KOCHI UNIVERSITY OF TECHNOLOGY

Social Design Engineering Series *SDES-2014-6*

Second thoughts

Tatsuyoshi Saijo
Institute of Economic Research, Hitotsubashi University
Research Center for Future Design, Kochi University of Technology
Center for Environmental Innovation Design for Sustainability, Osaka University

Yoshitaka Okano
Kochi University of Technology
Research Center for Social Design Engineering, Kochi University of Technology

28th April, 2015

School of Economics and Management
Research Center for Social Design Engineering
Kochi University of Technology

Second Thoughts[†]

Tatsuyoshi Saijo* and Yoshitaka Okano#

April 2015

* Institute of Economic Research, Hitotsubashi University,
*# Research Center for Future Design, Kochi University of Technology and
* Center for Environmental Innovation Design for Sustainability, Osaka University

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, social dilemma, cooperation, mechanism design

JEL classification: C72, C92, D74

[†] The Japan Society for the Promotion of Science has financial support for this research (JSPS KAKENHI Grant Number 24243028).

Tatsuyoshi Saijo: Corresponding author. Institute of Economic Research, Hitotsubashi University, Kunitachi, 186-8603 Tokyo Japan. E-mail: tatsuyoshisaijo@gmail.com. Phone: +81-42-580-8312. Fax: +81-42-580-8333.

1. Introduction

This paper is part of our endeavor to find a simple and natural solution to social dilemma. Since playing the social dilemma game once does not solve the problem, it is natural to add some stages before or after the game. This necessarily entails sequential rationality such as subgame perfection. 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 (2014) recently show, the experimental performance of these mechanisms is rather limited.[1]

On the other hand, Masuda, Okano, and Saijo (2014) designed a two stage mechanism called the minimum approval mechanisms to implement a Pareto efficient allocation in public provision problems when the number of players is two, and then showed that it works well even from early periods experimentally. They found that subjects understood the subgame perfection part well, and used the elimination of weakly dominated strategies (*EWDS*) instead of Nash equilibria at each subgames. In other words, they found an *affinity* among the mechanism, subgame perfection and *EWDS*. A basic question of the current paper is whether this approach works well beyond two players.

In an unpublished paper, Huang, Masuda, Okano, and Saijo (2014) designed a mechanism to solve a social dilemma when each player chooses cooperation (*C*) or defection (*D*). 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 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 two stages. 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,

[1] 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.

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 $y = \alpha m w$ 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 $\alpha m w + w = (\alpha m + 1)w$, while the contributor's payoff is $\alpha m w$. 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. (2014). If all participants are either *C* players or *D* players, the game ends. The payoff of a player in the former case is $\alpha n w$ 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.

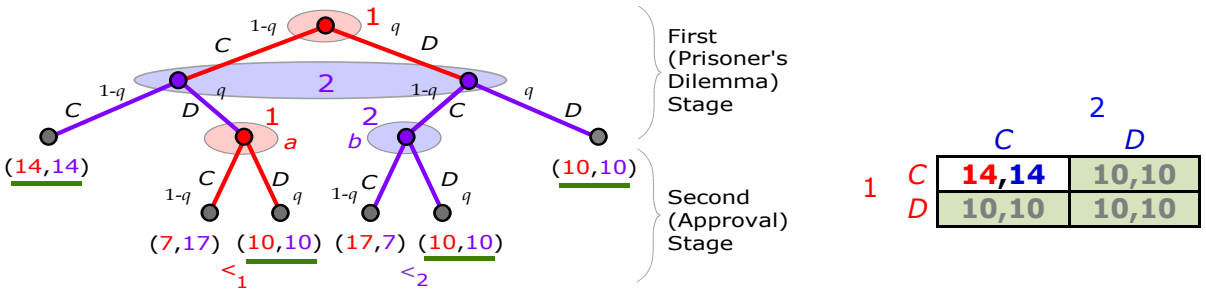


Figure 1: The *SAM* and its reduced normal form game when $n = 2$.

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. (2014) showed the following properties of the *SAM*.

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).*

"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*.

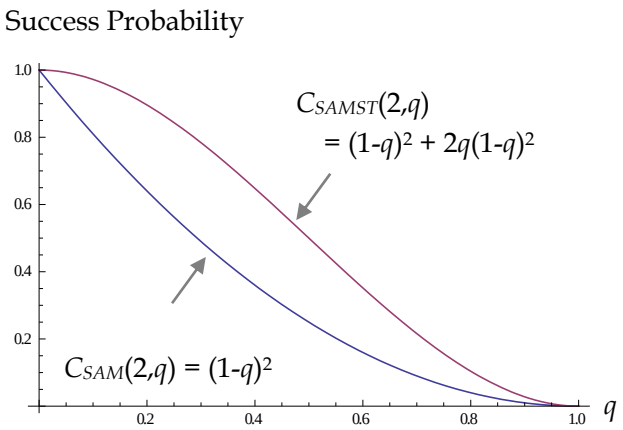


Figure 2: Success probability functions when $n = 2$.

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.[2] In order to improve the low cooperation rates in the early rounds in the experiment of Huang et al. (2014), 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 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

[2] Huang et al. (2014) 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.

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*.

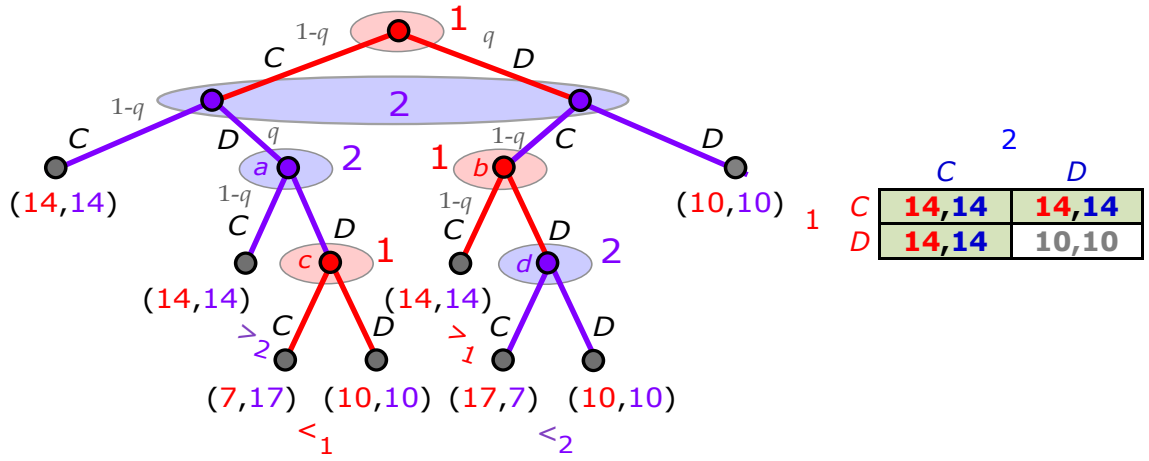


Figure 3: The *SAMST* and its reduced normal form game when $n = 2$.

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 decreases 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.

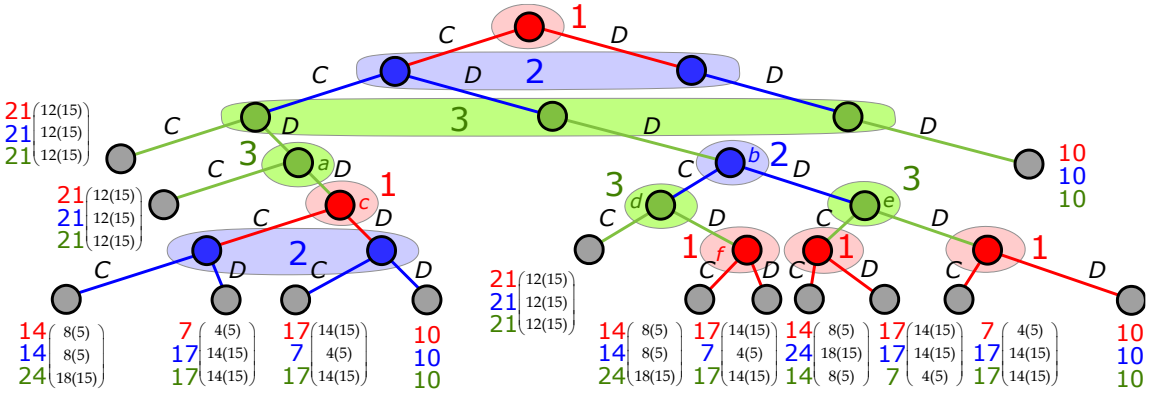


Figure 4.The *SAMST* when $n = 3$.

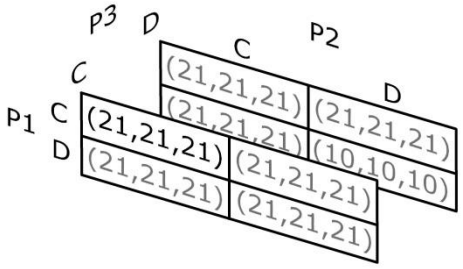
Moreover, consider subgame *e* where player 2 did not change his or her choice in subgame *b*. Since player 1 chooses *D* in the following subgames, player 3 chooses *D* in subgame *e* and the payoff is 10. Consider subgame *d*. In contrast, player 3, who can take advantage of the bonus effect in subgame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

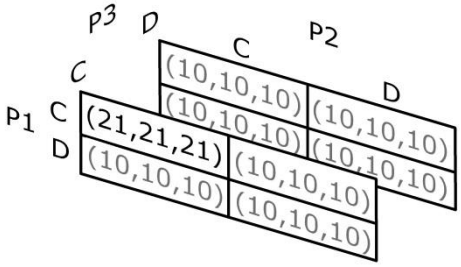
[3] We thank Xiaochuan Huang for indicating this fact.

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.

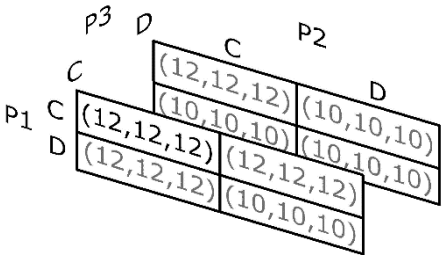
Let us next look at the case of $\alpha = 0.4$. Consider node *f* where player 1 chooses *D*. Then, player 3 at node *d* chooses *D* since 14 > 12. Knowing this fact, player 2 chooses *D* since 10 > 4. That is, the payoff in subgame *b* becomes (10,10,10). From the viewpoint of player 3, since α is too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



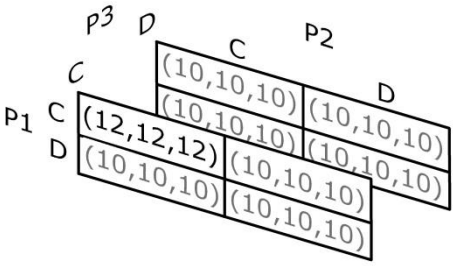
(a) $\alpha = 0.7$ with the *SAMST*.



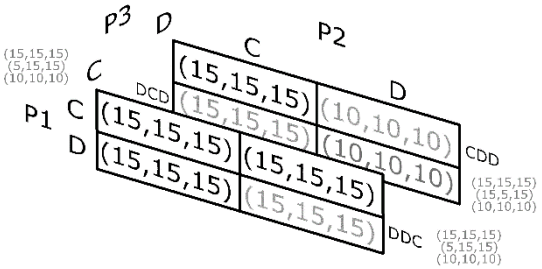
(b) $\alpha = 0.7$ with the *SAM*.



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



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



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

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

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.

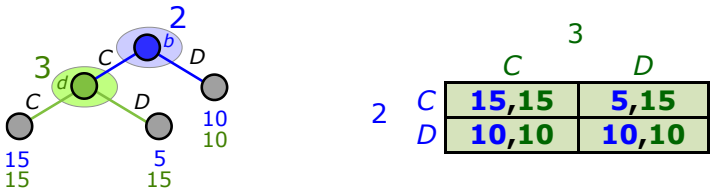


Figure 6. Subgame *b* and its reduced normal form game at node *b*.

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 $1/(n\text{-}1) > \alpha$.

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

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

Suppose $1 \le l < n$. 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 $0 \le l' < l$ 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 $1 \le m' \le m\text{-}1$, 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 $\alpha(l\text{-}1)w + w$. Since $\alpha nw - \alpha(l\text{-}1)w - w = \alpha\{n\text{-}(l\text{-}1)\}w - w$ and $n\text{-}(l\text{-}1) = m+1$,

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 $\alpha(l\text{-}2)w + w$ 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 $\alpha nw\text{-}\{\alpha(l\text{-}2)w + w\} = \{\alpha(n\text{-}l+2)\text{-}1\}w = \{\alpha(m+2)\text{-}1)\}w > 0$, the player chooses C. Since $\alpha nw\text{-}\{\alpha(l\text{-}2)w + w\} > 0$, $\alpha nw\text{-}\{\alpha(l\text{-}k)w + w\} > 0$ for all $2 \le k \le l$. 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 $(\alpha nw, \ldots, \alpha nw)$.

Suppose $\alpha < 1/(m+1)$. Then, the last D player chooses D and hence, the payoff of the penultimate player is $\alpha(l\text{-}1)w$ if the player chooses C. If the player chooses D, then the payoff is $\alpha(l\text{-}2)w + w$. Since $\alpha(l\text{-}2)w + w - \alpha(l\text{-}1)w = (1\text{-}\alpha)w > 0$, the player chooses D. Since $\alpha(l\text{-}k)w + w - \alpha(l\text{-}k+1)w = (1\text{-}\alpha)w > 0$ for all $2 \le k \le l$, 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 $(w, \ldots, w)$.

Take any α satisfying $1/n < \alpha < 1$ and $\alpha \notin \{1/(n\text{-}1), 1/(n\text{-}2), \ldots, 1/2\}$. Consider the case of $\alpha > 1/2$. Then, $\alpha > 1/2 \ge 1/(m+1)$ for all $m \ge 1$ and hence, the payoff outcome of every subgame other than $(D,D,\ldots,D)$ is $(\alpha nw, \ldots, \alpha nw)$: without loss of generality, consider player 1. The payoff in subgame $(C,D,D,\ldots,D)$ is αnw and that in subgame $(D,D,\ldots,D)$ is w. Since $\alpha nw > w$, C is better than D. Since $\alpha > 1/(m+1)$ for all $m \ge 1$, the outcome of the two subgames $(C,\cdot)$ and $(D,\cdot)$ is $(C,C,\ldots,C)$ where "$\cdot$" 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, $(C,C,\ldots,C)$ 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 "$\cdot$" 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 "$\cdot$" indicates that the number of C is not k, C weakly dominates D for all players and hence, $(C,C,\ldots,C)$ is the *SPEWDS* outcome.

Thus, if $\alpha \notin \{1/(n-1),1/(n-2),\ldots,1/2\}$, 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 $k=1,..,l-1$, 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, $\alpha nw - \{\alpha(\bar{l}-1)w + w\} = \{\alpha n - \alpha(\bar{l}-1) - 1\}w = \{\alpha(\underline{m}+1) - 1\}w = 0$, 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 αnw. 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,

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

Since the payoff outcome of the latter should be (*w*,*w*,..,*w*), each of the last $\bar{l}$ players in the former can obtain αnw 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 $(D,\underbrace{C,...,C}_{\underline{m}};\underbrace{D,...,D}_{\bar{l}-1})$ with the payoff outcome of subgame $(\underbrace{C,C,...,C}_{\underline{m}+1};\underbrace{D,...,D}_{\bar{l}-1})$. The latter payoff outcome should be ($\alpha nw,\ldots,\alpha nw$) and hence, the payoff of player 1 should be αnw. Since player 1 in the former should obtain αnw, 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*,..,*C*;*D*,…,*D*) is either αnw or *w*. Since the payoff outcome in this subgame can be either ($\alpha nw,\ldots,\alpha nw$) or (*w*,…,*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*. ∎

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 ${}_nC_{\underline{m}}$, and hence, the total number of all possible reduced normal form games at the beginning node is $(\bar{l}+1)^{{}_nC_{\underline{m}}}$, where ${}_nC_k$ 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 $1/4^{10}$.

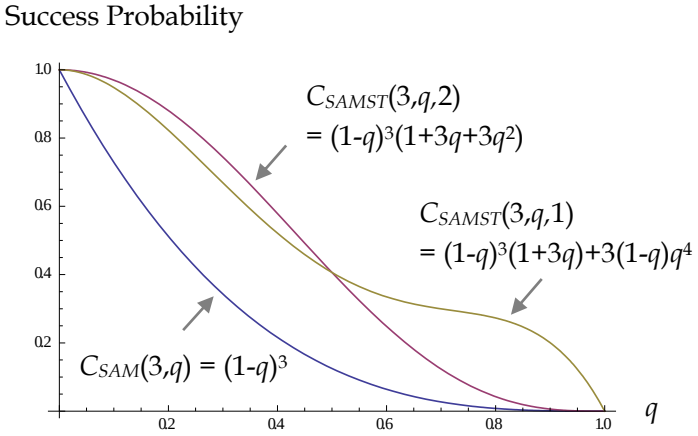


Figure 7. Success probability functions when $n = 3$.

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 $\alpha \notin \{1/(n-1), 1/(n-2), \ldots, 1/2\}$. Then, by summarizing the above argument, we have $C_{SAM}(n,q) = (1-q)^n$, and $C_{SAMST}(n,q,\bar{l}) = (1-q)^n \sum_{k=0}^{\bar{l}} {}_nC_k q^k + \sum_{k=\bar{l}+1}^{n-1} {}_nC_k (1-q)^{n-k} q^{2k}$.

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 $1 \le l \le n-1$, $C_{SAMST}(n,q,l) > C_{SAM}(n,q)$ on (0,1).

Proof. (i) Let $f(q) = (1-q)^n$, $g(q) = \sum_{k=0}^{\bar{l}} {}_nC_k q^k$ and $h(q) = \sum_{k=\bar{l}+1}^{n-1} {}_nC_k (1-q)^{n-k} q^{2k}$. Then, $f(0) = 1$. Since $f'(q) = -n(1-q)^{n-1}$, $f'(0) = -n$. Since $\bar{l} \in \{1, \ldots, n-1\}$ and $g(q) = {}_nC_0 q^0 + {}_nC_1 q^1 + \sum_{k=2}^{\bar{l}} {}_nC_k q^k = 1 + nq + \sum_{k=2}^{\bar{l}} {}_nC_k q^k$, $g(0) = 1$ and $g'(0) = n$. Since $h(q) = q^2 r(q)$, where $r(q) = \sum_{k=\bar{l}+1}^{n-1} {}_nC_k (1-q)^{n-k} q^{2(k-1)}$, $h'(0) = 0$. 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. ∎

The fact that $C_{SAMST}(n,q,\bar{l}) > C_{SAM}(n,q)$ on in (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 is 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, Okano and Saijo (2014) 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. (2014). Behavioral limitations of subgame-perfect implementation.

Huang, X., Masuda, T., Okano, Y., & Saijo, T. (2014). Simplified approval mechanism for social dilemmas. SDES-2014-7, Social Design Engineering Series, Kochi University of Technology.

Kalai, E. (1981). Preplay negotiations and the prisoner's dilemma. *Mathematical Social Sciences*, 1(4), 375-379.

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.

Okano, T., & Saijo, T. (2014). Second thoughts Experiment, in prep.

Varian, H. R. (1994). A Solution to the Problem of Externalities When Agents Are Well-Informed. *American Economic Review*, 84(5): 1278-93.