UTMD Working Paper

The University of Tokyo
Market Design Center

UTMD-024

FREE-RIDER PROBLEM AND SOVEREIGNTY PROTECTION

Hitoshi Matsushima
Department of Economics, the University of Tokyo,

First Version: January 28, 2022
This Version: April 14, 2022

FREE-RIDER PROBLEM AND SOVEREIGNTY PROTECTION*

Hitoshi Matsushima[1]

First Version: January 28, 2022

This Version: April 14, 2022

Abstract

This study investigates a free-rider problem in long-term relationships, where, due to its non-excludable nature, each player attempts to seek loopholes to impose the burden of cooperation on other players. The players establish a committee that demands that each player select an action as promised by a pre-set commitment rule, contingent on all players' announcements. We require the committee to protect player sovereignty in that no player is forced to carry out high cooperation levels against their will or receive future retaliation from the other players for their low commitment. We demonstrate a method called the cautious commitment rule, according to which the committee makes each player a promise that is not necessarily the same as, but close to, and not greater than, their announced upper limit. We show that by adopting this rule, the committee can solve the entire free-rider problem while adhering to sovereignty protection and rule sustainability. As an application, we investigate global warming and show that adopting the cautious commitment rule is crucial for solving the tragedy of the global commons that all countries in the world have long faced.

JEL Classification Numbers: C72, D74, D91, H41, H77, Q54.
Keywords: Free-Rider Problem, Global Commons, Sovereignty Protection, Rule Sustainability, Cautious Commitment Rule.
Word count: 9932 words

---

* This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS) and the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT). I am grateful to Professor Satoru Hibiki, Takeo Hoshi, and Tetsuji Okazaki for their useful comments and encouragements. All errors are mine

[1] Department of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi@e.u-tokyo.ac.jp

I. Introduction

This study investigates free-rider problems, such as the failure to voluntarily provide public goods, the tragedy of the commons, the global commons, and various social dilemmas, which are associated with the non-excludable nature of the commons. We assume that each player has a weak prosocial motive as a lexicographic preference. However, they are thoroughly selfish, and seek loopholes to pass the burden of cooperative efforts on to other players. Due to the non-excludable nature, a player may attempt to free-ride other players' efforts by deliberately continuing to be uncooperative and encouraging other players to disengage from the uncooperative player, and then rebuild cooperation on their own, excluding the uncooperative player. We formulate such strategic conflicts as an infinitely repeated game with perfect monitoring and explore the possibility of resolving conflicts through implicit collusion, that is, via subgame perfect equilibrium (SPE) behavior.

This study's central concern is clarifying how to achieve implicit collusion cooperatively in the free-rider problem while preserving player sovereignty and rule sustainability. To protect player sovereignty, the committee, which these relevant players establish to solve their free-rider problem, will prohibit the use of future retaliations against any player who is unwilling to cooperate. Such sovereignty protection inevitably narrows the range of SPEs in the repeated game; this range was originally quite wide, as the Folk theorem shows (Fudenberg and Maskin, 1986; Abreu, Dutta, and Smith, 1994). This study presents a new methodology for eliciting voluntary cooperation from all players while protecting such player sovereignty in a sustainable manner.

Players establish a committee in advance as this is the only place they can negotiate and make promises for cooperation. At the beginning of every period, the committee requires each player to announce the upper limit of a promise (commitment) they can keep. According to a pre-set commitment rule, the committee determines each player’s promise, depending on all players’ announcements. In this case, to protect player sovereignty, the committee never demands that any player select actions that are greater than their respective upper limits. Importantly, any player who keeps their promise is never retaliated against by other players in future periods. Hence, any player who behaves uncooperatively will never be retaliated against unless they promise high cooperation. However, once someone breaks their promise, the committee becomes dysfunctional, and players inevitably continue to behave uncooperatively thereafter. That is, a player is retaliated against practically only if they break their promise.

Based on this committee scenario, we define the concept of sovereignty protection as a constraint on SPEs and characterize the class of all SPEs that protect player sovereignty in the following manner. We first define a variant of the component game called the message game, where each player is forced to keep their promise. We then define a variant of the Nash equilibrium concept called the prosocial Nash equilibrium (PNE), where each player selects the maximal best response. As the characterization result, we show that with a sufficient discount factor, an SPE with sovereignty protection in the repeated game is equivalent to a PNE in the associated message game.

The range of PNEs in the message game depends crucially on the commitment rule the committee adopts. To solve the free-rider problem as comprehensively as possible, we propose the following method called the cautious commitment rule. The committee will demand that each player select their upper limit if they all announce the same upper limit as each other. Otherwise, the committee will demand that any player who announces an upper limit greater than the minimal upper limit select an action that is less than their respective upper limit at most some positive grid. We show that by adopting this commitment rule, the committee can incentivize all players to achieve full cooperation via the unique PNE in the message game, that is, via the unique SPE with sovereignty protection in the repeated game.

A particularly important point is that by setting a small enough grid, the committee can significantly prevent any player's elaborate attempt to free-ride other players. Due to the nonexcludable nature of the commons, a player may deliberately attempt to make very low promises, anticipating that the other players will renegotiate with each other excluding this uncooperative player. However, with a small enough grid, even if a player attempts to make such a low promise, other players can still enjoy high enough cooperation levels that they will hesitate to renegotiate. Thus, the cautious commitment rule is sustainable in that it is robust against renegotiation that excludes a player who purposefully takes uncooperative attitudes.

As an application of this study, we investigate global warming (global climate change, global commons) and show that adopting the cautious commitment rule is crucial for solving the tragedy of the global commons. This application will contribute greatly to improving the long stagnation of the COP (the international negotiation at the Conference of the Parties to the United Nations Framework Convention on Climate Change).[2] The COP has adopted a so-called pledge-and-review approach, where the COP member

[2] For details of the international negotiations on global climate change, see Victor (2001), Stern (2007), Nordhaus (1994, 2013), Wagner and Weitzman (2015), and Cramton et al. (2017).

countries first agree on a global emissions cap and then voluntarily promise their own emissions reduction levels without establishing self-enforcing treaties that will make their promises credible and cooperative. Many economics researchers, such as Stiglitz (2006), Cooper (2008), Stoft (2008), Cramton and Stoft (2012), Cramton, Ockenfels, and Stoft (2015), Weitzman (2014), MacKay et al. (2015), Nordhaus (2015), and Gollier and Tirole (2015), have pointed out that the long-term stagnation's major cause is the adoption of the pledge-and-review approach. As an alternative to this approach, they advocate ways for the COP members to negotiate and agree on a common carbon price rather than a global emissions cap. The replacement of global emissions cap with common carbon pricing will successfully change the rule of the game from a war of attrition to an easy-to-solve coordination format. This study's model is suitable for describing a consensus-building procedure such as the one they advocate.

More specifically, several economics researchers, including Cooper (2008), Stoft (2008), and MacKay et al. (2015), advocate the common commitment rule, according to which all countries in the COP agree on a minimal upper limit as the common carbon price and impose this price on their respective citizens. They showed that by adopting the common commitment rule, a reciprocal principle of "I will if you will" functions effectively, solving some aspects of the free-rider problem.

The common commitment rule is a polar case of the cautious commitment rule with a very large grid. However, due to this large grid setting, the common commitment rule fails to block the loophole that allows a player to free-ride on other players' renegotiations; that is, it fails to be robust against a player's temporary uncooperative attitude, and therefore, fails to be sustainable in the long run.

To make up for this shortcoming of the common commitment rule, there is a proposal to establish a Climate Club among only certain major countries and impose strong retaliations, such as trade sanctions and ostracism, on free-riding countries that are not members of this club (Nordhaus, 2015). This policy abandons sovereignty protection, socially excludes countries, leaves them behind the other countries, and provides major countries with a convenient excuse to strengthen the methods of sanctions that could spark international conflict. If the common commitment rule is adhered to, neither sovereignty protection nor rule sustainability can be established. From the perspective in this study, this incompatibility corresponds to the so-called Westphalia Dilemma addressed in Nordhaus (2015). We show that adopting the cautious commitment rule instead of the common commitment rule can provide a concrete path for overcoming the Westphalia dilemma. The cautious commitment rule can solve the entire free-rider problem under the Westphalian regime through a cautious version of the reciprocal principle such as "I will do better if you will do slightly better."

In order to solidify rule sustainability, we further modify the cautious commitment rule by making the size of the grid dependent on the number of uncooperative players. The modified cautious commitment rule is robust against renegotiations caused by a player who is deliberately demoralizing all players in achieving cooperation.

This study proposes a concrete path to seek achievement of various sustainable development goals centered on the three pillars of global environment, economic development, and fair social inclusion. As Ostrom (1990) and others say, it is generally important to build trust and reciprocity using sanctions and ostracism for solving the tragedy of the commons. However, trust and reciprocity sometimes conflict with fair social inclusion, and therefore, they must be moderately restricted. The cautious commitment rule and its modification balance trust and reciprocity with fair social inclusion by carefully setting the grid, making various sustainable development goals a reality.

This study's model on the commons is related to yet different from the network externality (Shapiro and Varian, 1999; Tirole, 2017). A network typically has the critical mass in the number of participants. Hence, the main concern of the network provider is to reach the critical mass even if they spend a lot of money temporally. In contrast, the commons have no such critical mass, which makes the sustainability of the solution method a substantial issue.

This study is also related to the development of open-source software (Lerner and Tirole, 2005). The contributors to this development typically have strong intrinsic preferences for prosocial behavior beyond their self-interests. Moreover, there are potential motivations that are specific to the software developers, such as signaling on their capabilities. In contrast, this study assumes neither strong intrinsic preferences nor problem-specific motivations. Instead, we pioneer a careful institutional design method for inducing cooperative behavior in literally selfish individuals. Each player in our model makes no choice other than a selfish best response and has no wishful thinking such as "You will as I will." We do not replace the Nash equilibrium with any ethical concept like the Kantian equilibrium (Roemer, 2010).

Existing literature on the applied theory of repeated games investigates the tragedy of the commons in implicit collusion, such as Barrett (1994), Harstad (2012), and Harrison and Lagunoff (2017). Above all, Harstad (2012) is closely related to this study, because both studies commonly restrict the use of future retaliations, alleviate the multiplicity of equilibria, and then focus only on simple and easy-to-understand equilibrium behaviors. Harstad considered only Markov equilibria addressed by Maskin and Tirole (2001), where equilibrium behavior depends only on the physical state. In contrast, this study does not allow equilibrium behavior to depend on the physical state. It imposes stricter restrictions to protect player sovereignty, and then derives the compatibility of uniqueness and cooperation.

The remainder of this paper is organized as follows. Section II illustrates the intuition of the logical core, and Section III models the free-rider problem as an infinitely repeated game with perfect monitoring. Section IV explains sovereignty protection by introducing commitment rules and various conditions regarding SPEs concerning the determination of promises and future retaliation. Section V shows the characterization result as the equivalence of the SPE in repeated games with sovereignty protection and the PNE in the message game. Section VI introduces the cautious commitment rule and describes a sufficient condition under which the cautious commitment rule achieves full cooperation as unique SPE behavior with sovereignty protection. Section VII models global warming as an application of this study. Section VIII considers further improvements of the commitment rule to strengthen rule sovereignty, and Section IX concludes.

II. Intuition

Consider three players who face voluntary provision of a public good that has the nonexcludable nature. Each player $i \in \{1,2,3\}$ simultaneously produces $a_i \in \{0,1,...,10\}$ units of the public good, where the cost of production per unit for its producer is 12 and the benefit from production per unit is 11 for every player. Hence, the payoff for each player $i \in \{1,2,3\}$ is given by

$$11 \times (a_1 + a_2 + a_3) - 12a_i .$$

Since the cost is greater than the benefit ($12 > 11$), $a_i = 0$ is strictly dominant for each player i. This implies failure to voluntarily provide the public good as an equilibrium behavior, that is, all players contribute nothing.

To overcome this failure, the committee, which consists of these three players, requires each player (member) $i \in \{1,2,3\}$ to voluntarily announce the upper limit of their promise $m_i \in \{0,1,...,10\}$. The committee then demands each player i produce their respective upper limit $a_i = m_i$. We assume in this section that any player keeps their promise. (We eliminate this assumption in the remaining parts of this study.) Since no self-enforcing device exists to make their announcements cooperative, $m_i = a_i = 0$ is still dominant, again implying failure to voluntarily provide the public good.

As a full-scale breakthrough, the committee will adopt the common commitment rule (Cooper, 2008; Stoft, 2008; MacKay et al., 2015) as follows. The committee demands each player i to produce, not their respective upper limit m_i, but the minimal upper limit $\min[m_1, m_2, m_3]$. Hence, each player $i's$ payoff is given by

$$11 \times 3 \times \min[m_1, m_2, m_3] - 12\min[m_1, m_2, m_3]$$

$$= 21\min[m_1, m_2, m_3] .$$

Clearly, $m_i = a_i = 10$ is weakly dominant for every player i, implying successful voluntary provision that brings each player the payoff 210. If the most uncooperative player increases their production level by some units (with cost per unit 12), they can increase all players' production levels by the same units (with benefit per unit $11 \times 3 = 33$). Since $33 > 12$, the reciprocal principle of "I will if you will" functions effectively, solving an aspect of the free-rider problem.

Unfortunately, the common commitment rule fails to solve another aspect of the free-rider problem; that is, it fails to be sustainable. Suppose that player 3 stubbornly sticks to announcing the lowest upper limit $m_3 = 0$. Then, irrespective of the other players' announcements, all players are forced to select only the worst level $\min[m_1, m_2, 0] = 0$, generating the worst common payoff $21 \min[m_1, m_2, 0] = 0$. To break this deadlock, the rest of the players will relaunch the committee and adopt the common commitment rule excluding player 3. The committee now demands only players 1 and 2 select $\min[m_1, m_2]$ but demand nothing from player 3. In this case, clearly, $m_i = a_i = 10$ is weakly dominant for both players 1 and 2, while player 3 selects $a_3 = 0$. Both players 1 and 2 obtain the payoff given by

$$11 \times 2 \times \min[m_1, m_2] - 12 \min[m_1, m_2] = 10 \min[m_1, m_2] = 100,$$

which is greater than the payoff 0 when player 3 joins the committee and behaves uncooperatively. However, due to the nonexcludable nature, player 3 also enjoys the benefit with no contribution, that is,

$$11 \times 2 \times \min[m_1, m_2] = 22 \min[m_1, m_2] = 220,$$

which is greater than the payoff 210 when player 3 joins the committee and makes the cooperative action choice $m_3 = a_3 = 10$. Hence, player 3 succeeds in free-riding by intentionally adopting an uncooperative attitude and encouraging the other players to renegotiate, excluding player 3 from the renegotiation. This implies the failure of the common commitment rule to be sustainable, that is, its failure to resolve the second aspect of the free-rider problem.

One way to eliminate such free rides is to force player 3 to behave cooperatively by exercising strong sanctions. The committee, which now consists of only player 1 and player 2, will take future retaliation against player 3 by using various means if player 3 behaves uncooperatively. The Climate Club proposal by Nordhaus (2015) is an example of this method. However, such methods of repairing the common commitment rule are not acceptable, because they neglect sovereignty protection, and create a hierarchical relationship between the committee and player 3 that is prone to social conflicts and inequality.

To solve the entire free-rider problem while protecting player sovereignty and rule sustainability, the committee will adopt a new rule termed the cautious commitment rule as follows. The committee will demand each player $i \in \{1,2,3\}$ produce their respective upper limits, that is, $a_i = m_i$, if $m_i = \min[m_1, m_2, m_3]$. Importantly, the committee will demand each player i produce not their upper limit but their upper limit minus one, that is, $a_i = m_i - 1$, if $m_i > \min[m_1, m_2, m_3]$. Hence, each player $i's$ payoff is given by

$$11 \sum_{j \in \{1,2,3\}} a_j - 12 a_i$$

$$= 11 \sum_{j \in \{1,2,3\}} \{m_j - \chi_j(m)\} - 12\{m_i - \chi_j(m)\},$$

where $\chi_j(m) \in \{0,1\}$ denotes the indicator such that

$$[\chi_j(m) = 0] \Leftrightarrow [m_i = \min[m_1, m_2, m_3]].$$

Note that full cooperation $m = a = (10,...,10)$ is a Nash equilibrium. By selecting 0 instead of 10, each player i can save the net cost $(12-11)\times 10 = 10$, but loses the benefit

$11 \times 2 = 22$ due to the one-unit decreases by the other players. Since $10 < 22$, each player is willing to select 10.

This full cooperation is not weakly dominant, and even multiple uncooperative Nash equilibria coexist. In fact, $m = a = (p,...,p)$ is a Nash equilibrium for every $p \in \{0,1,...,10\}$. Nevertheless, we can argue that only full cooperation $p = 10$ is a decent Nash equilibrium. We assume that each player has a prosocial motive as a lexicographic preference in that they always announce the maximal best response. This motive is very weak because each player makes no choice other than a selfish best response.

Due to such prosocial motives, each player can steadily conclude that only full cooperation is selfishly optimal as follows. Any player who is not the single player announcing the minimal upper limit $\min[m_1, m_2, m_3]$ is willing to announce $\min[m_1, m_2, m_3] + 1$, due to their selfish and prosocial motives. In this case, the single player who announces the minimal upper limit $\min[m_1, m_2, m_3]$ is willing to replace their announcement of $\min[m_1, m_2, m_3]$ with $\min[m_1, m_2, m_3] + 1$, if such a single player exists. By making this replacement, this single player spends an additional net cost $12 - 11 = 1$, but can obtain a greater additional benefit $11 \times 2 = 22 > 1$ from the one-grid increases by the other players. Thus, like climbing stairs, all players are incentivized to produce greater levels. Hence, we can conclude that full cooperation is the unique prosocial (decent) Nash equilibrium. Due to a weaker reciprocal principle such as "I will do better if you will do slightly better," the cautious commitment rule achieves full cooperation as a unique prosocial equilibrium behavior.[3,4]

Importantly, the cautious commitment rule goes beyond the limits of the common commitment rule. It thoroughly eliminates a player's free-riding caused by the anticipation in other players' renegotiations. Suppose that player 3 sticks to announcing $m_3 = 0$, while any other player $i \in \{1,2\}$ announces $m_i = 10$. Then, any other player $i \in \{1,2\}$ produces $a_i = m_i - 1 = 9$; this contrasts with the common commitment rule, which, if adopted in this case, results in no production. Hence, even if player 3 is uncooperative, players 1 and 2 can still achieve some level of cooperation and therefore, stop renegotiating while excluding player 3. As a result, player 3 stops taking such an uncooperative attitude, implying that the cautious commitment rule is sustainable, that is, it solves the second aspect of the free-rider problem. Consequently, by adopting the cautious commitment rule, the committee can solve the entire free-rider problem while adhering to both sovereignty protection and rule sustainability.[5]

III. The Model

---

[3] We can find a related concept in the unique implementation theory of social choice functions by Abreu and Matsushima (1992). They proposed a mechanism design method termed the tail-chasing competition, which successfully eliminates unwanted equilibria. The tail-chasing competition is like the climbing-chair method of this study.

[4] In the implementation theory literature, several studies such as Matsushima (2008a, 2008b) and Dutta and Sen (2012) showed that the presence of tiny prosocial motives such as a preference for honesty plays a crucial role in eliminating unwanted equilibria.

[5] In Section VIII, we will modify the cautions commitment rule to satisfy a stronger version of rule sustainability.

There exist $n \geq 2$ players who establish a committee. Let $N = \{1, 2, ..., n\}$ denote the set of all such players. Each player $i \in N$ has a set of actions defined as $A_i \equiv [0, \infty)$. Let $A \equiv \times_{i \in N} A_i$ and $a \equiv (a_i)_{i \in N} \in A$. These players have a long-term relationship modeled as an infinitely repeated game with perfect monitoring. In each period $t \geq 1$, each player $i \in N$ simultaneously announces a message $m_i(t) = m_i \in M_i$ as cheap talk to the committee and then simultaneously selects an action $a_i(t) = a_i \in A_i$, where M_i denotes player i's set of messages. Let $M \equiv \times_{i \in N} M_i$ and $m \equiv (m_i)_{i \in N} \in M$.

Each player i has an instantaneous payoff function $u_i : A \to R$. They receive the instantaneous payoff $u_i(a) \in R$ in period t when all players select the action profile $a(t) \equiv (a_i(t))_{i \in N} = a$. Note that the instantaneous payoff does not depend on the message profile $m(t) \equiv (m_i(t))_{i \in N}$. To specify the component game as a free-rider problem, we assume that $u_i(a)$ is increasing in a_j for all $i \in N$ and $j \neq i$, and there exists a strictly dominant action profile $\underline{a} = (\underline{a}_i)_{i \in N} \in A$ where for every $i \in N$ and $a_{-i} \equiv (a_j)_{j \neq i} \in A_{-i} \equiv \times_{j \neq i} A_j$,

$$u_i(\underline{a}_i, a_{-i}) > u_i(a) \text{ for all } a_i \in A_i \setminus \{\underline{a}_i\}.$$

We assume that $u_i(a)$ is decreasing in $a_i \in [\underline{a}_i, \infty)$ and increasing in $a_i \in [0, \underline{a}_i)$ for all $i \in N$. We further assume that there exists $\overline{p} \in [0, \infty)$, called the maximal common cooperation level, such that

$$\overline{p} > \underline{a}_i \text{ for all } i \in N,$$

and $\overline{a} \equiv (\overline{p}, ..., \overline{p})$ is Pareto-superior to $\underline{a}$, that is,

$$u_i(\overline{a}) > u_i(\underline{a}) \quad \text{for all } i \in N .$$

We regard $\overline{a}$ as these players' cooperation goal, and consider the possibility for them to achieve this goal constantly as unique equilibrium behavior in the repeated game with perfect monitoring.

A history up to period $t \geq 1$ is denoted by $h^t \equiv (m(\tau), a(\tau))_{\tau=1}^{t}$. Let h^0 denote the null history, and H denote the set of all histories. A strategy for player i is defined as $s_i = (\mu_i, \alpha_i)$, where $\mu_i : H \to M_i$ and $\alpha_i : H \times M \to A_i$. According to s_i, each player i announces the message $m_i(t+1) = \mu_i(h^t) \in M_i$ in period $t+1$, provided the history h^t occurs. Player i selects the action $a_i(t+1) = \alpha_i(h^t, m(t+1)) \in A_i$ in period $t+1$, provided the history h^t occurs and players announce $m(t+1) \in M$. Let us denote $\mu(h^t) \equiv (\mu_i(h^t))_{i \in N} \in M$ and $\alpha(h^t, m(t+1)) \equiv (\alpha_i(h^t, m(t+1)))_{i \in N} \in A$.

The payoff for player i in the repeated game is given by $\sum_{t=1}^{\infty} \delta^{t-1} u_i(a(t))$, where $\delta \in (0,1)$ denotes the discount factor. The payoff for player i generated by strategy profile s is given by

$$V_i(s) \equiv \sum_{t=1}^{\infty} \delta^{t-1} u_i(\alpha(h^{t-1}(s), \mu(h^{t-1}(s)))),$$

where we recursively define the histories generated by s as $h^0(s) \equiv h^0$, and

$$h^t(s) \equiv (h^{t-1}(s), \mu(h^{t-1}(s)), \alpha(h^{t-1}(s), \mu(h^{t-1}(s)))) \quad \text{for all } t \geq 1 .$$

Let $V_i(s \mid h^t)$ denote the payoff for player i generated by a strategy profile s after period $t+1$ when h^t occurs. Let

$$V_i(s \mid h^t, m(t+1)) \equiv u_i(a(t+1)) + \delta V_i(s \mid h^{t+1}),$$

where we denote $a(t+1)=\alpha(h^t, m(t+1))$. A strategy profile s is said to be an SPE if for every $h^t \in H$, every $i \in N$, and every strategy s_i' for player i,

$$V_i(s \mid h^t) \geq V_i(s_i', s_{-i} \mid h^t),$$

and

$$V_i(s \mid h^t, m) \geq V_i(s_i', s_{-i} \mid h^t, m) \text{ for all } m \in M.$$

IV. Sovereignty Protection

This study investigates SPEs that satisfy sovereignty protection from two points of view: that is, which action the committee demands that each player select, and how much future retaliations is restricted for incentivizing players to behave cooperatively. Throughout this study, we assume that

$$M_i = [0, \overline{p}] \subset A_i \text{ for all } i \in N.$$

We regard the message $m_i(t)$ as the upper limit of actions that player i can promise the committee that they will select in period t.

IV.A. Commitment Rule

We introduce a *commitment rule* as $g = (g_i)_{i \in N}$, where $g_i : M \to A_i$ for each $i \in N$. According to commitment rule g_i for player i and a message profile

$m(t) \in M$, the committee makes player i promise to select an action $a_i = g_i(m) \in A_i$ in period t.

Importantly, to protect player sovereignty, we will assume that the committee never demands each player i to select any action that is greater than the announced upper limit $m_i(t)$, and the strictly dominant action $\underline{a}_i$, in each period t, that is,

(1) $$g_i(m) \leq \max[m_i(t), \underline{a}_i] \text{ for all } i \in N \text{ and } m \in A.$$

This study permits interdependent commitment. That is, we permit the promise (commitment) $g_i(m)$ of player i to depend on the other players' messages m_{-i} as well as their own messages m_i. Due to this interdependence, the committee does not necessarily demand that player i select their announced upper limit m_i.

Let us fix an arbitrary positive integer $L \geq 1$. (We keep in mind that L is a large enough value.) We assume that

(2) $$m_i - g_i(m) \leq \frac{\bar{p}}{L} \text{ for all } m \in M \text{ and } i \in N.$$

Assumption (2) implies that with a large enough L, the promise $g_i(m)$ of each player i is not much different from their announced upper limit m_i.

IV.B. Future Retaliation

We regard the committee as the only place players can negotiate and make promises for cooperation. Once someone breaks the promise, the committee will malfunction thereafter. If the promises are kept, a steady cooperative relationship will be maintained.

Importantly, the committee prohibits each player from retaliating in the future against any other player who keeps their promise, even if they set their upper limit very low.

Based on these scenarios, this subsection introduces five conditions of a strategy profile s concerning sovereignty protection and the role of the committee as follows. We define $\hat{H} \subset H$ as the set of all histories h^t where every player keeps their promises up to period t, that is,

$$[h^t \in \hat{H}] \Leftrightarrow [a_i(\tau) = g_i(m(\tau)) \text{ for all } \tau \leq t \text{ and } i \in N].^6$$

Condition 1 implies that each player continues to keep their promises.

Condition 1: For every $t \geq 1$,

$$h^t(s) \in \hat{H}.$$

Condition 2 implies that once someone breaks their promises, the committee will malfunction and never cooperate, that is, all players continue to select the dominant actions, thereafter.

Condition 2: For every $h^t \notin \hat{H}$, $m \in M$, and $i \in N$,

$$\mu_i(h^t) = \underline{a}_i \text{ and } \alpha_i(h^t, m) = \underline{a}_i.$$

Condition 3 implies that each player i announces the same message, denoted by $\hat{m}_i$, across periods, provided all players kept their promises in the previous history of play.

---

[6] We can replace $=$ with $\geq$ without substantial change in this study.

Condition 3: For every $i \in N$, there exists $\hat{m}_i \in M_i$ such that

$$\mu_i(h^t) = \hat{m}_i \text{ for all } h^t \in \hat{H} .$$

Condition 4 implies that each player always keeps their promises irrespective of which messages are announced, if they kept their promises in the previous history of play.

Condition 4: For every $i \in N$ and $h^t \in \hat{H}$,

$$\alpha_i(h^t, m) = g_i(m) .$$

From Conditions 3 and 4, it follows that each player $i \in N$ announces the same message $\hat{m}_i$ and selects the same action $g_i(\hat{m})$ across periods if all players kept their promises in the previous history of play. This implies a restriction of future retaliation, because the level of their cooperation is never influenced by the levels of the promises in the previous history of play. With this restriction, any player is protected from being forced into unwilling promises due to fear of future retaliations. The committee never retaliates against any player who selects a low action level but does not promise any higher action level.

Finally, Condition 5 implies the presence of prosocial motive as a lexicographic preference such that each player is willing to enhance their announcement if this enhancement never decreases their instantaneous payoff.

Condition 5: For every $i \in N$,

$$u_i(g(m_i, \hat{m}_{-i})) < u_i(g(\hat{m})) \text{ for all } m_i > \hat{m}_i,$$

where $\hat{m} \equiv (\hat{m}_i)_{i \in N}$ is the message profile that was introduced in Condition 3.

A strategy profile s is said to *protect player sovereignty* if it satisfies Conditions 1, 2, 3, 4, and 5. It is well expressed by $\hat{m} \equiv (\hat{m}_i)_{i \in N} \in M$. Hence, we can write $\hat{m}$ instead of s for a strategy profile whenever it protects player sovereignty. According to a strategy profile with sovereignty protection $\hat{m}$, each player i always selects the same action across periods, which is given by

$$\hat{a}_i \equiv g_i(\hat{m}),$$

provided all players kept their promises in the previous history of play. Hence, from Conditions 2 and 3, we have

(3) $$V_i(s \mid h^t) = \frac{u_i(\hat{a})}{1-\delta} \text{ for all } h^t \in \hat{H},$$

and

(4) $$V_i(s \mid h^t) = \frac{u_i(\underline{a})}{1-\delta} \text{ for all } h^t \notin \hat{H},$$

where we denote $\hat{a} \equiv (\hat{a}_i)_{i \in N} \in A$.

V. Characterization

This section characterizes SPE strategy profiles with sovereignty protection. We define the *message game* as a variant of the component game where we regard M_i as the strategy space for each player $i \in N$ and assume that all players keep their promises.

We then define an equilibrium concept from a semantic viewpoint. That is, a message profile $m \in M$ is said to be a PNE *in the message game* if for every $i \in N$,

$$u_i(g(m)) \geq u_i(g(m_i', m_{-i})) \text{ for all } m_i' \in M_i,$$

and

$$u_i(g(m)) > u_i(g(m_i', m_{-i})) \text{ for all } m_i' > m_i.$$

A PNE is a special case of Nash equilibrium where each player selects the maximal best response message. The following theorem shows a characterization of SPEs with sovereignty protection, implying that, with any sufficient discount factor δ, an SPE with sovereignty protection in the repeated game is equivalent to a PNE in the associated message game.

Theorem 1: *A strategy profile with sovereignty protection* $\hat{m}$ *is an SPE if and only if* $\hat{m}$ *is a PNE, and for every* $i \in N$ *and* $m \in M$*,*

(5) $$u_i(\underline{a}_i, g_{-i}(m)) - u_i(g(m)) \leq \frac{\delta}{1-\delta}\{u_i(\hat{a}) - u_i(\underline{a})\}.$$

Proof: From Condition 4, for every $h^{t-1} \in \hat{H}$ and $m(t) \in M$, each player must have the incentive to keep their promises credibly, that is, to sincerely select the action $a_i(t) = g_i(m(t))$. We need to use future retaliation to make their commitments credible. Since limited use of future retaliation is expressed by (3) and (4), we can regard the inequalities in (5) as the necessary and sufficient condition for such credible commitments.

Suppose that the inequalities in (5) hold. Then, $\hat{m}$ is an SPE if and only if each player i has the incentive to announce $\hat{m}_i$ without the use of future retaliation. Note from Condition 5 that each player has a prosocial motive, that is, $\hat{m}_i$ is the maximal best response to $\hat{m}_{-i}$. Hence, the incentive constraint for each player i to announce $\hat{m}_i$ is equivalent to the PNE property of $\hat{m}$ in the message game.

Q.E.D.

The following proposition shows that to achieve both cooperation and sovereignty protection, it is necessary that the committee adopts an interdependent commitment rule.

Proposition 1: *Suppose that commitment rule g is independent, that is, $g_i(m)$ is independent of m_{-i} for all $i \in N$ and $m \in M$. Then, a strategy profile with sovereignty protection $\hat{m}$ is an SPE if and only if $L=1$, $\hat{m}=\underline{a}$, and*

$$g(m)=\underline{a} \text{ for all } m \in M.$$

Proof: From Theorem 1, the "if" part of Proposition 1 holds because $\underline{a}$ is a PNE in this case. Suppose that $\hat{m}$ is a PNE. Since players who keep their promises are never retaliated against in the future, it follows from Conditions 4 and 5 that any message profile $m \in M$ must satisfy the PNE property, that is, for every $m \in M$, $i \in N$, and $m_i' \in M_i$,

$$u_i(g_i(m_i), g_{-i}(m_{-i})) \geq u_i(g_i(m_i'), g_{-i}(m_{-i})),$$

where we write $g_i(m) = g_i(m_i)$ and $g_{-i}(m_{-i}) = (g_j(m_j))_{j \neq i}$. From (1) and these inequalities, we have

$$g_i(m_i) = g_i(m_i') \leq \underline{a}_i \text{ for all } i \in N, \ m_i \in M_i, \text{ and } m_i' \in M_i.$$

Hence, it follows from Theorem 4 and (5) that

$$g_i(m_i) = \underline{a}_i \text{ for all } i \in N \text{ and } m_i \in M_i,$$

which along with (2) implies $L=1$ and $\hat{m} = \underline{a}$.

Q.E.D.

VI. Cautious Commitment Rule

We specify an interdependent commitment rule $g = g^*$ as the *cautious commitment rule* as follows. We define $L+1$ real numbers $\{\beta^0,...,\beta^L\}$ as

$$\beta^l = \frac{\overline{p}l}{L} \text{ for all } l \in \{0,...,L\}.$$

For every $p \in (0, \overline{p}]$, we define $l(p) \in \{0,...,L-1\}$ as

$$\beta^{l(p)} < p \leq \beta^{l(p)+1}.$$

For every $m \in M$, we specify $g_i(m) = g_i^*(m)$ as

$$g_i(m) = \min[\max[\min_{j \neq i} m_j, \beta^{l(m_i)}], m_i]$$

$$\text{if } m_i > 0,$$

and

$$g_i(m) = 0 \qquad \text{if } m_i = 0.$$

According to the cautious commitment rule, if player i announces the minimal upper limit, that is, $m_i = \min_{j \in N} m_j$, the committee demands they select this upper limit, that is,

$$g_i(m) = m_i \qquad \text{if } m_i = \min_{j \in N} m_j.$$

If the upper limit m_i of player i is greater than but close to the minimal upper limit $\min_{j\in N} m_j$, the committee demands they select the minimal upper limit, that is,

$$g_i(m) = \min_{j\in N} m_j \qquad \text{if } \beta^{l(m_i)} \le \min_{j\in N} m_j < m_i .$$

If the upper limit m_i of player i is sufficiently greater than the minimal upper limit $\min_{j\in N} m_j$, the committee demands they select $\beta^{l(m_i)}$, which is smaller than but close to m_i, that is,

$$g_i(m) = \beta^{l(m_i)} \qquad \text{if } \min_{j\in N} m_j < \beta^{l(m_i)} .$$

Any player i who announces a message greater than $\min_{j\neq i} m_j$ is demanded to select $\max[\min_{j\neq i} m_j, \beta^{l(m_i)}]$, which is at most $\bar{p}/L$ less than m_i. Clearly, the cautious commitment rule $g = g^*$ satisfies all the assumptions in the previous sections.

One extreme case of the cautious commitment rule, which is given by $L=1$, corresponds to the *common commitment rule*, which many economic researchers have proposed to ease the difficulties of international negotiations concerning global warming (see Cramton et al., 2017). According to the common commitment rule, the committee always demands all players select the minimal upper limit, that is,

$$g_i(m) = \min_{j\in N} m_j \text{ for all } i\in N \text{ and } m\in M .$$

The common commitment rule has the following drawback in balancing rule sustainability and sovereignty protection. The committee demands nothing of any player i who sets high upper limits when some player $j\neq i$ sets a low upper limit. In this case, players other than player j prefer to renegotiate and change the common commitment rule so that they can cooperate with each other excluding player j. Due to the nonexcludable nature, player j can free-ride the other players' cooperative agreement. Hence, player j deliberately tries to make very low promises in anticipation of the other players renegotiating and excluding them. Hence, the common commitment rule fails to be sustainable. To avoid such free-riding, the committee has no choice but to apply future retaliation to those who make only low promises but keep their promises. However, this causes abandonment of sovereignty protection. An example is the proposal of the Climate Club in global warming.

Fortunately, by letting L be larger than 1, that is, by replacing the common commitment rule with a cautious commitment rule, we can avoid this type of free riding in many cases in the consistent manner with sovereignty protection. Even if a player makes a low promise, the other players can still enjoy sufficient cooperation within the range of the difference $\bar{p}/L$, stopping renegotiation.

The following theorem shows a sufficient condition under which the committee's adoption of the cautious commitment rule associated with an arbitrary fixed $L \geq 1$ succeeds in incentivizing all players to achieve the maximal common cooperation level $\bar{p}$ as a unique SPE behavior with sovereignty protection. Let us denote $a^l \equiv (\beta^l,..,\beta^l) \in A$ for all $l \in \{0,...,L\}$.

Theorem 2: *Consider the cautious commitment rule $g = g^*$ and an arbitrary strategy profile that protects player sovereignty $\hat{m}$, where we assume $\min_{i \in N} \hat{m}_i \geq \max_{i \in N} \underline{a}_i$. Suppose that for every $p \in [0, \bar{p}]$ and $p' \in [0, \bar{p}]$, if $p' > p \geq \max_{i \in N} \underline{a}_i$, then*

(6) $$u_i(a') > u_i(a) \text{ for all } i \in N,$$

where we denote $a = (p,...,p)$ *and* $a' = (p',...,p')$. *Suppose also that*

(7) $$u_i(\underline{a}_i, a^l_{-i}) \leq u_i(a^{l+1}) \text{ for all } i \in N \text{ and } l \in \{0,...,L-1\}.$$

Then, $\hat{m}$ *is an SPE if and only if the inequalities in (5) hold and* $\hat{m} = \overline{a}$.

Proof: Theorem 1 showed the necessity of the inequalities in (5). Suppose that $\hat{m}$ is a PNE, where we assume $\min_{i \in N} \hat{m}_i \geq \max_{i \in N} \underline{a}_i$. Let $l \in \{0,...,L\}$ denote the integer specified by

$$\beta^{l-1} - \frac{\overline{p}}{L} < \min_{j \in N} \hat{m}_j \leq \beta^{l-1}.$$

(Note from the definition of $l(\cdot)$ that $l(\beta^l) = l - 1$ holds.) Consider an arbitrary player $i \in N$. Suppose $\hat{m}_i > \min_{j \in N} \hat{m}_j$. Due to the cautious commitment rule $g = g^*$, player i can make their actions close to β^l, and therefore, can save their cost of action selections without changing the other players' promises. Due to their prosocial motives, (5), (7), and the specification of $g = g^*$, player i is willing to announce the message β^l and sincerely select the corresponding promise β^{l-1}. From these observations, a single player $j \in N$ must exist in this case, where we have

$$g_j(m) = \hat{m}_j \leq \beta^{l-1},$$

and

$$\hat{m}_i = \beta^l \text{ and } g_i(m) = \beta^{l-1} \text{ for all } i \neq j.$$

By replacing $\hat{m}_j$ with $m_j = \beta^l$, this single player j can change any other player's promise from β^{l-1} to β^l, and therefore, we have $g(m_j, \hat{m}_{-j}) = a^l$ instead of $g(\hat{m}) = (\hat{m}_j, a_{-j}^{l-1})$. From (7), player j is willing to announce $m_j = \beta^l$ instead of $\hat{m}_j$. This is a contradiction. Hence, no player announces greater than their minimal upper limit, that is, $\hat{m}_i = \hat{m}_1$ must hold for all $i \in N$.

Suppose that there exists $p < \overline{p}$ such that $\hat{m}_i = p$ for all $i \in N$. Note that

$$g_i(\hat{m}) = p \text{ for all } i \in N,$$

and there exists $l \in \{1, ..., L\}$ such that

$$\beta^{l-1} \leq p < \beta^l.$$

Due to the specification of $g = g^*$ and the prosocial motives, each player is willing to enhance their announcement from p to β^l because this enhancement never changes the promises of all players. This is a contradiction. Hence, we have shown that if $\hat{m}$ is a PNE and $\min_{i \in N} \hat{m}_i \geq \max_{i \in N} \underline{a}_i$, then $\hat{m} = \overline{a}$ must hold.

We can see that $\hat{m} = \overline{a}$ is a PNE as follows. From (7), it follows that for every $i \in N$ and $m_i \in [0, a_i^{L-1})$,

$$u_i(g(\overline{a})) = u_i(a^L) > u_i(m_i, a_{-i}^{L-1}) = u_i(g(m_i, \overline{a}_{-i})).$$

From (6), it follows that for every $i \in N$ and $m_i = p \in [a_i^{L-1}, a_i^L)$,

$$u_i(g(\overline{a})) = u_i(a^L) > u_i(p, ..., p) = u_i(g(m_i, \overline{a}_{-i})).$$

These inequalities imply the PNE property of $\hat{m} = \overline{a}$. Hence, we have shown that $\hat{m} = \overline{a}$ is the unique SPE with sovereignty protection and satisfies $\min_{i \in N} \hat{m}_i \geq \max_{i \in N} \underline{a}_i$.

Q.E.D.

The inequalities in (6) imply that a greater common cooperation level brings better welfare to every player. The inequalities in (7) imply a cautious sense of the reciprocal principle in that "I will do better if you will do slightly better," that is, every player is willing to keep pace with all other players if all other players raise the level of coordination by one grid $\bar{p}/L$. The inequalities in (7) and the presence of prosocial motives are necessary for proving that $\hat{m}=\bar{a}$ is the unique PNE. We will see more implications on (6) and (7) in the next section.

To understand Theorem 2, it might be helpful to consider a symmetric example where for every $i \in N$,

$$u_i(a) = f(a_i) + \sum_{j \neq i} a_j \quad \text{for all } a \in A,$$

and $u_i(a)$ is decreasing in $a_i \in [0,\infty)$ and increasing in $a_j \in [0,\infty)$ for all $j \neq i$. Note that $\underline{a}_i = 0$ for all $i \in N$. Fix an arbitrary $L \geq$, and assume that

(8) $$f(0) - f(\bar{p}) < \frac{\delta(n-1)\bar{p}}{1-\delta},$$

(9) $$f'(p) + n - 1 > 0 \quad \text{for all } p \in [0, \bar{p}],$$

and

(10) $$f(0) - f(\bar{p}) \leq \frac{(n-1)\bar{p}}{L}.$$

The inequalities in (8), (9), and (10) guarantee the inequalities in (5), (6), and (7), respectively. From Theorem 2, with the adoption of the cautious commitment rule, full cooperation $\hat{m} = \bar{a}$ is the unique SPE that protects player sovereignty. Note that all these inequalities are satisfied if the number of players n is large enough. Greater n means that more stakeholders are on the committee. Encouraging participation is the most effective way to achieve cooperation, because it makes the benefit from the other players' one-grid increases greater.

VII. Global Warming

This section investigates global warming as an application of this study. We explain that adopting the cautious commitment rule is crucial for resolving the tragedy of the global commons that all countries in the world have long faced. We clarify the circumstances under which the incentives in cooperative achievement are slightly constrained by showing how the inequalities in (6) and (7) in Theorem 2 are restrictive.

VII.A. The Model

Each country (player) voluntarily sets a carbon price for its citizens to reduce their CO_2 emissions. Let $\sigma_i \in (0,\infty)$ denote the population of each country $i \in N$. In reducing its CO_2 emissions by $x_i = x_i(t) \in [0,\infty)$ in period t, each country $i \in N$ needs to spend an economic cost given by $r_i(x_i) \in R$. We specify

$$r_i(x_i) = \frac{k_i x_i^2}{\sigma_i} \quad \text{for all } x_i \in [0,\infty),$$

where $k_i \in (0,\infty)$ denotes country $i's$ economic cost indicator in technological and ethical senses. The economic cost per capita in country i is expressed by

$$\frac{r_i(x_i)}{\sigma_i} = k_i(\frac{x_i}{\sigma_i})^2 .$$

Since it is decreasing in σ_i, each country i has a scale economy in population size σ_i. More importantly, it has a scale diseconomy in emissions reduction per capita ${x_i}/{\sigma_i}$.

The CO_2 emissions reduction in each period has a long-term impact on worldwide welfare. If worldwide CO_2 emissions are reduced by $\sum_{j\in N} x_j(t)$ in period t, each country $i \in N$ enjoys their respective environmental benefits $d_j \gamma^{\tau-t} \sum_{j\in N} x_j(t)$ in each future period $\tau \geq t$, where $\gamma \in (0,1)$, and $d_j \in (0,\infty)$ denotes the indicator of country $i's$ environmental interest. The world (the committee, or the COP) enjoys the sum of the environmental benefits across all countries $(\sum_{j\in N} d_j)\gamma^{\tau-t} \sum_{j\in N} x_j$ in each future period $\tau \geq t$.

The long-term impact of each country $i's$ emissions reduction $x_i(t)$ on worldwide welfare is expressed by

$$x_i(t)(\sum_{j\in N} d_j)\{1+\gamma\delta+(\gamma\delta)^2+\cdots\} - r_i(x_i(t))$$

$$= (\sum_{j\in N} d_j)\frac{x_i(t)}{1-\gamma\delta} - \frac{k_i x_i(t)^2}{\sigma_i} .$$

Hence, the worldwide optimal emissions reduction level of each country $i \in N$ is given by the solution of $\max_{x_i}[(\sum_{j\in N} d_j)\frac{x_i}{1-\gamma\delta} - \frac{k_i x_i^2}{\sigma_i}]$, that is,

$$x_i^* \equiv \frac{\sigma_i \sum_{j\in N} d_j}{2k_i(1-\gamma\delta)} .$$

Each country i voluntarily sets a domestic carbon price $a_i(t) = a_i \in (0, \infty)$ for the emissions of its citizens in each period t. Each citizen selects their reduction levels to minimize the economic cost $r_i(x_i)$ minus the savings on carbon price payments $a_i x$, that is, selects the solution of $\min_{x_i}[r_i(x_i) - a_i x_i]$, which is given by

$$\tilde{x}_i(a_i) \equiv \frac{\sigma_i a_i}{2k_i}.$$

We will define the *worldwide optimal carbon price* as

$$p^* \equiv \frac{\sum_{j \in N} d_j}{1 - \gamma\delta}.$$

Importantly, if any country i sets its domestic carbon price equal to this common price p^*, it can achieve the worldwide optimal reduction level, that is,

$$x_i^* = \tilde{x}_i(p^*) \quad \text{for all} \quad i \in N.$$

In this sense, the commonalty of the carbon prices across countries is crucial.

Since each country i considers only its own welfare, its instantaneous payoff in each period t can be the same as

$$d_i \sum_{\tau=0}^{t-1} (\gamma\delta)^\tau \sum_{j \in N} x_j(t-\tau) - r_i(x_i(t)),$$

where $\sum_{\tau=0}^{t-1} (\gamma\delta)^\tau \sum_{j \in N} x_j(t-\tau)$ denotes the stock of the previous and current emissions reductions. Moreover, note that the long-term impact of the emissions reduction $x_i(t) = x_i$ by country i in period t on its own welfare is expressed by

$$\frac{d_i}{1-\gamma\delta} x_i - r_i(x_i).$$

Hence, country i can maximize its own welfare by setting its carbon price equal to

$$\underline{a}_i \equiv \frac{d_i}{1-\gamma\delta},$$

where the resultant reduction level is given by

$$\underline{x}_i \equiv \tilde{x}_i(\underline{a}_i) = \frac{\sigma_i d_i}{2k_i(1-\gamma\delta)}.$$

Note that $\underline{a}_i < p^*$ and $\underline{x}_i < x_i^*$ for all $i \in N$.

Based on the above emissions scenario, we can specify each country $i's$ instantaneous payoff in the associated component game as

(11) $$u_i(a) = \frac{d_i}{1-\gamma\delta}\sum_{j\in N} x_j(a_j) - r_i(x_i(a_i)) \text{ for all } a \in A.$$

Clearly, $u_i(a)$ is decreasing in $a_i \in [\underline{a}_i, \infty)$, increasing in $a_i \in [0, \underline{a}_i)$, increasing in a_j for all $j \neq i$, and $\underline{a}_i$ is strictly dominant. Hence, we can regard the model of global warning in this section as an application of the analysis on the previous sections. As a consequence of Theorem 2, we can expect the adoption of the cautious commitment rule to play the crucial role in resolving the global warning issue.[7]

VII.B. Discussion

[7] This study implicitly assumes that carbon leakage caused by carbon price disparities across countries is suppressed by the introduction of Carbon Border Adjustment Mechanism (Gollier and Tirole, 2015; Blanchard and Tirole, 2021). With this, we can assume without any problems that each country's economic cost is independent of the other countries' carbon prices.

We investigate how restrictive the inequalities in (6) and (7) are in the model of global warming. We set the cooperative common carbon price $\bar{p}$ arbitrarily, where we assume $\bar{p} \geq \max_{i \in N} \underline{a}_i$.

1. Inequalities in (6)

We investigate the inequalities in (6), which implies that a greater common carbon price must bring better welfare to every country. From (11), if $a_i = p$ for all $i \in N$, we have

(12) $$u_i(a) = \frac{d_i p}{2(1-\gamma\delta)} \sum_{j \in N} \frac{\sigma_j}{k_j} - \frac{\sigma_i p^2}{4k_i} \text{ for all } i \in N.$$

By maximizing the right-hand side of (12) with respect to p, we have

$$p_i^{**} \equiv \left(\frac{d_i k_i}{\sigma_i}\right) \frac{\sum_{j \in N} \frac{\sigma_j}{k_j}}{1-\gamma\delta}.$$

Proposition 2: *The cooperative common carbon price* $\bar{p}$ *satisfies the inequalities in (6) if and only if*

$$\bar{p} \leq \min_{i \in N} p_i^{**} \equiv \{\min_{i \in N} \frac{d_i k_i}{\sigma_i}\} \frac{\sum_{j \in N} \frac{\sigma_j}{k_j}}{1-\gamma\delta}.$$

Moreover, $\bar{p} = p^*$ *satisfies the inequalities in (6) if and only if*

(14) $$\frac{d_i k_i}{\sigma_i} = \frac{d_1 k_1}{\sigma_1} \textit{ for all } i \in N.$$

Proof: Note that the right-hand side of (12) is increasing in p if and only if $p < p_i^{**}$. Hence, $\bar{p}$ satisfies (6) if and only if the equalities in (14) hold. Moreover, note that

$$[p^* = \min_{i \in N} p_i^{**}] \Leftrightarrow [p_i^{**} = p_1^{**} \text{ for all } i \in N]$$

$$\Leftrightarrow [\frac{d_i k_i}{\sigma_i} = \frac{d_1 k_1}{\sigma_1} \text{ for all } i \in N],$$

which implies the latter part of this proposition.

Q.E.D.

The committee can achieve the worldwide optimal carbon price p^*, and therefore, achieve the worldwide optimal emissions reduction level $\sum_{i \in N} x_i^*$, if and only if the equalities in (14) hold, that is, the indicator of environmental benefit per capita $\frac{d_i}{\sigma_i}$ multiplied by the indicator of economic cost k_i is the same across all countries. Otherwise, $\min_{i \in N} p_i^{**}$ is the maximal common carbon price that the world can achieve under state sovereignty, that is, under the Westphalian regime.

In this case, importantly, the country with the lowest $\frac{d_i k_i}{\sigma_i}$ will prevent more cooperative agreements on common carbon pricing. Given the homogeneity in k_i across countries, if a country has the lowest indicator of environmental benefit per capita $\frac{d_i}{\sigma_i}$, that is, it has the lowest environmental interest, it will not be attracted to achieve higher reduction goals.

More importantly, if a country i has a low economic cost indicator k_i, it will not be attracted to achieve higher reduction goals due to the scale diseconomy in emissions reduction per capita. In this case, its citizens are interested in lowering carbon prices, because they can reduce emissions without the help of carbon pricing. This may result in disruption of international cooperation.

Each country $i's$ economic cost indicator k_i reflects the level of its emissions reduction technology as well as the level of ethical propensity possessed by its citizens. In this respect, we can see that the spread of new emissions reduction technology around the world can lead to promoting better international cooperation. If a developed country possesses new emissions reduction technologies, the incentive constraint implied by (6) lowers the common carbon price that can be coordinated across countries, which in turn reduces the welfare of this developed country.

To overcome this tragedy, it is quite important for developed countries to support developing countries in introducing new emissions reduction technologies. In particular, the committee should consider using the Green Climate Fund to support technology transfers. Such assistance can promote not only international cooperation but also incentives in R&D investments for environmental technologies, accelerating the complete resolution of global warming due to technological progress.

To understand this point, let us revisit the example in Section II, and replace the cost of production per unit from 12 to 50. However, the cost 50 is greater than the social benefit $11 \times 3 = 33$. Therefore, no provision is socially optimal. Suppose that player 1 discovers a new technology to reduce the cost from 50 to 12. Then, player 1 still has no incentive to provide public goods because the cost 12 is greater than the private benefit 11. At this rate, player 1 has no incentive for investing in technological innovation. However, if this new technology is shared with every player, the adoption of the cautious commitment rule will enable the socially optimal public good provision, as explained in Section II. This technology sharing increases player 1's welfare from 0 to 210, which implies that player 1 is willing to invest in technological innovation assuming technology is shared.

However, things are different if the cause of low economic cost indicator is high ethical propensity of its citizens rather than its high emissions reduction technology. In this case, the disparity in economic cost indicator across countries can hardly be corrected. In contract with the disparity in emissions reduction technology, the improvement of ethical propensity in a country will inevitably disturb the pace of international cooperation.

This point is related to the licensing effect in experimental psychology, that is, the psychological tendencies to feel that ethical consumption activities have been granted permission to take non-social behaviors such as voting against environmental taxation thereafter (Monin and Miller, 2001; Mazar, et al., 2008; Mazar and Zhong, 2010). This is also related to the conflict between intrinsic motivations and extrinsic motivations, where the increase of environmental taxation as extrinsic motivations reduces intrinsic motivations to voluntarily refrain from consuming energy (Bénabou and Tirole, 2003).

Both cases are deeply associated with prejudices such as the bad image of punishment on taxation. Hence, it is quite important to dispel prejudice about environmental taxation or carbon pricing from the public, rather than looking for alternative approaches such as promoting ethical boycotts in economic activity.

2. Inequalities in (7)

Next, we investigate the inequalities in (7), which imply that every country is willing to keep pace with all other countries if all other countries raise the level of coordination on common carbon pricing by one grid $\bar{p}/L$. In the model of global warming, we can rewrite (7) as follows: for every $i \in N$ and $l \in \{0,...,L-1\}$,

(15) $$r_i(\underline{x}_i) - r_i(\tilde{x}_i(a_i^{l+1})) \leq \frac{d_i}{1-\gamma\delta}[\tilde{x}_i(a_i^l) - \underline{x}_i + \sum_{j\in N}\{\tilde{x}_j(a_j^{l+1}) - \tilde{x}_j(a_j^l)\}]\,.$$

This subsection additionally assumes that countries are symmetric, and then normalizes $(d_i, k_i, \sigma_i)_{i\in N}$ as

$$d_i = k_i = 1 \text{ and } \sigma_i = 1/n \text{ for all } i \in N\,.$$

Hence, the total population size is normalized to unity, that is, $\sum_{i\in N}\sigma_i = 1$, regardless of the number of participants in the committee.

For every $i \in N$, we have

$$\tilde{x}_i(a_i) \equiv \frac{a_i}{2n},\; x_i^* \equiv \frac{1}{2(1-\gamma\delta)},\; \underline{x}_i = \frac{1}{2n(1-\gamma\delta)},$$

$$p_i^* = p^* = \frac{n}{1-\gamma\delta}, \text{ and } \underline{a}_i \equiv \frac{1}{1-\gamma\delta}.$$

Hence, we can rewrite (15) as

$$\{\frac{n(l+1)}{L} + 1\}^2 \leq 2n(\frac{l+1}{L} + \frac{n-1}{L}) \text{ for all } l \in \{0,...,L-1\}\,.$$

By letting $w = \frac{l+1}{L}$, we can approximate the inequalities in (7), that is, the inequalities in (15), by the following inequalities:

(16) $$(nw+1)^2 - 2n(w + \frac{n-1}{L}) \leq 0 \text{ for all } w \in [0,1]\,.$$

Proposition 3: *The inequalities in (16) hold if and only if*

$$n \geq 1/2 + \sqrt{L + 1/4}.$$

Proof: The left-hand side of (16) is maximized at $w = 1/n$. Hence, we can replace the inequalities in (16) with $L \leq n(n-1)$, that is, $n \geq 1/2 + \sqrt{L + 1/4}$.

Q.E.D.

The greater L is, the more restrictive the inequalities in (7) are. The greater n is, the less restrictive the inequalities in (7) are. Hence, if the world is subdivided into smaller and more homogeneous regions and all these regions join the committee on behalf of their countries, the world can more easily overcome the tragedy of the global commons.

However, Proposition 2 in Section VII.B.1 indicates that if the world is subdivided into smaller but more heterogeneous regions, it will be more difficult for the world to build an agreement on high-level cooperation goals. Hence, the more the world is subdivided, the more likely each region is incentivized to achieve the cooperation goals and the less likely the world agrees on high-level cooperation goals.

VIII. Further Improvement

This study requires that commitment rules must be sustainable, that is, they must be robust against renegotiation. In this respect, the cautious commitment rule is superior to the common commitment rule. However, for a reason different from the viewpoint in the previous sections, the cautious commitment rule may remain vulnerable to renegotiation. Therefore, further improvements in the cautious commitment rule are needed as follows.

Let us reconsider the model of public good provision in Section II. Suppose that players 1 and 2 announce 10 each, while player 3 announces zero. Moreover, suppose that players 1 and 2 expect player 3 to stick to these zero announcements even if it is far from equilibrium behavior. According to the cautious commitment rule, players 1 and 2 would prefer to announce 1 instead of 10 due to their selfish motives, because they can save the cost by decreasing their provisions from 9 to 0. This reduces the morale of all players and causes only low cooperation to be achieved. In this case, if players 1 and 2 agree that the cause of this demoralization is player 3's insistence on uncooperative attitude, the possibility of their renegotiation with each other excluding player 3 may not be ruled out even in the cautious commitment rule.

To address this issue, we modify the cautious commitment rule in the following manner. Each player $i \in \{1,2,3\}$ produces less than m_i by the number of players who announce lower, that is,

$$g_i(m) = m_i \qquad \text{if } m_j \geq m_i \text{ for all } j \in \{1,2,3\},$$

$$g_i(m) = m_i - 1 \qquad \text{if there exists } j \neq i \text{ such that } m_j < m_i \text{ and } m_h \geq m_i \text{ for } h \notin \{i, j\},$$

and

$$g_i(m) = \max[m_i - 2, 0] \quad \text{if } m_j < m_i \text{ for all } j \neq i.$$

That is, a player who announces two or more can reduce their production by the number of other players who announce less. According to the same logic as in Section II, the full cooperation $\hat{m} = (10,10,10)$ is a PNE in the message game associated with the modified cautious commitment rule. The following proposition shows that there exists no other PNE.

Proposition 4: *In the model of Section II, the full cooperation* $\hat{m} = (10,10,10)$ *is the unique PNE in the message game associated with the modified cautious commitment rule.*

Proof: Suppose that $m \neq (10,10,10)$ is a PNE. Without loss of generality, we assume $m_1 \geq m_2 \geq m_3$. Note $m_3 < 10$. Clearly, player 3 produces $a_3 = m_3$. Due to the selfish and prosocial motives, player 2 must announce $m_2 = m_3 + 1$ and produce $a_2 = m_2 - 1 = m_3$. Moreover, due to the selfish and prosocial motives, if $m_3 < 9$, player 1 must announce $m_1 = m_3 + 2$ and produce $a_1 = m_1 - 2 = m_3$, while if $m_3 = 9$, player 1 must announce $m_1 = m_2 = m_3 + 1 = 10$ and produce $a_1 = m_1 - 1 = m_3 = 9$. In this case, player 3 prefers announcing $m_3 + 1$ instead of m_3 and producing $m_3 + 1$ instead of $a_3 = m_3$ with the additional net cost $12 - 11 = 1$, because player 2 increases their production levels from m_3 to $m_3 + 1$ and brings player 3 the additional benefit $11 > 1$. This contradicts the PNE property. Hence, $m_3 = 10$, that is, $m_1 = m_2 = m_3 = 10$, must hold.

Q.E.D.

Importantly, the modified cautious commitment rule is robust against renegotiation for this section's reason. Suppose again that players 1 and 2 announce 10 each, while player 3 announces zero. Suppose also that players 1 and 2 expect player 3 to stick to the zero announcement. In the modified cautious commitment rule, player 1 does not want to reduce their announcement, that is, does not want to save their net cost that is at most $(12-11)\times 10 = 10$, because player 2 lowers another unit of production level and reduces player 1's payoff by 11 (>10). Hence, even if someone deliberately adopts an uncooperative attitude, the rest can still maintain high morale and sufficient achievement of cooperation.

To be more precise, even if player 3 sticks to announcing 0, both players 1 and 2 are incentivized to produce greater levels as follows. Suppose that both players 1 and 2 announce 0, while player 3 purposefully sticks to announcing 0. Then, player 2 is willing to announce 2 instead of 0, still producing nothing. Given $(m_2, m_3) = (2, 0)$, player 1 is willing to announce 3 instead of 0, which now makes both players 1 and 2 produce 1 instead of 0. Given $(m_1, m_3) = (3, 0)$, player 2 is willing to announce 4 instead of 2, which makes both players 1 and 2 produce 2 instead of 1. Similarly, we can see that players 1 and 2 are willing to increase their production levels up to 10 as if they climb the stairs. Hence, with this modification, we can adopt the same logic of climbing-stair as the cautious commitment rule so that even if a player purposefully takes an uncooperative attitude and the other players recognize this fact, they will continue to have high morale and incentives to achieve cooperation.

From these observations, the modified cautious commitment rule is sustainable in a stricter sense, that is, it is also robust against renegotiation caused by decline in morale. The principle of "I will do better if you will do slightly better" is nested in the modified cautious commitment rule, which makes the rule more robust against renegotiation.

IX. Conclusion

Problem of the commons typically exist across multiple independent social groups, such as multiple countries, regions, communities, or individuals. Each social group may be self-governed in terms of solving their internal problems. However, any single group alone cannot solve such inter-group problems on behalf of the whole. Hence, the committee needs to be established as sustainable polycentric governance (Ostrom, 2010), but without hierarchy or social exclusion, to integrate all members' potential abilities without infringing on their sovereignty or compromising their self-governance. This study proposes a method for achieving such governance maintenance.

The committee requires each member (group) to express not a single promise but a scope of feasible promises. It is at the committee's discretion which promise within each member's scope is demanded. However, to protect their sovereignty, the committee must prohibit future retaliation against any member for expressing uncooperative scopes. Thus, it will be entirely at each member's discretion to decide what their scope is. Furthermore, for sustainable polycentric governance, the method of determining promises, that is, the commitment rule, must be robust against renegotiations to unavoidably eliminate members who purposefully remain uncooperative.

This study proposes the cautious commitment rule and its modification to solve the free-rider problem with well-balanced sovereignty protection and rule sustainability. The common commitment rule, that is, the reciprocal principle of "I will if you will," which many economics researchers have proposed as a better solution to global warming than the current regime, cannot balance sovereignty protection and rule sustainability. The cautious commitment rule, instead, incorporates the cautious version of reciprocal principle of "I will do better if you will do slightly better," or the climbing-stair method, and can solve the free-rider problem entirely by achieving cooperation as a unique equilibrium behavior while adhering to both sovereignty protection and rule sustainability.

The cautious commitment rule suggests a concrete way to achieve sustainable economic developments while maintaining fair social inclusion. Society must be resilient, that is, it never gets confused because of the presence of problematic players, and is open to any such player at any time so that they can return to a cooperative relationship. The climbing-stair method, which underpins the cautious commitment rule, successfully incentivizes all players to return to a cooperative relationship, no matter how uncooperative they are in the current situation.

As future research, it is important to scrutinize whether this study's arguments hold for a wider range of models. In particular, the model of global warming in Section VII should be extended to elucidate further policy implications because global warming is one of the most important cases that have defied spontaneous resolution over an extended period of time. Specific extensions include considering imperfect monitoring of whether promises have been kept, incomplete information about dynamic climate change and economic fluctuations, and incentives in environmental investment and funding.

The logic behind the unique achievement of cooperation relies on players' lexicographical prosocial motives. However, depending on contexts, various antisocial motives in players' minds can adversely affect their decision-making. Hence, to implement the cautious commitment rule in society, creating epistemological frames in which the prosocial motives dominate adversarial motives will play an auxiliary role.[8] It

[8] Matsushima (2021) investigates an epistemological aspect in implementation theory on this line.

is hoped that further consideration will be given to the role of such ethical motives in solving global commons.

This study presents a new direction for problem-solving as an alternative to state control and marketization. The atmosphere is non-excludable and cannot be privately or club-owned. In this case, it is important to change the rule of the game to common carbon pricing, taxation, and subsidies associated with the cautious commitment rule, rather than a war of attrition over emission allowances such as the pledge-and-review approach. The discussion of heterogeneity across countries in Section VII.B has shown that with this change, technological innovation in a developed country is facilitated by allowing all countries to benefit from this innovation. This incentive property is in line with the socialist principles such as “from each according to his ability, to each according to his needs” (Marx, 1890/1971; Roemer, 1986; Osborne and Rubinstein, 2020), rather than the capitalist principle of market competition. The cautious commitment rule is a unique decentralized system that protects sovereignty, does not rely on centralized control, does not exclude anyone, and can quell various social conflicts that cannot be solved by the market principle. The adoption of the cautious commitment rule will result in a successful social implementation of the socialist principle from the liberalist viewpoints rather than ending up as a mere idealist utopia.

University of Tokyo, Department of Economics

References

Abreu, Dilip, Prajit K. Dutta, and Lones Smith, "The Folk Theorem for Repeated Games: A NEU Condition," Econometrica, 62 (1994), 939–948. https://doi.org/10.2307/2951739

Abreu, Dilip, and Hitoshi Matsushima, "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," Econometrica, 60 (1992), 993–1008. https://doi.org/10.2307/2951536

Barrett, Scott, "Self-Enforcing International Environmental Agreements," Oxford Economic Papers, 46 (1994), 878–894. https://doi.org/10.1093/oep/46.Supplement_1.878

Bénabou, Roland, and Jean Tirole, "Intrinsic and Extrinsic Motivation," Review of Economic Studies 70 (2003), 489–520.

Blanchard, Olivier, and Jean Tirole. Major Future Economic Challenge. Republique Francaise, 2021.

Cooper, Richard N., "The Case for Charges on Greenhouse Gas Emissions," Harvard Project on International Climate Agreements, Belfer Center for Science and International Affairs, Harvard Kennedy School, Discussion Paper 08-10, 2008.

Cramton, Peter, David J.C. MacKay, Axel Ockenfels, and Steven Stoft, Global Carbon Pricing: The Path to Climate Cooperation. (Cambridge, MA: MIT Press, 2017). https://doi.org/10.7551/mitpress/10914.001.0001

Cramton, Peter, Axel Ockenfels, and Steven Stoft, "An International Carbon-Price Commitment Promotes Cooperation," Economics of Energy & Environmental Policy, 4 (2015), 51–64. https://doi.org/10.5547/2160-5890.4.2.aock

Cramton, Peter, and Steven Stoft, "Global Climate Games: How Pricing and a Green Fund Foster Cooperation," Economics of Energy & Environmental Policy, 1 (2012), 125–36. https://doi.org/10.5547/2160-5890.1.2.9.HELM

Dutta, Bhaskar, and Arunava Sen, “Nash Implementation with Partially Honest Individuals,” Games and Economic Behavior, 74 (2012), 154–169. https://doi.org/10.1016/j.geb.2011.07.006

Farrell, Joseph, and Eric Maskin, “Renegotiation in Repeated Games,” Games and Economic Behavior, 1 (1989), 327–360. https://doi.org/10.1016/0899-8256(89)90021-3

Fudenberg, Drew, and Eric Maskin, “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information,” Econometrica, 54 (1986), 533–556. https://doi.org/10.2307/1911307

Gollier Christian G., and T. Tirole, “Negotiating Effective Institutions against Climate Change,” Economics of Energy and Environmental Policy, 4 (2015), 5–27. https://doi.org/10.5547/2160-5890.4.2.cgol

Harrison, Rodrigo, and Roger Lagunoff, “Dynamic Mechanism Design for a Global Commons,” International Economic Review, 58 (2017), 751–782.

Harstad, Bard, “Climate Contracts: A Game of Emissions, Investments, Negotiations, and Renegotiations, Review of Economic Studies, 79 (2012), 1527–1557.

Lerner, Josh, and Jean Tirole, “The Economics of Technology Sharing: Open Source and Beyond,” Journal of Economic Perspectives 19 (2005), 99–120.

MacKay, David J.C., Peter Cramton, Axel Ockenfels, and Steven Stoft, “Price Carbon — I Will If You Will,” Nature, 526 (2015), 315–16. https://doi.org/10.1038/526315a

Marx, Karl, Critique of the Gotha Programme, Progress Publishers, Moscow (1890/1971).

Maskin, Eric, and Jean Tirole, “Markov Perfect Equilibrium I: Observable Actions,” Journal of Economic Theory, 100 (2001), 191–219.

Matsushima, Hitoshi, "Behavioral Aspects of Implementation Theory," Economics Letters, 100 (2008a), 161–164. https://doi.org/10.1016/j.econlet.2007.12.008

Matsushima, Hitoshi, "Role of Honesty in Full Implementation," Journal of Economic Theory, 139 (2008b), 353–359. https://doi.org/10.1016/j.jet.2007.06.006

Matsushima, Hitoshi, "Epistemological Implementation of Social Choice Functions," Discussion paper UTMD-13, University of Tokyo, 2021.

Mazar, Nina, On Amir, and Dan Ariely, "The Dishonesty of Honest People: A Theory of Self-Concept Maintenance," Journal of Marketing Research, 45 (2008), 633–634.

Mazar, Nina, and Cheng-Bo Zhong, "Do Green Products Make us Better People?," Psychological Science, 21 (2010), 494–498.

Monin, Benoit, and Dale T. Miller, "Moral Credentials and the Expression of Prejudice," Journal of Personality and Social Psychology, 81 (2001), 33–43.

Nordhaus, William, Managing the Global Commons: The Economics of Climate Change. (Cambridge, MA: MIT Press, 1994).

Nordhaus, William, "Life After Kyoto: Alternative Approach to Global Warming," mimeograph, Yale University, 2005.

Nordhaus, William, Climate Casino. (New Heaven, CT: Yale University Press, 2013).

Nordhaus, William, "Climate Clubs: Overcoming Free-Riding in International Climate Policy," American Economic Review, 105 (2015), 1339–70. https://doi.org/10.1257/aer.15000001

Osborne, Martin J., and Ariel Rubinstein, Models in Microeconomic Theory. (Cambridge, UK: Open Book Publishers, 2020).

Ostrom, Elinor, Governing the Commons: The Evolution of Institution for Collective Action. (Cambridge: UK: Cambridge University Press, 1990).

Ostrom, Elinor, "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," American Economic Review, 100 (2010), 641–72. https://doi.org/10.1257/aer.100.3.641

Roemer, John E., "Equality of Resources Implies Equality of Welfare," Quarterly Journal of Economics, 101 (1986), 751–84 https://doi.org/10.2307/1884177

Roemer, John E., "Kantian Equilibrium," Scandinavian Journal of Economics, 112 (2010), 1–24. https://doi.org/10.1111/j.1467-9442.2009.01592.x

Shapiro, Carl, and Hal R. Varian, Information Rules. (Harvard Business School Press, 1999).

Stern, Nicholas, The Economics of Climate Change: The Stern Review. (New York: Cambridge University Press, 2007).

Stiglitz, Joseph E., "A New Agenda for Global Warming," Economists' Voice, 3 (2006), 7. https://doi.org/10.2202/1553-3832.1210

Stoft, Steven, Carbonomics: How to Fix the Climate Change and Charge It to OPEC. (Nantucket, MA: Diamond Press, 2008).

Tirole, Jean, Economics for the Common Good. (New Jersey: Princeton University Press, 2017).

Victor, David, The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming. (Princeton, NJ: Princeton University Press, 2007).

Wagner, Gernot, and Martin Weitzman, Climate Shock: The Consequences of a Hotter Planet. (Princeton, NJ: Princeton University Press, 2015).

Weitzman, Martin L., "Can Negotiating a Uniform Carbon Price Help to Internalize the Global Warming Externality?" Journal of the Association of Environmental and Resource Economists, 1 (2014), 29–49. https://doi.org/10.1086/676039