



KOCHI UNIVERSITY OF TECHNOLOGY

Social Design Engineering Series

SDES-2019-6

Second thoughts of social dilemma in mechanism design

Tatsuyoshi Saijo

Research Institute for Humanity and Nature

Research Institute for Future Design, Kochi University of Technology

School of Economics and Management, Kochi University of Technology

Tokyo Foundation for Policy Research

13th June, 2019

School of Economics and Management

Research Institute for Future Design

Kochi University of Technology

KUT-SDE working papers are preliminary research documents published by the School of Economics and Management jointly with the Research Center for Social Design Engineering at Kochi University of Technology. To facilitate prompt distribution, they have not been formally reviewed and edited. They are circulated in order to stimulate discussion and critical comment and may be revised. The views and interpretations expressed in these papers are those of the author(s). It is expected that most working papers will be published in some other form.

Second Thoughts of Social Dilemma in Mechanism Design

Tatsuyoshi Saijo*

November 2018

Research Institute for Humanity and Nature, Kyoto, Japan and
Research Institute for Future Design, Kochi University of Technology, Kochi, Japan

Abstract

This paper shows that second thoughts are not an innocent device in our daily life, but is human wisdom that plays an important role in resolving problems such as social dilemmas. We design a simple mechanism to achieve Pareto efficiency in social dilemmas, and then compare the performance of this mechanism with and without second thoughts. First, second thoughts change the payoff structure of the game in favor of cooperation. Second, this mechanism is robust even when players deviate from a payoff maximizing behavior.

Keywords: second thoughts, subgame perfection, social dilemma, cooperation, mechanism design

JEL classification: C72, C92, D74

* The author thanks Yoshitaka Oakano for his helpful comments and suggestions. This research was supported by Scientific Research A (24243028 and 17H00980) and Challenging Exploratory Research (16K13354) of the Japan Society for the Promotion of Science; the Research Institute for Humanity and Nature (RIHN Project Number 14200122); and “Experimental Social Sciences: Toward Experimentally-based New Social Sciences for the 21st Century,” a project funded by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture of Japan.

1. Introduction

Why have we been using second thoughts? The second thoughts here refer to giving a player a chance to change the strategy in decision making after observing the strategies of others in a sequential game. We will show that second thoughts are not an innocent device in our daily life, but are human wisdom that plays an important role in resolving problems such as social dilemmas.

Consider a social dilemma game such as a prisoner's dilemma game. Saijo, Masuda and Yamakawa (2018) designed a two stage mechanism called the approval mechanism to implement Pareto efficient outcome when the number of players is two.¹ After having played a usual prisoner's dilemma, players can approve or reject the other's choice of cooperation (*C*) or defection (*D*) in the next approval stage. If both players approve the other's choice, the outcome is the result of the chosen strategies. However, if either rejects the other's choice, the outcome is the same as if they had mutually defected from the prisoner's dilemma. In theory, such an approval mechanism implements cooperation in elimination of weakly dominated strategies (*EWDS*), although this is not the case in the subgame perfect *Nash* equilibrium. They then showed that it works well even from early periods experimentally. They also found that subjects understood the subgame perfection part well, and used *EWDS* instead of *Nash* equilibria at each subgame.

In implementation literature such as Hurwicz and Schmeidler (1978), Hurwicz (1979) and Maskin (1999), they used *Nash* equilibria as the basic equilibrium concept, and then Moore and Repullo (1988), in their path breaking and influential paper, constructed mechanisms and found conditions on social goals to implement them in subgame perfect *Nash* equilibrium. However, as Fehr, Powell, and Wilkening (2015) recently show, the experimental performance of mechanisms implementing social goals with subgame perfect *Nash* equilibria is rather limited. On the other hand, Saijo et al. (2018) found an *affinity* among the mechanism, subgame perfection and *EWDS*, but not *Nash*.² A basic question is whether this affinity works well beyond two players.

Huang, Masuda, and Saijo (2017) extended the idea of the approval mechanism to include the cases with more than two players in a social dilemma game. In the first stage, each player chooses either *C* or *D*. Knowing the choices in the first stage, all *C* players can change from *C* to *D* in the second stage unless all choose *C*. If a player chooses *D* in the first stage, then the other *C* players will change to *D* in the second stage. Once players

¹ Saijo and Shen (2018) showed that the approval (or mate choice) mechanism works well in a class of quasi-dilemma games including prisoner's dilemma games.

² Varian (1994) constructed a simple mechanism called the compensation mechanism that implements a social goal in subgame perfect *Nash* equilibrium, but the experimental performance is limited as Andreoni and Varian (1999) showed.

understand this logic, no player would take D in the first stage, and hence the mechanism implements cooperation. They conducted a series of experiments and found that the performance of the mechanism is limited in early rounds if the number of players is at least three. In order to overcome this problem, we introduce second thoughts, as a new tool in implementation theory, avoiding complication of the mechanism.

Although second thoughts, allowing players to reconsider their decisions after observing them, have been widely used in our daily life, no theoretic analysis has been done. In the Huang et al. mechanism, we add one stage called the second thought stage between the social dilemma stage and the approval stage. All D players have a chance to change from D to C in the second thought stage unless all choose D in the first stage. After a player chose D in the first stage, the player notices that the other C players will change to D later. Understanding this logic, the D player changes to C in the second thought stage. What we find is that second thoughts in social dilemma works very well theoretically. First, second thoughts change the payoff structure of the game in favor of cooperation. Second, second thoughts make mechanisms robust even when players deviate from *EWDS*.

In the following, we show the two player case in section 2, the three player case in section 2, and then the general case in section 3. Section 4 is for further research.

2. The Simplified Approval Mechanism with Second Thoughts for $n = 2$.

Let $n \geq 2$ be the number of players, and each player has endowment $w > 0$. Each player must choose to contribute either the entire w for the production of a public good y or nothing. The production function of y is linear, namely $y = \alpha mw$ where $1 > \alpha > 1/n$, and m is the number of players who choose cooperation (C) (i.e., those that contribute the entire w). Hence, the payoff of a player who chooses no contribution (defection or D) is $\alpha mw + w = (\alpha m + 1)w$, while the contributor's payoff is αmw . We term a player who chooses C (D) as a C (D) player, respectively.

We consider a mechanism that has a new stage after the *PD* game, due to Huang et al. (2017). If all participants are either C players or D players, the game ends. The payoff of a player in the former case is αnw and that in the latter case is w . If the number of C players is at least one and at most $n-1$, then only C players can proceed to the second stage, in which they have the opportunity to change their decisions from C to D . This mechanism is called the *simplified approval mechanism* or the *SAM* in short. A natural behavioral procedure found in previous experiments on approval mechanisms is *subgame perfect elimination of weakly dominated strategies* (*SPEWDS*), which is also adopted, for example, in Kalai (1981). This requires two properties: (i) subgame perfection and (ii) that players do not choose weakly dominated strategies in each subgame and in the reduced normal form game.

Figure 1 illustrates the case of $n = 2$, $\alpha = 0.7$, and $w = 10$. Players 1 and 2 face a prisoner's dilemma game in the first stage. Knowing that player 2 chose D in the first stage in subgame a , player 1 proceeds to the second stage and faces a choice between C and D . Player 1 chooses D in subgame a since 10 dominates 7 , or $10 > 7$. Similarly, player 2 chooses D in subgame b . Then, as the reduced normal form game on the right hand side of Figure 1 shows, player 1 chooses C after eliminating weakly dominated strategy D . Similarly, player 2 also chooses C in the first stage and hence, (C,C) is the outcome. Hereafter, a strategy profile with parentheses, such as (C,C) , represents the choice in the reduced normal form game and a sequence of choices, such as CDC , shows a strategy path. Huang et al. (2017) showed the following properties of the SAM.

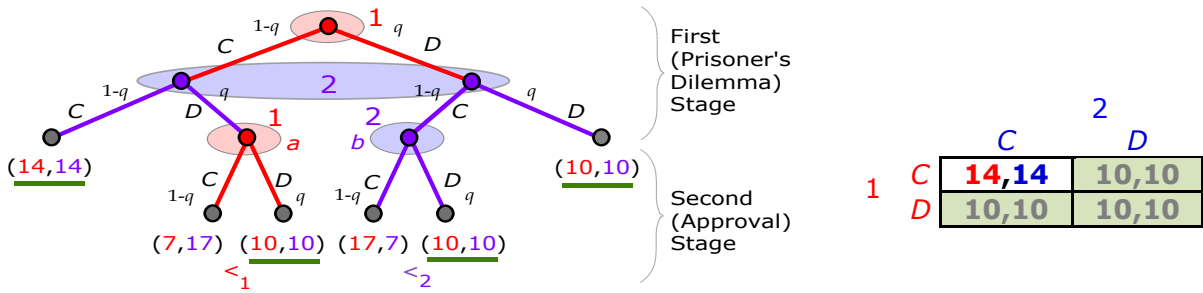


Figure 1: The SAM and its reduced normal form game when $n = 2$.

Proposition 1. (i) *The simplified approval mechanism implements cooperation in SPEWDS and (ii) the simplified approval mechanism cannot implement cooperation in subgame perfect equilibrium (SPE).*

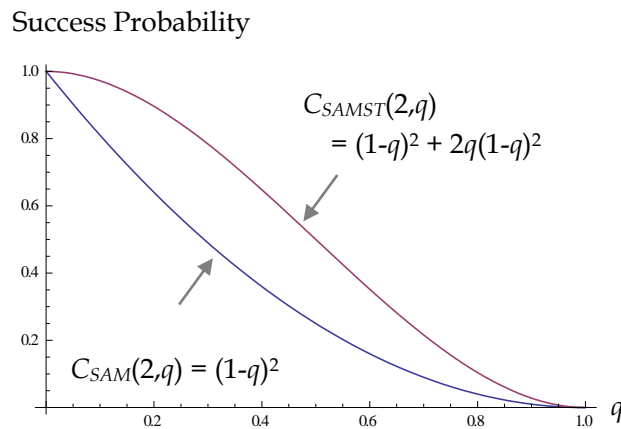


Figure 2: Success probability functions when $n = 2$.

“Cooperation” in Proposition 1 indicates that all players choose C in the reduced normal form game. As the reduced normal form game in Figure 1 shows, (D,D) is also a *subgame perfect equilibrium (SPE)* outcome and hence, the SAM cannot implement cooperation in *SPE*.

Thus far, we have supposed that every player chooses the alternative under *SPEWDS*, but we consider the cases where some player might deviate from it. For simplicity, we assume that every player deviates with a probability of q , with $0 < q < 1$ at each node. As shown in Figure 1, the success probability achieving $(14,14)$ that follows path CC is $(1-q)^2$. Let $C_{SAM}(n,q)$ be the *success probability function* of the SAM, where n is the number of players. Then, $C_{SAM}(n,q) = (1-q)^n$. Figure 2 shows the case of $n = 2$. The horizontal axis displays the probability of deviation and the vertical axis the probability of all players cooperating. Since $\partial C_{SAM}(2,0)/\partial q = -2$ and $\partial C_{SAMST}(2,0,1)/\partial q = 0$, the success probability of the SAM decreases as q rises around zero, whereas the success probability of the SAMST stays at a probability of one as q increases around zero. Next, fix any q . Since $C_{SAMST}(2,q,\cdot)$ is always higher than $C_{SAM}(2,q)$ because of $2q(1-q)^2$ except for $q = 0$ or 1 , the success probability of the SAMST is always better than that of the SAM excluding the end points. That is, the SAMST is relatively robust enough to handle deviation by players.

Huang et al. (2014) conducted experiments of the SAM with each group consisting of three subjects. In total, 63 subjects played the SAM for 15 periods. The groups were formed randomly in each period. The cooperation rates for the first four to seven periods were between 64.9% and 77.7% and they rose above 90% thereafter.³ In order to improve the low cooperation rates in the early rounds in the experiment of Huang et al. (2017), we introduce the one more stage called the second thought stage in the following manner. Every player chooses either C or D in the first stage simultaneously. If all players choose either C or D , the game ends. If the number of D players is at least one and at most $n-1$, then D players have the chance to change from D to C sequentially, knowing all the choices made in the first stage. The order of the choices of D players is determined exogenously, for example, based on the numbering assigned to players. By observing the choices of all D players in the second thought stage, C players can change their choices from C to D simultaneously except for the case when all D players change their choices in the second thought stage. This stage is called the third stage although the second thought stage might have several stages. When all D players change their choices, the game ends and the outcome is that all

³ Huang et al. (2017) used the *ex post* cooperation rate. For example, even though a player chose C in the first stage, this was not counted in the cooperation rate if that player changed his or her decision from C to D in the second stage.

players choose C. We call this the *simplified approval mechanism with second thoughts* (SAMST).

Figure 3 shows an example with $n = 2$, $\alpha = 0.7$, and $w = 10$. Consider subgame *a*. By observing player 1's choice C, player 2 (who has the chance to change his or her choice) must consider player 1's choice in subgame *c*. Since $10 > 7$, namely C is dominated by D, player 1 will choose D in subgame *c*. By understanding this fact, player 2 in subgame *a* thus chooses C since $14 > 10$. Therefore, the outcome in subgame *a* is (14,14), which differs from the outcome of the SAM. By applying the same argument, we have (14,14) in subgame *b*. That is, the outcomes except for that at (D,D) are (14,14), although (14,14) is achieved at (C,C) only in the SAM.

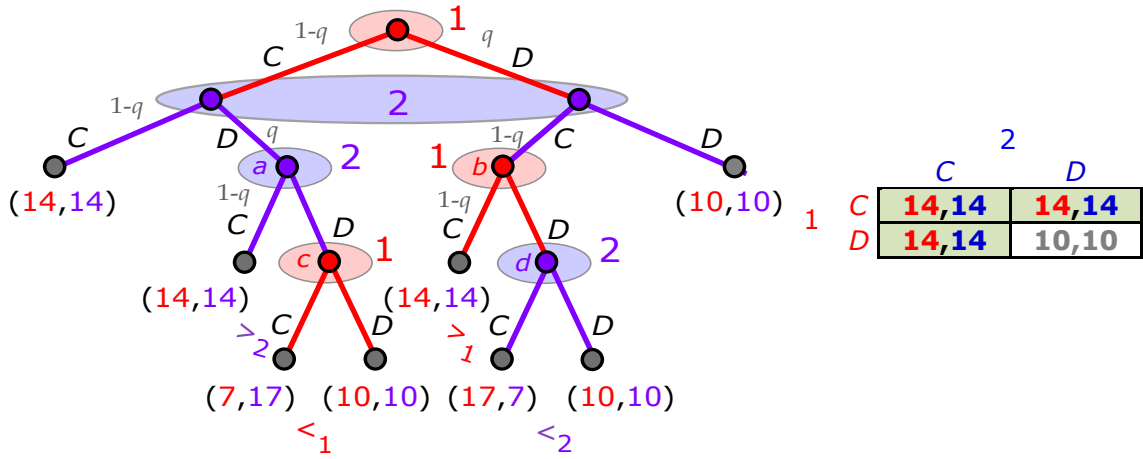


Figure 3: The SAMST and its reduced normal form game when $n = 2$.

When the number of players is two, each player chooses C in the second thought stage and thus the payoff outcome at (C,D) and (D,C) in the reduced normal form game is $(2\alpha w, 2\alpha w)$. Since the payoff outcomes at (C,C) and (D,D) are $(2\alpha w, 2\alpha w)$ and (w, w) , respectively, and $2\alpha w > w$, the SPE strategy profiles are (C,C), (C,D) and (D,C). That is, the SAMST implements cooperation in SPE if $n = 2$.

Consider that players deviate from SPEWDS. In contrast, as Figure 3 shows, three paths, namely CC, CDC, and DCC, achieve (14,14) when we use the SAMST. The probability of the paths CDC and DCC is $(1-q)q(1-q)$ and $q(1-q)(1-q)$, respectively. Hence, $C_{SAMST}(2,q) = (1-q)^2 + 2q(1-q)^2$. Since $\partial C_{SAMST}(2,0)/\partial q = 0$, the success probability of the SAM does not decrease as q rises around zero. As Figure 2 shows, $C_{SAMST}(2,q) > C_{SAM}(2,q)$ for all $q \in (0,1)$.

3. The Simplified Approval Mechanism with Second Thoughts for $n = 3$.

This section illustrates the three player case that basically contains problems that

should be handled for the general case. Figure 4 illustrates the SAMST with $n = 3$, $\alpha = 0.7$ (0.4 or 0.5), and $w = 10$. The bold face numbers show the payoffs with $\alpha = 0.7$ and the numbers in braces show the payoffs with $\alpha = 0.4$ and the numbers in parentheses in the braces show them with $\alpha = 0.5$. Since the entire tree is relatively large, we only show the subgames with CCC, CCD, CDD, and DDD, which are sufficient to understand the entire tree. Consider first the case of $\alpha = 0.7$. Look at subgame *a* where players 1 and 2 chose C, but player 3 chose D. Player 3, who faces the second thought stage, must consider what would happen in subgame *c*. Players who chose C in the first stage face a PD game in subgame *c* and hence, both choose D. In this sense, players who chose C in the first stage can *burden* players who chose D in the second thought stage, although this hurts every player. By understanding this fact, player 3 compares 21 with C and 10 with D. Since C dominates D, player 3 chooses C in subgame *a*. That is, player 3, who chose C at node *a*, can obtain the *bonus* from players 1 and 2, who chose C in the first stage. Therefore, the outcome of subgame CCD is (21,21,21).

Consider next subgame *b* where player 1 chose C but players 2 and 3 chose D. Players 2 and 3 face the second thought stage sequentially. Pay attention to the last nodes or the third stage where player 1 faces the choice between C and D. Although the number of players is one, players who arrive at the nodes face PD games and hence, they always choose D at each node.

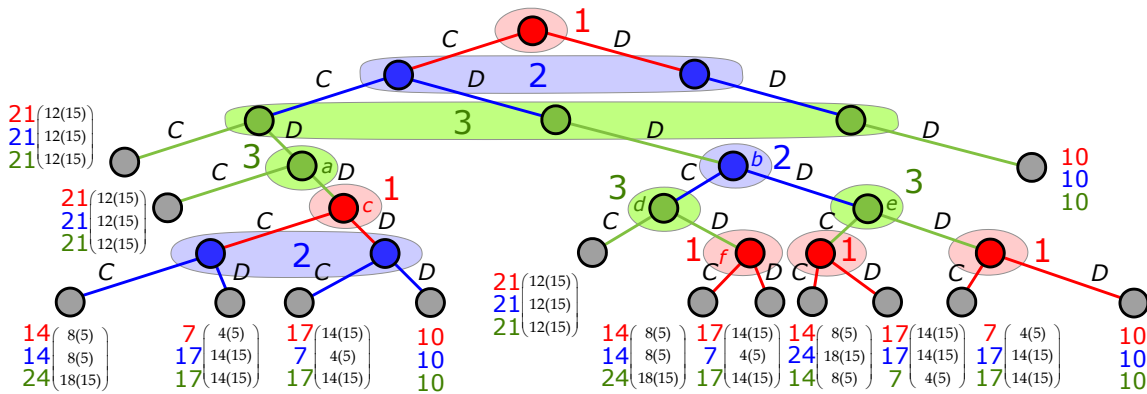


Figure 4. The SAMST when $n = 3$.

Moreover, consider subgame *e* where player 2 did not change his or her choice in subgame *b*. Since player 1 chooses D in the following subgames, player 3 chooses D in subgame *e* and the payoff is 10. Consider subgame *d*. In contrast, player 3, who can take advantage of the bonus effect in subgame *d*, chooses C since player 1 in the following

subgame will choose D if player 3 chooses D . That is, $21 > 17$. Knowing this process, player 2 chooses C since $21 > 10$. Therefore, all payoff outcomes other than (D,D,D) are $(21,21,21)$ and hence, the final outcome is (C,C,C) under *SPEWDS*, as the reduced normal form game in Figure 5-(a) shows. In contrast, as Figure 4-(b) shows, $(21,21,21)$ appears only in subgame CCC under the *SAM*.

The sequentiality of D players is important to implement cooperation.⁴ If nodes d and e were in the same information set, the payoff of player 3 from choosing C in the information set would be $(21,7)$ and the payoff from choosing D would be $(17,10)$; hence, both would survive by using the elimination of weakly dominated strategies.

		P3		P2	
		D	C	D	C
P1	C	(21,21,21)	(21,21,21)	(10,10,10)	(21,21,21)
	D	(21,21,21)	(21,21,21)	(21,21,21)	(21,21,21)

(a) $\alpha = 0.7$ with the *SAMST*.

		P3		P2	
		D	C	D	C
P1	C	(10,10,10)	(10,10,10)	(10,10,10)	(10,10,10)
	D	(21,21,21)	(10,10,10)	(10,10,10)	(10,10,10)

(b) $\alpha = 0.7$ with the *SAM*.

		P3		P2	
		D	C	D	C
P1	C	(12,12,12)	(10,10,10)	(10,10,10)	(12,12,12)
	D	(12,12,12)	(12,12,12)	(12,12,12)	(10,10,10)

(c) $\alpha = 0.4$ with the *SAMST*.

		P3		P2	
		D	C	D	C
P1	C	(10,10,10)	(10,10,10)	(10,10,10)	(10,10,10)
	D	(12,12,12)	(10,10,10)	(10,10,10)	(10,10,10)

(d) $\alpha = 0.4$ with the *SAM*.

		P3		P2	
		D	C	D	C
P1	C	(15,15,15)	(15,15,15)	(10,10,10)	(15,15,15)
	D	(15,15,15)	(15,15,15)	(15,15,15)	(15,15,15)

(15,15,15) DCD
 (5,15,15)
 (10,10,10)

 CDD
 (15,15,15)
 (15,5,15)
 (10,10,10)

 DDC
 (15,15,15)
 (5,15,15)
 (10,10,10)

(e) $\alpha = 0.5$ with the *SAMST*.

Figure 5. The reduced normal form games of the *SAMST* and *SAM* when $n = 3$.

⁴ We thank Xiaochuan Huang for indicating this fact.

Let us next look at the case of $\alpha = 0.4$. Consider node f where player 1 chooses D . Then, player 3 at node d chooses D since $14 > 12$. Knowing this fact, player 2 chooses D since $10 > 4$. That is, the payoff in subgame b becomes $(10,10,10)$. From the viewpoint of player 3, since α is too small, the player cannot take advantage of the bonus effect at node d .

As the reduced normal form game in Figure 5-(c) shows, the payoff outcome with subgames where two players choose C and one player chooses D is $(12,12,12)$ and the payoff outcome with subgames where one player chooses C and two players choose D is $(10,10,10)$, but the final outcome is still (C,C,C) under *SPEWDS*. In contrast, as Figure 5-(d) shows, $(12,12,12)$ appears only in subgame CCC under the *SAM*.

Consider the case when $\alpha = 0.5$. The payoff outcomes at paths $CDDCDD$ and $CDDCC$ in Figure 3 are $(15,5,15)$ and $(15,15,15)$, respectively. If this were the case, player 3 at node d would be indifferent between C and D . That is, both C and D survive by using the elimination of weakly dominated strategies at node d . This influences the decision of player 2 at node b . Figure 6 shows the reduced normal form game at node b excluding player 1.

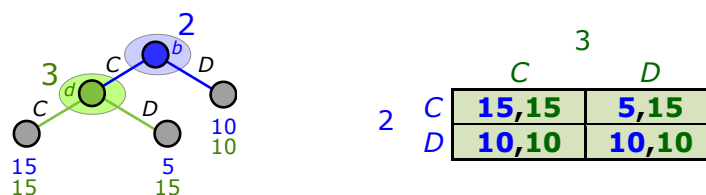


Figure 6. Subgame b and its reduced normal form game at node b .

Note that the payoff at (D,C) in Figure 6 should be $(10,10)$ when player 2 chooses D . That is, player 3's choice does not matter and hence, both C and D survive by using the elimination of weakly dominated strategies for player 2. To sum up, the payoff outcome in subgame b or CDD is $(15,15,15)$, $(15,5,15)$, or $(10,10,10)$. At $(15,5,15)$, player 1 changes from C to D at the third stage knowing that player 3 kept the choice at D . That is, only player who chose C is player 2. Similarly, the payoff outcome in subgame DCD or DDC is $(15,15,15)$, $(5,15,15)$, or $(10,10,10)$. Since the first D player in the second thought stage at DCD and DDC is player 1, the payoff outcome should be $(5,15,15)$ when player 1 changes from D to C at the second thought stage and the rest choose D . That is, there are $3^3=27$ reduced normal form games when $\alpha = 0.5$. Look at Figure 5-(e). Since there are three possible payoff outcomes, for example, at CDD , we write these payoff outcomes under CDD . Since one of them should be chosen at a reduced normal form game, we chose $(10,10,10)$ in Figure 5-(e). Similarly, three possible payoff outcomes at the left hand side of DCD , and at the right hand side of DDC .

Consider player 1. There is a case where both C and D survive by using the elimination of weakly dominated strategies for player 1: $(10,10,10)$ at CDD and $(15,15,15)$ at DCD and DDC . Then, both players 2 and 3 choose C and hence, the *SPEWDS* strategy profile of the reduced normal form game is (C,C,C) or (D,C,C) . Although player 1 chooses D at (D,C,C) , the player will change from D to C in the second thought stage; hence, the payoff outcome is $(15,15,15)$. The real problem is that player 1 cannot tell which reduced normal form game player 1 faces and hence, both C and D survive when player 1 must make a decision in the first stage.

Note also that the payoff at C and the payoff at D of player 1 are the same for all possible choices of players 2 and 3 in Figure 5-(e). In other words, player 1 cannot distinguish between C and D . We say that strategies A and B of a player are *indistinguishable* if the payoffs at A and B are the same for all possible choices of the other players.

Consider all possible reduced normal form games at each node. Strategy A *weakly rules* strategy B if A weakly dominates B in some reduced normal form game and strategies A and B are indistinguishable in the remainder of the reduced normal form games. Then, we can define a refinement of *SPEWDS* by using the *weak rule* instead of the weak dominance in the definition of *SPEWDS*, which we denote it backward elimination of weakly ruled strategies (*SPEWRS*).

Look at (c) in Figure 5 and consider (D,D,D) . This is also an *SPE* strategy profile, and hence, the *SAMST* cannot implement cooperation in *SPE* if $n = 3$. If $n > 3$, it is easy to find similar examples by choosing α to satisfy $1/(n-1) > \alpha$.

4. The Simplified Approval Mechanism with Second Thoughts

The following proposition shows that the *SAMST* implements cooperation in *SPEWDS* or *SPEWRS*.

Proposition 2. (i) If $\alpha \notin \{1/(n-1), 1/(n-2), \dots, 1/2\}$, the *SAMST* implements cooperation in *SPEWDS* and (ii) the *SAMST* implements cooperation in *SPEWRS*.

Proof. See Appendix 1.

Consider the meaning of inequality $\alpha > 1/(m+1)$, i.e., $m > (1/\alpha) - 1$. If $\alpha = 0.7$, $(1/\alpha) - 1 = 3/7$. That is, the minimum number \underline{m} of the C players in the first stage where the “bonus” effect is activated is at least one. If $\alpha = 0.4$, $\underline{m}(\alpha) = 2$, which shows that the cooperative outcome in subgame b in Figure 3 cannot be realized since only one C player is

in the first stage. In contrast, $\bar{l}(\alpha) = n - \underline{m}(\alpha)$ is the maximum number of D players in the first stage, thus leading to the cooperative outcome. The proof of Proposition 2 shows the following corollary.

Corollary 1. *If there is an indistinguishable player in a reduced normal form game at the beginning node, no other players are indistinguishable.*

Suppose that $(1/\alpha) - 1$ is an integer. Then, there are $\bar{l} + 1$ possible payoff outcome profiles in a subgame in the second thought stage. The total number of subgames in the second thought stage where the number of D players is \bar{l} is ${}_n C_{\bar{l}}$, and hence, the total number of all possible reduced normal form games at the beginning node is $(\bar{l} + 1)^n C_{\bar{l}}$, where ${}_n C_k$ is the number of k combinations from n players. Among these subgames, each player faces just one reduced form game in which C and D are indifferent. As Figure 5-(e) shows, the total number of all possible reduced normal form games at the node is $(2+1)^3=27$ when $n = 3$ and $\underline{m} = 1$. If $n = 5$ and $\underline{m} = 2$, it is 4^{10} . If this were the case, α must be $1/3$ and the chance of a player being indistinguishable would be $1/4^{10}$.

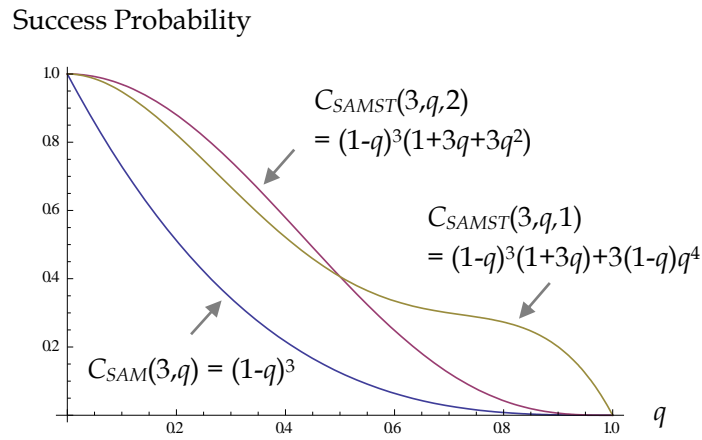


Figure 7. Success probability functions when $n = 3$.

Consider the case of $n = 3$. Although C_{SAMST} with $n = 2$ does not depend on \bar{l} , C_{SAMST} with at least three players depends on \bar{l} , which is determined by α . That is, C_{SAMST} is a function of n , q , and \bar{l} and thus, we write it as $C_{SAMST}(n, q, \bar{l})$. Consider the case of $\alpha = 0.4$. Then, $\bar{l}(0.4) = 1$ and there are two types of success paths. The first one is the success paths up to \bar{l} . These are CCC , $CCDC$, $CDCC$, and $DCCC$ and their probabilities are $(1-q)^3$, $(1-q)^2q(1-q)$, $(1-q)q(1-q)^2$, and $q(1-q)^3$, respectively. The second one is the success paths beyond \bar{l} . Look at node b in Figure 4. Although player 2 should choose D when $\alpha = 0.4$, the

player might choose C because of deviation. If player 3 also chooses C after player 2's choice induced by deviation, the path CDDCC is also a success path that has a probability of $(1-q)q^4$. Since there are two other paths of this kind, $C_{SAMST}(3,q,1) = (1-q)^3(1+3q)+3(1-q)q^4$. In contrast, if $\alpha = 0.7$, the probability of CDDCC is $(1-q)^3q^2$. That is, since $\bar{l}(0.7) = 2$, deviation in the second thought stage must lead players to choose D. Therefore, $C_{SAMST}(3,q,2) = (1-q)^3(1+3q+3q^2)$. Figure 7 shows this case. In order to avoid the indeterminacy case, let us assume $\alpha \notin \{1/(n-1), 1/(n-2), \dots, 1/2\}$. Then, by summarizing the above argument, we have $C_{SAM}(n,q) = (1-q)^n$, and $C_{SAMST}(n,q,\bar{l}) = (1-q)^n \sum_{k=0}^{\bar{l}} \binom{n}{k} q^k + \sum_{k=\bar{l}+1}^{n-1} \binom{n}{k} (1-q)^{n-k} q^{2k}$.

Proposition 3. (i) $\partial C_{SAM}(n,0)/\partial q = -n$ and $\partial C_{SAMST}(n,0,\bar{l})/\partial q = 0$ for all \bar{l} ; and (ii) for any $1 \leq l \leq n-1$, $C_{SAMST}(n,q,l) > C_{SAM}(n,q)$ on $(0,1)$.

Proof. See appendix 2.

The fact that $C_{SAMST}(n,q,\bar{l}) > C_{SAM}(n,q)$ on $(0,1)$ shows that the SAMST is always better than the SAM with respect to the success probability of cooperation. Since $\bar{l}(\alpha)$ is a non-decreasing function, roughly speaking, the success probability increases as α rises.

5. Concluding Remarks

Second thoughts are a powerful tool in implementation theory. They change the payoff structure of the game in favor of cooperation. Furthermore, the mechanism with second thoughts is robust even when players deviate from EWDS.

Our approach is different from the trend in implementation theory where finding some conditions on social goals such as social choice correspondences is a major research goal. Instead, we fix the social goal as Pareto efficiency in social dilemma, then construct a mechanism incorporating social dilemma. That is, introducing second thoughts in mechanisms implementing a social goal in some equilibrium concept is still an open question.

The validity of second thoughts should be confirmed in experiments. Although it is still an early stage, we started confirming that second thoughts make subjects cooperative in even early rounds in experiments.

References

- Andreoni, J., & Varian, H. R. (1999). Preplay Contracting in the Prisoners' Dilemma. *Proceedings of the National Academy of Sciences*, 96(19): 10933-8.
- Fehr, E., Powell, M., & Wilkening, T. (2015). Behavioral limitations of subgame-perfect implementation.
- Huang, X., Masuda, T., & Saijo, T. (2017). "Cooperation among behaviorally heterogeneous players in social dilemma with stay or leave decisions." University of Arizona Department of Economics working paper series 2017-16.
- Hurwicz, L. (1979). Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *The Review of Economic Studies*, 46(2), 217-225.
- Hurwicz, L., & Schmeidler, D. (1978). Construction of outcome functions guaranteeing existence and Pareto optimality of Nash equilibria. *Econometrica: Journal of the Econometric Society*, 1447-1474.
- Kalai, E. (1981). Preplay negotiations and the prisoner's dilemma. *Mathematical Social Sciences*, 1(4), 375-379.
- Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, 66(1), 23-38.
- Masuda, T., Okano, Y., & Saijo, T. (2014). The minimum approval mechanism implements the efficient public good allocation theoretically and experimentally. *Games and Economic Behavior*, 83, 73-85.
- Varian, H. R. (1994). A Solution to the Problem of Externalities When Agents Are Well-Informed. *American Economic Review*, 84(5): 1278-93.
- Saijo, T., Masuda, T. & Yamakawa, T. (2018). "Approval mechanism to solve prisoner's dilemma: comparison with Varian's compensation mechanism." *Social Choice and Welfare*, published on line.
- Saijo T., & Shen, J. (2018). "Mate choice mechanism for solving a quasi-dilemma," *Journal of Behavioral and Experimental Economics*, 72: 1-8.

Appendix 1: The proof of proposition 2

(i) Let m and l be the numbers of C and D players in the first stage, respectively. If $m = n$, the outcome is $(\alpha nw, \alpha nw, \dots, \alpha nw)$. If $l = n$, the outcome is (w, w, \dots, w) .

Suppose $1 \leq l < n$. Consider the choice of players who chose C in the first stage after observing the choices of D players in the second thought stage. Let $0 \leq l' < l$ be the number of D players who change their choices from D to C in the second thought stage. Since $\alpha(m'+l'+1)w < \alpha(m'+l')w + w$ for all $1 \leq m' \leq m-1$, where m' is the number of C players in the first stage who remains to choose C in the subgame after the second thought stage, D is better than C for any C player in the third stage after observing the choices in the second thought stage. That is, all players who chose C in the first stage choose D after the second thought stage.

Consider any strategy path on which at least one D player chose D again in the second thought stage. If this were the case, every C player after the second thought stage would choose D . Keeping this fact in mind, let us choose the youngest D player (e.g., by names or numbers assigned to players) who chose D in the second thought stage. Then, the subgame after the choice of the youngest D player is a sequential social dilemma game and hence, every D player after the choice chooses D .

We now identify the payoff outcome of every subgame constructed by the end nodes in the first stage. Choose any subgame except for the cases where all players chose C or all players chose D in the first stage. Suppose that every D player except for the last one changed his or her choice from D to C in the second thought stage. Consider next the choice of the last D player. If the player chooses C , the payoff is αnw , whereas if the player chooses D , the payoff is $\alpha(l-1)w + w$. Since $\alpha nw - \alpha(l-1)w - w = \alpha\{n-(l-1)\}w - w$ and $n-(l-1) = m+1$,

if $\alpha > 1/(m+1)$, then the last D player chooses C ;

if $\alpha = 1/(m+1)$, then the last D player is indifferent between C and D ; and

if $\alpha < 1/(m+1)$, then the last D player chooses D .

Suppose $\alpha > 1/(m+1)$. If the penultimate player chooses D , then the payoff is $\alpha(l-2)w + w$ since the last D player chooses D in the second thought stage and every C player in the first stage chooses D in the third stage. If the player chooses C , then it is αnw since the last player chooses C . Since $\alpha nw - \{\alpha(l-2)w + w\} = \{\alpha(n-l+2)-1\}w = \{\alpha(m+2)-1\}w > 0$, the player chooses C . Since $\alpha nw - \{\alpha(l-2)w + w\} > 0$, $\alpha nw - \{\alpha(l-k)w + w\} > 0$ for all $2 \leq k \leq l$. That is, the k -th player to last chooses C and hence, all D players choose C in the second thought stage and the payoff outcome is $(\alpha nw, \dots, \alpha nw)$.

Suppose $\alpha < 1/(m+1)$. Then, the last D player chooses D and hence, the payoff of the penultimate player is $\alpha(l-1)w$ if the player chooses C . If the player chooses D , then the payoff is $\alpha(l-2)w + w$. Since $\alpha(l-2)w + w - \alpha(l-1)w = (1-\alpha)w > 0$, the player chooses D . Since $\alpha(l-k)w + w - \alpha(l-k+1)w = (1-\alpha)w > 0$ for all $2 \leq k \leq l$, the k -th player to last chooses D and hence, no D players in the first stage change their decisions in the second thought stage and the payoff outcome is (w, \dots, w) .

Take any α satisfying $1/n < \alpha < 1$ and $\alpha \notin \{1/(n-1), 1/(n-2), \dots, 1/2\}$. Consider the case of $\alpha > 1/2$. Then, $\alpha > 1/2 \geq 1/(m+1)$ for all $m \geq 1$ and hence, the payoff outcome of every subgame other than (D, D, \dots, D) is $(\alpha nw, \dots, \alpha nw)$: without loss of generality, consider player 1. The payoff in subgame (C, D, D, \dots, D) is αnw and that in subgame (D, D, \dots, D) is w . Since $\alpha nw > w$, C is better than D . Since $\alpha > 1/(m+1)$ for all $m \geq 1$, the outcome of the two subgames (C, \cdot) and (D, \cdot) is (C, C, \dots, C) where “.” shows that at least one player’s choice is C . That is, player 1 is indifferent between the outcomes of subgames (C, \cdot) and (D, \cdot) . Therefore, C weakly dominates D for all players and hence, (C, C, \dots, C) is the *SPEWDS* outcome.

Consider next the case of $1/2 \geq 1/(k+1) > \alpha > 1/(k+2) \geq 1/n$. Consider player 1. Let “.” indicate that the number of C is k . Then, the payoff in subgame (C, \cdot) is αnw since $\alpha > 1/\{(k+1)+1\}$ and that in subgame (D, \cdot) is w since $1/(k+1) > \alpha$. That is, C is better than D . Since the outcome of the two subgames (C, \cdot) and (D, \cdot) is the same where “.” indicates that the number of C is not k , C weakly dominates D for all players and hence, (C, C, \dots, C) is the *SPEWDS* outcome.

Thus, if $\alpha \notin \{1/(n-1), 1/(n-2), \dots, 1/2\}$, the *SAMST* implements cooperation in *SPEWDS*.

(ii) Suppose $\alpha = 1/(m+1)$. Then, the last D player is indifferent between C and D since $\alpha nw = \alpha(l-1)w + w$. Suppose that the penultimate player chooses C . Then, the payoff of the penultimate player is αnw if the last D player chooses C and is $\alpha(l-1)w$ if the last D player chooses D . If the penultimate player chooses D , then the payoff is $\alpha(l-2)w + w$. Since $\alpha nw - \{\alpha(l-2)w + w\} = \alpha w > 0$, $\alpha nw > \alpha(l-2)w + w > \alpha(l-1)w$. That is, both C and D survive by using the elimination of weakly dominated strategies. Since $\alpha nw > \alpha(l-k-1)w + w > \alpha(l-k)w$ for all $k=1, \dots, l-1$, both C and D survive by using the elimination of weakly dominated strategies for all D players.

Let $\underline{m}(\alpha) = \lceil (1/\alpha) - 1 \rceil$ where $\lceil a \rceil$ is the smallest integer not less than a . Since $1/n < \alpha < 1$, $1 \leq \underline{m}(\alpha) \leq n-2$. Suppose that $(1/\alpha) - 1$ is an integer. Then, $\underline{m} = (1/\alpha) - 1$. The following case shows that there exists a player who is indifferent between C and D when the number of C players is \underline{m} or $\underline{m} - 1$. Consider two cases:

Case 1: Suppose that the number of C players is \underline{m} . Choose any player who is not a member of the C players. If the player chooses C , the payoff outcome is αnw . If the player chooses D ,

the maximum possible payoff is that all D players other than the player change from D to C and the player is the last D player since all C players change from C to D after the second thought stage. Then, the payoff is $\alpha(\bar{l}-1)w+w$ and hence, $\alpha n w - \{\alpha(\bar{l}-1)w+w\} = \{\alpha n - \alpha(\bar{l}-1)-1\}w = \{\alpha(\underline{m}+1)-1\}w=0$, where $\bar{l} = n - \underline{m}$ and $\bar{l} \geq 2$ since $n \geq \underline{m}+2$. That is, the payoff of C is the same as the payoff of D for the player.

Case 2: Suppose that the number of C players is $\underline{m}-1$. Choose any player who is not a member of the C players. If the player chooses C in the first stage, we show that the payoff outcome should be at least w . Since $\bar{l} \geq 2$, there must be at least one D player. If all D players change from D to C in the second thought stage, the C player obtains $\alpha n w$. If at least one D player chooses D in the second thought stage, the C player obtains at least w by changing from C to D after the second thought stage. If the player chooses D , the payoff is w . That is, the payoff of C can be the same as the payoff of D for the player.

Thus, there is a possibility that C and D are indistinguishable for some players. Let player 1 be such a player and suppose that the first \underline{m} players choose C . Then, since C and D are indistinguishable,

$$\begin{aligned} & \text{the payoff outcome of subgame } (\underbrace{C, C, \dots, C}_m; \underbrace{D, \dots, D}_{\bar{l}}) \\ & = \text{the payoff outcome of subgame } (D, \underbrace{C, \dots, C}_{\underline{m}-1}; \underbrace{D, \dots, D}_{\bar{l}}). \end{aligned}$$

Since the payoff outcome of the latter should be (w, w, \dots, w) , each of the last \bar{l} players in the former can obtain $\alpha n w$ by changing from D to C . That is, C weakly dominates D for the last \bar{l} players.

In contrast, compare the payoff outcome of subgame $(D, \underbrace{C, \dots, C}_m; \underbrace{D, \dots, D}_{\bar{l}-1})$ with the payoff outcome of subgame $(\underbrace{C, C, \dots, C}_{\underline{m}+1}; \underbrace{D, \dots, D}_{\bar{l}-1})$. The latter payoff outcome should be $(\alpha n w, \dots, \alpha n w)$ and hence, the payoff of player 1 should be $\alpha n w$. Since player 1 in the former should obtain $\alpha n w$, which is more than w , at least one player changes from D to C in the second thought stage, and hence, all C players who change from C to D after the second thought stage should obtain strictly more than w . Then, each of the same \underline{m} players obtains w by changing from C to D . That is, C weakly dominates D for the \underline{m} players. Thus, C and D are indistinguishable for player 1 and C weakly dominates D for the rest.

Suppose that C and D are indistinguishable for player 1. Then, there exists another reduced normal form game in the first stage where C weakly dominates D for player 1.

Since C and D are indistinguishable for player 1, the payoff of player 1 in subgame $(C, C, \dots, C; D, \dots, D)$ is either $\alpha n w$ or w . Since the payoff outcome in this subgame can be either $(\alpha n w, \dots, \alpha n w)$ or (w, \dots, w) , there is another reduced normal form game where C weakly dominates D for player 1.

Since the choice of a player who faces indistinguishability is arbitrary, C weakly rules D for all players. That is, the $SAMST$ implements cooperation in $SPEWRS$. ■

Appendix 2. The proof of proposition 3⁵

(i) Let $f(q) = (1-q)^n$, $g(q) = \sum_{k=0}^{\bar{l}} {}_n C_k q^k$ and $h(q) = \sum_{k=\bar{l}+1}^{n-1} {}_n C_k (1-q)^{n-k} q^{2k}$. Then, $f(0) = 1$.

Since $f'(q) = -n(1-q)^{n-1}$, $f'(0) = -n$. Since $\bar{l} \in \{1, \dots, n-1\}$ and

$$g(q) = {}_n C_0 q^0 + {}_n C_1 q^1 + \sum_{k=2}^{\bar{l}} {}_n C_k q^k = 1 + nq + \sum_{k=2}^{\bar{l}} {}_n C_k q^k, \quad g(0) = 1 \quad \text{and} \quad g'(0) = n. \quad \text{Since}$$

$$h(q) = q^2 r(q), \quad \text{where} \quad r(q) = \sum_{k=\bar{l}+1}^{n-1} {}_n C_k (1-q)^{n-k} q^{2(k-1)}, \quad h'(0) = 0. \quad \text{Since}$$

$$C_{SAMST}(n, q, \bar{l}) = f(q)g(q) + h(q),$$

$$\frac{\partial C_{SAMST}(n, 0, \bar{l})}{\partial q} = f'(0)g(0) + f(0)g'(0) + h'(0) = -n + n + 0 = 0.$$

(ii) By definition, since $C_{SAMST}(n, q, \bar{l})$ has a positive part in addition to $C_{SAM}(n, q)$ on $(0, 1)$, we have the result. ■

⁵ The author would like to thank Yoshitaka Okano for supporting the proof.