Casari [2009]).

Subjects will be paid privately based on their accumulated payoff at the end of the experiment. The conversion rate is 80 points in the experiment to one pound sterling.

2.2. Signaling through charitable-giving

Subjects in the Donate-N and Donate-P treatments decide how many points they wish to donate to the British Red Cross (an organization that supports people if a crisis strikes) at the onset of each round. The donation amounts in round t must not exceed their accumulated payoffs until that round in a given supergame. The maximum points subjects can donate in round 1 is 20. In case that a subject donates an amount larger than her payoff accrued in a given round, she receives a negative net payoff in that round. Subjects in the Commit-N and Commit-P treatments decide what percentage of a round's payoff they wish to donate to the Red Cross in each round. For example, suppose that a subject i committed to donate 10% of her payoff in round 4 and then (C, C) was realized. The subject's donation amount would be 2.5 (= 25×0.1) points in this case.

Note that the donation decisions are not revealed to anyone except the current-round interaction partners. It should also be worth noting that these donation decisions are not hypothetical. Subject's total donation amounts (rounded to the nearest pence) will be taken to the Red Cross after the experiment.

The reason that the commitment format was considered in addition to the "donate in advance" format is to see how the signaling value differs by whether subjects donate *outside* the

[3] While the continuation probability was 95% in Camera and Casari (2009) and also Kamei (2017), this study sets a little lower probability (90%) in order to avoid the duration of the experiment going too long considering that subjects in the present experiment have a charitable-giving stage in each round. Most of the experiment sessions lasted from around 45 minutes to 90 minutes (including payment to subjects).

current interaction or *using* the payoff from the current interaction. This setup will also give us a useful opportunity to check the robustness of findings to the charitable-giving format.

2.3. Post-interaction opportunity to punish

Subjects move on to the next round once they complete the prisoner's dilemma interaction in the N-N, Donate-N and Commit-N treatments. By contrast, subjects in the N-P, Donate-P and Commit-P treatments enter a peer-to-peer punishment stage. Once the subjects engage in the prisoner's dilemma game interaction and are informed of their interaction outcomes, they decide how many punishment points they wish to assign to their counterparts. Punishment activities are costly. For each punishment point a subject assigns, one point is deducted from her payoff and three points are deducted from the target's payoff. The punishment points must be an integer between 0 and 5.[4] Subjects are informed of total punishment points they received (not who punished them) at the end of the period.

2.4. Experimental procedure

The experiment, except the instructions and control questions, was computerized using the z-Tree software (Fischbacher, 2007). All the experiment sessions were conducted in the EXEC laboratory in the University of York from August 2017 to February 2019. As in Camera and Casari (2009) (and also Kamei (2017)), two sessions per treatment were conducted. Each session consisted of either 20 subjects (5 groups) or 16 subjects (4 groups). Recruiting messages were sent through *hroot* (Bock *et al.*, 2014) to all eligible subjects in the database; and subjects voluntarily registered for and participated in the experiment. The number of participants (sessions) in the six main treatments was 232 (12). A series of additional experiments were conducted exploratory to better explain subjects' behaviors because the key findings from the main experiments were surprising ones. The number of participants (sessions) in the additional experiments was 348 (18). All subjects joined only one session. The experiment procedure in the additional experiments was identical to that in the main experiment. The total number of subjects (sessions) was thus large: 580 (30).[5] The details of the additional experiments are reported in Sections 4 to 6.

[4] Camera and Casari (2009) used a simplified punishment technology in that each subject made binary choice regarding whether to reduce their partner's payoff by ten points by spending five points. As reported in Section 4, a comparison between the two control treatments (N-N and N-P) suggests that the finding on the impact of peer-to-peer punishment in Camera and Casari (2009) extends to the setup of this paper.

[5] The average supergame lengths were 10.4, 11.5, 11.9, 8.8, 10.2, and 7.7 rounds in the N-N, Donate-N, Commit-N, N-P, Donate-P, and Commit-P treatments, respectively. The average supergame lengths in the additional experiments were 7.7, 13.6, 6.1, 8.4, 6.7, 10.3, 10.7, 11.2 and 8.0 rounds in the Stag Hunt, Donate-N (Not Informed), Donate-P (Not Informed), Donate-N (One-Way), Donate-P (One-Way), Donate-P (Framing), Donate-P (1 point), Partner Choice, and Random Match Comparison treatments, respectively.

3. Related Literature on Signaling and Infinitely Repeated Games

The standard theory does not provide a point prediction on subjects' behaviors in the two control treatments. The experimental design is built on the framework of Camera and Casari (2009) and the threshold continuation probability above which mutual cooperation holds as an equilibrium outcome is 44.3%. Since a continuation probability is set as 90% in this study, mutual cooperation can be sustained only through community enforcement in the N-N treatment (e.g., Kandori, 1992). Specifically, the efficient outcome can be achieved, for example, if subjects strictly follow a grim trigger strategy where a subject continues to select cooperation until they confront at least one defection in the population (the subject starts selecting defection once she experiences one instance of defection by her partner). This also suggests that theoretically peer-to-peer punishment is not necessary to achieve the efficient outcome (see Camera and Casari [2009] for the discussions). Note, however, that various strategies can lead to a cooperative equilibrium in this study because the number of available strategies is not finite in an infinitely repeated prisoner's dilemma game.

For the same reason ($\delta = 0.900$), mutual cooperation holds as an equilibrium outcome also in the treatments with charitable-giving. The information on charitable-giving is theoretically not necessary. But how might the information on partners' charitable-giving behaviors affect one's decision to cooperate? There are two branches of experimental literature that can inform on this question, but they do not provide a definite prediction about the impact of information.

The first branch is the literature on infinitely repeated prisoners' dilemma games under random matching. Past research has shown that if there are no institutions to assist people's cooperation behaviors, the level of cooperation under random matching will not be high, even though mutual cooperation holds as an equilibrium outcome.[6] Yet past experiments have also indicated that cooperation can be sustained at high levels with forced disclosure of past action choices or with reputation mechanisms, because the information helps people select strategies that lead to the mutual cooperation equilibrium. From a theoretical viewpoint, Kandori (1992) studies the role of label (color) attached to each player, whose color is determined by their last-round action choice (C or D) and is only informed to the matched partners. He showed that mutual cooperation is easier to achieve with this mechanism because one's action in a given

[6] For example, when the community size was four, the average cooperation rate was 59.5% in Camera and Casari (2009) and it was 33.4% in Kamei (2017). Subjects' cooperation behaviors were extremely low in Duffy and Ochs (2009): the average cooperation rates were 14.9% when the community size was 6 and 7.5% when the community size was 14. See Dal Bó and Fréchette (2018) for a survey. Such low cooperation norms were also seen in an indefinitely repeated public goods game under random matching (Kamei, 2019). Heller and Mohlin (2018) theoretically discuss that in a population where some people do not maximize own payoff, say due to bounded rationality, the mechanism to support cooperation, such as the contagious equilibrium, may fail.

round affects not only her current-round partner but also her next-round partner through the label (see also Ellison [1994]). Stahl (2013) experimentally studied the exogenous color-coded reputation mechanism that Kandori (1992) proposed. The subjects' average cooperation rates more than doubled with the reputation mechanism.[7] To the knowledge of the author, however, all previous studies consider the transmission of information regarding their past action choices in the on-going interactions. This paper is the first to explore how costly signaling information (information on pro-social behaviors *outside* the prisoner's dilemma game) affects player's decisions to cooperate in an infinitely repeated prisoner's dilemma game under random matching. Although in the setup of this study, the labels subjects have are from their charitable-giving behaviors, the label – either a non-donor or a donor (and the size of donation amounts in the case for the donor) – is conveyed to their current-round interaction partners. Past experimental work suggests that a non-negligible fraction of people have stable preferences and thus on average, people's pro-social activities in one dimension are positively correlated with their pro-social activities in another dimension (e.g., Andreoni and Miller, 2002; Blanco *et al.*, 2007; Fisman *et al.*, 2007; Volk *et al.*, 2012). Hence, subjects in the charitable-giving treatments may achieve higher cooperation norms by effectively using the label as a signal of their cooperativeness and also utilizing discriminatory strategies.

The second branch which this paper can receive insights from is the broad experimental literature on the role of cheap talk and communication in people's strategic interactions. Prior research has found that letting people send a message directly to their interaction partners, regarding how they intend to play, encourages cooperation in prisoner's dilemma interactions (e.g., Duffy and Feltovich, 2002 and 2006). The same positive impact of cheap talk or messaging has also been seen in the context of coordination game (e.g., Cooper *et al.*, 1992; Charness, 2000; Blume and Ortmann, 2007; Blume *et al.*, 2017). In the present study, while mutual cooperation holds as an equilibrium outcome, subjects may not perceive the supergame as a coordination game since it may require high cognitive ability. Whether or not subjects perceive the experiment as a prisoner's dilemma game or a coordination game, past experiments suggest that directly sending a message as to their intended action choices could help improve cooperation. Such positive effects could be driven by the guilt a person feels when she breaks her word (e.g., Charness and Dufwenberg, 2006) and/or her preference for promise keeping (e.g., Vanberg, 2008). In the case of a coordination game, the positive effects could also be driven by material interests in achieving high earnings. Nevertheless, it is not clear how large the impact of a signal *indirectly* sent through a pro-social activity would be. Feelings of guilt when deviating from her indirect signal may be much attenuated relative to a direct signal.

[7] See also Camera and Casari (2009) and Kamei (2017).

4. Perverse Effects of Charitable-Giving

This section will overview the subjects' cooperation and charitable-giving behaviors. It reveals that the presence of the charitable-giving institutions undermines cooperation.

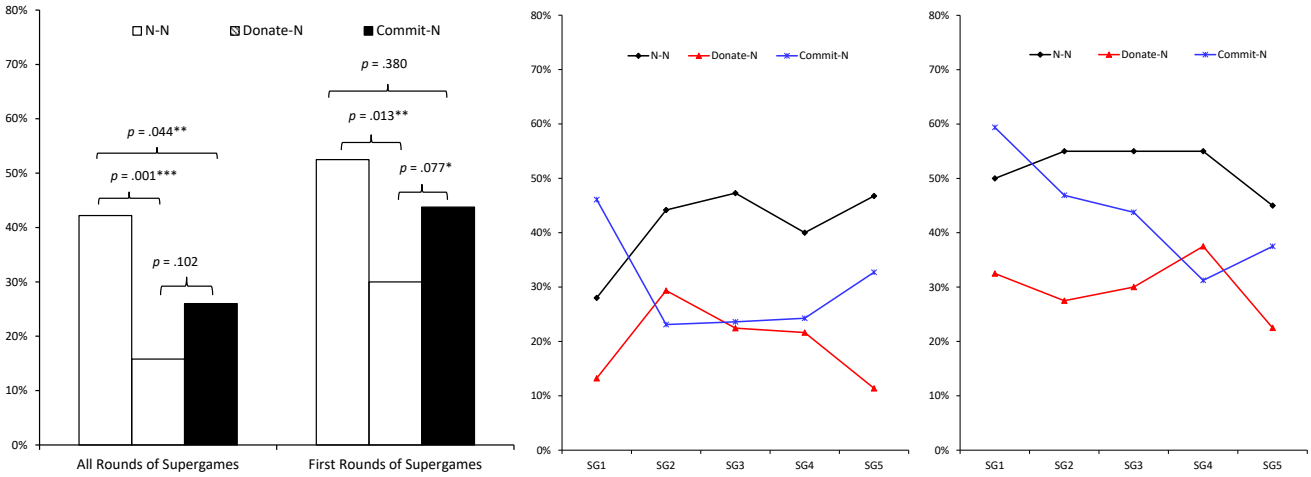
4.1. Perverse effects without peer-to-peer punishment

4.1.1. Average cooperation rate

As usual in indefinitely repeated prisoner's dilemma games experiment with random matching, subjects' cooperation behaviors were modest in the control treatment. Figure 2.a reports subjects' cooperation rates averaged across all supergames.[8] Subjects on average selected cooperation 42.2% of the time in the N-N treatment. The same rate was around ten percentage points higher – 52.5% if only round 1 action choices in the supergames are considered. The weaker cooperation behaviors after round 1 is reasonable, because some subjects often act on conditional cooperation or punishment strategies in indefinitely repeated interactions (e.g., Camera and Casari, 2009; Kamei, 2017). Although the random matching protocol is used, some subjects, if exploited by defectors, may engage in (blind) revenge by selecting defection in the following rounds for some duration.

Surprisingly, the charitable-giving institutions strongly undermined cooperation. This

Figure 2: *Average Cooperation Rates without Punishment*



(a) Avg. cooperation rate across all supergames

(i) All Rounds of Supergames

(ii) First Rounds of Supergames

(b) Avg. cooperation rate, supergame by supergame

Notes: As subjects' decisions to cooperate are binary, *p*-values (two-sided) in panel a were calculated based on subject random effects probit regressions with robust jackknife standard errors. All supergame data were used. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

[8] Average cooperation rates were calculated not only by using all rounds' cooperation decisions, but also by solely using their round 1 behaviors, because subjects' very first decisions can be interpreted as their inclinations to cooperate independent of experience (Dal Bó and Fréchette, 2018).

perverse effect was observed, whether subjects donated before interactions or they promised to donate through the on-going interaction payoffs. The average cooperation rates were 62.6% and 38.4% lower in the Donate-N and Commit-N treatments, respectively, than in the N-N treatment, when all rounds were considered. These decreases were 42.9% and 16.7% in the Donate-N and Commit-N treatments, respectively, when only round 1 behaviors were considered.

Result 1: *(i) The average cooperation rates were much lower in the Donate-N and Commit-N treatments than in the N-N treatment, whether subjects' action choices from all rounds or only the first rounds were considered.*

4.1.2. Across-supergame trends

Trends across supergames are also different between with versus without the charitable-giving institutions (Figure 2.b). The round 1 average cooperation rates stayed stable at around 50% across the supergames in the N-N treatments (panel b.ii). However, the subjects' likelihoods to select cooperation were less likely to decline over the rounds in the later than in the earlier supergames. For example, whereas in the first supergame the subjects' average cooperation rate across all rounds was 30.2% lower than their round 1 cooperation rate, the former was only 8.7% lower than the latter in the fifth supergame. As shown in panel b.i, the supergame average cooperation rates displayed an increasing trend, consistent with Camera and Casari (2009).[9] This trend is in clear contrast with the two treatments with charitable-giving. First, subjects' willingness to cooperate stayed at a low level across all the five supergames in the Donate-N treatment. Second, subjects' round 1 cooperation rate in the Commit-N treatment exhibited a decreasing trend, although the subjects on average selected cooperation more frequently than in the N-N treatment during the first supergame (panel b.ii). The session average cooperation rates in the Commit-N treatment stayed around 20% to 30% in the second to fifth supergames (panel b.i).

Result 1: (ii) *Subjects learned to cooperate from supergame to supergame in the N-N treatment, but not in the Donate-N or Commit-N treatment.*

4.1.3. Charitable-giving

How frequently did subjects use the charitable-giving opportunities? As shown in Table 2, subjects on average used these opportunities around 30% to 66% of the time. Nevertheless, subjects' willingness to donate fell after round 1. In addition, the percentage of those who made (committed) a donation and the average donation amounts (commitment percentages) both declined from

[9] The increase rate across supergames is, however, not significant due to high variance in individuals' behaviors according to a subject random effects probit regression with robust jackknife standard errors. Here, random effects, not fixed effects, were controlled for because a Hausman test did not reject the null that coefficients from random effect estimation are not systematically different from those from fixed effect estimation.

supergame to supergame in the Donate-N (Commit-N) treatment. If subjects were attempting to signal their future cooperation behaviors through charitable-giving, they might gradually give up doing so over time. However, even in the final supergame a large fraction of subjects still made donations. This suggests that subjects' acts of donation and/or the information of charitable-giving perversely undermined the community's cooperation norms across the experiment.[10]

Result 2: *Subjects' charitable-giving activities decreased gradually over time. However, a large fraction of subjects still made (committed) donations even in the final supergame in the Donate-N (Commit-N) treatment.*

Table 2: *Charitable-Giving Decisions in the Donate-N and Commit-N Treatments*

I. Donate-N treatment

		1st SG	2nd SG	3rd SG	4th SG	5th SG	across-supergame trends[#1]
All rounds	Average amount [points]	1.73	1.24	1.07	1.03	0.75	-.227*** (.004)
	Percentage of positive donations	56.9%	46.3%	31.3%	36.6%	44.7%	-.037*** (.001)
Round 1	Average amount [points]	4.20	1.70	1.45	1.60	1.15	-.62*** (.085)
	Percentage of positive donations	75.0%	55.0%	57.5%	57.5%	55.0%	-.038*** (.013)

II. Commit-N treatment

		1st SG	2nd SG	3rd SG	4th SG	5th SG	across-supergame trends[#1]
All rounds	Average commitment [%]	6.39	4.23	3.24	2.48	2.92	-.692*** (.038)
	Percentage of positive commitment	66.4%	56.5%	52.8%	44.2%	45.5%	-.051*** (.009)
Round 1	Average commitment [%]	7.53	4.00	3.53	3.63	3.44	-.856*** (.232)
	Percentage of positive commitment	78.1%	56.3%	59.4%	62.5%	62.5%	-.025 (.038)

Notes: Average points donated before interactions in the Donate-N treatment, and average commitment percentages in the Commit-N treatment.

[#1] Subject random effects linear regressions with standard errors clustered by session were conducted (the dependent variable is subject *i*'s donation decisions in round *t*, and the supergame number variables is the independent variable). Subject random effects were added to control for the panel structure because a Hausman test did not reject the null that the difference in coefficient estimate is not systematic between fixed versus random effects in each data.

*** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

4.2. Does peer-to-peer punishment alter the perverse effects of charitable-giving?

4.2.1. Incentive changes with punishment

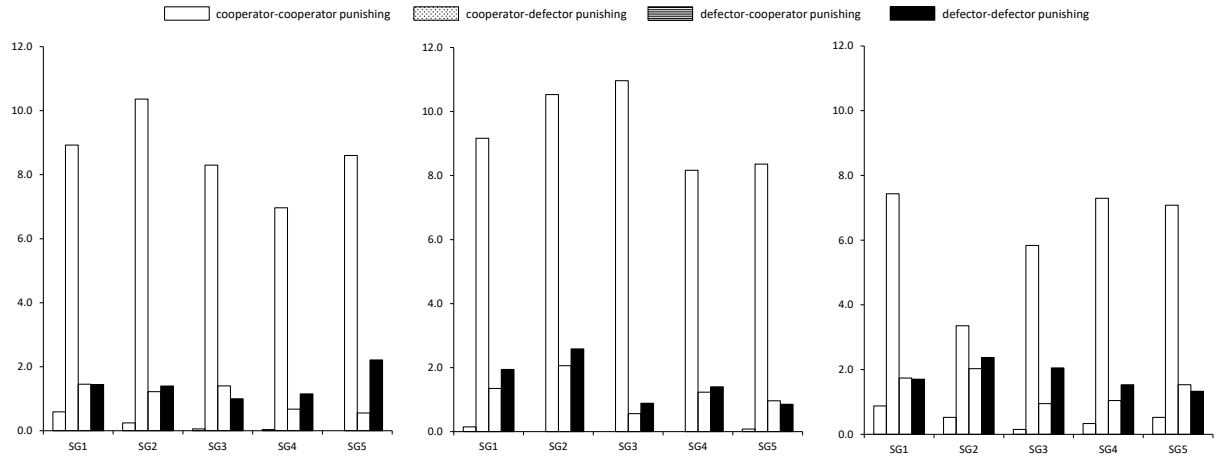
Let us now turn our attention to the results when subjects have an option to punish peers. Peer-to-peer enforcement of cooperation norms is commonly seen in the literature (e.g., Van

[10] Some readers may argue that the persistent donation activities reflect the subjects' true altruism. This is unlikely, however, because unconditional cooperators usually account for only a very small percentage of cooperation types in this kind of experiment. An analysis in Section 5 supports this view.

Lange *et al.* [2011] for a survey). For instance, the punishment treatments are designed built on the "Private monitoring with punishment" treatment in Camera and Casari (2009). Camera and Casari (2009) found that punishment opportunities significantly enhance cooperation.

The effects of punishment can be explained by incentive changes. For example, if punishment is not much mis-directed to cooperators or not heavily used for (blind) revenge of past punishment received, subjects' temptation to defect will be attenuated (see, e.g., Fehr and Gächter, 2000, 2002). A close look at the punishment data in fact revealed (a) that costly punishment of defectors by cooperators were much stronger than the three other types of punishment in all supergames, and as a result (b) that the average per round stage-game payoff matrices after punishment amounts are subtracted are transformed into stag-hunt games in each punishment

Figure 3: *Average Per Round Realized Stage Game Payoffs*



(a) N-P treatment (b) Donate-P treatment (c) Commit-P treatment

(I) Average per round punishment received

	C	D
C	24.79	1.18
D	21.55	7.74

(a) N-P treatment

	C	D
C	24.91	0.60
D	20.18	7.73

(b) Donate-P treatment

	C	D
C	24.66	1.71
D	23.37	7.25

(c) Commit-P treatment

(II) Average payoff matrices (raw player) after subtracting punishment received

	C	D
C	23.33	-0.62
D	19.07	6.87

(a) Donate-P treatment

	C	D
C	23.17	1.58
D	22.31	7.02

(b) Commit-P treatment

(III) Average payoff matrices (raw player) after subtracting both the punishment received and donation amounts

Notes: Numbers in panels II and III are the payoffs of row players averaged across the five supergames.

treatment (Figure 3).[11] Specifically, the outcome (C, C) in each stage game is a Pareto-dominant equilibrium, whereas (D, D) is a risk-dominant equilibrium, even if this game is played one time (panels II and III). Both cheap talk and communication are known to encourage players to select the Pareto-dominant equilibrium even in a strictly one-shot interaction of a stag-hunt game (e.g., Cooper *et al.*, 1992; Charness, 2000; Blume and Ortmann, 2007). Thus, one can expect that cooperation could be easier to evolve in the Donate-P and Commit-P treatments than in the N-P treatment, as subjects can signal their generosity through charitable-giving activities. It should be acknowledged, however, that discussions based on the average payoffs in Figure 3 may be misleading since our subjects may not perceive the stage-game structure as a stag-hunt game. Taking the payoff consequences of punishment into account without being myopic may require high cognitive ability and patience. Nevertheless, a bottom line here is that mutual cooperation could be easier to achieve with the help of punishment, compared with the three treatments studied in Section 4.1.

Result 3: *In the N-P, Donate-P and Commit-P treatments, (i) costly punishment of defectors by cooperators is much stronger than any other punishment scenario, and (b) costly punishment of cooperators by cooperators is significantly less common than any other punishment scenario.*

Sections 4.2.2 and 2.2.3 are devoted to an analysis on the effects of charitable-giving activities on cooperation in the three punishment treatments.

4.2.2. Average cooperation rate

The presence of the charitable-giving opportunities again markedly undermined subjects' willingness to cooperate (Figure 4.a). The average round 1 cooperation rates were around 28.4% and 48.0% lower in the Donate-P and Commit-P treatments, respectively, compared with the N-P treatment. The treatment effects when all data are considered are similar to the negative effects seen in round 1. The average overall cooperation rates were around 28.3% and 48.0% lower in the Donate-P and Commit-P treatments, respectively, compared with the N-P treatment.[12] This suggests that the perverse effects of charitable-giving are robust to having punishment opportunities.

Result 4: *The average cooperation rates were significantly lower in the Donate-P and Commit-*

[11] An analysis using all supergames' data found that in each punishment treatment, (a) the punishment strength is significantly weaker in the cooperator-cooperator punishing than in any other punishment scenario, and (b) the punishment strength is significantly stronger in the cooperator-defector punishing than in the defector-cooperator punishing. See Appendix Table A.1 for the detail.

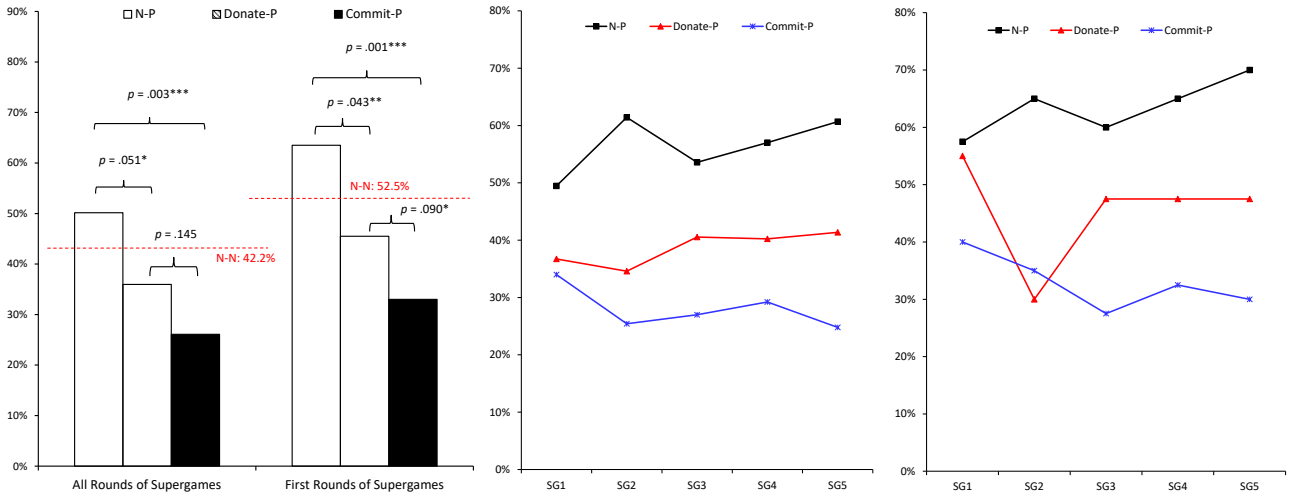
[12] As was the case for the treatments without punishment (Section 4.1), subjects' cooperation behaviors became weaker after round 1 in each punishment treatment, consistent with the idea that some subjects tend to withhold selecting cooperation for some duration if they encounter defection (Camera and Casari, 2009; Kamei, 2017).

P treatments than in the N-P treatment, whether subjects' action choices from all rounds or only the first rounds were considered.

4.2.3. Across-supergame trends

Across-supergame trends are next examined. Two clear findings emerged. First, when the charitable-giving institution is absent, the trend resembles the one in Camera and Casari (2009). As shown in Figure 4.b, subjects' average cooperation rates gradually increase from supergame to supergame.[13] Second, and by clear contrast, although in the first supergame subjects' initial willingness to cooperate was similar to that in the N-P treatment, their average cooperation rates hovered at low levels from the second supergame in the Donate-P treatment. In the Commit-P treatment, subjects persistently exhibited weak willingness to cooperate across all five supergames.

Figure 4: *Average Cooperation Rates with Punishment*



(a) Avg. cooperation rate across all supergames

(i) All Rounds of Supergames

(ii) First Rounds of Supergames

(b) Avg. cooperation rate, supergame by supergame

Notes: *p*-values (two-sided) were calculated based on subject random effects probit regressions with robust jackknife standard errors. All supergame data were used. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

4.2.4. Charitable-giving

Subjects did use the charitable-giving opportunities in the Donate-P and Commit-P treatments (Table 3). However, in the Donate-P treatment, the average amounts donated were decreased across supergames. Similar to the Donate-N treatment, subjects' charitable-giving activities were more common in round 1 than in the other rounds in each supergame. This again implies that some subjects might have given up using the charitable-giving activities for signaling purposes after trying for some time in the Donate-P treatment. Nevertheless, subjects made

[13] The increase rate is not significant because of high variance in individuals' behaviors.

positive donations around 38.7% of the time even in the final supergame of the Donate-P treatment. Likewise, in the Commit-P treatment, while around 55% to 72.5% of subjects committed to make positive donations in the first rounds, this percentage became lower over time in each supergame. It did, however, remained at 48.2% even in the final supergame. These results pose a puzzle as to why lower cooperation norms prevailed in the Donate-P and Commit-P treatments even though the positive charitable-giving information was sent in many cases.

Result 5: *Subjects' charitable-giving activities became less and less common after round 1 in each supergame. However, a large fraction of subjects made (committed) positive donations even in the final supergame in the Donate-P (Commit-P) treatment.*

Table 3: *Charitable-Giving Decisions in the Donate-P and Commit-P Treatments*

I. Donate-P treatment

		1st SG	2nd SG	3rd SG	4th SG	5th SG	across-supergame trends[#1]
All rounds	Average amount [points]	1.54	.97	.90	.86	.71	-.205*** (.077)
	Percentage of positive donations	51.0%	47.5%	43.1%	36.1%	38.7%	-.037* (.022)
Round 1	Average amount [points]	3.20	1.65	1.33	1.20	1.15	-.455*** (.130)
	Percentage of positive donations	75.0%	55.0%	57.5%	60.0%	55.0%	-.035* (.020)

II. Commit-P treatment

		1st SG	2nd SG	3rd SG	4th SG	5th SG	across-supergame trends[#1]
All rounds	Average commitment [%]	4.59	4.35	4.33	4.11	3.69	.314 (.310)[#2]
	Percentage of positive donations	62.5%	48.6%	55.9%	57.1%	48.2%	-.007 (.027)
Round 1	Average commitment [%]	4.10	4.20	3.23	4.30	6.13	.415 (.286)
	Percentage of positive donations	72.5%	55.0%	57.5%	60.0%	57.5%	-.025* (.015)

Notes: Average points donated before interactions in the Donate-P treatment, and average commitment percentages in the Commit-P treatment. [#1] Subject random effects linear regressions with standard errors clustered by session were conducted (the dependent variable is subject *i*'s donation decisions in round *t*, and the supergame number variables is the independent variable). Subject random effects were added to control for the panel structure because in all regressions except (#2), a Hausman test did not reject the null that the difference in coefficient estimate is not systematic between fixed versus random effects in each data. For (#2), the estimate is .336** (.146) if fixed effects, instead of random effects, are included. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

4.2.5. Did subjects respond to incentive changes with punishment?

Analyses so far found that the charitable-giving institutions undermine cooperation, regardless of the option to punish peers. This is contrary to the initial expectation that cooperation could be easily achieved with punishment (Figure 3). The discrepancy between this expectation and the subjects' actual behaviors may mean that subjects did not anticipate the consequences of punishment and as a result that their way of using the charitable-giving opportunities did not depend on the presence of the punishment institution. Another possibility is that subjects recognized that (C, C) is an equilibrium outcome of the stage game, but thought that it would be

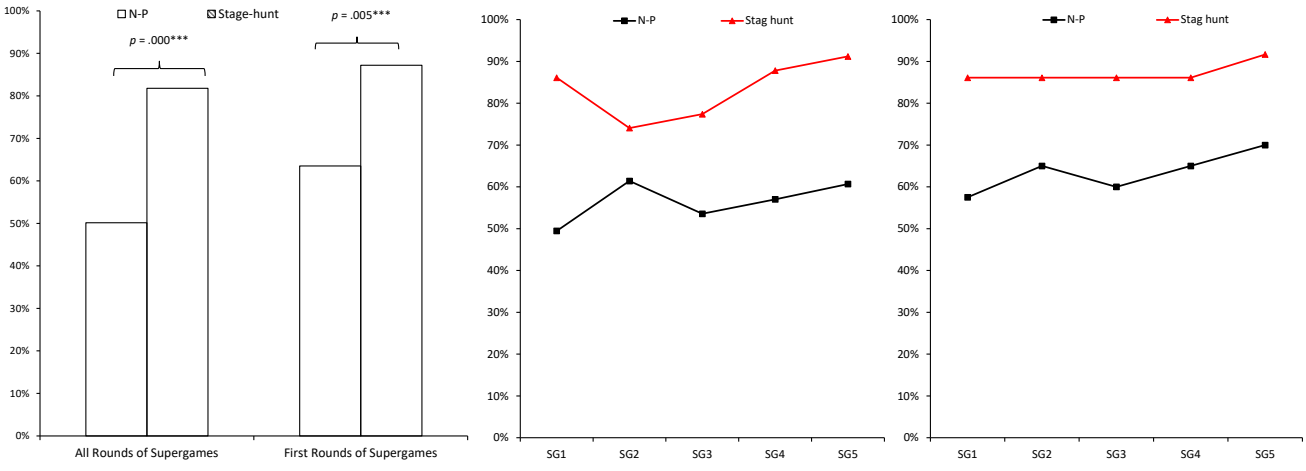
too risky to select cooperation since (D, D) is risk-dominant. If one of the two possibilities is correct, it would explain why the charitable-giving information did not encourage cooperation. In order to check how subjects perceived the incentive structure with punishment, an additional treatment (two sessions), called "Stag-Hunt," were conducted using the stage game payoff matrix as shown in Figure 5. The payoffs used in this matrix can be obtained by rounding the payoff matrix of panel II.a in Figure 3. Except for the change in the payoff matrix and the absence of the punishment opportunities, all the other design pieces are identical to the N-P treatment.

Figure 5: *Payoff Matrix in the Stag Hunt treatment*

		Player 2	
		C	D
Player 1	C	25, 25	1, 22
	D	22, 1	8, 8

Figure 6 reports the average cooperation rates in the Stag Hunt treatment. The data of the N-P treatment are also shown for a comparison. It clearly indicated that subjects were significantly more likely to select cooperation in the Stag Hunt than in the N-P treatment (Figure 6.a). The difference in the average cooperation rate is remarkable, 31.6 percentage points overall (23.7

Figure 6: *Subjects' Decisions to Select C in the Stag Hunt Treatment*



(a) Avg. cooperation rate across all supergames

(b) Avg. cooperation rate, supergame by supergame

Notes: *p*-values (two-sided) in panel a were calculated based on subject random effects probit regressions with robust jackknife standard errors. All supergame data were used. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

percentage points only when round 1 action choices are compared).[14] The differences in the subjects' cooperation behavior persisted across the supergames (Figure 6.b). It can be thus concluded that selecting C is not at all risky in the Stag Hunt treatment. It should be acknowledged that some readers may find the clear difference between the Stag Hung and N-P treatments obvious, since, as already discussed, the N-P treatment has two stages (an interaction stage and a punishment stage), but the Stag hunt treatment has a degenerated single stage in each round. Nevertheless, this result meaningfully confirms that subjects did not perceive the stage-game structure of the N-P treatment as a stag hunt game even after gaining experiences. This suggests that potential drivers behind Result 4 are likely to be the same as those behind Result 1.

5. Charitable-Giving Acts and People's Decisions to Cooperate

5.1. The impact of charitable-giving act in itself

In the author's view, there are two potential causes for the perverse effects of charitable-giving. The first possibility is that subjects' acts of charitable-giving per se undermine their willingness to cooperate. The second possibility is that the information of charitable-giving negatively, not positively, influences the behaviors of those involved.

The first possibility can be formulated by using the so-called "self-regulatory depletion" hypothesis. Psychologists have consistently found that humans have limited resources to deal with self-regulatory activities, such as exercising self-control, coping with stress and dealing with conflicts between prosocial versus selfish motivations (e.g., Baumeister *et al.*, 1994; Muraven and Baumeister, 2000; DeWall *et al.*, 2008). Subjects' charitable-giving decision-making may consume such mental energy, rendering them unable to resist the temptation to defect for an immediate gain in the prisoner's dilemma interaction.[15] Recently, economists have also proposed a similar concept in relation to altruistic acts. For instance, Ashraf and Bandiera (2017) use the term "altruistic capital" and define it as "an asset that enables individuals to internalize the effect of their actions on others" (page 70). They argue that people can increase, not deplete, the altruistic capital by exercising altruistic acts. The self-regulatory depletion hypothesis just discussed is not inconsistent with this capital accumulation idea, because one can reasonably assume that humans accumulate the altruistic capital over a longer time horizon rather than in the short timeframe of the experiment. Parallel to the capital accumulation idea, the self-regulation theory also discusses that the resources are like muscle: individuals can gradually strengthen the resources by using them, over a long time span.

[14] This difference is much larger, compared with the corresponding difference between the N-N and N-P treatments (Figure 4).

[15] Using a simple cake-eating problem, Ozdenoren *et al.* (2012) theoretically illustrated a dynamic self-regulatory depletion path when an agent with limited resources decides consumption over time (also see Kamei (2012)).

Hypothesis A: *Self-regulatory resources deplete if subjects are engaged in charitable-giving activities. Thus, charitable-giving acts per se undermine cooperation.*

By contrast, an alternative possibility is that the transmission of the charitable-giving information drives the perverse effects. First, some people may be opportunistic and strategic. For example, subjects may donate positive amounts as bait aiming to induce their partners to select cooperation, even though they plan to select defection. Second, mis-coordination could easily happen with the additional information of charitable-giving due to beliefs. For example, high donors may expect low donors to select defection. Thus, while conditionally cooperative individuals may signal their cooperativeness through charitable-giving acts, they may tend to select defection if they are matched with persons whose willingness to donate is lower than their own. In this sense, successful coordination may be difficult.[16] On the other hand, subjects who made/committed a small donation may become cautious about action choices of high donors, because they might expect high donors to select defection if they are interacting with low donors to avoid a risk of being exploited. Hence, the low donors may also tend to select defection, even when matched with a high donor.

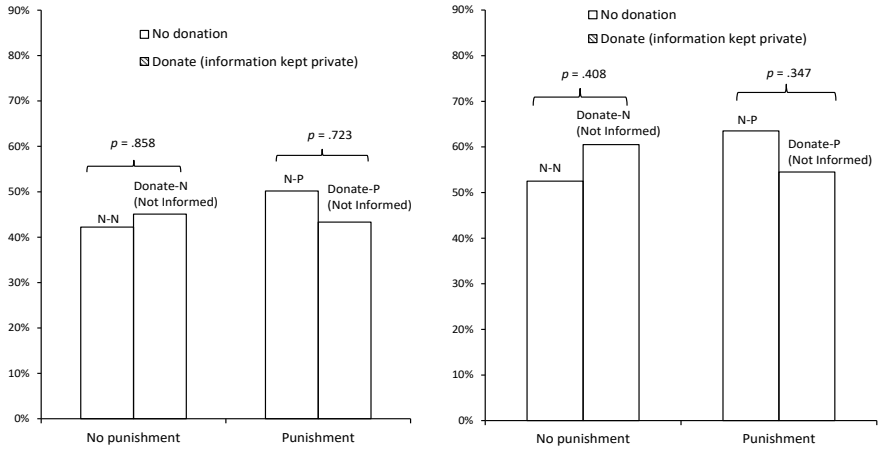
Hypothesis B: *The transmitted information of charitable-giving decisions makes mutual cooperation difficult.*

In order to explore which hypothesis is more appropriate, additional experiments were conducted so that subjects make charitable-giving decisions but the donation decisions are not informed to anyone. Considering that the perverse effects of charitable-giving were observed in both charitable-giving formats (Section 4), the additional treatments used the "donate in advance" format as the framework of the additional treatments. Two treatments – one without punishment, and the other with punishment, were conducted. The two treatments were called "Donate in advance, No Punishment (Not Informed)," dubbed Donate-N (Not Informed), and "Donate in advance, Punishment (Not Informed)," dubbed Donate-P (Not Informed). Except for the donation amounts being private information, these two treatments are identical to the Donate-N and Donate-P treatments. Two sessions per treatment (four sessions in total) were conducted with the same subject pool and the recruiting procedure.

As shown in Figure 7, subjects' average cooperation rates were slightly higher in the Donate-N (Not Informed) than in the N-N treatment (panel I). Especially, in the fifth supergame, the round 1 average cooperation rate was 12.5 percentage points higher in the former

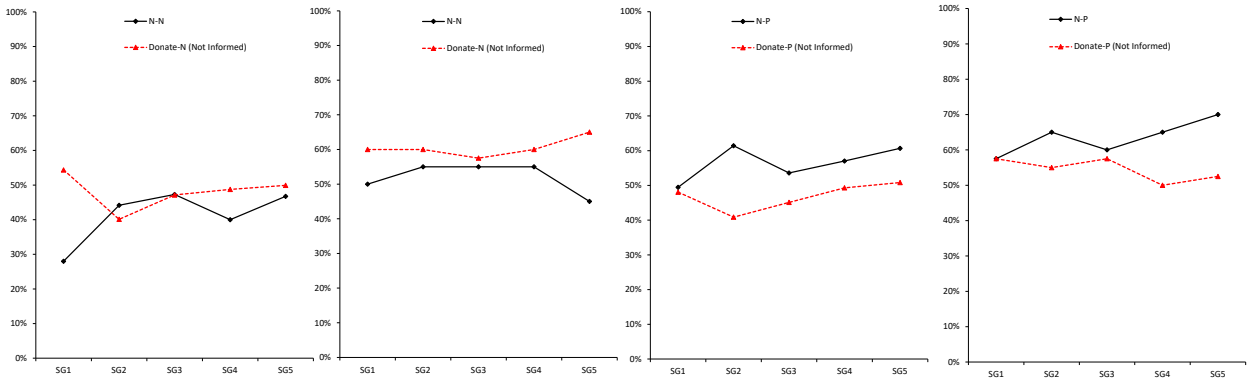
[16] Coordination could even become difficult if it is assumed that a player exhibits some interdependent preferences, such as inequity aversion (e.g., Fehr and Schmidt, 1999). For example, an inequity averse high donor, when matched with a low donor, may prefer to obtain a higher payoff than the low donor in the prisoner's dilemma interaction, since the loss due to donation is larger for the high donor than that for the low donor, other things being equal.

Figure 7: *Subjects' Cooperation Behaviors when Donation Amounts are Private Information*



(a) All Rounds of Supergames (b) First Rounds of Supergames

(I) Average cooperation rate across all supergames



(i) All Rounds of Supergames (ii) First Rounds of Supergames (i) All Rounds of Supergames (ii) First Rounds of Supergames

(A) Donate-N and Donate-N (Not Informed) treatments (B) Donate-P and Donate-P (Not Informed) treatments

(II) Average cooperation rate, supergame by supergame

Notes: *p*-values (two-sided) in panel I were calculated based on subject random effects probit regressions with robust jackknife standard errors. All supergame data were used. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

than in the latter treatment. However, the willingness of subjects in the former treatment to cooperate quickly decreased. As a result, the difference in the supergame average cooperation rate was only 6.6 percentage points in the fifth supergame. By contrast, subjects' average cooperation rates were slightly lower in the Donate-P (Not Informed) than in the N-P treatment (panel I). In the fifth supergame, the difference in the subjects' round 1 average cooperation rate was 17.5 percentage points. However, subjects' cooperation rates did not decline much in the Donate-P (Not Informed) treatment. As a result, the difference in the supergame average cooperation rate between the Donate-P (Not Informed) treatment and the control N-P treatment became less than ten percentage points in that supergame (panel II.B.i). In sum, the overall

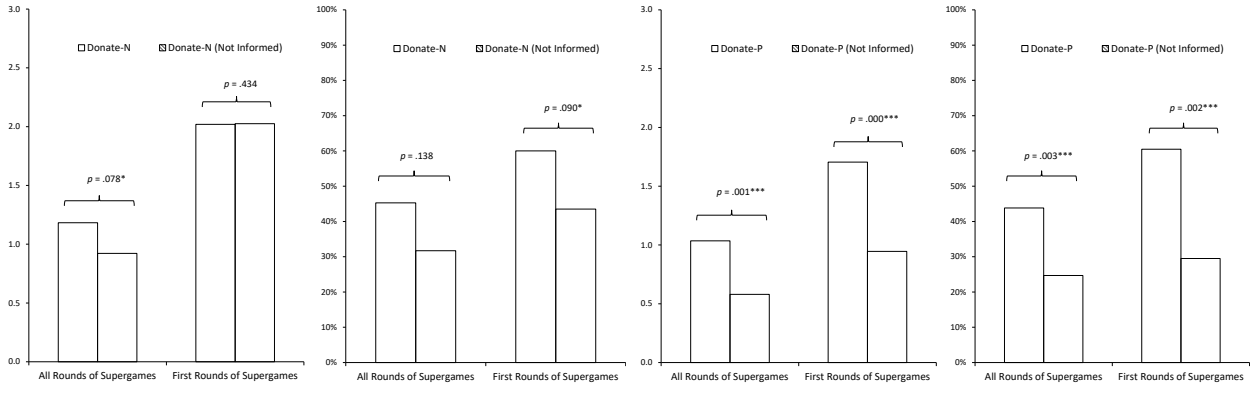
cooperation behaviors were not significantly different between the control treatments and the corresponding additional treatments, whether subjects had an option to punish peers or not (panel I). These results suggest that Hypothesis B is the more appropriate reason behind the negative impact of charitable-giving activities in the main treatments.

Result 6: *Hypothesis A was rejected. A likely factor that drives the negative effects of charitable-giving is the transmitted information of people's charitable-giving decisions.*

These additional treatments can also be used to test whether subjects changed their charitable-giving behaviors when donation amounts were informed to peers. If they attempt to signal their generosity, either aiming to achieve mutual cooperation or strategically exploiting naïve cooperators through inducing them to select cooperation, the donation amounts would be larger under the visibility condition than otherwise. In addition, some subjects may enjoy some satisfaction (e.g., esteem) from their high pro-social acts being seen by partners. In that case, stronger charitable-giving behavior would be predicted with the information transmission.

Figure 8 reports the size of subjects' donation decisions by the information condition. The results show that the information transmission enhances subjects' charitable-giving activities. First, subjects were more likely to be engaged in charitable-giving in the Donate-N than in the Donate-N (Not Informed) treatment, although the difference is not significant at the 5% level (panel A). Second, subject's charitable-giving activities were significantly stronger in the Donate-P than in the Donate-P (Not Informed) treatment (panel B). The second observation is consistent with Result 3 which showed that achieving mutual cooperation would be easier with

Figure 8: *Donation Amounts and Information Transmission*



Notes: *p*-values (two-sided) in panels i were calculated based on subject random effects tobit regressions with robust jackknife standard errors because a non-negligible fraction of decisions were zeros. *p*-values (two-sided) in panels ii were calculated based on subject random effects probit regressions with robust jackknife standard errors because the dependent variable is binary. All supergame data were used. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level. See Appendix Table A.2 for subjects' donation decisions by supergame.

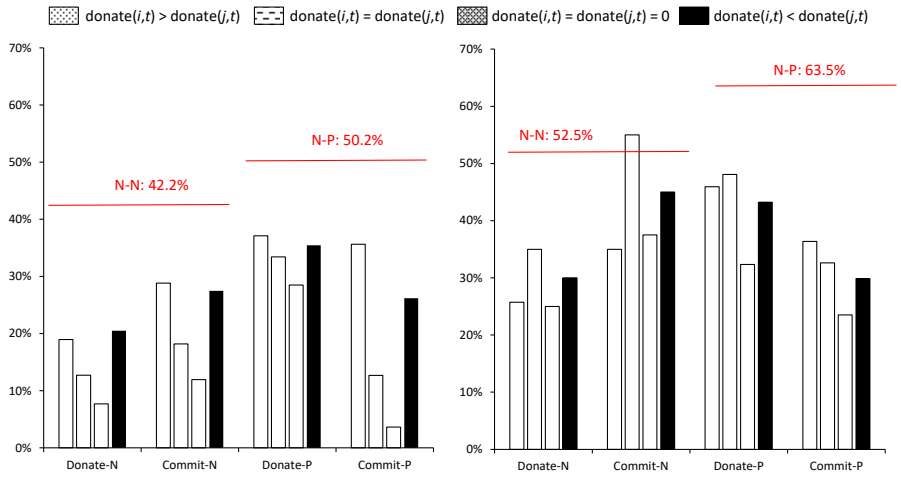
punishment. Subjects' charitable-giving motives and the impact of information will be further examined in Section 5.2.

Result 7: *Subjects spent significantly larger amounts for donation and also made such positive donation significantly more frequently in the Donate-P than in the Donate-P (Not Informed) treatment.*

5.2. Charitable-giving information and people's behaviors

The analysis in Section 5.1 indicated that Hypothesis B is more appropriate to explain the negative effects of charitable-giving. This section is devoted to a further analysis on how the charitable-giving information affected subjects' decisions to cooperate. For this purpose, the data are classified into the three categories (whether own donation amounts are larger than, equal to, or smaller than matched partners' donation amounts) to examine how subjects' cooperation behaviors differ by the category. Figure 9 reports the calculation results. Three meaningful patterns are found. First, a high donor was less likely to select cooperation when matched with a low donor, compared with the level of cooperation in the control, N-N or N-P treatment.[17] The second pattern seen here is that a subject's cooperation rate stayed at a low level even when matched with a person who displayed a higher willingness to donate than his own. This suggests that low donors were not persuaded by the charitable-giving information of high donors. Third, a

Figure 9: *Average Cooperation Rates by the Donation Pattern*



(a) All Rounds of Supergames (b) First Rounds of Supergames

Notes: donate(i,t) and donate(j,t) indicate subject i's and subject i's partner j's donation amounts or percentages in round t. In each of the Donate-N, Commit-N, Donate-P and Commit P treatments there are no significant differences in the frequency of selecting cooperation across the donation patterns for almost all comparisons (see Appendix Table A.3 for the test results).

[17] As discussed in Section 5.1, there are two potential factors that could explain this result: (i) high donors might have had tendencies to select defection, forming pessimistic beliefs on the low donors' cooperation behaviors, and/or (ii) high donors might have made donation decisions strategically to induce partners to select cooperation.

subject's cooperation rate remained low when he was matched with someone whose willingness to donate was the same as his own.[18] These results imply that charitable-giving information undermines cooperation uniformly across the three situations.

Result 8: *Regardless of whether the partner donated larger, lower, or the same amount as i, i was less likely to select cooperation when she received the charitable-giving information than otherwise.*

Next, in order to study how subjects' donation decisions and the charitable-giving information affected their cooperation decisions in details, a regression analysis was conducted. As shown in Table 4, those with higher inclinations to donate in a given round were more likely than the others to select cooperation in that round in all the treatments. This suggests that despite Result 8, subjects' donation decisions were still reliable signals of cooperativeness. In addition, the information of opponents' charitable-giving activities did affect high donors' decisions to cooperate. When subject *i* donated more than her partner *j* did in a given round, the larger the difference in the donation amount between *i* and *j*, the more likely *i* was to select defection (see the positive deviation in Table 4). This suggests that high donors' decisions to cooperate were positively correlated with the matched low donors' inclinations to donate. On the other hand, low donors' cooperation decisions were not

Table 4: *Subjects' Donation Decisions and Decisions to Cooperate*

Dependent variable: A dummy which equals 1(0) when subject *i* selects to cooperate (defect)

Treatment:	Donate-N (Not Informed)	Donate-P (Not Informed)	Donate-N	Commit-N	Donate-P	Commit-P	Donate-P (Framing)	Donate-P (1 point)
Independent variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i's donation decision in round *t*	.113*** (.047)	.113*** (.035)	.362*** (.069)	.125** (.060)	.265*** (.082)	.115*** (.027)	1.298*** (.244)	.381*** (.048)
Positive deviation (= max{*i*'s donation decision minus *j*'s donation decision in round *t*, 0})	---	---	-.352*** (.102)	-.109** (.052)	-.338*** (.088)	-.099*** (.024)	-.939*** (.214)	-.381*** (.054)
Abs. negative deviation (= max{*j*'s donation decision minus *i*'s donation decision in round *t*, 0})	---	---	.040 (.028)	.006 (.007)	.024 (.048)	.006 (.006)	.312* (.157)	.036 (.034)
Constant	-.528** (.209)	-.271*** (.159)	-1.988*** (.216)	-1.010*** (.216)	-.691*** (.235)	-1.323*** (.214)	-.989*** (.210)	-1.167*** (.178)
# of observations	2,720	1,220	2,300	1904	2,040	1,540	1,952	2,060
Prob > F	.0055***	.0062***	.000***	.2435	.0111**	.0012***	.0001***	.0000***

Notes: Subject random effects probit regressions with robust jackknife standard errors. The numbers in parentheses are the standard errors. The realized length of supergame $k-1$ (= 0 for the first supergame) was included as an independent variable as a control (the estimates were omitted to conserve space). The Donate-P (1 point) and Donate-P (framing) treatments are additional treatments (see Section 6). *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

[18] This cannot be explained by the mis-coordination hypothesis. This result may mean that having additional information may make people more strategically-minded.

affected by the matched high donors' inclinations to cooperate (see the abs. negative deviation in Table 4).

Subjects were less likely to cooperate when the charitable-giving institution was present than otherwise from the first round in almost all situations (Figure 9.b). How did subjects' experiences in prior rounds affect the negative impact of information? This sub-question was explored by restricting data to the observations in which history variables can be defined (Table 5). Three kinds of regressions were performed, again by dividing the data into three situations – whether a decision-maker *i* donated more than *i*'s matched partner *j* in a given round (panel A), donated less than *j* in a given round (panel B), or donated the same amount as *j* in a given round (panel C). The dependent variable is subjects' decisions to cooperate in round *t*, and independent variables include the fraction of cases in which their matched partners selected cooperation in a given situation so far (their own experiences). The estimated coefficients on the experience variables (variables (a), (b) and (c) in Table 5) are all positive, and are significant in almost all specifications. This implies that independent of the donation balance between the two players, the more frequently *i*'s partners selected defection in the past for a given situation, the less likely *i* is to choose cooperation when the same scenario happens. This suggests that subjects' reciprocal responses to the negative experiences deepened the negative impact of information seen in round 1 over time.

Result 9: *(i) Those with higher inclinations to donate were more likely than the others to select cooperation. (ii) Charitable-giving information affected high donors' cooperation behaviors: high donors' decisions to cooperate were positively correlated with their matched low donors' inclinations to donate. (iii) In each of the three donation patterns (whether subject i donated higher, lower, or the same amount as j), the more frequently i's partners selected defection in the past, the less likely i was to select cooperation when the same situation happened in the current round.*

Table 5: *Subjects' Experiences and Decisions to Cooperate*

Dependent variable: A dummy which equals 1(0) when subject *i* select to cooperate (defect)

(A) *i*'s decision to cooperate in round *t* of supergame *k* when *i*'s partner *j* donated more than *i* in that round

Treatment:	Donate-N	Commit-N	Donate-P	Commit-P	Donate-P (Framing)	Donate-P (1 point)
Independent variable:	(1)	(2)	(3)	(4)	(5)	(6)
i's donation decision in round *t*	.581***	.117	.176	.134***	.324***	---
	(.211)	(.134)	(.155)	(.048)	(.119)	
j's donation decision in round *t*	-.022	-.000	-.063	.008	-.078	---
	(.046)	(.010)	(.087)	(.017)	(.064)	
(a) Fraction of cases in which *i*'s partner	1.441***	1.558***	1.209***	.605	1.767***	1.864***
donated more than *i* and selected cooperation	(.442)	(.321)	(.414)	(.599)	(.303)	(.492)
Constant	-2.653***	-1.595***	-1.144**	-1.697***	-1.517***	-1.458***
	(.550)	(.328)	(.485)	(.368)	(.346)	(.447)
# of observations	498	534	439	431	516	243
Prob > F	.0013***	.0009***	.0367 **	.0564*	.0000***	.0013***

(B) *i*'s decision to cooperate in round *t* of supergame *k* when *i*'s partner *j* donated less than *i* in that round

Treatment:	Donate-N	Commit-N	Donate-P	Commit-P	Donate-P (Framing)	Donate-P (1 point)
Independent variable:	(1)	(2)	(3)	(4)	(5)	(6)
i's donation decision in round *t*	-.015	.016	-.169*	.013	-.033	---
	(.097)	(.010)	(.098)	(.012)	(.047)	
j's donation decision in round *t*	.153	.122	.446***	.065**	.291***	---
	(.140)	(.089)	(.142)	(.031)	(.096)	
(b) Fraction of cases in which *i*'s partner	.890**	1.522***	1.199***	.654	1.149***	1.223*
donated less than *i* and selected cooperation	(.337)	(.459)	(.399)	(.505)	(.313)	(.626)
Constant	-1.799***	-1.834***	-.482	-.991***	-1.245***	-.823**
	(.394)	(.421)	(.452)	(.285)	(.356)	(.312)
# of observations	516	544	475	450	523	254
Prob > F	.0199**	.0027***	.0003***	.0124**	.0003***	.1667

(C) *i*'s decision to cooperate in round *t* of supergame *k* when *i*'s partner *j* donated the same amount as *i*

Treatment:	Donate-N	Commit-N	Donate-P	Commit-P	Donate-P (Framing)	Donate-P (1 point)
Independent variable:	(1)	(2)	(3)	(4)	(6)	(5)
i's and *j*'s donation decision in round *t*	.319**	.167	-.068	.223	.412**	.976***
	(.156)	(.161)	(.221)	(.342)	(.157)	(.265)
(c) Fraction of cases in which *i*'s partner donated	1.790***	2.521***	1.809***	2.372	1.470***	1.939***
the same amount as *i* and selected cooperation	(.350)	(.625)	(.307)	(1.906)	(.370)	(.298)
Constant	-2.579***	-2.560***	-1.527***	-3.278*	-2.255***	-1.780***
	(.350)	(.682)	(.317)	(1.869)	(.393)	(.253)
# of observations	819	450	683	309	556	1120
Prob > F	.0000***	.0008***	.0000***	.4014	.0000***	.0000***

Notes: Subject random effects probit regressions with robust jackknife standard errors. The numbers in parentheses are the standard errors. Observations when *j* donated more than *i* and variable (a) is defined were used in panel A. Observations when *j* donated less than *i* and variable (b) is defined were used in panel B. Observations when *j* donated the same amount as *i* and variable (c) is defined were used in panel C. The realized length of supergame $k - 1$ (= 0 for the first supergame) was included as an independent variable as a control (the estimates were omitted to conserve space). *** Significant at the 1% level. ** Significant at the 5% level. * Significant at the 10% level.

5.3. Structural estimation of subjects' strategy choices

One may wonder what exactly subjects' strategy choices changed by having the charitable-giving information. This question can be explored by utilizing the recent structural estimation approach to identify the distribution of subjects' strategy choices. A two-step approach was undertaken based on the maximum likelihood method developed by Dal Bó and Frechétte (2011). In the first step, exactly the same set of strategies adopted in Dal Bó and Frechétte (2011) was considered to examine how subjects' likelihoods to select cooperative strategies were affected by the presence of the charitable-giving institution. In the second step, two new strategies that are conditional upon the charitable-giving information were included to estimate the percentage of subjects that acted according to the information.

In the first step, the strategies considered are "Always Defect" (AD), "Always Cooperate" (AC), "Grim Trigger" (GT), "Tit for Tat" (TFT), "Win Stay Lose Shift" (WSLS), "Trigger Strategy with 2 Periods of Punishment" (T2).[19] Two clear patterns emerged (Table 6.A). First, subjects were more likely to act according to the AD strategy when the charitable-giving information was sent to each other. The percentages of the AD subjects are significantly larger in the Donate-N treatment than in the N-N treatment, and in the Donate-P and Commit-P treatments than in the N-P treatment.[20] Second, the percentage of those who act according to the most generous strategy – AC strategy – is much smaller when their partners' charitable-giving information is available than otherwise. These two patterns meaningfully suggest that subjects were more uncooperative in terms of strategy choices in the Donate-N and Commit-N (Donate-P and Commit-P) treatments than in the N-N (N-P) treatment.

The analysis in the first step did not consider a strategy taken based on the charitable-giving information. Thus, a question remains – what fraction of subjects conditioned their action choices on the charitable-giving information? In the second step, two strategies where a subject acts according to the donation pattern in their pair were newly considered. The first strategy is called the "Coordination with Positive Donation" (CP) strategy. A CP subject i is assumed to select cooperation in round t if both i and i's round t partner donated positive amounts (committed positive percentages). This strategy can be thought as the simplest form to use donation as a signal of cooperativeness. In addition, considering that the difference in donation acts between pair mates might also affect their action choices, the second strategy, called the "Coordination with Similar Donation Acts" (CS) strategy, was also considered. A CS subject i is assumed to select cooperation in round t if (a) both i and i's round t partner donated positive amounts (committed positive percentages), and (b) the absolute difference in the donation

[19] See Dal Bó and Frechétte (2011) for the detail.

[20] The differences in the percentage are significant at $p < .001$ between the Donate-N and N-N treatments, and at $p = .0395$ ($p < .0001$) between the Donate-P (Commit-P) and N-P treatment, according to two-sided t tests.

amount (commitment percentage) within the pair is less than or equal to two.[21] Panel B of Table 6 reports the estimation results.[22] It first reveals that 20.7%, 21.0%, 23.6%, and 39.7% of subjects acted according to either the CP or CS strategy in the Donate-N, Commit-N, Donate-P and Commit-P treatments, respectively.[23] This underscore the significance of possible mis-coordination with the charitable-giving information, strengthening Result 6 (reported in Section 5.1), since CP and CS subjects would select defection if confronted with mis-aligned charitable-giving information. Recall, especially, that in these four treatments, subjects did not make any donation on average 35% to 70% of the time (Tables 2 and 3), inducing the CP and CS subjects to frequently make uncooperative choices in the experiment.

Further, panel B also indicates that even after incorporating the two donation strategies, the percentage(s) of the AD subjects in the Donate-N treatment (Donate-P and Commit-P treatments) is (are) much higher compared with the N-N (N-P) treatment. This may mean that, as conjectured in Section 5.2 (see Result 8 and footnote 18), having the additional charitable-giving information may make people more strategically-minded, driving them to act according to the unconditional defection strategy.

Table 6: *Structural Estimation of Subjects' Strategy Choices*

(A) Estimation based on the strategies used in Dal Bó and Frechétte (2011)

Strategy	(i) Treatment without punishment			(ii) Treatment with punishment		
	N-N	Donate-N	Commit-N	N-P	Donate-P	Commit-P
AD	.406 (.068)	.685 (.035)	.475 (.045)	.254 (.073)	.477 (.080)	.685 (.078)
AC	.162 (.072)	.000 (.134)	.000 (.081)	.248 (.90)	.144 (.112)	.150 (.076)
GT	.029 (.056)	.000 (.000)	.060 (.000)	.127 (.086)	.047 (.062)	.024 (.074)
TFT	.308 (.028)	.286 (.000)	.393 (.070)	.218 (.071)	.260 (.043)	.080 (.042)
WSLS	.054 (.091)	.000 (.124)	.000 (.132)	.047 (.085)	.000 (.076)	.000 (.049)

[21] There is a clear difference in the donation pattern by the donation format. When both own and partner's donation amounts were positive, the absolute differences in the donation amounts were less than or equal to two points for the majority of the cases in the Donate-N and Donate-P treatments. By contrast, the differences in the commitment percentage within pairs were quite diverse in the Commit-N and Commit-P treatments. See Appendix Figure A.1 for the cumulative distribution in each treatment.

[22] The AC and WSLS strategies were removed from the set of strategies in the Donate-N and Commit-N treatments since the percentages of subjects that acted according to these strategies were estimated to be zero. The GT and WSLS strategies were removed from the set of strategies in the Donate-P and Commit-P treatments since the estimation found that almost no subjects acted according to these two strategies. See Panel A of Table 6.

[23] Some subjects may be more discriminatory compared with the definition of the CS strategy just used. As an additional analysis, the distributions of subjects' strategy choices were estimated by re-defining condition (b) of the CS strategy so that the absolute difference in the donation amount (commitment percentage) within the pair is less than or equal to one, however showing similar results to Table 6.B: 24.8%, 19.5%, 23.7%, and 40.9% of subjects acted according to either the CP or CS strategy in the Donate-N, Commit-N, Donate-P and Commit-P treatments, respectively. See Appendix Table A.4 for the detail.

T2	.040 (.049)	.029 (.000)	.072 (.000)	.106 (.043)	.072 (.000)	.062 (.000)
Gamma	.611	.483	.582	.611	.674	.541

(B) Estimation by having two donation strategies

Strategy	(i) Treatment without punishment	
	Donate-N	Commit-N
AD	.510 (.036)	.334 (.043)
GT	.000 (.103)	.060 (.094)
TFT	.265 (.000)	.321 (.067)
T2	.017 (.139)	.075 (.184)
CP	.122 (.028)	.061 (.066)
CS	.085 (.089)	.149 (.058)
Gamma	.469	.568

Strategy	(ii) Treatment with punishment	
	Donate-P	Commit-P
AD	.392 (.076)	.399 (.067)
AC	.146 (.089)	.126 (.122)
TFT	.179 (.059)	.000 (.064)
T2	.047 (.066)	.077 (.024)
CP	.236 (.055)	.272 (.052)
CS	.000 (.087)	.126 (.083)
Gamma	.596	.495

Note: The numbers in parentheses are bootstrapped standard errors.

Result 10: *(i) Subjects were more likely to select uncooperative strategies when they received the charitable-giving information than otherwise. (ii) 20.7%, 21.0%, 23.6%, and 39.7% of subjects took discriminatory strategies based on the donation patterns seen in their pairs, undermining cooperation norms in the community.*

6. Discussions: When Could the Charitable-Giving Institution Encourage Cooperation?

The conclusion from the analyses in Sections 4 and 5 is that the charitable-giving information significantly undermines one's willingness to cooperate, perhaps due to mis-coordination, but does this mean that humans cannot cooperate with strangers through investment in pro-social activities outside their on-going dilemma interactions? What might instead encourage players to use charitable-giving institutions positively, aiming to achieve mutual cooperation? One possible explanation for the failure of cooperation in the main experiments is that subjects had limited cognitive ability, with which they behaved myopically, succumbing to the strong temptation to defect when they received mis-aligned information. If this interpretation is appropriate, additional coordination devices would be required for successful cooperation to happen. As is usual for this kind of experimental work, budget constraints preclude testing all potential causes. However, considering that the key finding was a surprising one, it is insightful to explore factors that might reverse the damaging effects of charitable-giving by collecting more data. Hence, this study conducted additional experiments, using the framework of the Donate-P treatment solely for cost reasons.

One straightforward way to guide players' use of the charitable-giving information is to add some framing by emphasizing the signaling value of charitable-giving activities. For example, we can explain to subjects that high donors are more likely to cooperate (Result 9). Recall that (C, C) was an equilibrium outcome in the degenerate stage-game payoff matrix if punishment is considered, even if the interaction is one-shot (Figure 3). If Results 1 and 4 happened due to subjects' limited cognitive ability, providing the additional positive information may alter subjects' decisions to cooperate. Notice also that in reality, charitable-giving programs are often linked to economic transactions in online marketplaces, such as the eBay giving works.[24] While these programs are known to potentially magnify mutual cooperation online among anonymous users, the marketplaces intensively advertise the positive aspects of joining the charitable-giving program.[25] This kind of advertisement may influence users' behaviors.

To study the role of framing, an additional treatment (two sessions, each with 20 subjects, as in the Donate-P treatment), called the "Donate-P (Framing)" treatment, was conducted. The Donate-P (Framing) treatment is identical to the Donate-P treatment, except that subjects were informed of some positive aspects seen in the Donate-P treatment. The following explanations were included in the instructions and were read aloud to participants:

"[…]… your donation decision will be informed to your matched partner; and it may affect the partner's decision."

"We conducted this experiment before. The data of the past experiment sessions suggest that **the information on donation decisions can instill the participants' confidence and trust.** On the one hand**,** participants were more likely to select Y when they were matched with partners that donated positive amounts than when they were matched with those who did not donate. Further, those who made positive donation were more likely to select Y than those who did not do so."[26]

[24] It should be acknowledged that economic transactions on eBay can be modeled using a sequential-move investment game rather than a prisoner's dilemma game. However, this example is useful in the context of this study.

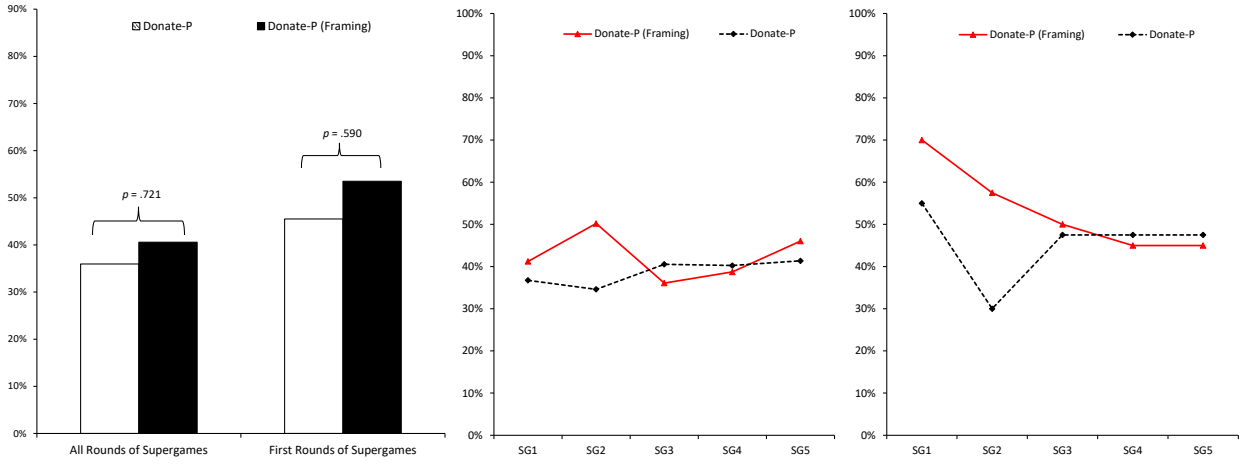
[25] For example, eBay's website in the USA explains that Charity listings "tend to sell more often and at higher prices because buyers are more willing to purchase items that benefit a good cause." (https://charity.ebay.com/help/seller [accessed on April 26, 2018]). Its website in the United Kingdom emphasizes that "donation information is displayed in the item description, instilling buyer confidence and trust." (http://pages.ebay.co.uk/help/sell/nonprofit.html [accessed on April 26, 2018]). eBay occasionally advertises the charitable-giving program heavily through media outlets including their blog with some data. For example, the eBay blog on December 2nd, 2014 explained: "eBay giving works listings have a 30% higher chance of selling" and "People feel good about making a purchase that has the added benefit of a tax deduction, and that helps others in need. So, by giving a portion to charity, which results in a deduction for you, you're actually elevating the exposure of your charity, and your item, increasing the chances that it will sell. Plus, charity listings command 2-6% higher prices than non-charity listings. Essentially, people are willing to spend a little bit more for a good cause. I think that makes this a win-win-win-win-win?" (https://community.ebay.com/t5/eBay-for-Business/eBay-Giving-Works-What-It-Is-How-It-Works-and-How-You-Can/ba-p/26163534 [accessed on August 26, 2018])

[26] As in the other treatments, neutral framing was used throughout except for this description. Cooperate (defect) was labelled as Y (Z) in the instruction.

Providing subjects the additional information significantly increased subjects' donation amounts. The per round average donation amounts in the Donate-P (framing) and Donate-P treatments were 2.23 and 1.04 points per subject, respectively.[27,28] However, the framing had only limited impact on subjects' cooperation decisions (Figure 10). Although, in the first supergame, the subjects' round 1 average cooperation rate was 15 percentage points higher in the Donate-P (framing) than in the Donate-P treatment (panel b.ii), the subjects' high willingness to cooperate diminished quickly over time within that supergame (panel b.i). Also, the round 1 average cooperation rate dropped rapidly from supergame to supergame: in the fourth and fifth supergames, it was even lower compared with the Donate-P treatment (panel b.ii). The levels of supergame average cooperation rates were similar between the Donate-P and Donate-P (framing) treatments in each supergame (panel b.i). In sum, subjects' average cooperation rates were somewhat higher in the Donate-P (framing) than in the Donate-P treatment (panel a); however, the increases are not significant. These results suggest that providing the selective positive information from past sessions was not enough to reverse the effects of the charitable-giving institution.

Result 11: *The charitable-giving institution did not encourage cooperation under random matching even if subjects were informed of the evidence on positive correlations between players' donation amounts and their likelihood to cooperate.*

Figure 10: *Subjects' Cooperation Behaviors in the Donate-P (Framing) treatment*



(a) Avg. cooperation rate across all supergames

(i) All Rounds of Supergames

(ii) First Rounds of Supergames

(b) Avg. cooperation rate, supergame by supergame

[27] The difference is significant at two-sided $p < .001$, according to a subject random effects tobit regressions with robust jackknife standard errors. Here, the tobit regression was used as subjects in the Donate-P and Donate-P (framing) treatments selected zero as donation amounts around 47% of the time.

[28] As was the case for the Donate-P treatment (Table 3), the per subject average donation amounts decreased from supergame to supergame in the Donate-P (framing) treatment, according to subject random effects tobit regressions with robust jackknife standard errors. The decrease rates were significant at two-sided $p < .0001$ and =.097, respectively, both when the data from round 1 and from all rounds were used.

Notes: *p*-values (two-sided) in panel a were calculated based on subject random effects probit regressions with robust jackknife standard errors. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame.

One may assume that if mis-coordination between two parties drove Results 1 and 4, then the negative effects may be partly due to a by-product of the experimental design that the choice space is rich for subjects' donation decisions. To address this concern, two sessions of an additional treatment that used a simplified donation format, called "Donate-P (1 point)," were conducted. Subjects in this treatment decided whether or not to donate one point to the Red Cross at the onset of each round (i.e., binary choice).[29] The rest of the design piece is identical to the Donate-P treatment.

Figure 11 shows subjects' cooperation behaviors in the Donate-P (1 point) treatment. The behaviors were very similar to those in the Donate-P treatment. Nevertheless, coordination among subjects may have been easier in the Donate-P (1 point) than in the Donate-P treatment. Donors' average cooperation rate when matched with another donor was 67.0%, more than 15 percentage points higher than the average cooperation rate in the N-P treatment. However, not everyone selected to donate, even though the donation amount was only one point. For example, around 43.9% of subjects chose to donate in round 1.[30] The donors' average cooperation rates were 29.3% when matched with a non-donor. Non-donors consistently showed low inclinations to cooperate. Their average cooperation rate when matched with a donor (non-donor) was 30.8% (25.7%). This suggests that the low cooperation norms in the Donate-P (1 point) treatment can be explained by the transmission of charitable-giving information and the low types' weak cooperative behaviors – similar findings to the original Donate-P treatment. This paper thus conclude that the rich choice space in the donation actions in the original treatments was not the cause behind the perverse effects of charitable-giving.

Result 12: *Result 4 was not caused by the setup that subjects could donate any amount. Subjects' cooperation behaviors in the Donate-P (1 point) treatment were statistically indistinguishable from those in the Donate-P treatment.*

Results 11 and 12 imply that a stronger coordination device might be required to give players enough incentives to invest in charitable-giving activities for achieving mutual cooperation.

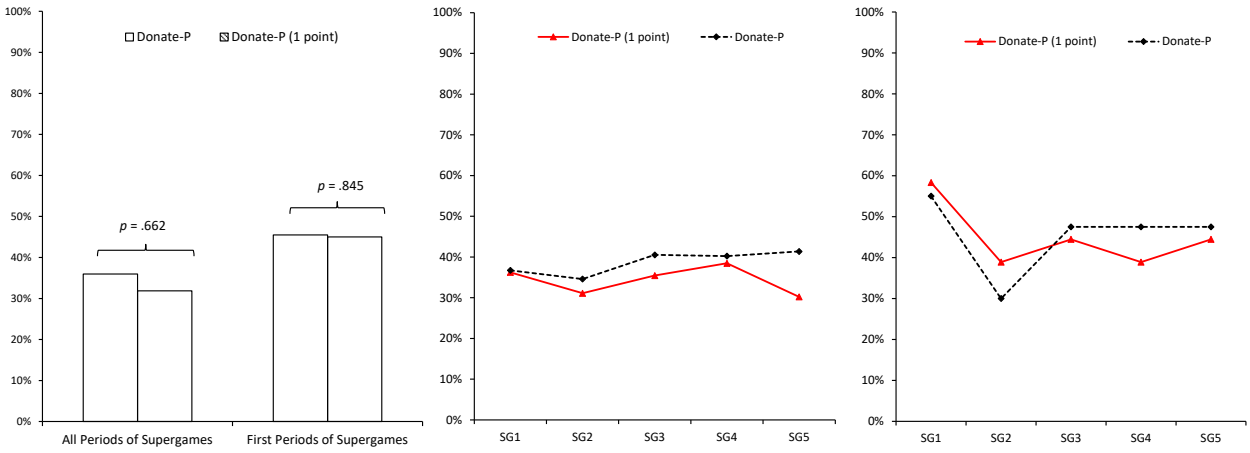
---

[29] There is a charitable-giving program run by Amazon seemingly related to the costly signaling hypothesis, although the decision-maker who donates is Amazon, not users that use the online marketplace. The company operates Amazonsmile – the website on which users can buy the same products as Amazon's main website. If a user purchases a product there, Amazon donates 0.5% of the net purchase price (https://www.amazon.com/gp/help/customer/display.html?nodeId=201365340 [accessed on January 23, 2019]).

[30] The percentage of donors decreased within supergames: the percentage of instances in which subjects made a donation in the experiment was 28.8%. In addition, the percentage of donors decreased from supergame to supergame, according to a subject random effects probit regression with robust jackknife standard errors. These dynamics are consistent with the earlier discussions that some subjects in the original treatments may have given up utilizing the charitable-giving opportunities gradually over time.

One likely candidate would be to let players choose with whom they interact with based on the charitable-giving information. Such partner choice may not only mitigate mis-coordination

Figure 11: *Subjects' Cooperation Behaviors in the Donate-P (1 point) Treatment*



(a) Avg. cooperation rate across all supergames

(i) All Rounds of Supergames

(ii) First Rounds of Supergames

(b) Avg. cooperation rate, supergame by supergame

Notes: *p*-values (two-sided) in panel a were calculated based on subject random effects probit regressions with robust jackknife standard errors. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame.

between the two parties, but in the presence of multiple potential partners also give players strong motives to build a reputation for trustworthy or cooperative behaviors. This hypothesis can be formulated based on prior theoretical and experimental studies. From a theoretical perspective, for example, Zahavi (1975) discusses that humans could engage in costly activities that confer handicaps for survival if such activities signal the true quality of themselves (also see Grafen [1990] and Boone [1998]). Such costly signaling has theoretically been shown to be beneficial for players when forming advantageous partnerships or alliance with high types in dilemma interactions (e.g., Gintis *et al.*, 2011; Roberts, 1998). In addition, prior experiments have shown that players' partner selection can deepen cooperation norms if the selection is made based on their past action choices in the on-going interactions: letting players choose own partners creates competition for trustworthy partners (e.g., Page *et al.*, 2005; Kamei and Putterman, 2017; Coricelli *et al.*, 2004), and partner choice per se may also enhance players' inclination to cooperate (Kamei, 2019). One could then imagine that, even if the only available information is potential partners' past donation activities, competition for trustworthy partners could change players' ways of utilizing the charitable-giving institution, since subjects' inclinations to donate are positively related to their cooperation decisions in prisoner's dilemmas interactions (as seen in Result 9(i)).

To fill the gap between the findings of the present paper and the research that advertises the role of costly signaling, one final experiment was conducted to examine the potential impact

of partner choice. Incorporating the partner selection procedure into our experiment framework is not an easy task, however. Since the group size in the original experiments was four, the number of potential partners would be too small if subjects are allowed to select partners in the same group size. If partner selection is allowed within such a small group, the matching may become too similar to the fixed matching protocol. In the standard experiment with partner choice, there are usually a number of potential partners. The experiment was therefore re-designed by increasing the group size and then incorporating the partner selection procedure in Kamei and Putterman (2017). Two sessions were conducted, each with 20 subjects.[31] The new treatment is called the "Partner Choice" treatment (Table 1). The Partner Choice treatment consists of five supergames, like in the other treatments. At the onset of a given supergame, subjects were randomly assigned to a group of ten and the group composition stayed the same throughout the supergame. Each round had four stages. In the first stage, all subjects simultaneously decided how many points to donate to the Red Cross. The computer then calculated average donation amounts so far in a given supergame for each player. In the second stage, the ten subjects were randomly divided into two subsets of five. A subject's partner in this round was one of the five persons in the set that the subject did not belong to. Every subject was then informed of each potential partner's average donation amount and ranked their potential five player matches from the first to the fifth.[32] Pairs were formed based on the ordering of the sums of rank numbers (see the instruction in the Appendix; also see Kamei and Putterman [2017]). In the third stage, each pair interacted in the prisoner's dilemma stage game described in Figure 1. In the final stage, subjects had a punishment stage, as in the other punishment treatments. Subjects repeated this interaction subject to the random termination rule (with a continuation probability of 90%).

One difficulty in running this additional treatment is the long duration of the experiment. As acknowledged in Kamei and Putterman (2017), the partner selection procedure makes the experiment significantly longer. The expected length of the Partner Choice treatment is 50 rounds (= 1/(1–0.9)×5), whereas the number of rounds was 40 in Kamei and Putterman (2017).[33] Further, the Partner Choice treatment is arguably more complex than in Kamei and Putterman (2017) due to the random termination rule and charitable-giving, which may make the experiment even longer. For this reason, an additional requirement was put into place to avoid

[31] Two sessions were conducted for each treatment also in Kamei and Putterman (2017).
[32] This setup increased the "broadcast efficiency" of the signal – which the anthropologists and biologists define as "the number of signal observers attracted per unit signaling effort" (Smith and Bliege Bird [2000]). They argue that higher broadcast efficiency promotes competition among signalers.
[33] Kamei and Putterman (2017) aimed to have the duration of sessions less than two hours, partly because having lengthy sessions may make experiment data less clean due to some fatigue on the subjects' side.

making the experiment too long: subjects can interact for the five supergames, but only up to 90 minutes.[34]

As the group size was ten in the Partner Choice treatment, two sessions of a special control treatment, called the "Random Match Comparison" treatment (Table 1), were conducted to identify the treatment effect. There were five supergames in this treatment. At the onset of each supergame, subjects were randomly assigned into a group of ten subjects. In each round, they were randomly divided into five pairs of two subjects. A given round consisted of three stages – the donation decision in the first stage, the prisoner's dilemma interaction in the second stage and the punishment activities in the third stage.[35] Like in the Donate-P (Not Informed) treatment, subjects' donation amounts were kept private.

Without the partner choice mechanism, subjects' average cooperation rates remained at a low level (Figure 12). These were 44.0% when only round 1 behaviors are considered, and 37.8% when decisions in all rounds are considered. These rates were much lower, compared with the N-P and Donate-P (Not Informed) treatments. The weaker cooperation behaviors in the Random Match Comparison treatment were consistent with the theory. Standard game theoretic models predict that under random matching, the larger the group size is, the more difficult it is for players to sustain cooperation, because the threshold discounting factor (the degree of patience) above which mutual cooperation hold as an equilibrium outcome is increasing with the group size (e.g., Kandori, 1992).

Partner choice magnified subjects' inclinations to donate. The per round average donation amount in the Partner Choice treatment was 2.78 points, more than 2.5 times that in the Random Match Comparison treatment (1.05 points).[36] In addition, partner choice altered the functioning of the charitable-giving institution (Figure 12). Subjects' achieved much higher cooperation norms in the Partner Choice than in the Random Match Comparison treatment (panel a).[37] The

[34] It takes a total of around 30 minutes in reading instructions before the experiment and in paying subjects after the experiment. It should be acknowledged that this additional requirement may decrease the impact of partner choice, because it may give subjects an impression that their interactions are finitely repeated. If this concern is relevant, this requirement could decrease subjects' willingness to cooperate. This is acceptable for the purpose of this paper because this means the Partner Choice treatment can provide a conservative estimate for the impact of partner choice. The numbers of supergames subjects played were five for one session and two for the other session.

[35] Although an analysis in Section 5 found that donation activities per se do not affect subjects' decisions to cooperate, the donation stage was included to make the number of decisions in the Random Match Comparison treatment similar to that of the Partner Choice treatment.

[36] The difference is significant at two-sided $p < .001$, according to a subject random effects tobit regression with robust jackknife standard errors, although the average donation amounts decreased from supergame to supergame in the Partner Choice treatment, similar to the Donate-P treatment. As high cooperation norms were well sustained in the Partner Choice treatment (Figure 12), the decreasing trend of donation amounts may mean that in later supergames, subjects did not need to strongly signal their willingness to cooperate through donation activities.

[37] p-values (two-sided) in Figure 12 were calculated using the data from all the supergames the subjects played. It should be acknowledged that this comparison might not be precise since not all subjects played the five supergames (see footnote 34). Considering that the subjects in all the four sessions in the Partner Choice and Random Match

positive effect of the charitable-giving institution stayed stable from supergame to supergame (panel b). Successful matching between like-minded individuals drove this positive effect. First, subjects on average gave better rank numbers to those with higher inclinations to donate (panel c). The mutual ranking procedure generated pairs of those whose own and partners' donation amounts were positively correlated (panel d.i). The donation amounts made in the experiment were clear indicators of subjects' decisions to cooperate (panel d.ii). For example, pairs with the lowest sum of rank numbers in their group (the "first pair" bar in panel d.ii) selected cooperation around 90% of the time on average in the first rounds, and more than 80% of the time even when data in all rounds were considered. These outcomes suggest that even when the only available information is people's charitable-giving activities, partner choice could be a powerful device to reverse the effects of charitable-giving from negative to positive.

As in Section 5.3, subjects' strategy choices were also estimated using the maximum likelihood method, revealing strong impact of partner choice on limiting subjects' selection of uncooperative strategies (see Appendix Table A.5 for the detail). While the percentage of subjects estimated to have acted according to the AD strategy is more than 50% in the Random Match Comparison treatment, it is only 20 to 25% in the Partner Choice treatment. Instead, more than 40% of subjects in the latter treatment were estimated to have acted according to the AC strategy.

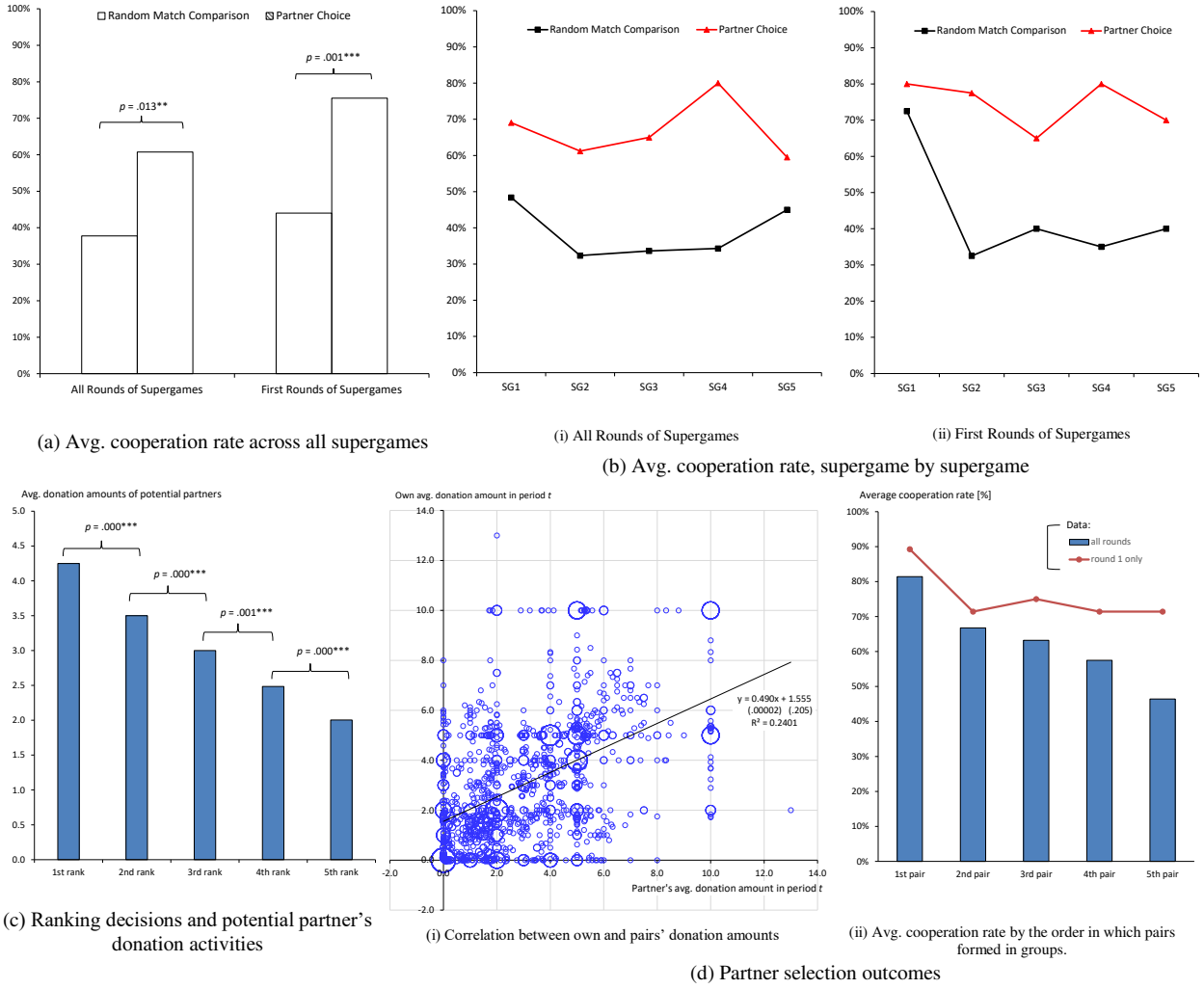
Result 13: *Subjects achieved by far higher cooperation norms in the Partner Choice than in the Random Match Comparison treatment. Subjects in the Partner Choice treatment on average gave better rank numbers to those who made larger donations. As a result, pairs were formed between subjects whose donation amounts were similar to each other. The higher a pair's inclination to donate, the stronger the cooperation norm that they achieved in the prisoner's dilemma.*

There are other bi-directional partner selection procedures (e.g., Gale-Shapley stable marriage mechanism [Bayer, 2011]; second price auction [Coricelli *et al.*, 2004]). It would be useful to test the robustness of Result 13 using an algorithm other than the one in Kamei and Putterman (2017). Although subjects chose with whom they interact in the additional sessions, a simpler form of sorting, such as automatic pair formation based on the sizes of subjects' donation activities, may also have a similar effect (e.g., Gächter and Thöni, 2005; Gunnthorsdottir *et al.*, 2007).

Comparison treatments played the first two supergames, *p*-values were also calculated using the data from the two supergames. The additional estimation found that the differences in average cooperation rate between these two treatments are significant at p (two-sided) = .023 and .008 when the data from all rounds and from round 1, respectively, are used in the supergames.

Of course, mechanisms other than partner choice or sorting may encourage players to cooperate via transmission of the charitable-giving information too. One possibility could be to make costly pro-social activities directly benefit the community members (e.g., in the form of contributing to a public good), since then the community members may reciprocate to those who take such pro-social actions. This reciprocation may help people build cooperative relationships more easily within the prisoner's dilemma interactions.

Figure 12: *Partner Choice and Decisions to Cooperate*



(a) Avg. cooperation rate across all supergames

(i) All Rounds of Supergames

(ii) First Rounds of Supergames

(b) Avg. cooperation rate, supergame by supergame

(c) Ranking decisions and potential partner's donation activities

(i) Correlation between own and pairs' donation amounts

(ii) Avg. cooperation rate by the order in which pairs formed in groups.

(d) Partner selection outcomes

Notes: *p*-values (two-sided) in panel a were calculated based on subject random effects probit regressions with robust jackknife standard errors. In the regressions, the last supergame length was controlled for observations in the second to fifth supergame. *p*-values (two-sided) in panel c were calculated based on individual random effects tobit regressions with bootstrapped standard errors. The size of each circle in panel d.i indicate its frequency. The numbers in parentheses in the linear equation (OLS) in panel d.i are robust standard errors. In panel d.ii, for example, the 1st pair means a pair of subjects whose sum of rank numbers was the lowest in their group (i.e., the pair formed first in the ranking stage in a given group).

Another possibility is to relax the anonymity condition. In this study, in order to assess the impact of charitable-giving information in a controlled manner, subjects were not given any

identified information. Since some people may care about their social image (e.g., Ariely *et al.*, 2009; Bénabou and Tirole, 2006; Linardi and McConnell, 2011), lifting confidentiality or introducing some identified information may effectively limit strategic behaviors and encourage them to invest in charitable-giving for improving cooperation (e.g., Andreoni and Petrie, 2004). Further, providing a promise of public recognition or symbolic prize based on charitable-giving activities could strengthen such a positive impact (e.g., Lacetera and Macis, 2010; Karlan and McConnell, 2014). How people's concerns for social image interact with costly signaling would be an interesting avenue for future research.

Finally, needless to say, the presence of social norms that nurture common goals could alter people's behaviors. Mis-coordination and strategic behaviors could likely be driven by conflicting interests among people. Past experiments have demonstrated that humans' cooperation types are heterogeneous (e.g., Fischbacher *et al.*, 2001; Kamei, 2014). Recently, Fehr and Williams (2018) showed that, given an opportunity to form a consensus regarding normatively appropriate behaviors, a culture of universal cooperation quickly emerges in the community through efficient institutional choices. Such norm consensus opportunities may help reverse the negative effects of charitable-giving information by effectively mitigating mis-coordination between high and low donors and discouraging strategic behaviors.

7. Conclusion

Prior research in the social sciences consistently proposes that cooperation can evolve among non-kins with the help of costly signaling of own ability to cooperate. It was not, however, clear how the pro-social information in itself affects people's cooperation behaviors. In the framework of an indefinitely repeated two-player prisoner's dilemma game with random matching, this study let subjects make donation decisions to the British Red Cross in each round. The charitable-giving information was conveyed to their matched partners. To our surprise, the charitable-giving information significantly undermined cooperation. This negative effect was robust to the donation format used, and was commonly observed regardless of whether subjects had a post-interaction punishment stage in each round. Nevertheless, subjects' donation decisions remained still signals of their cooperative tendencies: subjects' inclinations to donate were positively correlated with their likelihood to select cooperation.

A series of additional experiments helped shed light on the perverse effect of charitable-giving. First, the key cause behind the perverse effect is likely the charitable-giving information sent to peers. Subjects' acts of charitable-giving per se did not undermine their willingness to cooperate, implying that these acts in itself did not deplete subjects' self-regulatory resources. Second, the perverse effect of charitable-giving was not alleviated even when subjects were

provided with strong framing that emphasized the signaling value of charitable-giving, or simplified the charitable-giving process.

All these results consistently suggest that charitable-giving information alone is not enough to improve cooperation among strangers under random matching. As such, a stronger coordination device may be required to reverse the effect of costly signaling to positive. As demonstrated, if subjects have the potential to choose with whom they deal, remarkably, the charitable-giving information can have a positive effect. The data clearly indicated successful matching among those with similar inclinations to donate. This resonates with the view in anthropology and biology that the endogenous group formation process may play an important role in making costly signaling work.

This paper is also related to the experimental work on cheap talk and communication. Prior research states that letting people *directly* send signals regarding their intended action choices helps improve cooperation under certain conditions. The positive impact of a direct signal has been seen in the context of both prisoner's dilemma games and coordination games (e.g., Cooper *et al.*, 1992; Charness, 2000; Duffy and Feltovich, 2002 and 2006; Blume and Ortmann, 2007; Blume *et al.*, 2017). So, why did the charitable-giving information have a negative effect under random matching in this study? There is a stark difference between this study and these prior studies: unlike in the prior experiments, subjects in this study did not directly send a signal to their partners. Thus, the subjects may not have felt guilty for deviating behaviorally from the signal just implied by their charitable-giving activities, which they may or may not have viewed as a promise.

This paper also contributes to the literature on infinitely repeated prisoner's dilemma game with random matching. People's cooperation behaviors are known to be modest when there are no institutions that help assist their cooperation behaviors, even if mutual cooperation holds as an equilibrium outcome under infinite repetition. Recent experiments suggest that cooperation can evolve with forced disclosure of past action choices or with reputation mechanisms based on the own action choices (e.g., Stahl, 2013; Camera and Casari, 2009; Kamei, 2017). Unlike these studies, to the knowledge of the author, this paper is the first attempt to study how subjects' dilemma interactions are affected by the information of partners' pro-social behaviors *outside* the on-going dilemmas. The experiments suggest that making such indirect signals available to each other may make them more conservative and perversely induce them to choose uncooperative strategies under random matching.

REFERENCES

Andreoni, James, and John Miller, 2002. "Giving according to Garp: an experimental test of the consistency of preferences for altruism." *Econometrica* 70: 737-753.

Andreoni, James, and Ragan Petrie, 2004. "Public goods experiments without confidentiality: a glimpse into fund-raising." *Journal of Public Economics* 88: 1605-1623.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99(1): 544-55.

Ashraf, Nava, and Oriana Bandiera, 2017. "Altruistic capital." *American Economic Review* 107 (5): 70-75.

Baumeister, Roy, Todd Heatherton, and Dianne Tice, 1994. *Losing control: How and why people fail at self-regulation*. San Diego, CA, USA: Academic Press.

Bayer, Ralph-C, 2011. "Cooperation in Partnerships: The Role of Breakups and Reputation." The University of Adelaide School of Economics Research Paper No. 2011-22.

Bénabou, Roland, and Jean Tirole, 2006. "Incentives and prosocial behavior." *American economic Review* 96(5): 1652-1678.

Blanco, Mariana, Dirk Engelmann, and Hans Theo Normann, 2011. "A within-subject analysis of other-regarding preferences." *Games and Economic Behavior* 72: 321-338.

Bliege Bird, Rebecca and EricAlden Smith, 2005. "Signaling Theory, Strategic Interaction, and Symbolic Capital." *Current Anthropology* 46(2): 221-248

Blume, Andreas, and Andreas Ortmann, 2007. "The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria." *Journal of Economic Theory* 132(1): 274-290.

Blume, Andreas, Peter Kriss, and Roberto Weber, 2017. "Pre-play communication with forgone costly messages: experimental evidence of forward induction." *Experimental Economics* 20: 368-395.

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch, 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71: 117-120.

Boone, James, 1998. "The evolution of magnanimity: when is it better to give than to receive?" *Human Nature* 9: 1-21.

Camera, Gabriele, and Marco Casari, 2009. "Cooperation among Strangers under the Shadow of the Future." *American Economic Review* 99: 979-1005.

Charness, Gary, 2000. "Self-Serving Cheap Talk: A Test Of Aumann's Conjecture." *Games and Economic Behavior* 33: 177-194.

Charness, Gary, and Martin Dufwenberg, 2006. "Promises and partnership." *Econometrica* 74 (6): 1579-1601.

Coricelli, Giorgio, Dietmar Fehr, and Gerlinde Fellner, 2004. "Partner Selection in Public Goods Experiments." *Journal of Conflict Resolution* 48(3): 356-378.

Cooper, Russell, Douglas DeJong, Robert Forsythe, and Thomas Ross, 1992. "Communication in Coordination Games." *Quarterly Journal of Economics* 107: 739-771.

Dal Bó, Pedro, and Guillaume Fréchette, 2018. "On the Determinants of Cooperation in Infinitely Repeated Games: A Survey." *Journal of Economic Literature* 56(1): 60-114.

Darwin, Charles, 1874. *The Descent of Man, and Selection in Relation to Sex*.

DeWall, Nathan, Roy Baumeister, Matthew Gailliot, and Jon Maner, 2008. "Depletion Makes the Heart Grow Less Helpful: Helping as a Function of Self-Regulatory Energy and Genetic Relatedness." *Personality and Social Psychology Bulletin* 34(12): 1653-1662.

Duffy, John, and Nick Feltovich, 2002. "Do Actions Speak Louder Than Words? An Experimental Comparison of Observation and Cheap Talk." *Games and Economic Behavior* 39: 1-27.

Duffy, John, and Nick Feltovich, 2006. "Words, deeds, and lies: Strategic behavior in games with multiple signals." *Review of Economic Studies* 73: 669-688.

Elfenbein, Daniel, Ray Fisman, and Brian Mcmanus, 2012. "Charity as a Substitute for Reputation: Evidence from an Online Marketplace." *Review of Economic Studies* 79(4): 1441-68.

Ellison, Glenn, 1994. "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching." *Review of Economic Studies* 61: 567-588.

Fehr, Ernst, and Simon Gächter, 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90(4): 980-994.

Fehr, Ernst, and Simon Gächter, 2002. "Altruistic punishment in humans." *Nature* 415: 137-140.

Fehr, Ernst, and Klaus Schmidt, 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(3): 817-868.

Fehr, Ernst, and Tony Williams, 2018. "Social Norms, Endogenous Sorting and the Culture of Cooperation." Department of Economics, University of Zurich. Working Paper No. 267

Fehrler, Sebastian, and Wojtek Przepiorka, 2016. "Choosing a partner for social exchange: Charitable giving as a signal of trustworthiness." *Journal of Economic Behavior & Organization* 129: 157-171.

Fischbacher, Urs, 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10: 171-178.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr, 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71(3): 397-404.

Fisman, Raymond, Shachar Kariv, and Daniel Markovits, 2007. "Individual preferences for giving." *American Economic Review* 97: 1858-1876.

Gächter, Simon, and Christian Thöni, 2005. "Social Learning and Voluntary Cooperation among Like-minded People." *Journal of the European Economic Association* 3: 303-314.

Gintis, Herbert, Eric Smith, and Samuel Bowles, 2001. "Costly Signaling and Cooperation." *Journal of Theoretical Biology* 213: 103-119.

Grafen, Alan, 1990. "Biological signals as handicaps." *Journal of Theoretical Biology* 144: 517-46.

Gunnthorsdottir, Anna, Daniel Houser, and Kevin McCabe, 2007. "Disposition, history and contributions in public goods experiments." *Journal of Economic Behavior & Organization* 62(2): 304-315.

Hawkes, Kristen, and Rebecca Bliege Bird, 2002. "Showing off, handicap signaling, and the evolution of men's work." *Evolutionary Anthropology* 11(2): 58-67.

Heller, Yuval, and Erik Mohlin, 2018. "Observations on Cooperation." *Review of Economic Studies* 85(4): 2253-2282.

Kamei, Kenju, 2012. "Self-regulatory strength and dynamic optimal purchase." *Economics Letters* 115(3): 452-454

Kamei, Kenju, 2014. "Conditional Punishment." *Economics Letters* 124(2): 199-202.

Kamei, Kenju, 2017. "Endogenous Reputation Formation under the Shadow of the Future." *Journal of Economic Behavior & Organization* 142: 189-204.

Kamei, Kenju, 2019. "Cooperation and Endogenous Repetition in an Infinitely Repeated Social Dilemma." *International Journal of Game Theory* 48: 797-834.

Kamei, Kenju, and Louis Putterman, 2017. "Play it Again: Partner Choice, Reputation Building and Learning from Finitely-Repeated Dilemma Games." *Economic Journal* 127(602): 1069-95.

Kandori, Michihiro, 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59: 63-80.

Karlan, Dean, and Margaret McConnell, 2014. "Hey look at me: The effect of giving circles on giving." *Journal of Economic Behavior & Organization* 106: 402-412.

Lacetera, Nicola, and Mario Macis, 2010. "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme." *Journal of Economic Behavior & Organization* 76: 225-237.

Linardi, Sera, and Margaret McConnell, 2011. "No excuses for good behavior: Volunteering and the social environment." *Journal of Public Economics* 95(5-6): 445-454.

Mailath, George, and Larry Samuelson, 2006. *Repeated Games and Reputations: Long-Run Relationships*. Oxford University Press USA.

Muraven, Mark, and Roy Baumeister, 2000. "Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?" *Psychological Bulletin* 126(2): 247-259.

Ozdenoren, Emre, Stephen Salant, and Dan Silverman, 2012. "Willpower and the Optimal Control of Visceral Urges." *Journal of European Economic Association* 10(2): 342-368.

Page, Talbot, Louis Putterman, and Bulent Unel, 2005. "Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry, and Efficiency." *Economic Journal* 115(506): 1032-1053.

Roberts, Gilbert, 1998. "Competitive altruism: from reciprocity to the handicap principle." *Proceedings of the Royal Society of London B: Biological Sciences* 265:427-431.

Smith, Eric, and Rebecca Bliege Bird, 2005. "Costly signaling and prosocial behavior." In: *Moral Sentiments and Material Interrests: The Foundations of Cooperation in Economic Life* (edited by Gintis, H., Bowles, S., Boyd, R., and Fehr, E.), pp. 115-148. MIT press.

Smith, Eric, and Rebecca Bliege Bird, 2000. "Turtle hunting and tombstone opening: public generosity as costly signaling." *Evolution and Human Behavior* 21(4): 245-261

Stahl, Dale, 2013. "An Experimental Test of the Efficacy of a Simple Reputation Mechanism to Solve Social Dilemmas." *Journal of Economic Behavior & Organization* 94: 116-124.

Van Lange, Paul, Bettina Rockenbach, and Toshio Yamagishi (editors), 2011. *Reward and Punishment in Social Dilemmas*. Oxford University Press.

Volka, Stefan, Christian Thöni, and Winfried Ruigrok, 2012. “Temporal stability and psychological foundations of cooperation preferences.” *Journal of Economic Behavior & Organization* 81(2): 664-676.

Zahavi, Amotz, 1975. “Mate Selection - A Selection for a Handicap.” *Journal of Theoretical Biology* 53: 205-214.

Zahavi, Amotz, 1977. “Reliability in communication systems and the evolution of altruism.” In *Evolution Ecology* (edited by B. Stonehouse and C.M. Perrins), pages 253-259. Macmillan Press: London.