UTMD Working Paper
The University of Tokyo
Market Design Center

UTMD-058

Preventing Global Catastrophes

Hitoshi Matsushima
University of Tokyo

October 17, 2023

Preventing Global Catastrophes[1]

Hitoshi Matsushima[2]

October 17, 2023

Abstract

To prevent global catastrophes that would cause irreversible and enormous damage to the humanity and environment, we should not expect that only philanthropy and tacit collusion suffice without taking any measures. While global citizens are under a lack of strong coercive power, it is necessary to carefully design explicit negotiation procedures that make effective use of the limited social order. In doing so, it is necessary to design institutional rules that are robust against unforeseen circumstances where many citizens happen to be irrational and adhere to uncooperative attitudes. We show a possibility that, under a constraint of sovereignty protection, there are commitment rules in a global negotiation forum that can uniquely elicit incentives for cooperative behavior from agents while coping with such unforeseen circumstances.

Keywords: Global Catastrophe, Commitment rule, Robustness against Unforeseen Circumstances, Sovereignty Protection, Uniqueness

JEL Classification: C72, DD91, H41, H77, Q54

---

[1] Research for this study was financially supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government, as well as by the Center of Advanced Research in Finance, the Market Design Center, and the Chair of Social Common Capital at the University of Tokyo. We are grateful to Mr. Kenjiro Asami for his helpful comments and suggestions. All errors are mine.

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

1. Introduction

We examine the design of social institutions to prevent global catastrophes. The following recent experience of not making a desired progress in solving the climate change problem are behind the motivation of this study.

The climate change problem implies that excessive CO2 (or greenhouse gases) emissions will catastrophically and irreversibly damage global citizens' living environments, species diversity, and ecosystems (Nordhaus, 1994, 2005, 2013; Victor, 2001; Stern, 2007; Wagner and Weitzman, 2015; Cramton et al., 2017; Tirole, 2017; Blanchard and Tirole, 2021). The international community has been working for a long time to resolve this global issue. For example, the UN provided the Conference of the Parties (COP) as a forum for international negotiation about global emission reduction. The UN also formulated the Sustainable Development Goals (SDGs) that supports campaign activities that stimulate environmental concerns. Thereafter, global citizens became more concerned about future generations, species diversity, and ecology, as well as how much information and knowledge about environmental damage has been acquired. They also became more tolerant of their effort for contributing to emission reductions through their lifestyles and behavior patterns, such as recycling habits, consumption boycotts, and interest in ESG investing, as well as the technological progress of energy and circular engineering.

Despite a long period of awareness-raising activities, ongoing international negotiations, and a growing public interest in environmental issues, they have so far failed to produce sufficient results. For example, although the COP has been proceeding with negotiations, the negotiations are proving to be extremely difficult. Our bitter experience can be summerized:

1. Agents (citizens, countries, or companies) recognize climate change as a serious problem and have latent prosocial motives, but succumb to their selfish motives and are unable to escape the free-rider problem by relying solely on self-help.

2. Even if agents have a long-term relationship as members of the international community and are gradually able to monitor their emission reductions to a large extent albeit imperfectly, behavior patterns to sanction free riders were not generated spontaneously.
3. We sometimes fail to predict actual agents' reduction behaviors due to the emergence of unforeseen contingencies that are not considered in advance but influence their behavioral attitudes.
4. Agents change their own specific commitments midway through when such unforeseen circumstances arise. This change discourages other agents to reduce emissions one after another.

From the above, we do not have high hopes for tacit collusion to be spontaneously established and provide a solution to a global catastrophe. Surely, in a repeated game, a behavior pattern that voluntarily penalizes free riders can be described as a noncooperative equilibrium, even when monitoring is imperfect and the state fluctuates over time. Counting on this, the COP did not seriously attempt to install any explicit incentive scheme. However, in the repeated game, there are multiple heterogeneous equilibria, and no satisfactory foundation has been provided for which of them can be realized (Aumann and Shapley, 1976; Fudenberg and Maskin, 1986; Abreu, 1988; Farrell and Maskin, 1989; Fudenberg et al., 1994; Barrett, 1994, 2003; Kandori and Matsushima, 1998; Finus, 2001; Matsushima, 2004; Dutta and Radner, 2004, 2009; Dal Bó and Fréchette, 2018; Sugaya, 2022). The experience of the confusion surrounding climate change exposes this equilibrium selection as a serious problem that cannot be pegged. Both from theoretical and empirical standpoints, it is mere wishful thinking to believe that a convenient equilibrium will be chosen.

Nor is the cooperative game approach, which presupposes a broad-based binding agreement on players' activities, appropriate for this resolution. Rather, we need to limit the binding force to a scope that infringes as little as possible on the inherent sovereignty. Specifically, binding force should be limited at most to the extent that it is grounded in a

commonly recognized social order such as the Westphalian system. In order to make a broad agreement binding, we must use means of individual sanctions such as boycotts. An example in climate change is the response measure such as the Carbon Border Adjustment Mechanism (Nordhaus, 2015) to address the carbon leakage that makes the gain from free riding unduly high. The problem is that allowing such individual sanctions excessively may be in effect a violation of citizen sovereignty (Cramton et al., 2017). Hence, in terms of consistency with the protection of citizen sovereignty, there are limits to solving the free-rider problem through such individual sanctions.

Because of the limited availability of coercive force in the international community, the COP has long neglected to set up any explicit negotiation procedure with incentive concerns by taking the pledge-review approach. However, this neglect by the COP is wrong. We show that even under such limited enforcement power, it is possible to design institutions that are robust against unforeseen circumstances, preventing global catastrophes with the help of philanthropy and tacit collusion. We investigate an abstract single-period model of global catastrophes and demonstrate a method of designing an explicit negotiation procedure that satisfies the following requirements:

i. Agents behave cooperatively as unique equilibrium behavior.
ii. Even in unforeseen circumstances where non-negligible number of agents happen to be irrational and adhere to uncooperative attitudes, many of the remaining agents are still willing to behave cooperatively.
iii. Each agent is not forced by others to make decisions about the content of their own commitment. They can offer to change what they have committed at any time if necessary.

We assume that while agents are not forced to make specific commitments, they can commit in advance to a particular negotiation procedure to determine their commitments collectively, which we call a commitment rule, to the extent that they are in accordance with the social order. Here, the social order is defined as a combination of sovereignty

protection and adherence to commitments. That is, the combination of the norm that any agent is not forced to make commitments that they do not like, and the norm that any agent does not silently break their own commitment. This definition is modeled after the international order of the Westphalian system.

We consider how to discourage free riding by explicitly incorporating a mechanism to link each agent's tolerance for their own commitment to the other agents' commitments. We show a positive result, such that there exists a commitment rule that satisfies the above-mentioned requirements, i.e., the requirements i, ii, and iii.

Following McKay et al. (2015), we let agents announce their respective upper limit of commitments that they can tolerate, and then decide what they actually commit within the range below their upper limit (sovereignty protection). That is, McKay et al. (2015) proposed a commitment rule according to which each agent's actual commitment is tied to the other agents' upper limits. McKay et al. (2015) designed the rule so that as an agent raises their own upper limit, the actual commitment levels of the other agents would increase in tandem. Under adherence to commitments, McKay et al. (2015) proposed that this linkage would have a synergetic effect that discourages the temptation to free-ride and moves everyone toward cooperation. See also Cooper (2008), Cramton and Stoft (2012), Cramton, Ockenfels, and Stoft (2015), and Cramton et al (2017).

However, the commitment rule designed by McKay et al. (2015) is inadequate, because it fails to meet the requirements for appropriate handling of unforeseen circumstances where a non-negligible number of agents happen to be irrational and adhere to the uncooperative attitudes. Thus, a commitment rule must be redesigned so that even if such uncooperative agents exist, the remaining agents keep their commitments close to their upper limits. Moreover, a commitment rule must be redesigned so that even if such unforeseen circumstances arise, (a large proportion of) the remaining agents have no incentive to change their upper limits downward. Neither of these requirements is met by the rule of McKay et al. (2015).

We propose a new design method of commitment rule, which satisfies all the above requirements, as follows. Each agent's commitment is lowered as the number of

uncooperative agents increases. However, to deal with the unforeseen circumstances, we must always keep the lowering width sufficiently small. To reduce this lowering width, we will categorize the number of uncooperative agents, and set the commitment rule so that the lowering width depends only on which category the number of uncooperative agents belongs to. This design method makes it possible to keep the lowering widths as small as possible while dealing with the unforeseen circumstances. We show that if the number of agents is sufficiently large and they have prosocial motives in a minimal (i.e., lexicographical) sense, cooperative behavior can be explained as unique equilibrium behavior, and it is robust against the unforeseen circumstances.

Thus, the novelties of this paper are as follows:

1) The resolution of the free-rider problem in global catastrophes is considered through a design of commitment rules under sovereignty protection and adherence to commitments.
2) We demonstrate a design method of commitment rule to incentivize agents to behave cooperatively as unique equilibrium behavior and make their cooperation robust against unforeseen circumstances where a non-negligible number of agents happen to be irrational and adhere to the uncooperative attitudes.

We further investigate a dynamic in which agents repeatedly negotiate about catastrophe prevention. We do not consider dynamic resource management of the global commons such as Harstad (2012) and Harrison and Lagunoff (2017), but we instead formulate the dynamic as an repeated game whose component game is given by our single-period model with commitment rule. We assume that the next round of negotiation is inevitably postponed if someone violates the social order. We then argue that this inevitable postponement provides an incentive for rouge agents to voluntarily maintain the social order. We also consider the possibility that the next round of negotiation is artificially postponed if some agents behave uncooperatively. We then argue that this artificial

postponement serves to help a commitment rule to meet the robustness requirement against the unforeseen circumstances.

Moreover, we consider the possibility of artificially changing the shape of commitment rule depending on the past history of play. That is, if some agents are found to adhere to the uncooperative attitudes from the past history of play, the commitment rule will be temporally changed to cover only the remaining agents. We then argue that with such history-dependence, the uniqueness of cooperative behavior is restored even in the unforeseen circumstances.

The remainder of this paper is organized as follows. Section 2 defines the model and commitment rules. Section 3 shows the main theorems. Section 4 discusses about our results. Section 5 considers the dynamic model. Finally, section 6 concludes this study.

2. Commitment Rule

Let $N = \{1,...,n\}$ denote the set of all agents. Each agent $i \in N$ has a set of actions $A_i = [0,1]$ and a utility function $u_i : A \to R$, which is specified as:

$$u_i(a) = \sum_{j \in N} a_j - ca_i \text{ for all } i \in N \text{ and } a \in A,$$

where $A \equiv \underset{i \in N}{\times} A_i$, $a \equiv (a_i)_{i \in N} \in A$, c is a real number that is considered as the (constant marginal) cost, and

$$1 < c < n.$$

Each agent $i's$ action $a_i \in A_i$ implies their voluntary contribution to prevent global catastrophes, which has positive externality to all agents' welfares. Thus, $\sum_{j \in N} a_j$ expresses the expected gain (relative to the cost c) resulting from all agents' contributions.

These agents experience the following free-rider problem. As $c > 1$, in the strategic game defined as a triple $(N, A, (u_i)_{i \in N})$, any agent $i \in N$ prefers to select zero as a

dominant strategy. As $c < n$, each agent prefers to increase their action level if the other agents simultaneously increase their action levels by the same amount.

To overcome this free-rider problem, we explore a commitment rule $\alpha = (\alpha_i)_{i \in N}$, where $\alpha_i : M \to A_i$ for each $i \in N$, and $M \equiv \underset{i \in N}{\times} M_i$. Each agent $i \in N$ contains a set of messages $M_i \in [0,1]$. A message $m_i \in M_i$ announced by agent i defines the upper limit of commitments that they can tolerate. The action level $\alpha_i(m) \in A_i$ of agent i implies the commitment of action selection that agent i must keep. Each agent $i's$ commitment $\alpha_i(m)$ depends not only on their own message m_i but also on the other agents' messages m_{-i}.

Following McKay et al. (2015), we assume that $\alpha_i(m)$ is nondecreasing in the n-dimensional vector $m = (m_1, ..., m_n)$. We have in mind that the commitment rule should be set so that if an agent allows a higher commitment by raising their own upper limit, the commitments imposed on other agents are higher in tandem.

We assume adherence to commitments in that every agent $i \in N$ will keep their commitment $\alpha_i(m) \in A_i$. Therefore, if agents announce a message profile $m \in M$, the resultant action profile is given by $a = \alpha(m) \in A$. Notably, we can permit each agent to change their upper limit to lower their commitment at their discretion. They are forbidden only to silently breaking their commitment. If a player does not want to keep the commitment, they should just let everyone know by replacing their upper limit with a lower one.

To maintain the social order, each agent sacrifices their own self-interest to some degree. Each agent could make a higher commitment and raise the bar for other agents' commitments as well, while themselves benefiting selfishly by silently breaking their own commitment. However, if their agreements are violated in this manner, the social order is disturbed, and the future risk of conflict in their society generally increases. Therefore, we assume that when agents agree to a commitment rule, they are accepting adherence to commitments. See Subsections 4.3 and 5.1.1 for further discussions.

We have in mind that n is fixed sufficiently large, and therefore, $\frac{c}{n}$ is fixed close to zero. Since each agent earns the utility $n-c$ from the full cooperation in society, we can consider the catastrophe as a tremendous damage (large n) to each agent relative to their cost c.

Fix an arbitrary triple (ε, w, z), where $\varepsilon > 0$ is a positive real number, $w > 2$ is a positive integer, and $z \in \{1, ..., w-1\}$. For convenience of arguments, we assume that n is an integer multiple of w. We have in mind that ε is close to zero, w is large but less than n, and $\frac{z}{w}$ is close to zero.

Based on the triple (ε, w, z), we require a commitment rule α to satisfy the following four requirements: sovereignty protection (SP), virtual upper limits (VUL), uniqueness (U), and robustness (R). SP implies that each agent's commitment does not exceed their announced upper limit.

Sovereignty Protection (SP): For every $i \in N$ and $m \in M$,

$$\alpha_i(m) \leq m_i.$$

VUL implies that the commitment rule always makes each agent commit to a level close to their announced tolerance.

Virtual Upper Limits (VUL): For every $i \in N$ and $m \in M$,

$$\alpha_i(m) \geq m_i - \varepsilon.$$

We define the commitment game as a triple (N, M, v), where $v = (v_i)_{i \in N}$, $v_i : M \to R$ for each $i \in N$, and

$$v_i(m) = u_i(\alpha(m)) = \sum_{j \in N} \alpha_j(m) - c\alpha_i(m) \text{ for all } i \in N \text{ and } m \in M.$$

We assume that each agent announces the maximal best response. An interpretation is that while each agent is prosocial in a minimal (i.e., lexicographical) sense, their prosocial motives are always outweighed by their selfish motives. Another interpretation is that although an explicit "stick-carrot" mechanism is mounted, its effects are limited and can be only likened to the prosocial motives such as lexicographic preferences. See Subsection 4.4 for further discussions. Thus, a message profile $m \in M$ is said to be a Nash equilibrium in the commitment game if for every $i \in N$,

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

and

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

We define the maximal message profile as $\bar{m} \equiv (\bar{m}_i)_{i \in N} \in M$ by

$$\bar{m}_i = 1 \text{ for all } i \in N.$$

U implies that $\bar{m}$ is the unique Nash equilibrium and it achieves the full cooperation.

Uniqueness (U): The maximal message profile $\bar{m}$ is the unique Nash equilibrium in the commitment game, and

$$\alpha_i(\bar{m}) = 1 \text{ for all } i \in N.$$

We further define a Nash equilibrium for each subset of agents $\tilde{N} \subset N$ in the commitment game as a message profile $m \in M$ so that every agent who belongs to $\tilde{N}$ selects the maximal best response to m, whereas any other agent (irrationally) adheres to zero: for each $i \in \tilde{N}$,

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

and

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

and for each $i \in N \setminus \tilde{N}$,

$$m_i = 0.$$

R implies that even if a non-negligible number of agents (i.e., $N \setminus \tilde{N}$) happen to be irrational and adhere to uncooperative attitudes (i.e., even in the unforeseen circumstances), a large proportion of the remaining agents (i.e., $\bar{N} \subset \tilde{N}$) are willing to behave cooperatively. Recall that we have in mind that w is sufficiently large, and $\frac{z}{w}$ is close to zero.

Robustness (R): Consider an arbitrary subset $\tilde{N} \subset N$, where

$$\frac{l}{w} \leq \frac{|\tilde{N}|}{n} < \frac{l+1}{w} \text{ for some } l \in \{z+1, ..., w\}$$

A Nash equilibrium m for $\tilde{N}$ and a subset $\bar{N} \subset \tilde{N}$ exist, such that

$$\frac{|\bar{N}|}{n} = \frac{l}{w},$$

and

$$m_i = 1 \text{ for all } i \in \bar{N}.$$

VUL and R are similar in that they deal with cases where there are uncooperative agents as an unforeseen circumstance. However, VUL and R are essentially different for the following reason. VUL requires a level of cooperation close to agents' acceptable upper limits to be always achieved, irrespective of their message profile. In contrast, R deals with a possibility that each agent would change their upper limit as a countermeasure to the unforeseen circumstances. Hence, R requires that even if a non-negligible number of agents happen to be irrational and adhere to the uncooperative attitudes (unforeseen circumstances), a large proportion of the remaining agents will maintain their cooperation as equilibrium behavior without changing behaviors in response.

3. Main Theorems

The following theorem shows a necessary and sufficient condition for the existence of a commitment rule that satisfies SP, VUL, U, and R. For this proof, we specify a commitment rule α^* and show that the necessary and sufficient condition implies that α^* satisfies SP, VUL, U, and R.

We define $\delta : \{0,...,w\} \to R$ as follows. Let $\delta(0) \equiv 0$. Recursively, for each integer $x \in \{1,...,w-1\}$, we define $\delta(x)$ by the following equation:

$$(\frac{w-x+1}{w}n-1)\{\delta(x)-\delta(x-1)\} = \{1-\delta(x-1)\}(c-1).$$

Note that $\delta(x)$ is increasing in $x \in \{0,...,w-1\}$. Let $\delta(w) \equiv \delta(w-1)$. We can show the calculated values as follows: for every $x \in \{1,...,w-1\}$,

$$\delta(x) = 1 - \prod_{x'=1}^{x}(1 - \frac{c-1}{\frac{w-x'+1}{w}n-1}). \tag{1}$$

For each $m \in M$, the number of agents whose messages are less than one is denoted by:

$$y(m) \equiv \left|\{i \in N \mid m_i < 1\}\right| \in \{0,...,n\}.$$

We specify $x(m) \in \{0,...,w\}$ as follows:

$$x(m) = 0 \qquad \text{if } y(m) = 0,$$

and for each $x \in \{1,...,w\}$,

$$x(m) = x \qquad \text{if } \frac{x-1}{w} < \frac{y(m)}{n} \le \frac{x}{w}.$$

By using $x(m) \in \{0,...,w\}$, we classify the message space M into $w+1$ categories. Each category $x \in \{0,...,w\}$ includes all message profiles m such that the number of agents whose messages are less than one is between $\frac{(x-1)n}{w}$ and $\frac{xn}{w}$ (i.e., $\frac{x-1}{w} < \frac{y(m)}{n} \le \frac{x}{w}$). We consider $\delta(x)$ as the lowering width of agents' commitments when the message profile m belongs to the category x (i.e., $x(m) = x$).

Based on $(\delta(\cdot), x(\cdot))$, we specify the commitment rule α^* as follows. For every $i \in N$ and $m \in M$, let

$$\alpha_i^*(m) = \max[m_i - \delta(x(m)), \min_{j \in N} m_j, m_i - \varepsilon],$$

that is,

$$\alpha_i^*(m) = m_i - \delta(x(m)) \quad \text{if } m_i - \delta(x(m)) \geq \max[\min_{j \in N} m_j, m_i - \varepsilon],$$

$$\alpha_i^*(m) = \min_{j \in N} m_j \quad \text{if } \min_{j \in N} m_j \geq m_i - \min[\delta(x(m)), \varepsilon],$$

and

$$\alpha_i^*(m) = m_i - \varepsilon \quad \text{if } m_i - \varepsilon \geq \max[\min_{j \in N} m_j, m_i - \delta(x(m))].$$

If a message profile changes from category $x-1$ to category x, the increase in the lowering width of an agent's commitment is given as $\delta(x) - \delta(x-1)$. However, the lowering width is limited by ε and the minimal upper limit $\min_{j \in N} m_j$. If a message profile changes but the category remains unchanged, the lowering width remains the same.

Theorem 1: A commitment rule exists that satisfies SP, VUL, U, and R, if and only if:

(2) $\delta(w - z) \leq \varepsilon$.

Proof: The 'if' part of Theorem 1 is proved by showing that, under the inequality (2), the commitment rule α^* satisfies SP, VUL, U, and R. Clearly from its specification, α^* satisfies SP and VUL.

We show that α^* satisfies U as follows. Consider the maximal message profile $\bar{m} = (1, ..., 1)$ and show that it is a Nash equilibrium. Note that $x(\bar{m}) = 0$, $\delta(x(\bar{m})) = \delta(0) = 0$, and $\alpha_i^*(\bar{m}) = 1$ for all $i \in N$. Suppose that agent 1 selects $m_1 < 1$ instead of $\bar{m}_1 = 1$. Note that $x(m_1, \bar{m}_{-1}) = 1$, $\delta(x(m_1, \bar{m}_{-1})) = \delta(1) = \dfrac{c-1}{n-1}$, $\alpha_1^*(\bar{m}) = m_1$, and

$$\alpha_i^*(m) = \max[1 - \delta(x(m_1, \bar{m}_{-1})), m_1] \text{ for all } i \neq 1.$$

If

$$m_1 \geq 1-\delta(x(m_1,\bar{m}_{-1})),$$

then any agent's commitment decreases from 1 to m_1. Hence, agent 1 has the gain $(c-1)(1-m_1)$ and the loss $(n-1)(1-m_1)$. As $n>c$, we have

$$(n-1)(1-m_1)>(c-1)(1-m_1),$$

which implies that agent 1 decreases its utility. Next, if

$$m_1 < 1-\delta(x(m_1,\bar{m}_{-1})),$$

then any other agent's commitment decreases from 1 to $1-\delta(x(m_1,\bar{m}_{-1}))=1-\frac{c-1}{n-1}$. Hence, agent 1 has the gain $(c-1)(1-m_1)$ and the loss $(n-1)\frac{c-1}{n-1}$. Since

$$(n-1)\frac{c-1}{n-1}=c-1\geq(c-1)(1-m_1),$$

agent 1 decreases their utility. We can make the same argument even if we replace agent 1 with any other agent. Therefore, we have proved that $\bar{m}$ is a Nash equilibrium. By definition, we obtain $\alpha_i^*(\bar{m})=1$ for all $i\in N$.

We show that $\bar{m}$ is the unique Nash equilibrium as follows. Consider an arbitrary message profile $m\in M\setminus\{\bar{m}\}$. There exists an agent $i\in N$ such that $m_i=\min_{j\in N} m_j<1$. Suppose that m is a Nash equilibrium. We can show that any other agent $j's$ message must be commonly equal to $\min[1,m_i+\delta(x(m)),m_i+\varepsilon]$ as follows. Note that if agent j selects this message, their commitment is equal to m_i. If $m_j<\min[1,m_i+\delta(x(m)),m_i+\varepsilon]$, then we have $\alpha_j^*(m)=m_i$, and therefore, agent j has an incentive to increase their message up to $\min[1,m_i+\delta(x(m)),m_i+\varepsilon]$ due to minimal prosociality. Next, if $1>m_j>\min[1,m_i+\delta(x(m)),m_i+\varepsilon]$, then we have

$$\alpha_j^*(m_j)=\max[m_j-\delta(x(m)),m_j-\varepsilon]>m_i.$$

In this case, agent j has an incentive to lower their message because $x(m)$ is unchanged. Moreover, if $m_j = 1$ for all $j \neq i$, each agent $j \neq i$ has an incentive to lower their message: $y(m) = 1$ only changes to 2, and therefore, $x(m)$ and $\delta(x(m))$ are unchanged. (Recall that $w < n$ and n is an integer multiple of w.) Accordingly, any other agent message is commonly equal to $\min[1, m_i + \delta(x(m)), m_i + \varepsilon]$, their commitment is equal to m_i, and it is less than their upper limit. In this case, however, agent i has an incentive to increase their message because any other agent's commitment increases simultaneously. This is a contradiction. Thus, we have proved that α^* satisfies U.

We show that α^* satisfies R as follows. Consider an arbitrary subset $\tilde{N} \subset N$, where we assume that $\frac{l}{w} n \leq \left|\tilde{N}\right| < \frac{(l+1)}{w} n$ for some $l \in \{z+1, ..., w\}$. We can select a subset $\bar{N} \subset \tilde{N}$ where $\left|\bar{N}\right| = \frac{l}{w} n$. We specify a message profile m as

$$m_i = 0 \text{ for all } i \in N \setminus \tilde{N},$$

$$m_i = 1 \text{ for all } i \in \bar{N},$$

and

$$m_i = \delta(x(m)) \text{ for all } i \in \tilde{N} \setminus \bar{N}.$$

Note that $y(m) = \frac{(w-l)}{w} n$ and $x(m) = w - l$. Note from the inequality (1) that $\delta(x(m)) \leq \varepsilon$. We can prove that this message profile is a Nash equilibrium for $\tilde{N}$ as follows. Note that no agent $i \in \tilde{N} \setminus \bar{N}$ influences the category. Hence, they prefer to set their commitment equal to zero; because of (2), their maximal best response is $\delta(x(m))$. Next, note that any agent $i \in \bar{N}$ can influence the category. In other words, by selecting their message to be less than one, they can change the category from $x(m) = w - l$ to $x(m) + 1 = w - l + 1$. This change decreases the commitment of any other agent in $\bar{N}$ from $1 - \delta(x(m))$ to $1 - \delta(x(m) + 1)$, that is, by the amount of

$$\{\delta(x(m)+1)-\delta(x(m))\} == \frac{\{1-\delta(x(m))\}(c-1)}{\frac{l}{w}n-1}.$$

As $\left|\bar{N}\right| = \frac{l}{w}n$, this change decreases the agent $i's$ utility by the amount of

$$(\left|\bar{N}\right|-1)\{\delta(x(m)+1)-\delta(x(m))\} = \{1-\delta(x(m))\}(c-1).$$

This equality implies that agent i have no incentive to deviate, because they only earn $\{1-\delta(x(m))\}(c-1)$ from deviation and prefer the maximal message. Hence, we have proved that the specified message profile is a Nash equilibrium for $\tilde{N}$, and therefore, α^* satisfies R. From these observations, we have proved the 'if' part of Theorem 1.

Next, we show the proof of the 'only if' part of Theorem 1 as follows. Suppose that a commitment rule α exists that satisfies the SP, VUL, U, and R. Note that any commitment rule derived from this commitment rule and a permutation on N also satisfies these requirements. Moreover, any commitment rule derived from a weighted sum of these commitment rules also satisfies these requirements. Thus, without loss of generality, we can assume that the commitment rule α is symmetric in that for every permutation $\mu: N \to N$ and $m \in M$,

$$\alpha_i(m) = \alpha_{\mu(i)}(m') \text{ for all } i \in N,$$

where we denote $m' = (m'_j)_{j\in N}$ and $m'_{\mu(i)} = m_i$ for all $i \in N$.

Fix an arbitrary $l \in \{z+1,...,w\}$. Let $\tilde{N} = \{1,...,\frac{l}{w}n\}$. From R, $\bar{N} = \tilde{N}$ must hold, and therefore, we have a Nash equilibrium $\tilde{m}^l$ for $\tilde{N}$ such that

$$\tilde{m}^l_i = 1 \text{ for all } i \in \tilde{N},$$

and

$$\tilde{m}^l_i = 0 \text{ for all } i \in N \setminus \tilde{N}.$$

From the symmetry of α, if an agent $\frac{l}{w}n \in \tilde{N}$ announces zero instead of one, they earn $\alpha_1(\tilde{m}^l)(c-1)$ from this deviation. Hence, each of the other agents in $\tilde{N}$ (i.e., each of the

$\frac{l}{w}n-1$ agents) must decrease their commitments at least by $\frac{\alpha_1(\tilde{m}^l)(c-1)}{\frac{l}{w}n-1}$. Hence, we have

$$\alpha_1(\tilde{m}^l_{\_ln/w},0) \le \alpha_1(\tilde{m}^l) - \frac{\alpha_1(\tilde{m}^l)\{c-1\}}{\frac{l}{w}n-1} = \alpha_1(\tilde{m}^l)\{1-\frac{c-1}{\frac{l}{w}n-1}\}.$$

Since $\alpha_1(m)$ is nondecreasing, we have

$$\alpha_1(\tilde{m}^l) \le \alpha_1(\tilde{m}^{l+1}_{\_(l+1)n/w},0).$$

Moreover, we have

$$\alpha_1(\tilde{m}^w_{\_wn/w},0) = \alpha_1(\overline{m}_{\_n},0) \le \alpha_1(\overline{m})\{1-\frac{c-1}{\frac{w}{w}n-1}\} = 1-\frac{c-1}{n-1}.$$

From these observations, we have

$$\alpha_1(\tilde{m}^l_{\_ln/w},0) \le \prod_{x'=1}^{w-l+1}(1-\frac{c-1}{\frac{w-x'+1}{w}n-1}) = 1-\delta(w-l+1),$$

and therefore,

$$\delta(w-z) \le 1-\alpha_1(\tilde{m}^{z+1}_{\_(z+1)n/w},0).$$

From $\tilde{m}^{z+1}_1 = 1$ and VUL, the inequality (2) must hold.

From these observations, we have completed the proof of Theorem 1.

Q.E.D.

To help understanding Theorem 1, we consider an infinite sequence $(c(n))_{n=1}^{\infty}$, where $c(n)>1$ and there exists $\rho \in [0,1]$ such that

$$\lim_{n\to\infty}\frac{c(n)}{n} = \rho.$$

We assume that n is fixed sufficiently large and $\frac{c}{n}$ is approximated by ρ. The following theorem states that irrespective of (ε, w, z), there exists a commitment rule that satisfies SP, VUL, U, and R.

Theorem 2: For a sufficiently large n, there exists a commitment rule that satisfies SP, VUL, U, and R, if

$$1-\prod_{x'=1}^{w-z}\{1-\frac{w}{w-x'+1}\rho\}<\varepsilon. \tag{3}$$

For a sufficiently large n, there is no commitment rule that satisfies SP, VUL, U, and R, if

$$1-\prod_{x'=1}^{w-z}\{1-\frac{w}{w-x'+1}\rho\}>\varepsilon. \tag{4}$$

Proof: Since

$$\lim_{n\to\infty}\{1-\prod_{x'=1}^{w-z}(1-\frac{c-1}{\frac{w-x'+1}{w}n-1})\}=1-\prod_{x'=1}^{w-z}\{1-\frac{w}{w-x'+1}\rho\},$$

it follows from (1) that if the inequality (3) holds, then for a sufficiently large n, the inequality (2) also holds. Moreover, if the inequality (4) holds, then for a sufficiently large n, the inequality (2) does not hold. Hence, from Theorem 1, Theorem 2 can be proved.

Q.E.D.

To help understanding the case with $\rho=0$, i.e., a special case with inequality (4), we specify a commitment rule $\hat{\alpha}$, which is a simpler version of the commitment rule α^*, as follows. For every $i\in N$ and $m\in M$,

$$\hat{\alpha}_i(m)=\max[m_i-\frac{x(m)}{w}\varepsilon,\min_{j\in N}m_j],$$

that is,

$$\hat{\alpha}_i(m)=m_i-\frac{x(m)}{w}\varepsilon \qquad \text{if } m_i-\frac{x(m)}{w}\varepsilon\geq\min_{j\in N}m_j,$$

and

$$\hat{\alpha}_i(m) = \min_{j \in N} m_j \qquad \text{if } m_i - \frac{x(m)}{w}\varepsilon < \min_{j \in N} m_j .$$

If the category rises by one unit, the lowering width $m_i - \hat{\alpha}_i(m)$ rises by the constant amount $\frac{\varepsilon}{w}$ within the range defined by the minimal upper limit $\min_{j \in N} m_j$. Clearly, $\hat{\alpha}$ satisfies SP, VUL, and $\hat{\alpha}_i(\bar{m}) = 1$ for all $i \in N$.

Similarly to α^*, we can show that $\hat{\alpha}$ satisfies R. Consider an arbitrary message profile $m \in M$ where $x(m) \leq w-1$ and $y(m) = \frac{x(m)n}{w}$. The $(\frac{x(m)n}{w}+1)-th$ deviant, whose gain from deviation is $(1-\frac{x(m)\varepsilon}{w})(c-1)$, will change the category, and therefore, will be penalized from the remaining $\frac{w-x(m)}{w}n-1$ agents' commitment reductions by the amount of $\frac{\varepsilon}{w}$ for each. Since n is sufficiently large and $\rho = 0$, we have

$$(5) \qquad \frac{\varepsilon}{w}\{\frac{w-x(m)}{w}n-1\} \geq (1-\frac{x(m)\varepsilon}{w})(c-1),$$

because

$$\lim_{n\to\infty}\frac{1}{n}[\frac{\varepsilon}{w}\{\frac{w-x(m)}{w}n-1\}-(1-\frac{x(m)\varepsilon}{w})(c-1)]$$

$$=\lim_{n\to\infty}[\frac{\varepsilon}{w}\{\frac{w-x(m)}{w}-\frac{1}{n}\}-(1-\frac{x(m)\varepsilon}{w})(\frac{c}{n}-\frac{1}{n})]$$

$$=\frac{\varepsilon}{w}\{\frac{w-x(m)}{w}\}-(1-\frac{x(m)\varepsilon}{w})\rho$$

$$=\frac{\varepsilon}{w}\{\frac{w-x(m)}{w}\}>0.$$

The inequality (5) discourages such deviations. This observation implies that $\hat{\alpha}$ satisfies R.

We can also show that $\hat{\alpha}$ satisfies U. Note that $\bar{m}$ is a Nash equilibrium; from (5) and $x(\bar{m}) = 0$, we have

$$\frac{\varepsilon}{w}(n-1) \geq c-1.$$

Consider an arbitrary $m \in M \setminus \{\bar{m}\}$, where an agent $i \in N$ exists such that $m_i = \min_{j \in N} m_j < 1$. Suppose that m is a Nash equilibrium. Due to minimal prosociality, in the same manner as α^*, any other agent's message must be commonly equal to $\min[1, m_i + \frac{x(m)}{w}\varepsilon]$, which is greater than m_i. Accordingly, their commitment must be equal to $m_i = \min_{j \in N} m_j$, which is less than their announced upper limit. However, in this case, agent i has an incentive to increase their message because any other agent's commitment increases simultaneously. This is a contradiction. Thus, $\bar{m}$ is the unique Nash equilibrium, that is, $\hat{\alpha}$ satisfies U.

From the above, we have shown that if $\rho = 0$ and n is sufficiently large, the commitment rule $\hat{\alpha}$ satisfies SP, VUL, U, and R.

4. Discussion

4.1. Robustness

R requires that even if a nonnegligible number of agents irrationally adhere to the uncooperative attitudes, a large proportion of the remaining agents are willing to behave cooperatively. As a more stringent requirement, we introduce exact robustness (ER), which requires that even if a nonnegligible number of agents irrationally adhere to the uncooperative attitudes, all the remaining agents are willing to behave cooperatively. Fix an arbitrary positive real number $\lambda \in (0,1)$, which has a similar role to $\frac{z+1}{w}$ in the definition of R.

Exact Robustness (ER): Consider an arbitrary subset $\tilde{N} \subset N$, where it is assumed that

$$\frac{|\tilde{N}|}{n} \geq \lambda .$$

There exists a Nash equilibrium m for $\tilde{N}$ such that

$$m_i = 1 \text{ for all } i \in \tilde{N} .$$

The following proposition states that if we replace R with ER, we can no longer prevent global catastrophes through commitment rule design, even if we do not require U.

Proposition 1: If

$$\varepsilon < (1-\varepsilon)(c-1)\ln\frac{1}{\lambda}, \tag{6}$$

then, for a sufficiently large n, there exists no commitment rule α that satisfies the SP, VUL, and ER.

Proof: Suppose that a commitment rule α exists that satisfies the SP, VUL, and ER. According to the same logic as in the proof of the "only if" part of Theorem 1, we can assume without loss of generality that the commitment rule α is symmetric.

From ER, $\bar{m}$ must be a Nash equilibrium for $\tilde{N} = N$. If agent n announces 0 instead of 1, it follows from VUL that they earn at least $(1-\varepsilon)(c-1)$ from this deviation, and each of the other agents must decrease their commitments at least by $\frac{(1-\varepsilon)(c-1)}{n-1}$. Therefore, their commitment must be at most $1-\frac{(1-\varepsilon)(c-1)}{n-1}$. Next, consider $\tilde{N} = \{1,2,...,n-1\}$ and the Nash equilibrium for $\tilde{N}$ where every agent in $\tilde{N}$ selects the maximal message 1. If agent $n-1$ announces 0 instead of 1, they earn at least $(1-\varepsilon)(c-1)$ from this deviation, and each of the other agents in $\tilde{N}$ must decrease their

commitments at least by $\frac{(1-\varepsilon)(c-1)}{n-2}$. Since α is nondecreasing, their commitment must be at most $1-\frac{(1-\varepsilon)(c-1)}{n-1}-\frac{(1-\varepsilon)(c-1)}{n-2}$. Recursively, for each $l \in \{2,...,n-1\}$, consider $\tilde{N}=\{1,2,...,n-l\}$ and the Nash equilibrium for $\tilde{N}$ where every agent in $\tilde{N}$ selects the maximal message 1. If agent $n-l$ announces zero instead of one, each of the other agents in $\tilde{N}$ must decrease their commitments at least by $\frac{(1-\varepsilon)(c-1)}{n-l-1}$. Hence, their commitment must be at most $1-\sum_{l'=0}^{l}\frac{(1-\varepsilon)(c-1)}{n-l'-1}$.

From VUL, for each $l \le (1-\lambda)n$, $\sum_{l'=0}^{l}\frac{(1-\varepsilon)(c-1)}{n-l'-1} \le \varepsilon$ must hold. For a sufficiently large n, we can approximate $\sup_{l\le(1-\lambda)n}\sum_{l'=0}^{l}\frac{(1-\varepsilon)(c-1)}{n-l'-1}$ by $(1-\varepsilon)(c-1)\ln\frac{1}{\lambda}$, which is greater than zero. (Note that $\sup_{l\le(1-\lambda)n}\sum_{l'=0}^{l}\frac{1}{n-l'-1}$ is approximated by $\ln(\frac{n-1}{\lambda n-2})$.)

From the above observations, given a sufficiently large n, $\max_{m\in M}\{m_1-\alpha_1(m)\}$ is approximated by $(1-\varepsilon)(c-1)\ln\frac{1}{\lambda}$ or more. Hence, $(1-\varepsilon)(c-1)\ln\frac{1}{\lambda}\le\varepsilon$ must hold. However, this notion contradicts the inequality (6). Hence, we have proved Proposition 1.

Q.E.D.

For a commitment rule to satisfy ER, the remaining sane agents must reduce their commitments by at least $(1-\varepsilon)(c-1)$ in total for each additional deviant. If n is sufficiently large, then each agent's reduction could be large, and each agent therefore needs to reduce their commitment by $(1-\varepsilon)(c-1)\ln\frac{1}{\lambda}$ or more in the worst-case scenario. With the inequality (6), this is a contradiction of VUL.

From Proposition 1, if $\varepsilon > 0$ is close to zero and n is sufficiently large, there is no commitment rule α that satisfies SP, VUL, and ER. If $\lambda > 0$ is close to zero and n is sufficiently large, there is no commitment rule α that satisfies SP, VUL, and ER. Hence, we can consider Proposition 1 as an impossibility result. However, in contrast to ER, R only requires commitment reductions for a limited number of deviants. Thus, by weakening ER to R, we can save dramatically on commitment reductions, making a robust commitment rule successfully consistent with VUL.

4.2. Role of Minimal Upper Limits

The commitment rules designed in this study such as $\hat{\alpha}$ and α^* impose on any agent a commitment not to go below the minimum of the announced upper limits across all agents $\min_{j \in N} m_j$. This dependence of a commitment rule on this minimal upper limit (MUL) will play a crucial role in satisfying U (uniqueness) if this rule satisfies VUL (virtual upper limit).

To understand this point, we first define the unanimity rule $\tilde{\alpha}$: for each $i \in N$, $\tilde{\alpha}_i(\overline{m}) = 1$, and

$$\tilde{\alpha}(m) = 0 \text{ for all } m \neq \overline{m}.$$

The unanimity rule $\tilde{\alpha}$ does not depend on MUL and does not satisfy VUL. Nevertheless, it satisfies U; due to minimal prosociality, the maximal message profile $\overline{m}$ is the unique Nash equilibrium.

If we limit the scope to commitment rules that satisfy VUL, then MUL will have a crucial role for U. We specify another commitment rule $\alpha^\dagger$ as follows: for each $i \in N$, $\alpha_i^\dagger(\overline{m}) = 1$, and

$$\alpha_i^\dagger(m) = \max[m_i - \varepsilon, 0] \text{ for all } m \neq \overline{m}.$$

This rule is a modification of the unanimity rule $\tilde{\alpha}$ and satisfies VUL, but does not depend on MUL. Surely, the maximal message profile $\overline{m}$ is a Nash equilibrium. However, the

message profile $\hat{m}$ specified by $\hat{m}_i = \varepsilon$ for all $i \in N$ is also another Nash equilibrium, failing to satisfy U.

We further specify a commitment rule $\hat{\alpha}^{\dagger}$ by:

$$\hat{\alpha}_i^{\dagger}(m) = \max[m_i - \varepsilon, \min_{j \in N} m_j] \text{ for all } m \in M \text{ and } i \in N.$$

The commitment rule $\hat{\alpha}^{\dagger}$ is a modification of $\alpha^{\dagger}$, which depends on MUL and satisfies SP and VUL. Importantly, $\hat{\alpha}^{\dagger}$ satisfies U; due to minimal prosociality, many agents prefer to select messages that are greater than MUL. This also motivates the agent who announces MUL to increase their message, because many agents' commitments are increased simultaneously. Thus, agents can ascend MUL like climbing stairs. Consequently, the maximal message profile is the only equilibrium that will survive through this stair-climbing procedure.

Finally, we consider the commitment rule $\alpha^{\dagger\dagger}$ that McKay et al. (2015) demonstrated as the common commitment rule, which is defined to emulate the lowest price guarantee clause and to assign the minimal upper limit to every agent as their commitment:

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

Note that both the unanimity rule $\tilde{\alpha}$ and the common commitment rule $\alpha^{\dagger\dagger}$ satisfies U. Both rules do not satisfy VUL. While the common commitment rule $\alpha^{\dagger\dagger}$ depends on MUL, the unanimity rule $\tilde{\alpha}$ does not depend on it. Thus, the dependence on MUL is non-essential for satisfying U in the common commitment rule $\alpha^{\dagger\dagger}$.

See the Table to understand the difference between all the commitment rules investigated in this study. We can consider $\hat{\alpha}^{\dagger}$ as a hybrid of $\alpha^{\dagger}$ and $\alpha^{\dagger\dagger}$. We can consider $\hat{\alpha}$ as a modification of $\hat{\alpha}^{\dagger}$ to satisfy R.

The Table: Various Commitment Rules

	SP	VUL	U	R	MUL
$\tilde{\alpha}(m)=0$ for $m\neq\bar{m}$ (Unanimity)	Y	N	Y	N	N
$\alpha_i^{\dagger\dagger}(m)=\min_{j\in N} m_j$ (Common Commitment)	Y	N	Y	N	Y
$\alpha_i^{\dagger}(m)=\max[m_i-\varepsilon,0]$ for $m\neq\bar{m}$	Y	Y	N	N	N
$\hat{\alpha}_i^{\dagger}(m)=\max[m_i-\varepsilon,\min_{j\in N} m_j]$ for $m\neq\bar{m}$	Y	Y	Y	N	Y
$\alpha_i^{*}(m)=\max[m_i-\delta(x(m)),\min_{j\in N} m_j, m_i-\varepsilon]$	Y	Y	Y	Y	Y
$\hat{\alpha}_i(m)=\max[m_i-\frac{x(m)}{w}\varepsilon,\min_{j\in N} m_j]$	Y	Y	(Y)	(Y)	Y

4.3. Adherence to Commitments

We have assumed adherence to commitments so that all agents keep their own commitments unless they publicly offer to change their own commitments. In this subsection, we weaken this assumption and consider a robustness against unforeseen circumstances where there exist rogue agents who do not want to uphold the social order.

Suppose that the society is mixed with such rogue agents who declare their upper limits to be one but actually choose zero. Then, the fear arises that the sincere agents who uphold the social order will no longer have sufficient incentive backing for cooperative behavior.

However, if the catastrophe is sufficiently enormous, the commitment rules such as $\hat{\alpha}$ and α^* are resistant to the appearance of such rogue agents. In fact, as long as each agent is expected to keep their commitment with a positive (but less-than-one) probability, the positive result implied by the second part of Theorem 2 remains valid for a sufficiently

small ρ. Even when it is known in advance who the rogue agents are, the sincere agents do not lose their incentives to cooperate; the mere lowering of the commitments of the sincere agents, which is the mechanism included in $\hat{\alpha}$ and α^* for the satisfaction of R, is also a sufficient penalty for those who are sincere but turned uncooperative even in this situation.

4.4. Prosocial Motives

We have assumed that any agent's prosocial motive is outweighed by their selfish motive. However, if there is a high level of awareness of being engaged in resolving the catastrophe issue, the prosocial motive can be stronger and even outweigh the selfish motive (Hirschman, 1970; Matsushima, 2008; Abeler et al., 2019; Hart and Zingales, 2019).

We can expect the following two effects of stronger prosocial motives. One, even if the catastrophe is not severely damaging, the commitment rule can be complemented by this stronger prosocial motive to resolve the free-rider problem. The other is the case where the prosocial motive is enhanced by concerns not only about humanity but also about biodiversity and ecological crises. In this case, it could be expected to further enhance the goal of catastrophe control beyond one.

5. Dynamics

We have investigated the single-period model. In this section, we consider a continuous time horizon in which the agents are repeatedly faced with negotiations at a fixed interval $\Delta > 0$. In this section, we argue about postponement of the next round of negotiation, history-dependent commitment rule design, and history-dependent equilibrium behavior.

5.1. Postponement

In the case of ongoing free-rider problems surrounding catastrophe prevention, the continued use of the commitment rule will allow the problem to be resolved over the long term. Importantly, explicit consideration of long-term relationships makes problem-solving easier in the following two ways, both of which relate to the possible postponement of the next round of negotiation.

5.1.1. Adherence to Commitments Revisited

First, it avoids the contingency of having rogue agents who do not adhere to their commitments (i.e., do not uphold the social order). A rogue agent can earn a gain $c-1$ without the other agents' commitment reductions, by declaring upper limit one as a lie but actually selecting zero so that other companies are not aware of it. However, if such a disruptive behavior is discovered, it will be difficult to adopt the commitment rule with a nonchalant face at the next round, because this rule was designed on the premise of the social order maintenance. Thus, inevitably, it becomes impossible to take measures to prevent catastrophes for a certain period. Because of this inevitable postponement of the next round of negotiation, a selfish incentive to keep one's word will sprout even for those rouge agents who have no ethical hesitation in disrupting the social order.

To clarify this point, we consider the following continuous time horizon with discount rate $\varphi > 0$. The negotiations are held and adopt the commitment rule $\hat{\alpha}$ every fixed time interval Δ (unless there are special circumstances explained later). The utility at each round is discounted by the discount factor $\delta \equiv \exp(-\varphi\Delta) \in (0,1)$. If there exists a rouge agent who silently breaks their commitment, this breach of trust is immediately discovered, and therefore, the next round of renegotiation is inevitably postponed from Δ to $\Delta + t$ later. In this case, the future payoff, which is defined as the discounted sum of the utilities in the future rounds, is changed from $\frac{\delta}{1-\delta}(n-c)$ to $\frac{\beta\delta}{1-\delta}(n-c)$, where we denote

$\beta \equiv \exp(-\varphi t) \in (0,1)$. In this case, the rogue agent has the instantaneous gain $c-1$ from deviation and has the future loss given by

$$\frac{\delta}{1-\delta}(n-c)-\frac{\beta\delta}{1-\delta}(n-c)=\frac{(1-\beta)\delta}{1-\delta}(n-c).$$

Hence, the rogue agent hesitates to silently break their own commitment if and only if

$$c-1\le\frac{(1-\beta)\delta}{1-\delta}(n-c).$$

This inequality is approximated by

$$\lim_{n\to\infty}\frac{c(n)-1}{n-c(n)}=\frac{\rho}{1-\rho}\le\frac{(1-\beta)\delta}{1-\delta},$$

that is,

(6) $$\rho\le\frac{(1-\beta)\delta}{1-\delta+(1-\beta)\delta}.$$

Note that if the fixed time interval Δ is short enough, only a slight postponement (small t) will be sufficient for such rouge agents' adherence to commitments, because the right-hand side of (6) is close to one.

5.1.2. Robustness and Uniqueness

Second, an artificially designed postponement device will help with the incentive effects inherent in the commitment rule $\hat{\alpha}$ concerning R, i.e., the robustness against unforeseen circumstances where a non-negligible number of agents happen to be irrational and adhere to the uncooperative attitudes. If the agents select the message profile $m \in M \setminus \{\bar{m}\}$ and sincerely adhere to their commitments, then the next round will be artificially postponed by the time interval $t(m) \ge 0$, which we define by the following equation:

$$\exp(-\varphi t(m))=1-\gamma\frac{x(m)}{w},$$

where we assume $\gamma \in (0,1)$. If the category of the message profile is increased by one, the future payoff is further discounted by $\frac{\gamma}{w}$. With this artificial postponement device, we can maintain the robustness in the substantial sense even if the commitment rule $\hat{\alpha}$ fails to satisfy R.

The $(\frac{x(m)n}{w}+1)-th$ deviant (agent i), whose gain from deviation is $\hat{\alpha}_i(m)(c-1)$, will be penalized by the future loss generated by the artificial postponement, which is given by $\frac{\gamma}{w}\frac{\delta}{1-\delta}(n-c)$, as well as the remaining $\frac{w-x(m)}{w}n-1$ agents' commitment reductions by the amount of $\frac{\varepsilon}{w}$ for each. Hence, the incentive constraint we need to require for R can be replaced with:

$$\frac{\varepsilon}{w}\{\frac{w-x(m)}{w}n-1\}+\frac{\gamma}{w}\frac{\delta}{1-\delta}(n-c)\geq \hat{\alpha}_i(m)(c-1).$$

Since n is sufficiently large, we have

$$\frac{\{w-x\}\varepsilon}{w^2}+\frac{\gamma}{w}\frac{\delta}{1-\delta}(1-\rho)\geq \hat{\alpha}_i(m)\rho \text{ for all } x\in\{0,...,w-z-1\}.$$

Hence, the commitment rule $\hat{\alpha}$ can maintain the robustness in the substantial sense if

$$\rho<\frac{\gamma\delta}{(1-\delta)w+\gamma\delta}. \tag{7}$$

Note that if δ is fixed close to one, we can maintain the robustness irrespective of $\rho\in[0,1)$, because the right-hand side of (7) is close to one irrespective of degree of postponement γ. In this case, the robustness requirement is not met by the design of the commitment rule, but entirely by the ingenuity of the artificial postponements.

On the other hand, the achievement of U still relies on the ingenuity of the design of the commitment rule. Consider the commitment rule $\hat{\alpha}$ with $\varepsilon=0$. Without artificial postponement device, the maximal message profile $\bar{m}$ is never a Nash equilibrium. With the artificial postponement device, if the inequality (7) holds, that is,

$$\frac{\gamma}{w}\frac{\delta}{1-\delta}(n-c) \geq c-1,$$

then $\bar{m}$ is a Nash equilibrium in the corresponding dynamic model. However, it is not unique; all agents announcing zero is also another one.

To restore U, we consider using $\hat{\alpha}^{\dagger}$ specified in Subsection 4.2 and adopting the artificial postponement device. We assume the inequality (7) and

(8) $\rho < \varepsilon$.

Given a sufficiently large n, it follows from (8) that

$$\varepsilon(n-1) \geq c-1,$$

which guarantees U irrespective of whether an artificial postponement device is installed. Since the inequality (8) is less restrictive than the inequality (2), by utilizing the artificial postponement device, we can dramatically extend the range of catastrophe problems that we can resolve.

However, we should not overestimate the ingenuity of such artificial postponement devices. Artificial postponements bring social costs because the catastrophe cannot be stopped for that period. Artificial postponements are thwarted by renegotiation, because the social order is still maintained for that period. While the commitment rule $\hat{\alpha}$ can be replaced with a simpler one such as $\hat{\alpha}^{\dagger}$, devising artificial postponements still requires detailed design using the categorization of message profiles as the design of the commitment rule $\hat{\alpha}$.

5. 2. History-Dependent Commitment Rules

We have required a commitment rule to satisfy R, which implies that many agents are willing to behave cooperatively as an equilibrium behavior even if non-negligible number of agents happen to be irrational and adhere to the uncooperative attitudes. However, we did not require the uniqueness of this equilibrium behavior in such accidental cases. In fact, all rational agents committing to action zero by announcing lower upper limits is another Nash equilibrium outcome. The reason for the failure of uniqueness in the unforeseen

circumstances is that since the minimal upper limit inevitably stays at zero, the commitment rules (such as $\hat{\alpha}$) lose the impetus to derive the uniqueness (see Subsection 4.2).

However, if agents can foresee who will be irrational and adhere to action zero from the past history of play, we can dramatically change the situation in the following manner. Suppose that, based on the past history of play, the agents in a subset $N' \subset N$ are judged to be irrational and adhere to the uncooperative attitudes. In this case, instead of the commitment rule $\hat{\alpha}$, we adopt a modified version, which is specified as follows. Select a subset $\tilde{N} \subset N$ so that $\left|\tilde{N}\right|$ is a integer multiple of w and its members are rational (i.e., $\tilde{N} \bigcap N' = \phi$). We then replace the minimal upper limit $\min_{j \in N} m_j$ with $\min_{j \in \tilde{N}} m_j$ for each agent $i \in \tilde{N}$, which eliminates the irrational agents' upper limits (i.e., zero) from the minimal upper limit assessment for each agent in $\tilde{N}$. In contrast, we do not make this replacement for any agent in $N \setminus \tilde{N}$. We do not make any replacement concerning the categorization for all agents at all. This modification successfully restores the function of the minimal upper limit among $\tilde{N}$. We can therefore show in the same manner as U (with no irrational agent) that all agents in $\tilde{N}$ behave cooperatively as the unique equilibrium behavior.

If these irrational agents are freed from the spell of uncooperative attitudes, we can expect them to make the maximal best response that is greater than zero. This would state that they are no longer irrational. Thus, in the end, we can revert back to the original commitment rule $\hat{\alpha}$.

5.3. History-Dependent Behavior

We have assumed that agents do not consider the past history of play to determine their announcement behaviors at each round of negotiation. If we remove this assumption, we are faced with the multiplicity of equilibrium behaviors as the Folk theorem indicates.

Not only on the subject of this paper, but in general, whether multiple equilibria can be eliminated by introducing explicit negotiation procedures is an important open question. For example, Matsushima (2012) considered a repeated prisoners' dilemma where a long-term binding side-payment contract is negotiated between agents, as follows. First, the agents agree that a particular strategic profile is the goal to be realized. Next, they agree to a long-term contract that individually punishes only the agent who deviates from this strategy at the final round. Matsushima (2012) then proved that there exists a strategy profile that achieves the full cooperation and is the unique subgame perfect equilibrium in the repeated prisoners' dilemma associated with the long-term side-payment contract.

In a repeated game in which each agent has the means to sanction other agents individually, there is an equilibrium that can prevent catastrophes even in situations where agents are not at all concerned about this catastrophe. In such situations, they do not see the catastrophe as a problem, nor do they even recognize it as a free rider problem. Despite this seemingly hopeless situation, Abreu (1988) presented a way to forcefully prevent the catastrophe by defining an individual penalty code for each agent and creating a mechanism as a tacit collusion to individually punish each agent who does not exercise the penalty codes for other agents.

Such rather means-less methods of accomplishment are not well supported by experiments. Relevant literature includes Kayaba et al. (2020), which experimentally considered situations where monitoring accuracy is imperfect, and reported that subjects tend to reinforce behaviors that sanction others even beyond their self-interest as monitoring accuracy increases. This reinforcement is, however, not motivated by the fact that they would otherwise be sanctioned by others, but rather by a growing anger against violators (Matsushima, 2019).

6. Conclusion

We have experienced many crises in the past such as financial crises, pandemic crises, and international conflict crises, and we have been able to apply the lessons learned to some

extent. However, global catastrophes are devastating and irreversible, so that they must be strictly prevented before they occur. Therefore, in this study, we examined the institutional design that can achieve appropriate responses to unforeseen circumstances and stable coordination without infringing on citizen sovereignty.

We demonstrated an explicit mechanism as a commitment rule, whereby if an agent increases its tolerance for their own commitment, the level of actual commitments of other agents is increased in tandem. We proved that the well-designed commitment rule spurs all agents to voluntary cooperative behavior if they correctly perceive the global catastrophe to be severely damaging. This function is not dented by the emergence of irrational agents who persist in uncooperative attitudes, or by the emergence of rogue agents who do not uphold the social order. These properties are further strengthened by considering dynamic aspects.

It is important to develop future research in various directions as follows. We assumed that negotiations take place only in one place as a global negotiation forum, and that in principle all agents participate in it. It should be considered that only some agents come together to establish a local negotiation forum, separately from the global one. Such local negotiation forums can use coercion by local communities. We could view the global forum as a place in which only local representatives participate. Thus, a global system should be considered whereby the procedure to global consensus building is hierarchical thereby utilizing local consensus-building ability for global consensus building.

We assumed that agents are homogeneous. We should analyze situations where agents are heterogeneous and vary over time in their types, because such heterogeneity is considered as one of the main obstacles in resolving the tragedy of the global commons. Hence, we need a model in which the pattern of fluctuations in types is assessed as a stochastic phenomenon.

We need a more in-depth discussion about how to define a dynamic model. Activities that prevent catastrophe could be effective when they accumulate through time. If there is sufficient accumulation in the past, the severity of damage will temporarily decrease. Such

path-dependence should be explicitly analyzed by considering the global catastrophe problem as a dynamic resource management of the global commons.

There are various incentive issues concerning ex-ante investments that were not discussed in this study; that is, investments in preventing global catastrophes, investments in technological innovation to reduce the cost, investments in improving sustainable lifestyles, investments in early detection of suspected catastrophes, and investments that each agent make in saving only themselves from the catastrophe that has occurred. These will be the possible subjects of future research, beyond the scope of this study.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019): Preference for Truth-Telling, Econometrica 87 (4), 1115–1153.

Abreu, D. (1988): "On the Theory of Infinitely Repeated Games with Discounting," Econometrica 56 (2), 383-396.

Aumann, R. and L. Shapley (1976): "Long Term Competition - A Game Theoretic Analysis," mimeo.

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

Barrett, S. (2003): "Environment and Statecraft: The Strategy of Environmental Treaty-Making, Oxford: Oxford University Press.

Blanchard, O. and J. Tirole (2021): Major Future Economic Challenge. Republique Francaise.

Cooper, R. (2008): "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.

Cramton, P., D. MacKay, A. Ockenfels, and S. Stoft (2017): Global Carbon Pricing: The Path to Climate Cooperation, Cambridge, MA: MIT Press.

Cramton, P., A. Ockenfels, and S. Stoft (2015): "An International Carbon-Price

Commitment Promotes Cooperation," Economics of Energy & Environmental Policy 4, 51–64.

Cramton, P. and S. Stoft (2012): "Global Climate Games: How Pricing and a Green Fund Foster Cooperation," Economics of Energy & Environmental Policy 1, 125–36.

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

Dutta, P. and R. Radner (2004): "Self-Enforcing Climate-Change Treaties," Proceedings of the National Academy of Sciences 101, 5174–9.

Dutta, P. and R. Radner (2009): "A Strategic Analysis of Global Warming: Theory and Some Numbers," Journal of Economic Behavior and Organization 71, 187–209.

Finus, M. (2001): Game Theory and International Environmental Cooperation, Finus, Cheltenham: Edward Elgar.

Farrell, J. and E. Maskin (1989): "Renegotiation in Repeated Games," Games and Economic Behavior 1, 327–360.

Fudenberg, D., D. Levine, and E. Maskin (1994): "The Folk Theorem with Imperfect Public Information," Econometrica 62, 997-1039.

Fudenberg, D. and E. Maskin (1986): "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," Econometrica 54, 533–556.

Harrison, R. and R. Lagunoff (2017): "Dynamic Mechanism Design for a Global Commons," International Economic Review 58, 751–782.

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

Hart, O. and L. Zingales (2019), "Companies Should Maximize Shareholder Welfare Not Market Value," forthcoming in the Journal of Law, Finance, and Accounting.

Hirschman, A. (1970): Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States, Cambridge, MA: Harvard University Press.

Kandori, M. and H. Matsushima (1998): Private Observation, Communication and Collusion, Econometrica 66(3), 627-652.

Kayaba, Y., H. Matsushima, and T. Toyama (2020): "Accuracy and Retaliation in Repeated

Games with Imperfect Private Monitoring: Experiments," Games and Economic Behavior 120, 193-208.

MacKay, D., P. Cramton, A. Ockenfels, and S. Stoft (2015): "Price Carbon — I Will If You Will," Nature 526, 315–16.

Matsushima, H. (2004): "Repeated Games with Private Monitoring: Two Players," Econometrica 72(3), 823-852.

Matsushima, H. (2008): "Role of Honesty in Full Implementation," Journal of Economic Theory 139, 353–359.

Matsushima, H. (2012): "Finitely Repeated Prisoner's Dilemma with Small Fines: The Penance Contract," Japanese Economic Review 63, 333-347.

Matsushima, H. (2019); "Behavioral Theory of Repeated Prisoner's Dilemma: Generous Tit-For-Tat Strategy," B. E. Journal of Theoretical Economics 20 (1).

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

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

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

Nordhaus, W. (2015): "Climate Clubs: Overcoming Free-Riding in International Climate Policy," American Economic Review 105, 1339–70.

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

Sugaya, T. (2022): "The Folk Theorem in Repeated Games with Private Monitoring," The Review of Economic Studies 89, 2201-2256.

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

Victor, D. (2001): The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming, Princeton, NJ: Princeton University Press.

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