The Role of Visibility on Third Party Punishment Actions

for the Enforcement of Social Norms

Kenju Kamei[#]

Department of Economics and Finance, Durham University

Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk

Abstract:

This paper presents results from a prisoner's dilemma game experiment with a third party punisher. Third party punishment was frequently observed, in line with previous studies. Despite the prevalence of punishment, having one third party punisher in a group did not make one's defection materially unbeneficial because of the weak punishment intensity observed. When a third party player's action choice was made known to another third party player in a different group, however, third party punishment was sufficiently strong to transform the dilemma's incentive structure into a coordination game, through which cooperation norms can be effectively enforced.

JEL classification: C92, D01, H49

Keywords: experiment, cooperation, dilemma, third party punishment, social norms

[#] I thank John Hey for his hospitality when I conducted the experiment in the University of York. I also thank Mark Wilson (an IT manager at the University of York) for his support in setting the computers in the experimental sessions. I also thank the editor, Joseph E. Harrington, and anonymous referee for valuable comments. This project was funded by Durham University Business School, and the Kyoto University Institute of Economic Research foundation.

1. Introduction

One well-known and consistent finding in recent decades is that some people display other-regarding preferences, such as inequity aversion, when interacting with others. A large body of experimental research has shown that even third parties, who are not directly involved in the relevant interactions, frequently impose punishment when they encounter unfair economic behavior in dilemma games (e.g., Fehr and Fischbacher, 2004; Lergetporer *et al.*, 2014; Kamei, forthcoming).

Most research in this area to date found that while third party punishment is frequently observed, it is much weaker than direct punishment (e.g., Fehr and Fischbacher, 2004). This paper experimentally studies how the visibility of third parties' punitive actions may affect their punishment behaviors. This research question is motivated by past work proposing that the visibility of actions enhances people's pro-social behavior through image motivation (see, e.g., Bénabou and Tirole (2006) for a theoretical model, and Ariely *et al.* (2009) for experimental evidence on charitable giving). It is also motivated by the research which suggests that increasing the visibility of actions within a group may affect people's altruistic tendencies in the ongoing interactions (see, e.g., Sell and Wilson (1991) for the impact of individualized, instead of aggregate, information on voluntary contributions to public goods, and Kamei and Putterman (2015) for direct higher-order punishment in a public goods game). High visibility of action choices may trigger social effects, such as shame and pride (e.g., Bowles and Gintis, 2005), potentially influencing third parties' punishment behaviors.

In the experiment, there are two players that engage in a one-shot prisoner's dilemma game with each other (PD players, hereafter), and a third party player who decides how to impose sanctions on the PD players in each group. The results demonstrate that the punishment intensity on a norm violator is much stronger when each punisher's action choice is made known to another punisher, than when the punitive actions are kept private. Moreover, in the high visibility condition, the third party players almost completely refrain from perverse punishment of cooperators. These findings suggest that raising the visibility of third parties' punitive actions can be a powerful device for disciplining their sanctioning activities.

2. Experimental Design and Hypotheses

This study is based on a prisoner's dilemma game with a third party player (Fehr and Fischbacher, 2004). There are two treatments, namely the "Standard" and "Visibility" treatments, implemented using a between-subjects design. Experimental points are converted into pounds sterling at a rate of five points to £1.

2.1. The Standard Treatment

At the onset of this treatment, subjects are randomly assigned to a group of three so that each group has two PD players and one third party player. There are two stages. In Stage 1, PD players are each endowed with 25 points and simultaneously decide whether to send 10 points to their counterparts. If a subject sends 10 points to her counterpart, the counterpart receives 30 (= 3×10) points and the remaining 15 points become the sender's payoff. If the subject does not send 10 points, she retains the full endowment as her payoff. Hereafter, we call a subject who sends (does not send) 10 points a "cooperator" ("defector"). The third party player in each group is not involved in the prisoner's dilemma interaction.

In Stage 2, each third party player receives an endowment of 40 points and makes punishment decisions for their respective group. Punishment points assigned to each PD player must be an integer between 0 and 20. For each punishment point assigned to a target, one point is deducted from the third party player's payoff and three points are deducted from the target's payoff. Each punisher makes the following four decisions using the strategy method (punishers make decisions before being informed of the first stage outcome):

Scenario CC: Punishment points targeted at a **c**ooperator who interacted with another **c**ooperator;

Scenario DC: Punishment points targeted at a **d**efector who interacted with a **c**ooperator;

Scenario CD: Punishment points targeted at a **c**ooperator who interacted with a **d**efector;

Scenario DD: Punishment points targeted at a **d**efector who interacted with another **d**efector.

After third party players complete four decisions, their choices corresponding to the realized PD players' sending decisions are applied.

Standard theory predicts no punitive behaviors of third parties because punishment is privately costly. As shown in Appendix C.1, however, the inequity aversion model of Fehr and Schmidt (1999) suggests that (a) a third party player *i* punishes a defector in Scenario DC if *i* exhibits sufficiently strong aversion to disadvantageous inequality (i.e., $\alpha_i > 1 - \frac{\beta_i}{2}$), and (b) *i* even punishes a cooperator in Scenario CC if *i* exhibits much stronger aversion to disadvantageous inequality (i.e., $\alpha_i > 2$).

2.2. The Visibility Treatment

The Visibility treatment is identical to the Standard treatment, except that each third party player is randomly and anonymously paired with another punisher in a different group, akin to an enforcement team, and their respective punishment behavior is made known to the partner. This visibility condition is common knowledge to all subjects. Even though two third parties are put in a team, they act independently to make punishment decisions toward different PD players. There is real-world relevance here: for instance, individuals who work in public enforcement usually share reports with other officers working in the same role in the event of encountering law violators.

If *i* has (non-strategic) image motivation (e.g., Ariely *et al.*, 2009; Bénabou and Tirole, 2006), *i* may punish a cooperator less in Scenario CC and punish a defector more in Scenario DC in the Visibility than in the Standard treatment. As we can reasonably assume that a non-trivial fraction of subjects are concerned about their image, we can formulate the following hypotheses in the paper:

Hypothesis 1: *Punishment strength in Scenario CC is weaker in the Visibility than in the Standard treatment.*

Hypothesis 2: *Punishment strength in Scenario DC is stronger in the Visibility than in the Standard treatment.*

The Fehr-Schmidt model also predicts Hypothesis 1, because as illustrated in Appendix C.2, inequality averse *i*'s punishment strength in Scenario CC would be positively correlated with *i*'s beliefs on her matched punisher *j*'s punishment strength in this scenario. This implies that *i* may refrain from engaging in punishment of cooperators in Scenario CC, considering that usually only a minority of subjects commit such anti-social punishment (thus *i* would form a belief that her counterpart is less likely to punish cooperators). The model does not, however, predict Hypothesis 2. This is because (a) *i*'s disutility resulting from inequality with someone in *j*'s group increases if *i* attempts to match her punishment with *j*'s strength in Scenario DC (note that $\alpha \geq \beta$), and (b) when *i* makes punishment decisions in Scenario DC, *j* does not necessarily confront with Scenario DC(CD) – he may also confront with Scenario CC or DD with some probability (see the Appendix for the detail).

In the experiment, the identities of all subjects are not disclosed in order to measure the pure impact of high visibility on third parties' punishment behaviors.

2.3. Experimental Procedure

Four sessions were conducted at the EXEC laboratory at the University of York in December 2015 and February 2016. A total of 96 students (48 students per treatment) participated in the experiment. No subjects participated in more than one session. All experiments except instructions were programmed using the z-Tree software (Fischbacher, 2007). The instructions and verbal explanations in the experiment were neutrally framed.

3. Results

PD players' cooperation rates were the same for the two treatments at 71.9% (23 out of 32 PD players). This implies that PD players did not expect changes in visibility would affect third parties' behaviors. Third parties' punishment behaviors, however, were very different between the treatments.

3.1. Punishment Decisions of Third Party Players

The pattern of punishment replicates that of past research. Fig. 1 indicates that third party punishment is common, and that both its frequency and strength are much higher in Scenario DC than in any other scenario.[1] This pattern resonates with the idea that people are inequality-averse and that third party players attempt to mitigate income inequality by inflicting punishment.

A comparison between the treatments provides two intriguing findings, each of which supports our hypothesis (Fig. 1). First, third party players are significantly *less likely* to impose punishment on a cooperator in the Visibility than in the Standard treatment, regardless of

[1] The frequency of punishment in Scenario DC is significantly higher than that in Scenarios CC and CD in the Visibility treatment (Appendix Table A.1). The punishment intensity in Scenario DC is significantly stronger than that in Scenarios CC and CD in both the treatments (Appendix Table A.2).

whether her partner is a cooperator or a defector (panel(A)). Second, the average punishment points imposed on a defector in Scenario DC are 6.63 in the Visibility treatment, which is *more than double* of those realized in the Standard treatment (3.00 points) [panel(B)].

Result 1: *Raising visibility on punitive actions not only nearly eliminates punishment of cooperators completely, but it also substantially enhances punishment of defectors in Scenario DC.*

3.2. Incentive Changes with Punishment

Third party punishment in the Visibility treatment changes the incentive structure that a PD player faces. Fig. 2 shows the payoff matrix of the PD player after deducting the average punishment amounts. In the Standard treatment, the average realized payoff matrix is still a prisoner's dilemma because punishment in Scenario DC is not sufficiently strong (panel(I)). In the Visibility treatment, however, the average realized matrix is a *coordination game,* where both mutual cooperation and mutual defection are Nash Equilibria (NEs), because of Result 1 (panel(II)).

As explained, the fractions of cooperators were identical across the two treatments. One may wonder then how PD players' cooperation would evolve in each treatment if they repeated the interactions with different players. Assuming a random matching protocol, we can explore this question by calculating PD player *i*'s per-period expected payoff when selecting to cooperate or defect, based on the realized fraction of cooperators and average punishment strength (Table 1). First, in the Standard treatment, the expected payoff when a subject chooses to cooperate is 13.7% *lower* than when she chooses to defect. This suggests that PD players' interactions could converge to mutual defection in repeated interactions, consistent with the analysis in Fig. 2. Second, and by sharp contrast, in the Visibility treatment, the expected payoff from cooperating is 19.4% *higher* than that from defecting. This suggests that, with the aid of higher visibility, PD players' interactions could converge to the mutual cooperation equilibrium through repetition.

Result 2: *(i) The incentive structure that PD players face is a coordination game in the Visibility treatment. (ii) The expected payoff of a PD player in the Visibility treatment is higher when she selects to cooperate than otherwise.*

4. Conclusions

This paper showed that when the punitive action of a third party player is made known to another punisher, punishment not only becomes better targeted, but punishment of a norm violator also becomes stronger. We further demonstrated that the high punishment visibility can change the incentive structure for PD players, through which mutual cooperation becomes an equilibrium for the monetary payoff and selecting to cooperate becomes a materially beneficial strategy. These findings underscore the importance of high visibility in regulating third parties' effective sanctioning activities.

REFERENCES

Ariely, D., Bracha, A., Meier, S., 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99: 544-55.

Bénabou, R., Tirole., J., 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96: 1652-1678.

Bowles, S., Gintis, H., 2005. "Prosocial emotions," in L. Blume, S. Durlauf (Eds.), The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow, Oxford University Press, Oxford: 337-367.

Fehr, E., Fischbacher, U., 2004. "Third-party Punishment and Social Norms." *Evolution and Human Behavior* 25: 63-87.

Fehr, E., Schmidt, K., 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114: 817-68.

Fischbacher, U., 2007. “z-Tree: Zurich Toolbox for Ready-made Economic Experiments.” *Experimental Economics* 10: 171-178.

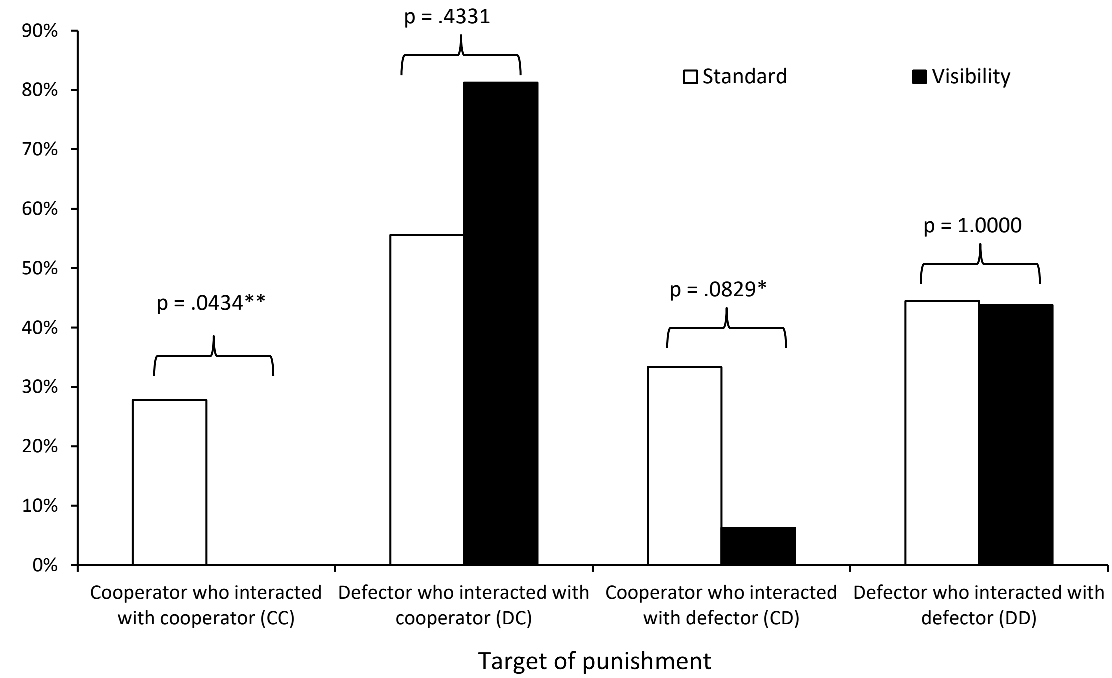
Kamei, K., forthcoming. “Group size effect and over-punishment in the case of third party enforcement of social norms.” *Journal of Economic Behavior and Organization*.

Kamei, K., Putterman, L., 2015. “In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation.” *Journal of Economic Behavior and Organization* 120: 145-159.

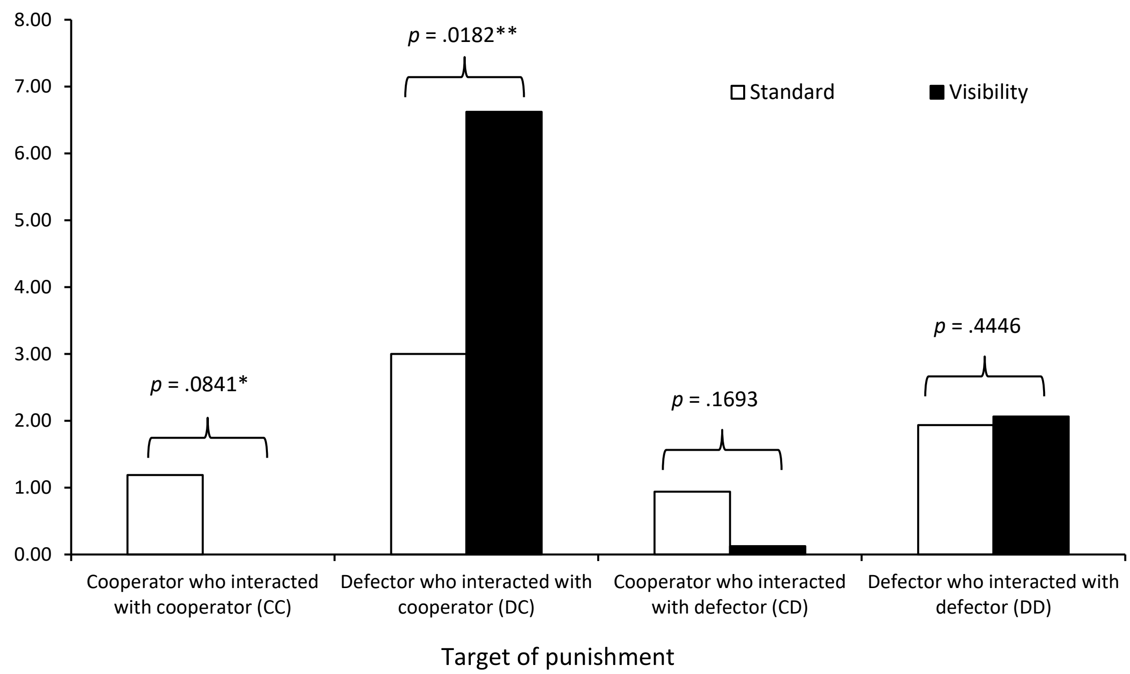
Lergetporer, P., Angerer, S., Glätzle-Rützler, D., Sutter, M., 2014. “Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation.” *Proceedings of the National Academy of Sciences* 111: 6916-6921.

Sell, J., Wilson, R., 1991. “Levels of Information and Contributions to Public Goods.” *Social Forces* 70: 107-124.

Fig. 1: *Frequency of Punishment, and Punishment Strength*



(A) Frequency of punishment



(B) Average punishment points received by PD players

Note: *p*-values in panels (A) and (B) are for two-sided Fisher's exact tests and two-sided Mann-Whitney tests, respectively.

Fig. 2: *Incentive Changes with Punishment*

	Cooperate	Defect
Cooperate	41.44	12.19
Defect	46.00	19.19

(I) the Standard treatment

	Cooperate	Defect
Cooperate	45.00	14.63
Defect	35.13	18.81

(II) the Visibility treatment

Notes: The payoffs of the row player (PD player) when punishment amounts are subtracted from the Stage 1 payoffs. The shaded cells indicate NEs.

Table 1: *Average Expected Payoff of Subject i when Choosing to Cooperate or Defect*

i's action choice: *i*'s counterpart *j*'s action choice:	Cooperate	 Cooperate [Scenario CC] (1)	 Defect [Scenario CD] (2)	Defect	 Cooperate [Scenario DC] (3)	 Defect [Scenario DD] (4)
(i) Subject *i*'s payoff in Stage 1	----	45	15	----	55	25
(ii) Subject *i*'s payoff in Stage 1 *minus* average punishment amount subject *i* would receive						
Standard treatment	----	41.44	12.19	----	46.00	19.19
Visibility treatment	----	45.00	14.63	----	35.13	18.81
(iii) Subject *i*'s expected payoff after Stage 2 [line (ii) × percentage of cooperators or defectors for columns (1) to (4)]						
Standard treatment	**33.21**	29.78	3.43	**38.46**	33.06	5.40
Visibility treatment	**36.46**	32.34	4.11	**30.54**	25.25	5.29

Note: The numbers in bold are *i*'s expected payoffs when *i* chooses to cooperate or defect.

Not for Publication

Supplementary Online Appendix for Kamei, 2018,

"The Role of Visibility on Third Party Punishment Actions

for the Enforcement of Social Norms"

Kenju Kamei

Department of Economics and Finance, Durham University

Email: kenju.kamei@gmail.com, kenju.kamei@durham.ac.uk

Table of Contents

Appendix A: Additional Tables

Table A.1: *The Differences in the Frequency of Third Party Punishment between Scenarios (supplementing Fig. 1(A) of the paper)*

a. The Standard Treatment

	Scenario CC	Scenario DC	Scenario CD	Scenario DD
Scenario CC	---	.3603	1.0000	.5322
Scenario DC	---	---	.5487	.7773
Scenario CD	---	---	---	.7572
Scenario DD	---	---	---	---

b. The Visibility Treatment

	Scenario CC	Scenario DC	Scenario CD	Scenario DD
Scenario CC	---	.0014***	1.0000	.0287**
Scenario DC	---	---	.0073***	.3918
Scenario CD	---	---	---	.1074
Scenario DD	---	---	---	---

Notes: Two-sided Fisher's exact tests. The numbers in these tables are two-sided *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Table A.2: *The Differences in Average Third Party Punishment Point between Scenarios (supplementing Fig. 1(B) of the paper)*

(I) The Standard Treatment

	Scenario CC	Scenario DC	Scenario CD	Scenario DD
Scenario CC	---	.0394**	.6045	.1716
Scenario DC	---	---	.0367**	.4188
Scenario CD	---	---	---	.0481**
Scenario DD	---	---	---	---

(II) The Visibility Treatment

	Scenario CC	Scenario DC	Scenario CD	Scenario DD
Scenario CC	---	.0007***	.3173	.0088***
Scenario DC	---	---	.0009***	.0021***
Scenario CD	---	---	---	.0149**
Scenario DD	---	---	---	---

Notes: Two-sided Wilcoxon signed ranks tests. The numbers in these tables are two-sided *p*-values.
*, **, and *** indicate significance at the 10 percent level, at the 5 percent level and at the 1 percent level, respectively.

Appendix B: Instructions Used in the Experiment

Loaded words, such as “cooperate” and “punish,” were avoided in the instructions.

B.1. The Standard treatment

[The following are the instructions used in the Standard treatment. The instructions were read aloud to subjects by the researcher at the onset of the experiment:]

At the beginning of this experiment, all members in your group are randomly assigned identification numbers 1, 2 and 3. That is, each member is assigned either number with a probability of 1/3 (= 33.3%). The same numbers will not be assigned to 2 members in a group. We call a subject who is assigned number k “player k.” In each group, 3 members are named as player 1, player 2 and player 3. This experiment consists of 2 phases.

Phase 1

In this phase, each of player 1 and player 2 is assigned an endowment of 25 points, and simultaneously decides whether or not to send 10 points to each other. If player 1 sends 10 points to player 2, the 10 points will be tripled and becomes earnings of player 2. Likewise, if player 2 sends 10 points to player 1, the 10 points will be tripled and becomes earnings of player 1.

There are 4 possible situations.

(a) Both player 1 and player 2 send 10 points to their counterparts. In this situation, each player obtains 25 – 10 + 3 × 10 = 45 points.

(b) Player 1 sends 10 points to player 2, but player 2 does not send 10 points to player 1. In this situation, the earnings of player 1 are 25 – 10 = 15 points. The earnings of player 2 are 25 – 0 + 3 × 10 = 55 points.

(c) Player 2 sends 10 points to player 1, but player 1 does not send 10 points to player 2. In this situation, the earnings of player 1 are 25 – 0 + 3 × 10 = 55 points. The earnings of player 2 are 25 – 10 = 15 points.

(d) Neither player 1 nor player 2 sends 10 points to her counterpart. In this situation, player 1 and player 2 each obtain earnings of 25 – 0 = 25 points.

You are not allowed to communicate with anyone during the decision stage. As indicated in the calculations above, your own earnings will be maximized when you do not send 10 points but your counterpart sends 10 points. However, if both player 1 and player 2 send 10 points to each other, the total earnings of the 2 players will be maximized and will be 45 × 2 points = 90 points;

and each player obtains 45 points as earnings. Your earnings will be minimized if you send 10 points to your counterpart but your counterpart does not send 10 points to you.

While players 1 and 2 decide whether or not to send 10 points, player 3 is asked to answer how many persons among players 1 and 2 in his/her group (= 0, 1, 2) he or she thinks will send 10 points to their counterparts. The response of player 3 to this question will not affect his or her earnings in the experiment.

Any questions?

Instructions for Phase 2:

In this phase, player 3 is given an opportunity to reduce earnings of players 1 and 2. In this phase, player 3 is assigned an endowment of 40 points.

Each reduction point player 3 allocates to reduce someone's earnings **reduces player 3's earnings by 1 point** and **reduces that individual's earnings by 3 points**. The reduction points to each player (player 1 or player 2) must be an integer. They must also be less than or equal to 20.

Player 3 will be asked to make decisions for the following four scenarios:

(a) how many reduction points player 3 would like to assign to a player that sent 10 points to his or her counterpart when the counterpart also sent 10 points to that player

(b) how many reduction points player 3 would like to assign to a player that did not send 10 points to his or her counterpart when the counterpart sent 10 points to that player

(c) how many reduction points player 3 would like to assign to a player that sent 10 points to his or her counterpart when the counterpart did not send 10 points to that player

(d) how many reduction points player 3 would like to assign to a player that did not send 10 points to his or her counterpart when the counterpart also did not send 10 points to that player

The reduction points to a player must be an integer. They must also be less than or equal to 20. One of the four decisions will be applied for player 1 and player 2 based on the two players' actual sending decisions. Player 3 will be informed of the actual sending decisions of player 1 and player 2 in Phase 1 at the end of the experiment.

How to calculate earnings:

Earnings of player 1 are calculated as:

Player 1's earnings in Phase 1

minus

Reduction amounts received from player 3

If the earnings are negative, then, the earnings will be 0. Earnings of player 2 are calculated with the same formula.

Player 3 obtains earnings of:

40 – reduction points assigned to player 1 – reduction points assigned to player 2.

Comprehension Questions:

1. At the beginning of the experiment, you are assigned either players 1, 2 or 3. Answer the following questions.

(a) What is the probability that you are assigned a role of player 1?

[ ]

(b) What is the probability that you are assigned a role of player 3?

[ ]

2. Suppose that player 1 in a group sends 10 points to player 2, and player 2 does not send 10 points to player 1. What are the interim earnings of player 1? What are the interim earnings of player 2?

[ ]

3. Suppose that player 3 decides to impose 3 reduction points to a member. How many points are deducted from the earnings of the target? How many points are deducted from the earnings of player 3?

Any questions? Once all questions are answered, we will start the experiment.

[Subjects were asked to answer these control questions. After that, the experimenter explained the answers using a whiteboard in order to make sure that the subjects understood the experiment fully.]

B.2. The Visibility treatment

[The following are the instructions used in the Visibility treatment. The instructions were read aloud to subjects by the researcher at the onset of the experiment:]

At the beginning of this experiment, you will be randomly broken into 2 groups, Group A and Group B. Individuals put into Group A will be further broken into sets of 2 individuals and will be given identification numbers of 1 or 2. Individuals put into Group B will also be further broken into sets of 2 individuals and will be given identification numbers of 3 or 4. You will be put into Group A and Group B with a probability of 66.7% and 33.3%, respectively. We call a subject whose identification number is k "player k." This experiment consists of 2 phases.

In Phase 1, player 1 and player 2 are randomly matched and interact with each other. Player 3 and player 4 are randomly matched and act as a team. Player 3 is assigned to a set of 2 individuals in Group A (player 1 and player 2) and observes their interaction while player 4 observes the interaction of another set of 2 individuals from Group A. Each set of 2 individuals in Group A is observed only by one individual in Group B. In Phase 2, player 3 and player 4 are asked to make some decisions related to their observations in Phase 1. Thus, each interaction unit in this part consists of three participants: two members (player 1 and player 2) of Group A and one member (either player 3 or 4) of Group B. The following are the details of Phase 1 and Phase 2.

Phase 1

In this phase, each of player 1 and player 2 is assigned an endowment of 25 points, and simultaneously decides whether or not to send 10 points to each other. If player 1 sends 10 points to player 2, the 10 points will be tripled and becomes earnings of player 2. Likewise, if player 2 sends 10 points to player 1, the 10 points will be tripled and becomes earnings of player 1.

There are 4 possible situations.

(a) Both player 1 and player 2 send 10 points to their counterparts. In this situation, each player obtains 25 – 10 + 3 × 10 = 45 points.

(b) Player 1 sends 10 points to player 2, but player 2 does not send 10 points to player 1. In this situation, the earnings of player 1 are 25 – 10 = 15 points. The earnings of player 2 are 25 – 0 + 3 × 10 = 55 points.

(c) Player 2 sends 10 points to player 1, but player 1 does not send 10 points to player 2. In this situation, the earnings of player 1 are 25 – 0 + 3 × 10 = 55 points. The earnings of player 2 are 25 – 10 = 15 points.

(d) Neither player 1 nor player 2 sends 10 points to her counterpart. In this situation, player 1 and player 2 each obtain earnings of 25 – 0 = 25 points.

You are not allowed to communicate with anyone during this decision stage. As indicated in the calculations above, your own earnings will be maximized when you do not send 10 points but your counterpart sends 10 points. However, if both player 1 and player 2 send 10 points to each other, the total earnings of the 2 players will be maximized and will be 45 × 2 points = 90 points; and each player obtains 45 points as earnings. Your earnings will be minimized if you send 10 points to your counterpart but your counterpart does not send 10 points to you.

While players 1 and 2 decide whether or not to send 10 points, players 3 and 4 are asked to answer how many persons among players 1 and 2 in their interaction unit (= 0, 1, 2) they think will send 10 points to their counterparts. The response of player 3 and player 4 to this question will not affect his or her earnings in the experiment.

Any questions?

Instructions for Phase 2:

In this phase, player 3 and player 4 are given an opportunity to reduce earnings of players 1 and 2 in their interaction units. (The 2 persons whose earnings player 3 can reduce are different from the 2 persons whose earnings player 4 can reduce.) At the end of their reduction decisions, player 3 will be informed of what reduction points his partner (i.e., player 4) assigned to players 1 and 2 in her interaction unit along with her payoff consequences; likewise, player 4 will also be informed of what reduction points her partner (i.e., player 3) assigned to players 1 and 2 in his interaction unit along with his payoff consequences.

In this phase, player 3 and player 4 are each assigned an endowment of 40 points. Each reduction point a player allocates to reduce someone's earnings **reduces that individual's earnings by 1 point** and **reduces the target's earnings by 3 points**.

Specifically, if you are assigned either player 3 or player 4, you will be asked to make reduction decisions for the following four scenarios (player 3 and player 4 do not communicate prior to deciding on the reduction points):

(a) how many reduction points you would like to assign to a player that sent 10 points to his or her counterpart when the counterpart also sent 10 points to that player

(b) how many reduction points you would like to assign to a player that did not send 10 points to his or her counterpart when the counterpart sent 10 points to that player

(c) how many reduction points you would like to assign to a player that sent 10 points to his or her counterpart when the counterpart did not send 10 points to that player

(d) how many reduction points you would like to assign to a player that did not send 10 points to his or her counterpart when the counterpart also did not send 10 points to that player

The reduction points to each player must be an integer. They must also be less than or equal to 20. One of the four decisions will be applied for player 1 and player 2 based on the two players' actual sending decisions. Player 3 and player 4 will be informed of the actual sending decisions of player 1 and player 2 in Phase 1 at the end of the experiment.

Once both players 3 and 4 press the "OK" button to submit their reduction amounts, they will be informed of what reduction points their partner assigned along with the payoff consequences.

How to calculate earnings:

Earnings of player 1 are calculated as:

Player 1's earnings in Phase 1

minus

Reduction amounts received from player 3 or 4

If the earnings are negative, then the earnings will be 0. Earnings of player 2 are calculated with the same formula.

Earnings of player 3 are calculated as:

40 – reduction points assigned to player 1 – reduction points assigned to player 2.

Earnings of player 4 are calculated with the same formula.

Comprehension Questions:

1. At the beginning of the experiment, you are assigned either players 1, 2, 3 or 4. Answer the following questions.

(a) What is the probability that you are assigned a role of player 1?
[ ]

(b) What is the probability that you are assigned a role of player 3 or player 4?
[ ]

2. Suppose that player 1 in a group sends 10 points to player 2, and player 2 does not send 10 points to player 1. What are the interim earnings of player 1? What are the interim earnings of player 2?

[ ]

3. Suppose that player 4 decides to impose 3 reduction points to a member. How many points are deducted from the earnings of the target? How many points are deducted from the earnings of player 4?

Any questions? Once all questions are answered, we will start the experiment.

[Subjects were asked to answer these control questions. After that, the experimenter explained the answers using a whiteboard in order to make sure that the subjects understood the experiment fully.]

[Subjects answered open-ended questions at the end of the experiment. Although these instructions contained much information, no subjects commented that the instructions were unclear.]

Appendix C: Theoretical Analysis based on the Fehr-Schmidt (1999) model

In this part of the Appendix, I summarize how the inequity-averse preference model by Fehr and Schmidt (1999) predicts the punishment behaviors of third party players. The Fehr-Schmidt (1999) utility function is given as follows: for a list of *n* players' material payoffs (*x*), player *i* receives the following utility:

$$U_i(x) = x_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{x_i - x_j, 0\}. \quad (0)$$

Here, x_i is the material payoff of player *i*, $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. α_i indicates player *i*'s aversion to disadvantageous inequality, while β_i indicates player *i*'s aversion to advantageous inequality. In the theoretical analyses below, I use a continuous interval for *i*'s punishment activities for simplicity, although a discrete interval {0, 1, …, 20} is used as the choice space in the experiment.

C.1. The Standard Treatment

As explained below, for the Standard treatment, in which a third party player *i*'s action choice is never revealed to her matched punisher, (a) *i* imposes positive punishment points on a cooperator in Scenario CC when α_i is sufficiently large; (b) *i* never punishes a cooperator in Scenario CD, while *i* punishes a defector if $\alpha_i + \frac{\beta_i}{2}$ is sufficiently large in Scenario DC, and (c) *i* never punishes a defector in Scenario DD.

(a) *i*'s Punishment Behavior in Scenario CC

Third party player *i* makes punishment decisions with the strategy method (see Section 2 of the paper). Because of this procedure, *i* will impose the same punishment points on two cooperators and receive the following utility if this scenario happens:

$$U_i(x) = (40 - 2P_{cc}) - \alpha_i \frac{1}{2} \cdot 2 \cdot \max\{(45 - 3P_{cc}) - (40 - 2P_{cc}), 0\} - \beta_i \frac{1}{2} \cdot 2 \cdot \max\{(40 - 2P_{cc}) - (45 - 3P_{cc}), 0\}, \quad (1)$$

where $x = (45 - 3P_{CC}, 45 - 3P_{CC}, 40 - 2P_{CC})$. Here, P_{CC} is punishment points from *i* to a cooperator. Equation (1), by simplifying them, reduces to:

$$U_i(x) = \begin{cases} (40 - 5\alpha_i) + (\alpha_i - 2)P_{cc}, \text{if } P_{cc} < 5. \\ (40 + 5\beta_i) + (-2 - \beta_i)P_{cc}, \text{if } P_{cc} \geq 5. \end{cases} \quad (2)$$

Equation (2) suggests that *i* will punish (will not punish) a cooperator if $\alpha_i > 2$ ($\alpha_i < 2$).

Some subjects are known to have high enough α_i that $\alpha_i > 2$. Fehr and Schmidt (1999) estimated that about 30% of individuals have (α, β) = (0, 0), about 30% of them have (α, β) = (0.5, 0.25), 30% of them have (α, β) = (1, 0.6), and the rest, 10%, have (α, β) = (4, 0.6) (also see Fehr and Schmidt [2010]).

(b) *i*'s Punishment Behavior in Scenarios DC and CD

In this scenario, three players' payoffs are given by: $x = (15 - 3P_{CD}, 55 - 3P_{DC}, 40 - P_{CD} - P_{DC})$, where P_{CD} is punishment points from *i* to the cooperator and P_{DC} is punishment points from *i* to the defector in the *i*'s group. The third party player will receive the following utility based on Equation (0):

$$U_i(x) = 40 - P_{CD} - P_{DC} - \alpha_i \tfrac{1}{2}\max\{(15 - 3P_{CD}) - (40 - P_{CD} - P_{DC}), 0\} - \alpha_i \tfrac{1}{2}\max\{(55 - 3P_{DC}) - (40 - P_{CD} - P_{DC}), 0\} - \beta_i \tfrac{1}{2}\max\{(40 - P_{CD} - P_{DC}) - (15 - 3P_{CD}), 0\} - \beta_i \tfrac{1}{2}\max\{(40 - P_{CD} - P_{DC}) - (55 - 3P_{DC}), 0\}. \quad (3)$$

Equation (3), by simplifying them, reduces to:

$$U_i(x) = 40 - P_{CD} - P_{DC} - \tfrac{\alpha_i}{2}\max\{-25 - 2P_{CD} + P_{DC}, 0\} - \tfrac{\alpha_i}{2}\max\{15 - 2P_{DC} + P_{CD}, 0\} - \tfrac{\beta_i}{2}\max\{25 + 2P_{CD} - P_{DC}, 0\} - \tfrac{\beta_i}{2}\max\{-15 + 2P_{DC} - P_{CD}, 0\}. \quad (4)$$

Equation (4) means we need to consider three cases to analyze *i*'s punishment behavior.

<u>Case 1</u>: $15 - 2P_{DC} + P_{CD} > 0$

In this case, $25 + 2P_{CD} - P_{DC} > 0$ because $25 + 2P_{CD} - P_{DC} > 25 + 2P_{CD} - \frac{15}{2} - \frac{P_{CD}}{2} = \frac{35}{2} + \frac{3P_{CD}}{2} > 0$. Thus, Equation (4) can be simplified as follows:

$$U_i(x) = 40 - \frac{15\alpha_i}{2} - \frac{25\beta_i}{2} + \left(-1 - \frac{\alpha_i}{2} - \beta_i\right)P_{CD} + (-1 + \alpha_i + \frac{\beta_i}{2})P_{DC}. \quad (5)$$

As $\frac{dU_i}{dP_{CD}} = -1 - \frac{\alpha_i}{2} - \beta_i < 0$, i never punishes a cooperator in Scenario CD. By contrast, i will punish (will not punish) a defector in Scenario DC if $\alpha_i + \frac{\beta_i}{2} > 1$ ($\alpha_i + \frac{\beta_i}{2} < 1$).

Case 2: $15 - 2P_{DC} + P_{CD} \leq 0$ and $25 + 2P_{CD} - P_{DC} > 0$.

In this case, Equation (4) can be simplified as follows:

$$U_i(x) = 40 - 5\beta_i + \left(-1 - \frac{\beta_i}{2}\right) P_{CD} + (-1 - \frac{\beta_i}{2})P_{DC}. \tag{6}$$

As $\frac{dU_i}{dP_{CD}} < 0$ and $\frac{dU_i}{dP_{DC}} < 0$, $P_{CD} = 0$ and $P_{DC} = 7.5$ as $15 - 2P_{DC} + P_{CD} \leq 0$. Note that this case happens as the boundary between Case 1 and Case 2 ($15 - 2P_{DC} + P_{CD} = 0$).

Case 3: $15 - 2P_{DC} + P_{CD} \leq 0$ and $25 + 2P_{CD} - P_{DC} \leq 0$.

This case never happens as $P_{DC} \leq 20$ in this experiment.

These analyses show that i never punishes a cooperator in Scenario CD, while i will punish (will not punish) a defector in Scenario DC if $\alpha_i + \frac{\beta_i}{2} > 1$ ($\alpha_i + \frac{\beta_i}{2} < 1$).

(c) *i*'s Punishment Behavior in Scenario DD

As is similar to Section C.1(a) above, since the strategy method is used in the experiment, i will impose the same punishment points on two defectors and receive the following utility if this scenario happens:

$$U_i(x) = (40 - 2P_{DD}) - \alpha_i \frac{1}{2} \cdot 2 \cdot \max\{(25 - 3P_{DD}) - (40 - 2P_{DD}), 0\} - \beta_i \frac{1}{2} \cdot 2 \cdot \max\{(40 - 2P_{DD}) - (25 - 3P_{DD}), 0\}, \tag{8}$$

where $x = (25 - 3P_{DD}, 25 - 3P_{DD}, 40 - 2P_{DD})$. Here, P_{DD} is punishment points from i to a defector in Scenario DD. Equation (8) can be simplified as follows:

$$U_i(x) = 40 - 15\beta_i - (2 + \beta_i)P_{DD}. \tag{9}$$

As $\frac{dU_i}{dP_{DD}} = -(2 + \beta_i) < 0$, i never punishes a defector in Scenario DD.

C.2. The Visibility Treatment

When a third party player *i*'s action choice is made known to another third party player *j*, *i* behaves differently from how she does in the Standard treatment described in Section C.1. This is because *i* makes punishment decisions, taking the payoffs of five other players (two PD players in the *i*'s group, and *j* and two PD players in the *j*'s groups) into account. The differences can be summarized as below:

(I) *i*'s punishment strength in Scenario CC is on average positively correlated with her belief on *j*'s punishment strength in that scenario. In addition, *i* refrains from engaging in anti-social punishment activities in Scenario CC if *i* believes that *j* does not perversely punish a cooperator or only mildly does so.

(II) When deciding on punishment strength on a defector in Scenario DC, *i* assigns sizable, but lower punishment points than one allocated in the other group by the matched third party *j*'s, provided that *i* is sufficiently strongly inequality-averse. On average, we have a positive correlation between *i*'s punishment of a defector in Scenario DC and *i*' beliefs on *j*'s punishment of a defector in that scenario. The average punishment strength in this case is smaller than the average punishment strength seen in Scenario DC in the Standard treatment.

The following is the summary of the theoretical analyses.

(a) *i*'s Punishment Behavior in Scenario CC

Suppose that third party player *i* believes that the probabilities that Scenarios CC, DC (CD) and DD are realized are *y*, *z*, and *w*, respectively ($y + z + w = 1$). Then, *i* will maximize the following expected utility based on the probability distribution and Equation (0), by selecting P_{CC}:

$$E[U_i(x)] = y \cdot [U_i(x)]_{\text{CC in } j\prime\text{s group}} + z \cdot [U_i(x)]_{\text{DC(CD) in } j\prime\text{s group}} + w \cdot [U_i(x)]_{\text{DD in } j\prime\text{s group}}$$

Note that *i* only knows situations in the *j*'s group stochastically. This is a more complicated optimization problem than the one studied in Section C.1(a). Thus, for simplicity, I assume that the following conditions hold for *i*'s beliefs:

Assumption A: (i) $P'_{CD} = 0$; (ii) $P'_{DC} \leq 7.5$; (iii) $P'_{DD} = 0$.

Here, P'_{CD} (P'_{DC}) is subject *i*'s belief on her matched third party player *j*'s punishment points on a cooperator (defector) in Scenario CD (DC) in *j*'s group. P'_{DD} is subject *i*'s belief on *j*'s punishment points on a defector in Scenario DD in *j*'s group. Assumption A is reasonable. First, the cooperator in Scenario CD receives the lowest payoff, 15 points, and $\alpha_j \geq \beta_j$. Thus, an inequality averse third party player would not inflict punishment on a cooperator in Scenario CD. Second, if $P'_{DC} = 7.5$ and $P'_{CD} = 0$, *j* receives the same payoff as the defector in his group in Scenario DC. If *j* imposes more than 7.5 punishment points, he receives a higher payoff than the defector. Thus, an inequality averse third party player would not inflict more than 7.5 punishment points on the defector in Scenario DC. Third, two defectors in Scenario DD receive lower payoffs than *j* (25 < 40). If an inequality averse third party player inflicts punishment on the defector in Scenario DD, his utility would decrease.

In the analysis, I consider the case where $P_{CC} \leq 5$. This is because when *i* inflicts five punishment points on each cooperator in Scenario CC ($P_{CC} = 5$), three players in *i*'s group receive the same payoff (45 – 3 × 5 = 40 – 2 × 5 = 30). With these setups, we can express third party player *i*'s expected utility as follows:

$$
\begin{aligned}
U_i(x) = {} & 40 - 2P_{CC} - \frac{2\alpha_i}{5}(5 - P_{CC}) \\
& -y \cdot \left[\tfrac{2\alpha_i}{5}\max\{5 - 3P'_{CC} + 2P_{CC}, 0\} + \tfrac{2\beta_i}{5}\max\{-5 + 3P'_{CC} - 2P_{CC}, 0\} + \tfrac{\alpha_i}{5}\max\{2P_{CC} - 2P'_{CC}, 0\} + \right. \\
& \quad \left. \tfrac{\beta_i}{5}\max\{2P'_{CC} - 2P_{CC}, 0\}\right] \\
& -z \cdot \left[\tfrac{\alpha_i}{5}\max\{15 + 2P_{CC} - 3P'_{DC}, 0\} + \tfrac{\beta_i}{5}\max\{-15 - 2P_{CC} + 3P'_{DC}, 0\} + \tfrac{\beta_i}{5}(25 - 2P_{CC}) + \tfrac{\alpha_i}{5}\max\{2P_{CC} - \right. \\
& \left. P'_{DC}, 0\} + \tfrac{\beta_i}{5}\max\{P'_{DC} - 2P_{CC}, 0\}\right] \\
& -w \cdot \left[\tfrac{2\beta_i}{5}(15 - 2P_{CC}) + \tfrac{\alpha_i}{5}2P_{CC}\right] \qquad (10)
\end{aligned}
$$

Here, P'_{CC} is subject *i*'s belief on her matched third party player *j*'s punishment of a cooperator in Scenario CC. We are interested in the sign of $\frac{dU_i}{dP_{CC}}$ in order to derive *i*'s optimal punishment schedule. Equation (10) suggests that we need to consider nine cases.

<u>Case 1</u>: $5 - 3P'_{CC} + 2P_{CC} \leq 0$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} \geq P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \frac{6\beta_i}{5} + \text{z} \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 2</u>: $5 - 3P'_{CC} + 2P_{CC} \leq 0$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} < P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \frac{6\beta_i}{5} + \text{z} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 3</u>: $5 - 3P'_{CC} + 2P_{CC} \leq 0$ and $15 + 2P_{CC} - 3P'_{DC} < 0$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \frac{6\beta_i}{5} + \text{z} \cdot \frac{6\beta_i}{5} + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 4</u>: $5 - 3P'_{CC} + 2P_{CC} > 0$ and $P_{CC} < P'_{CC}$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} \geq P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{z} \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 5</u>: $5 - 3P'_{CC} + 2P_{CC} > 0$ and $P_{CC} < P'_{CC}$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} < P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{z} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 6</u>: $5 - 3P'_{CC} + 2P_{CC} > 0$ and $P_{CC} < P'_{CC}$ and $15 + 2P_{CC} - 3P'_{DC} < 0$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} + y \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{z} \cdot \frac{6\beta_i}{5} + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 7</u>: $P_{CC} \geq P'_{CC}$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} \geq P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} - y \cdot \frac{6\alpha_i}{5} + \text{z} \cdot \left(-\frac{4\alpha_i}{5} + \frac{2\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 8</u>: $P_{CC} \geq P'_{CC}$, $15 + 2P_{CC} - 3P'_{DC} \geq 0$ and $2P_{CC} < P'_{DC}$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} - y \cdot \frac{6\alpha_i}{5} + \text{z} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right) + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

<u>Case 9</u>: $P_{CC} \geq P'_{CC}$ and $15 + 2P_{CC} - 3P'_{DC} < 0$.

In this case, $\frac{dU_i}{dP_{CC}}$ is calculated as follows:

$$\frac{dU_i}{dP_{CC}} = -2 + \frac{2\alpha_i}{5} - y \cdot \frac{6\alpha_i}{5} + \text{z} \cdot \frac{6\beta_i}{5} + \text{w} \cdot \left(-\frac{2\alpha_i}{5} + \frac{4\beta_i}{5}\right).$$

These calculations suggest that the optimality conditions depend on *i*'s belief on the probability distribution of scenarios (*y*, *z*, *w*). For this simulation, I assume that subjects' beliefs are correct (i.e., the same as the realized probability distribution in the experiment). In the experiment, the percentage of cooperators (defectors) was 72% (28%) – see Section 3 of the paper. Thus, I set: $y = .72 \times .72 \approx .52$, $z = .72 \times .28 \times 2 \approx .40$, $w = .28 \times .28 \approx .08$.

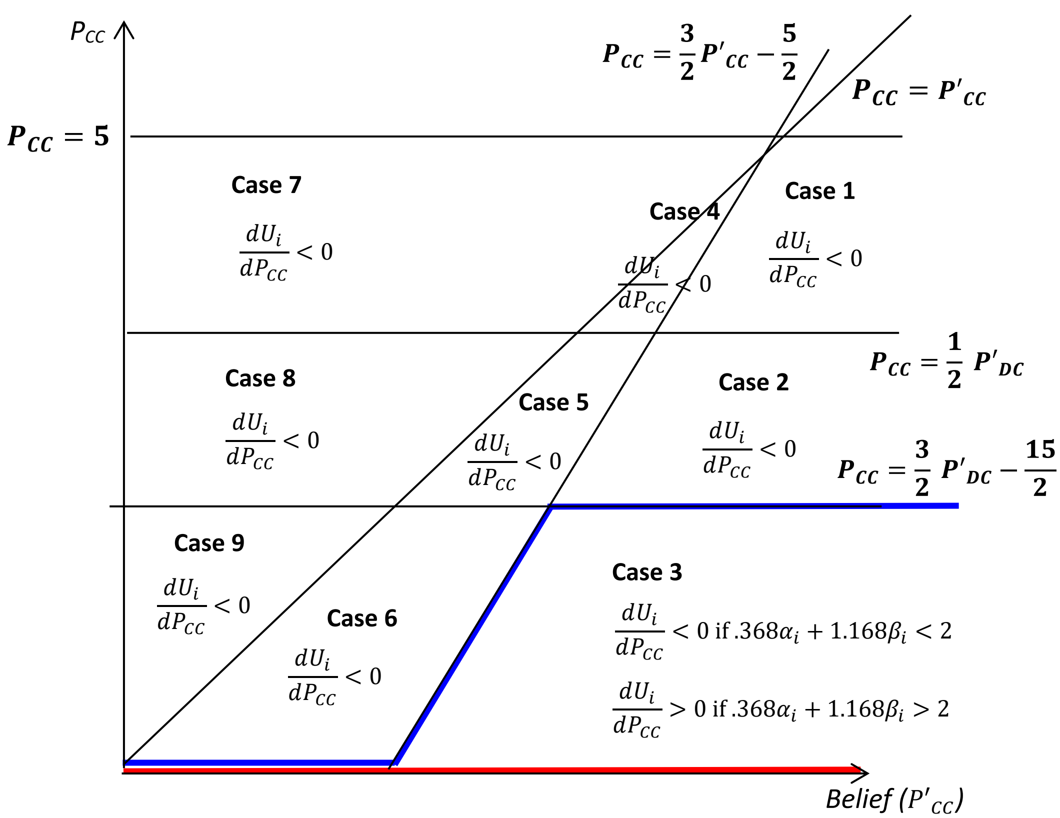
Under this assumption on *y*, *z* and *w*, numerical values of dU_i/dP_{CC} are summarized as in the following table:

	dU_i/dP_{CC}	Sign of dUi/dP_{CC}
Case 1	$-2 + .048\alpha_i + .848\beta_i$	$dU_i/dP_{CC} < 0$
Case 2	$-2 + .208\alpha_i + 1.008\beta_i$	$dU_i/dP_{CC} < 0$
Case 3	$-2 + .368\alpha_i + 1.168\beta_i$	$dU_i/dP_{CC} < 0$ if $.368\alpha_i + 1.168\beta_i < 2$ $dU_i/dP_{CC} > 0$ if $.368\alpha_i + 1.168\beta_i > 2$
Case 4	$-2 - .368\alpha_i + .432\beta_i$	$dU_i/dP_{CC} < 0$
Case 5	$-2 - .208\alpha_i + .592\beta_i$	$dU_i/dP_{CC} < 0$
Case 6	$-2 - .048\alpha_i + .752\beta_i$	$dU_i/dP_{CC} < 0$

Case 7	$-2 - .576\alpha_i + .224\beta_i$	$dU_i/dP_{CC} < 0$
Case 8	$-2 - .416\alpha_i + .384\beta_i$	$dU_i/dP_{CC} < 0$
Case 9	$-2 - .256\alpha_i + .544\beta_i$	$dU_i/dP_{CC} < 0$

These considerations are summarized in Figure C.1 (*i*'s optimal punishment schedule). This suggests that (a) *i*'s punishment of a cooperator in Scenario CC is on average positively correlated with her belief on *j*'s punishment of cooperators in that scenario and (b) *i* will refrain from such anti-social punishment even though *i* is strongly inequality averse if *i* believes that *j* does not perversely punish a cooperator or only mildly does so in Scenario CC.

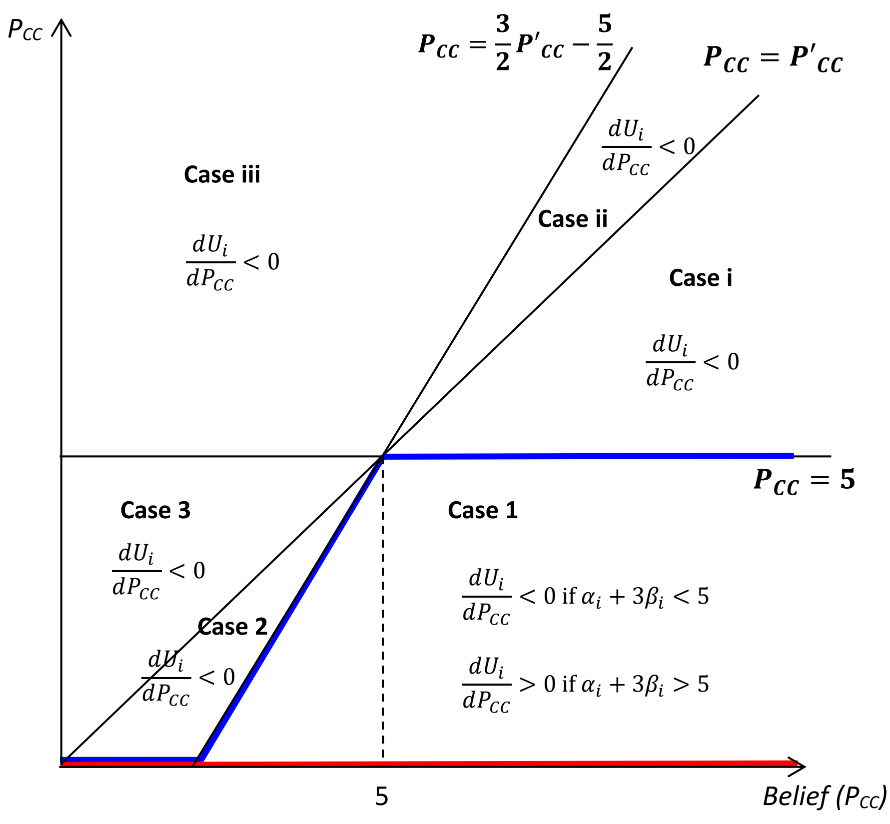
Figure C.1: The sign of $\frac{dU_i}{dP_{CC}}$ and the optimal punishment schedule of third party player *i*



Notes: The blue connected line indicates *i*'s optimal punishment schedule if $.368\alpha_i + 1.168\beta_i > 2$. The red line indicates *i*'s optimal punishment schedule if $.368\alpha_i + 1.168\beta_i < 2$. Note that $.368\alpha_i + 1.168\beta_i > 2$ holds if (α, β) = (4, 0.6).

I also conducted a theoretical analysis regarding *i*'s punishment behaviors in Scenario CC, assuming that *i* is concerned about income inequality with the five players (two PD players in *i*'s groups, and *j* and two PD players in the *j*'s group) on condition that the same Scenario CC is realized in the other group (equivalent to the situation where $y = 1$, $z = 0$ and $w = 0$). This analysis was conducted because *i* may care particularly about *j*'s punishment decision and its consequence in the same scenario that *i* is in. This additional simulation generates a qualitatively similar result (*i*'s punishment in Scenario CC is positively correlated with her belief on *j*'s punishment in Scenario CC; and *i* refrains from such anti-social punishment when *j* does not engage in anti-social punishment). See Figure C.2. The detailed calculations are omitted to conserve the space.

Figure C.2: Sign of $\frac{dU_i}{dP_{CC}}$ and optimal punishment schedule of third party player *i* when $y = 1$



Notes: The blue connected line indicates *i*'s optimal punishment schedule if $\alpha_i + 3\beta_i > 5$. The red line indicate *i*'s optimal punishment schedule if $\alpha_i + 3\beta_i < 5$.

(b) *i*'s Punishment Behavior in Scenarios DC and CD

As in the analysis in Section C.2(a), we can derive *i*'s optimal punishment schedule based on the expected utility maximization. As before, I use *y*, *z*, and *w* ($y + z + w = 1$) to refer to third party player *i*'s beliefs regarding the probabilities that CC, DC (CD) and DD, respectively, are realized. I denote *i*'s punishment points imposed on a defector (cooperator) in Scenario DC (CD) as P_{DC} (P_{CD}). In this analysis, I will impose the following assumptions:

Assumption B: (i) $P'_{CD} = 0$; (ii) $P'_{DC} \leq 7.5$; (iii) $P'_{DD} = 0$; (iv) $P'_{CC} \leq 5$.

As in Assumption A, Assumption B consists of reasonable conditions. For simplicity, we set $P_{CD} = 0$. Notice that the cooperator in Scenario CD receives the lowest payoff (15 points) in *i*'s group.

Under these setups, *i*'s expected utility can be written as follows:

$$E[U_i(x)] = y \cdot [U_i(x)]_{\text{CC in } j'\text{s group}} + z \cdot [U_i(x)]_{\text{DC(CD) in } j'\text{s group}} + w \cdot [U_i(x)]_{\text{DD in } j'\text{s group}}$$

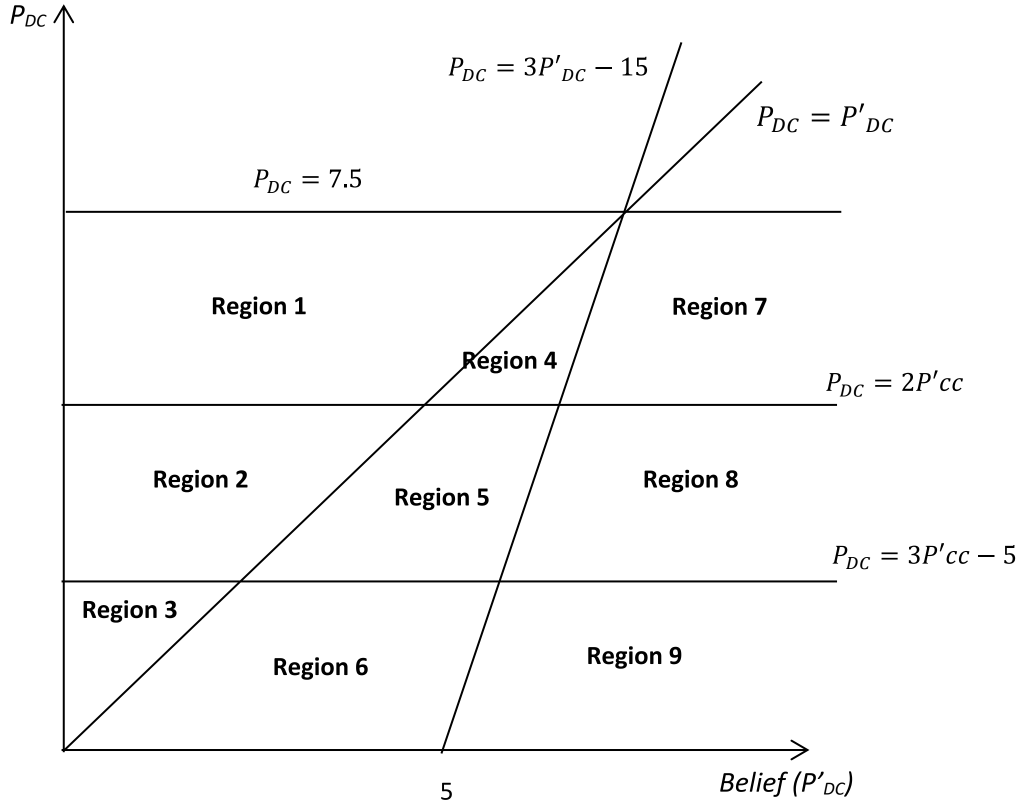
$$= (40 - P_{DC}) - \frac{\alpha_i}{5}(15 - 2P_{DC}) - \frac{\beta_i}{5}(25 - P_{DC})$$

$$-z \cdot \Big[\frac{\alpha_i}{5}\max\{P_{DC} - P'_{DC}, 0\} + \frac{\beta_i}{5}\max\{-P_{DC} + P'_{DC}, 0\} + \frac{\alpha_i}{5}\max\{15 + P_{DC} - 3P'_{DC}, 0\} +$$
$$\frac{\beta_i}{5}\max\{-15 - P_{DC} + 3P'_{DC}, 0\} + \frac{\beta_i}{5}(25 - P_{DC})\Big]$$

$$-y \cdot \Big[\frac{\alpha_i}{5}\max\{-2P'_{CC} + P_{DC}, 0\} + \frac{\beta_i}{5}\max\{2P'_{CC} - P_{DC}, 0\} + \frac{2\alpha_i}{5}\max\{5 + P_{DC} - 3P'_{CC}, 0\} +$$
$$\frac{2\beta_i}{5}\max\{-5 - P_{DC} + 3P'_{CC}, 0\}\Big]$$

$$-w \cdot \Big[\frac{\alpha_i}{5}P_{DC} + \frac{2\beta_i}{5}(15 - P_{DC})\Big].$$

The form of this expected utility function suggests that we need to consider nine regions (see the figure on the next page). The derivatives (dU_i/dP_{DC}) are summarized in the table on next page.



	dU_i/dP_{DC}
Region 1	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{2\alpha_i}{5}-\frac{\beta_i}{5}\right]-y\frac{3\alpha_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 2	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{2\alpha_i}{5}-\frac{\beta_i}{5}\right]-y\left[-\frac{\beta_i}{5}+\frac{2\alpha_i}{5}\right]-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 3	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{2\alpha_i}{5}-\frac{\beta_i}{5}\right]+y\frac{3\beta_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 4	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]-y\frac{3\alpha_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 5	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]-y\left[-\frac{\beta_i}{5}+\frac{2\alpha_i}{5}\right]-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 6	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}-z\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]+y\frac{3\beta_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 7	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}+z\frac{3\beta_i}{5}-y\frac{3\alpha_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$

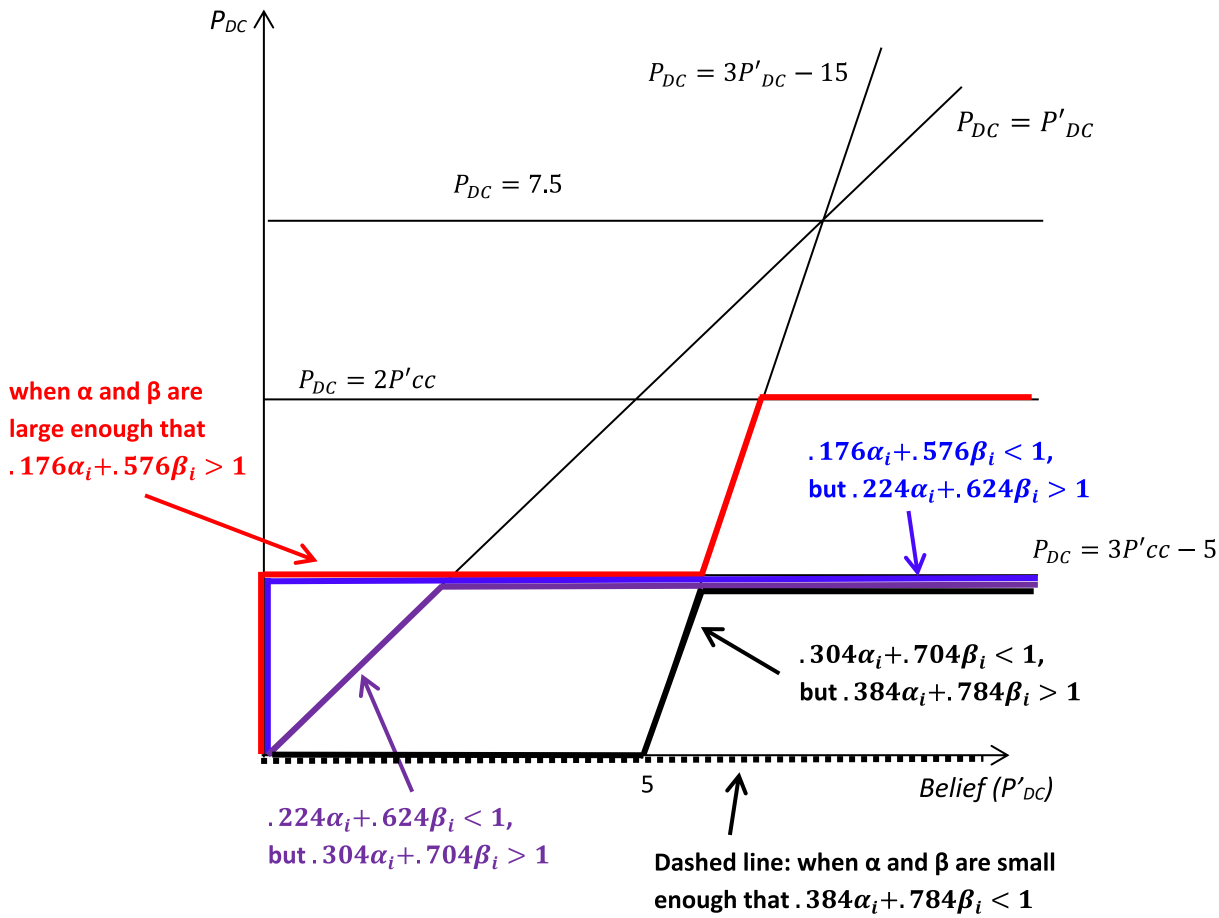
Region 8	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}+z\frac{3\beta_i}{5}-y\left[-\frac{\beta_i}{5}+\frac{2\alpha_i}{5}\right]-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$
Region 9	$-1+\frac{2\alpha_i}{5}+\frac{\beta_i}{5}+z\frac{3\beta_i}{5}+y\frac{3\beta_i}{5}-w\left[\frac{\alpha_i}{5}-\frac{2\beta_i}{5}\right]$

As in the analysis in Section C.2(a), I assume that subjects' beliefs are correct, and use: y = .52, z =.40, and w =.08 for this simulation. Then, the derivatives dU_i/dP_{CC} are numerically calculated as in the following table:

	dU_i/dP_{DC}	Sign of dU_i/dP_{DC}
Region 1	$-1-.088\alpha_i+.312\beta_i$	$dU_i/dP_{DC} < 0$
Region 2	$-1+.016\alpha_i+.416\beta_i$	$dU_i/dP_{DC} < 0$
Region 3	$-1+.224\alpha_i+.624\beta_i$	$dU_i/dP_{DC} < 0$ if $.224\alpha_i+.624\beta_i<1$ $dU_i/dP_{DC} > 0$ if $.224\alpha_i+.624\beta_i>1$
Region 4	$-1-.008\alpha_i+.392\beta_i$	$dU_i/dP_{DC} < 0$
Region 5	$-1+.096\alpha_i+.496\beta_i$	$dU_i/dP_{DC} < 0$
Region 6	$-1+.304\alpha_i+.704\beta_i$	$dU_i/dP_{DC} < 0$ if $.304\alpha_i+.704\beta_i<1$ $dU_i/dP_{DC} > 0$ if $.304\alpha_i+.704\beta_i>1$
Region 7	$-1+.072\alpha_i+.472\beta_i$	$dU_i/dP_{DC} < 0$
Region 8	$-1+.176\alpha_i+.576\beta_i$	$dU_i/dP_{DC} < 0$ if $.176\alpha_i+.576\beta_i<1$ $dU_i/dP_{DC} > 0$ if $.176\alpha_i+.576\beta_i>1$
Region 9	$-1+.384\alpha_i+.784\beta_i$	$dU_i/dP_{DC} < 0$ if $.384\alpha_i+.784\beta_i<1$ $dU_i/dP_{DC} > 0$ if $.384\alpha_i+.784\beta_i>1$

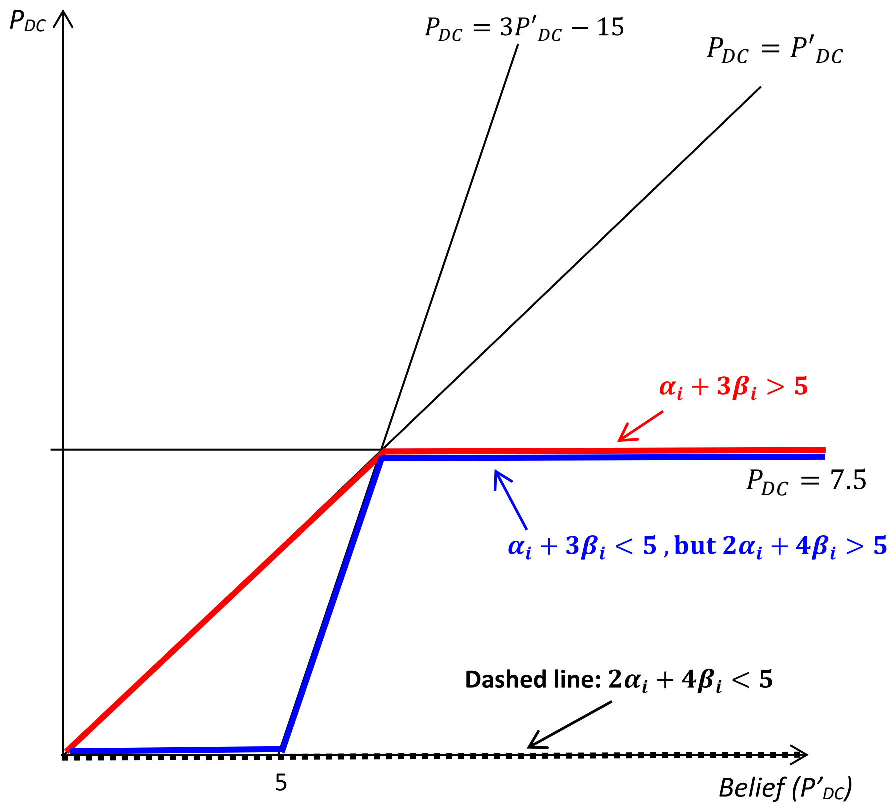
Based on these calculations, we can find *i*'s optimal punishment schedule as in Figure C.3 (next page):

Figure C.3: Sign of $\frac{dU_i}{dP_{DC}}$ and optimal punishment schedule of third party player *i*



I also conducted a theoretical analysis regarding *i*'s punishment behaviors in Scenario DC, assuming that *i* is concerned about income inequality with the five players (two PD players in *i*'s group, and *j* and two PD players in the *j*'s group) on condition that the same Scenarios DC and CD are realized in the other group (equivalent to the situation where $y = 0$, $z = 1$ and $w = 0$). This analysis was performed because *i* may care particularly about *j*'s punishment decision and its consequence in the same scenarios which *i* is dealing with. This additional simulation generates qualitatively a similar result (*i*'s punishment of a defector in Scenario DC is positively correlated with her belief on *j*'s punishment in Scenario DC). See Figure C.4. The detailed calculations are omitted to conserve the space.

Figure C.4: Sign of $\frac{dU_i}{dP_{DC}}$ and optimal punishment schedule of third party player *i* when z = 1



Regardless of the assumption on beliefs, although we observe the positive relationship between *i*'s punishment of a defector and *i*'s belief regarding *j*'s punishment of the defector in *j*'s group (Figures C.3 and C.4), the punishment strength is on average smaller in the Visibility than in the Standard treatment (Section C.1(b)). Thus, the Fehr-Schmidt model does not support Hypothesis 2 in the paper. One reason is that when *i* increases her punishment of a defector in Scenario DC aiming to decrease inequality with *j*, *i*'s disutility due to the inequality with others in the *j*'s group could increase (for example, the payoff between *i* and the defector in *j*'s group under Scenario DC widens). Note that *i*'s behindness aversion is at least as strong as her aheadness aversion ($\alpha_i \geq \beta_i$). Another reason is that third party player *i* in the Visibility treatment makes punishment decisions, assuming that *j* confronts with Scenario DC only with probability *z* (*j* may confront with Scenario CC or DD) – see the analysis above and Figure C.3.

References:

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

Fehr, Ernst and Klaus Schmidt, 2010. On Inequity Aversion: A Reply to Binmore and Shaked. *Journal of Economic Behavior & Organization* 73: 101-108.