HEIDELBERG UNIVERSITY
DEPARTMENT OF ECONOMICS

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Solving Dilemma Games with Evolving Conditional Commitments

Jörg Oechssler

AWI DISCUSSION PAPER SERIES NO. 770
January 2026

Solving Dilemma Games with Evolving Conditional Commitments[*]

Joerg Oechssler[†]
Department of Economics, University of Heidelberg

– Very preliminary –

January 27, 2026

Abstract

I study a formal mechanism that can sustain Pareto optimality in a new and very broad class of dilemma games. In the absence of a central authority that could enforce multilateral agreements, the mechanism is based on binding unilateral commitments, which condition a player's (possibly multidimensional) contribution on other players' contributions. I show that unexploitable better response dynamics converge to Pareto optimal contributions when the game is played recurrently.

Keywords: Public Goods; climate treaties; conditional contributions.
JEL-Classification: C72;D82; H41

[*] I thank Jobst Heitzig for many discussions in the early stages of this project. I also thank Subhasish Chowdhury, Christian Hilbe, Andis Sofianos, Georg Nöldeke, Christoph Vanberg, Georg Weizsäcker, and participants at the Workshop Algorithmic and Evolutionary Learning Dynamics and Markets, Maastricht University, ASSET Rabat, and at seminars at the universities of Heidelberg and St. Gallen for comments.

[†] oechssler@uni-hd.de

1 Introduction

In this short paper I suggest a mechanism designed to solve dilemma games, games which are ubiquitous in economics, politics, international relations, and related fields. Currently, probably the most prominent example are international climate treaties that aim to limit CO_2 growth, a task where the difficulties are plentiful. Countries may not agree on the urgency of the issue and they may have different actions they can choose from to mitigate emissions. But the biggest difficulty lies in the absence of a *central* authority that can enforce agreements.

Thus, the only hope for stable international treaties is when they are self-enforcing. The conditional contribution mechanism proposed here is designed to achieve this. Each player promises to contribute certain (vectors of) actions under the condition that other players' contributions meet the thresholds set by the player. What is required is that each player can make internal and *unilateral* commitments (e.g. through national laws).[1] Given these unilateral commitments, if a government of one country decides to deviate from its promise, it knows full well that the mechanism would trigger automatic actions by all other countries making the deviation unattractive.[2]

The conditional contribution mechanism (CCM) suggested here is designed to work under voluntary participation and incomplete information (by players and the mechanism designer) about other players' preferences. Both requirements seem desirable for many applications. For example, in the context of international climate treaties, no country has complete information about other countries' costs or their willingness to pay for various mitigation measures.

The mechanism is intended for the repeated play of dilemma games and works as follows. In each period, all players submit two statements of the form "We will contribute action vector a_i if other players contribute at least A_{-i} in total." The mechanism then picks one action profile that is compatible with at least one statement for all players. If no such feasible action profile exists, it will pick a default action a^0, which is the "business-as-usual" (Harstad, 2024) Pareto dominated Nash equilibrium of the dilemma game. The

[1] As Heitzig (2019) points out, this type of binding mutually conditional commitments is known from an important current example: the US National Popular Vote Interstate Compact (NPVIC) (Bennett and Bennett, 2001; Muller, 2007) that aims at repairing the deficiencies of the US electoral college and electing the winner of the national popular vote for president of the US.

[2] On a smaller scale, refundable deposits, contractual arrangements, or even crypto smart contracts can achieve unilateral commitments.

mechanism will then adjust all statements so that they agree with the chosen action profile (see below for details) and announce this as feedback to all players.

The reason why the mechanism requires two conditional statements from each player is that this makes it possible for players to increase their conditional commitments without risking the status quo. With the first statements they can fix the status quo and with the second they can suggest better alternatives, which - when feasible for all - would make everyone better off. With just one conditional statement, the mechanism could get stuck at contribution levels, which – while being better than a^0 – are still not Pareto optimal.

Simplified versions of the suggested conditional contribution mechanism have been studied theoretically by Reischmann and Oechssler (2018), Heitzig (2019), and Oechssler et al. (2022).[3] An interesting alternative are contractive mechanisms studied by Healy and Mathevet (2012). They have desirable theoretical properties but may be too complex is some situations. There is of course also a very extensive literature on mechanisms for public goods (Vickrey, 1961, Clarke, 1971, Groves and Ledyard, 1977) and conditional contributions (Guttman, 1978, 1986).[4] Experimentally, the CCM has been studied by Reischmann and Oechssler (2018), Oechssler et al. (2022), Gürdal et al. (2024), and Casari et al. (2025), where the last show that the mechanism works for groups as large as 15 players.[5] That it actually also works in the field has been shown (albeit at a small scale) in a Ukraine fund raiser, where we used a slightly different version under the name “You contribution squared” and raised more than €60.000.[6]

To generalize those earlier contributions, I allow for a much broader class of dilemma games. Most of the literature is concerned with (linear) public good games or multi-person prisoners’ dilemmas. A seminal paper by Dawes (1980) defines dilemma games as N-person games with two actions, defect and cooperate, where defect is strictly dominant and if all players cooperate, this is better for everyone. To account for the complicated action spaces e.g. in climate agreements, I allow for multidimensional action spaces $A_i \subseteq \mathbb{R}_+^k$. For example, in an international climate policy context, $a_{i,1}$ might be country i’s emissions mitigation, $a_{i,2}$ and $a_{i,3}$ might be its investments into a climate adaptation fund

[3] Somewhat overoptimistically, we called the last paper the “general case” because it generalized the binary case. But it was not general at all since it considered a linear public good game.

[4] See e.g. Oechssler et al. (2022) for a more systematic literature review.

[5] MacKay et al. (2015) and Schmidt and Ockenfels (2021) study related mechanisms that are particularly geared towards reducing CO2 emissions through carbon pricing.

[6] This fundraiser was organized by Georg Weizsäcker and the current author. The idea was that participants promise to donate X€ if at least X people would also promise to donate at least X€. See https://yourcontributionsquared.eu/en/ for details.

and research into renewable energy, and $a_{i,4}$ a binary variable indicating its agreement to ban deforestation. Obviously, it makes sense to allow for any number of actions rather than just two.[7] The definition of dilemma games I propose is close to the one adopted by Peña and Nöldeke (2023) but is generalized to the case of more than two actions. Accordingly, a game is a dilemma game if (1) there is a Pareto dominated Nash equilibrium and (2) players' actions (or contributions) exert a weakly positive externality on all other players.[8] This allows for all the usual dilemmas, like public good games, tragedy of the commons, and multi-person prisoners' dilemmas. But it also allows for situations where players are intrinsically motivated to contribute some positive (but insufficient) amount on their own. Finally, it allows for stag-hunt coordination games with Pareto dominated equilibria.

The CCM mechanism studied in this paper is particularly suitable when the dilemma game is (finitely) repeated like the different rounds of climate conferences. In the one-shot play of the CCM, there are many Nash equilibria, including the undesirable one, in which the dominated default action is played every time. In the repeated play of the mechanism, players have the chance to learn. I assume that they play better responses to the previous round. But they are also somewhat forward looking and anticipate that other players may change their strategies too. Thus, among the better responses, they choose better responses that are unexploitable by other players, in the sense that feasible outcomes that are worse than the status quo are excluded.

The main result of the paper is that the resulting unexploitable better response process will converge to Pareto optimal states. This holds even when I allow for players who do not care at all for the contributions of others and if other players do not know about their existence. Thus, despite the fact that no player knows how much other players care about contributions, the process will eventually reach a Pareto optimal state.

2 Dilemma games

Let I be a finite set of $N \geqslant 2$ players playing a normal form game. For each $i \in I$, let $A_i \subseteq \mathbb{R}^k_+$ be a finite, non-empty set of possible actions (or contributions), with the vector

[7] In a strict literal sense, a dilemma refers to only two actions but it is not uncommon to be used in situations with more than two actions. See e.g. the Oxford English Dictionary's (2025) definition: "A choice between two (or, loosely, several) alternatives, which are or appear equally unfavourable."

[8] Liebrand (1983, p. 135) defines social dilemmas as situations in which, "by the very act of choosing a strategy with negative externalities, the ultimate outcome can be called deficient." Note that one can always reverse the sign of the action space to change positive into negative externalities.

of zero contributions included, $0 \in A_i$. An action profile $a = (a_i)_{i \in I} \in A$ denotes an action for each player. As usual I write $a = (a_i, a_{-i}) \in A$ when composing action profiles and write $a < a'$ if $a_i \leqslant a'_i, \forall i$ and $a \neq a'$.

For each $i \in I$, let $\succcurlyeq_i$ denote i's preferences on A, which I assume to be complete and transitive. Let $\succ_i$ and $\sim_i$ be its strict and symmetric parts. I say that a' is a Pareto improvement over a, $a' \succ_I a$, if $a' \succcurlyeq_i a, \forall i \in I$ and $a' \succ_j a$ for at least one j. An action profile a is Pareto optimal if there is no $a' \in A$ that is a Pareto improvement over a. I focus on Pareto optimality rather than some welfare measure like the sum of utilities since I do not want to take a stance on cardinal utility and in particular on interpersonal comparisons of utilities.[9] In most applications, it will be plausible to assume that players only care about the aggregate contributions of all other players, $A_{-i} := \sum_{j \neq i} a_j$. Formally, $(a_i, a_{-i}) \sim_i (a_i, a'_{-i})$ if $A_{-i} = A'_{-i}$. That is, players do not care who makes those contributions as long as someone makes them. Of course, there may be some situations where players do care who makes which contributions. At the cost of more cumbersome notation, this could be accommodated by making preferences depend on the whole vector a_{-i}. However, it would make the application of the CCM mechanism rather unwieldy in practice as players would have to provide very detailed and long lists of conditions. I hence restrict my analysis to the case where players care about other players' cumulative contributions.

In a seminal paper, Dawes (1980) defines dilemma games as N-person games with two actions, defect (here denoted as 0) and cooperate (denoted as 1), $A_i = \{0, 1\}$. Furthermore, defect is strictly dominant (and hence the unique Nash equilibrium), $(1, a_{-i}) \succ_i (a'_i, a_{-i}), \forall a_{-i}$ and $a'_i \neq 1$, and finally, if all players cooperate, this is better for everyone, $(1, 1, ..., 1) \succ_I (0, 0, ..., 0)$.

The generalization I propose is very close to the one used by Peña and Nöldeke (2023),[10] except that rather than just having two actions, defect and cooperate, in my setting I allow cooperation to be gradual, where choosing to "contribute more" corresponds to a larger a_i. For this to make sense, actions spaces in a dilemma game need to be (partially) ordered such that the following condition is satisfied.

PosExternality Players' actions (or contributions) exert a weakly positive externality on all other players i,

$$(a_i, a'_{-i}) \succsim_i (a_i, a_{-i}) \text{ if } A'_{-i} \geq A_{-i}.$$

[9] This is also the reason why I focus on pure strategies rather than mixed ones.

[10] Related definitions of binary dilemma games are discussed in Kollock (1998) and Nowak (2012). Peña and Nöldeke (2023) contain a very through discussion of the various definitions in the literature.

This is of course satisfied in the canonical case of public goods. If others' actions exert a negative externality, like in the tragedy of the commons, one can simply reverse the sign of all actions. For example, rather than deciding to pollute, the action would be to avoid pollution.

In many dilemma games, zero contribution is in fact a dominant strategy as Dawes (1980) assumes. However, important strategic aspects of dilemma games are preserved if I assume that there is a Nash equilibrium a^0 in which players contribute something positive (Harstad (2024) calls this the "business-as-usual" outcome). This Nash equilibrium acts as a threat point if all agreements fail.[11] A crucial ingredient of dilemma games is that there is a Pareto optimal outcome that dominates the Nash equilibrium a^0. But I can allow for players who do not care at all for others' contributions, $(a_i, a'_{-i}) \sim_i (a_i, a_{-i}), \forall a_i, a_{-i}, a'_{-i}$. For simplicity, I assume that these players have one best action profile a_i^*, $(a_i^*, a_{-i}) \succ_i (a_i, a_{-i}), \forall a_{-i}, a_i \neq a_i^*$ and let I^* denote the set of these players. In the following I assume that players in I^* always choose their best action profile. For them there is thus no room for improvement, which gives rise to the following definition.

Definition 1 *An outcome a' is a strict* Pareto improvement over a if a' is a Pareto improvement over a that is strict for all players $i \in I \backslash I^*$.*

Note that if there exists an $a' \succ_{I \backslash I^*} a^0$, then there exists a Pareto optimal a^* such that $a^* \succ_{I \backslash I^*} a^0$ since A is finite.

Pareto There exists a (Pareto optimal) outcome a^* that is a strict* Pareto improvement over some pure Nash equilibrium a^0.

I can now define the class of dilemma games.

Definition 2 *A normal form game is a* dilemma game *if it satisfies Conditions PosExternality and Pareto.*

The class of dilemma games describes well the issues I deal with in this paper. It describes a society that is stuck at a focal or status quo equilibrium a^0 but a Pareto optimal outcome a' exists that is preferred by all players. The class of games includes public good games, common pool resource games, the tragedy of the commons, prisoners'

[11] Recall that a^0 is supposed to be status quo point. Thus, it seems plausible to assume that players are aware of the Nash equilibrium a^0 even though they are not assumed to know others' preferences.

dilemma games but also games with multiple Pareto ranked Nash equilibria (e.g. the stag-hunt game, bank runs, weakest link games Riedl et al. (2016), or minimum effort games Van Huyck et al. (1990)) and many other games.

Table 1 shows some simple one-dimensional examples of dilemma games. Apart from the prisoners' dilemma, the most important examples are public good games. The linear case has been studied in Oechssler et al. (2022). But more generally, all baseline models on global public goods discussed in Buchholz and Sandler (2021) and Harstad (2024) fit into this framework. Another important class are common pool resource games and, if we ignore consumers' welfare, oligopoly games, like Cournot or Bertrand oligopolies. Tullock contests Tullock (1980) also count if again one reverses the sign of the actions to satisfy Condition PosExternality.

Table 1: Examples of one-dimensional dilemma games

Game	Description
Prisoners' dilemma	$A_i = \{0,1\}$, where $a_i = 0$ is defect and $a_i = 1$ is cooperate, $a^0 = (0,0)$ and Pareto optimal action profile is $(1,1)$.
Public good games	Linear public good game with $n \geq 2$ players. Contributions $a_i \in A_i \subset [0,1]$, MPCR is θ_i, payoff function $u_i(a) = 1 - a_i + \theta_i \sum_{j=1}^{n} a_j$, with $\sum_{i=1}^{n} \theta_i > 1$ $a^0 = (0,0,....,0)$ and Pareto optimal action profile is $(1,1,...,1)$.
Common resource game Cournot oligopoly	$u_i(a) = g(a_i, \sum_i a_i)$, where g is decreasing in second argument, overfishing, tragedy of the commons, NATO burden-sharing.
Tullock contests	$u_i(a_i, a_j) = \frac{a_i^r}{a_i^r + a_j^r}$, $r > 0$.
Travellers' dilemma	Basu (1994)
Games with multiple Pareto ranked equilibria	stag-hunt games, minimum effort game (Van Huyck et al., 1990),

Dilemma games can however also have multiple Pareto ranked equilibria, like the minimum effort game Van Huyck et al. (1990) or (some) stag-hunt games e.g.

	stag	hare
stag	$8, 8$	$0, 6$
hare	$6, 0$	$4, 4$

namely those, where players wants the other players to play stag even if they themselves plan to play hare (this implies that Condition PosExternality is satisfied). These are stag-hunt games where cheap-talk has theoretically no value (see Aumann, 1990).[12]

3 The conditional contribution mechanism

I define the Conditional Contribution Mechanism (CCM) as $G^{CCM} := (M^{CCM}, g^{CCM})$, where M^{CCM} describes the mechanism's message space and $g^{CCM} : M^{CCM} \mapsto A$ describes the mechanism's outcome function. In other words, the mechanism's outcome is an action profile for the underlying dilemma game with a given Nash equilibrium profile a^0. A player's message consists of *two* conditional statements that tie their action to others' actions. Each statements is of the form "I am willing to contribute a_i to the public good if others' contributions are at least A_{-i}". The set of statements is denoted by M_i. Since players are allowed to send two conditional statements, the message space in the CCM is given by $M^{CCM} := \prod_{i=1}^{n} M_i^{CCM}$, where $M_i^{CCM} := M_i \times M_i$. As a special case, this allows players to free-ride completely, by stating $((0, A_{-i}), (0, A_{-i}))$, or to unconditionally contribute an action a_i, by stating $((a_i, 0), (a_i, 0))$.

For a message profile $m \in M^{CCM}$, the outcome $g^{CCM}(m)$ of the CCM is then determined as follows.

1. Let $A^{CCM}(m) \subseteq A$ be the set of *feasible* outcomes for a message profile $m \in M^{CCM}$, that is the set of outcomes that is compatible with at least one of the two conditional statements for each player,

$$a \in A^{CCM}(m) \Leftrightarrow \; \forall \, i \in I, \exists l_i \in \{1, 2\} \text{ s.t. } a_i = a_i^{l_i} \text{ and } A_{-i} \geq A_{-i}^{l_i}. \tag{1}$$

[12] Experiments show that cheap-talk can help to coordinate on the payoff-dominant equilibrium (Charness, 2000, Clark et al., 2001, Duffy and Feltovich, 2002) in symmetric stag-hunts. This does not work so well anymore in asymmetric stag-hunts, though (Agranov, 2024).

2. If $A^{CCM}(m)$ is not empty, the mechanism will select one outcome $a^{'} \in A^{CCM}(m)$ using some arbitrary probability distribution on $A^{CCM}(m)$ that has full support, i.e., chooses each element with positive probability.

3. If $A^{CCM}(m)$ is empty, the outcome of the mechanism is the vector of default actions $a^{'} = a^0$.

4. The mechanisms *automatically* changes *both* conditions for each player to agree with the chosen outcome, $((a'_i, A'_{-i}), (a'_i, A'_{-i}))$. I call this the adjusted message for each player.

5. The mechanisms supplies all players with the chosen outcome in period t, which is denotes by $g(m^t_i, m^t_{-i})$, and with feedback about the adjusted message profile m^t.

Steps (4) and (5) are the key innovations and require some discussion. Once a feasible new action vector a' is chosen, there is always the risk of backsliding in the following periods. This is why the mechanism adjusts all messages to $((a'_i, A'_{-i}), (a'_i, A'_{-i}))$ in step (4) and this is the only feedback players receive in step (5). Given these adjusted messages, players know they risk dropping back to a^0 if they choose messages in the next period that would lower their contributions. But why should players agree to submit themselves to such a mechanism (recall that we require voluntary participation)? The reason is that the adjusted messages do not require them to contribute more than they have already agreed to in the current period. And even though the new messages establish a default for the next period, this default is only relevant if players do not change their message. Since players are always free to change their messages in the next period, the mechanism does not restrict their possible actions in any way. The only effects the adjusted messages have is on the beliefs players have about other players' behavior, and, by setting a default, it works like a nudge.

However, in the next section, I will make the assumption that players use the adjusted messages as input in their best- or better reply process. This assumption is new and it is crucial, so it requires some discussion. Why would players assume that other players will stick to a message even though they know that these messages were possibly adjusted by the mechanism? The first obvious reason is that players have nothing else to base their beliefs on as the mechanism does not report the original messages chosen by players as feedback. Furthermore, as argued above, players will never have to contribute more with the adjusted messages than with their original messages so that it seems reasonable for them to accept the adjusted messages as the status quo. Of course, as with any best response dynamic, what is assumed is that players are somewhat myopic but this is an assumption for which there is ample experimental evidence (see e.g. Healy, 2006).

4 Dynamic behavior in the CCM

In this section, I analyze the properties of the CCM under dynamic considerations, where the assumptions on the dynamic behavior follow closely Oechssler et al. (2022).[13] In the dynamic model, players play the same CCM game recurrently over several periods in fixed groups. I follow the literature (in particular, Cabrales and Serrano, 2011) and assume that players are myopic and treat the *adjusted messages* of the other players from the previous period, as reported back by the mechanism in step 4, as a prediction of the other players' messages in the next period.

There is evidence (see e.g. Healy, 2006) that players' behavior in recurrent public good mechanisms can be well described using best response dynamics.[14] In my case, given that the CCM mechanism reacts in a very discontinuous fashion,[15] I find it most plausible if players react only to the most recent information from the previous period.[16]

Given the evidence that players contribute in public good games even when it is a dominant strategy not to (see e.g. Ledyard (1995)) and thus do seem to not fully exploit their strategic advantages, I allow players to simply choose a *better* response (abbreviated as BR). Formally, given the status quo outcome $g(m_i^t, m_{-i}^t)$ and the adjusted message profile m^t, a message m_i^{t+1} is a *better response* for player i if player i weakly prefers all possible new outcomes $g(m_i^{t+1}, m_{-i}^t)$ to the status quo,

$$g(m_i^{t+1}, m_{-i}^t)) \succeq_i g(m_i^t, m_{-i}^t). \tag{2}$$

Definition 3 *In Better Response Dynamics (BRD) each player i switches in period $t+1$ to a message m_i^{t+1} that is a better response. If several better responses exist, all of them*

[13] Since the class of games considered here is much broader and the mechanisms differ, the results of Oechssler et al. (2002) do not apply, of course.

[14] There is also evidence from strategically similar Cournot games that many subjects are well described by best response behavior (see e.g. Huck et al. 1999).

[15] Since the message space is finite, the word "discontinuous" is not meant literally here. Instead it describes the fact that a small change in one agent's message can change the outcome from full contribution to zero contribution or the other way round.

[16] Of course, some agents will be more sophisticated and forward looking. I try to account for that by introducing the concept of unexploitability below.

are chosen according to some arbitrary probability distribution that has full support.

Definition 4 *Given an outcome $a^t = g(m_i^t, m_{-i}^t)$, a message m_i^{t+1} is called* exploitable at a^t *if there is any $m_{-i}^{t+1} \in M_{-i}$ such that there exists a possible new feasible outcome $g(m_i^{t+1}, m_{-i}^{t+1}) = a^{t+1} \in A^{CCM}(m^{t+1})$, with $a^{t+1} \prec_i a^t$. A message m_i^{t+1} is called unexploitable at a^t if it is not exploitable.*

In words, a message of player i is unexploitable at a^t if there is no chance that after any deviation of other players, player i is worse off than in the previous period. Note that all possible message profiles m_{-i} of other players are considered. One could argue, since I assume a BRD model, that I should only consider profiles of better responses of other players at this point. However, since I allow for the general case, in which players have no information on the preferences of other players, players cannot tell whether a certain message of another player is a better response. Therefore, from a player's perspective it seems prudent to account for all possible choices.

When I combine the BRD with the requirement that messages be unexploitable, I get the following.

Definition 5 *An unexploitable better response dynamic (UBRD) is a BRD dynamic with the restriction that players only choose unexploitable messages.*

Unexploitable better responses always exist since the status quo message (after adjustment through the mechanism) is a better response and it is unexploitable. It is a better response because it yields the same outcome if other players do not adjust their messages. And it is unexploitable because either other players weakly increase A'_{-i}, which would weakly increase i's payoff, or they decrease A'_{-i} in at least one dimension, in which case there would be no feasible outcome and the process would revert to a^0.[17]

The UBRD defines a Markov chain on the (finite) state space M^{CCM}. By standard results (see e.g. Karlin and Taylor, 1975, p.64) states (i.e. message profiles) can be partitioned into transient profiles and recurrent profiles.

Definition 6 *A recurrent class of UBRD is a set of message profiles, which, if ever reached by the dynamics, is never left and which contains no smaller set with the same property.*

[17] Recall that for a message to be exploitable it must result in a *feasible* outcome for the message profile.

Lemma 1 *If the UBRD has reached an action profile a^t that is a strict* Pareto improvement over a^0, then each further step of the UBRD will be a weak Pareto improvement.*

Proof. Suppose in period t the mechanism has chosen a new action profile a^t that is a strict* Pareto improvement over a^0. Then it must adjust the conditions to $((a_i^t, A_{-i}^t), (a_i^t, A_{-i}^t))$ for each player. I claim that players will choose the next message m_i^{t+1} such that

(A) the status quo a^t remains feasible,

(B) and no a^{t+1} such that $a^{t+1} \prec_i a^t$, for any i, becomes feasible (including a^0).

If claims (A) and (B) hold, the mechanism can choose only action profiles a^{t+1} in period $t+1$ that are weakly better than a^t for all players.

To prove claim (A), suppose a^t became infeasible. This can only happen if at least one player $i \in I \backslash I^*$ switches to messages $(a_i^1, A_{-i}^1), (a_i^2, A_{-i}^2)$ such that both conditional statements in equation (1) are violated. That is, either $a_i^{l_i} \neq a_i^t$ or $\neg \left(A_{-i}^{l_i} \leq A_{-i}^t\right)$ for both statements l_i. A unilateral deviation by i to some message with $a_i^{l_i} \geq a_i^t, a_i^{l_i} \neq a_i^t$, and $A_{-i}^{l_i} \leq A_{-i}^t$, would not be a better response as player i would be the only player contributing more. A unilateral deviation by i to some some message with $a_i^{l_i}$, which would contribute less in at least one dimension, i.e. $\neg \left(a_i^{l_i} \geq a_i^t\right)$ or with some $A_{-i}^{l_i}$ such that $\neg \left(A_{-i}^{l_i} \leq A_{-i}^t\right)$, would imply that $A^{CCM}(m)$ were empty and a^0 would be chosen by the mechanism. But this would not be a better response for the deviating player since a^t was already a strict* Pareto improvement over a^0.

To prove claim (B), suppose that some a^{t+1} became feasible such that $a^{t+1} \prec_i a^t$ for player i. However, note that player i can always prevent a move to a^{t+1} by simply not changing the conditions $((a_i^t, A_{-i}^t), (a_i^t, A_{-i}^t))$. Thus any change by player i that would make a^{t+1} feasible would be exploitable. □

I am now ready to state the main theorem of my paper.

Theorem 1 *Any outcome of a recurrent class of the CCM under UBRD is Pareto optimal.*

Proof. Suppose a is an outcome of a recurrent class but not Pareto optimal. I will show a contradiction by showing that from a there is a path with strictly positive probability to some a' that is a strict* Pareto improvement over a^0. I will show that the UBRD can move to a' with positive probability. Furthermore, from a' the UBRD process cannot return to a, which proves the desired contradiction that a cannot be an outcome of a recurrent class.

Case 1) Suppose that a is already a strict* Pareto improvement over a^0. Given that a was chosen by the mechanism, the current default conditions are $((a_i, A_{-i}), (a_i, A_{-i}))$. Since a is not Pareto optimal, there exists some a' that is a Pareto improvement over a. Consider the following pair of messages,

$$\left((a_i, A_{-i}), (a'_i, A'_{-i})\right),$$

where again $A'_{-i} = \sum_{j \neq i} a'_j$. I claim that those messages are unexploitable best replies for all $i \in I \backslash I^*$. They are best replies because the outcome will not change if one player deviates unilaterally. They are unexploitable because the only new feasible outcome that can be chosen is weakly preferred by all players. To see this, note that if player i has to contribute a'_i other player need to contribute at least A'_{-i}. If others contribute exactly A'_{-i}, then i is weakly better off since a' is a Pareto improvement over a. If others contribute more than A'_{-i}, this is even better for i due to Condition PosExternality in the definition of a dilemma game.

Thus, these messages are messages are unexploitable best replies and are hence chosen with strictly positive probability and, again with strictly positive probability, the mechanism will select a' as the new outcome. Given that a' is a Pareto improvement over a and therefore a strict* Pareto improvement over a^0, by the Lemma, the UBRD cannot return to a.

Case 2) If a is not a strict* Pareto improvement over a^0, there are some players $i \in I \backslash I^*$, for which violating the binding condition is an unexploitable BR, which would return the process to a^0. By Condition Pareto in the definition of a dilemma game, there exists an a' that is a strict* Pareto improvement over a^0. With positive probability the process would move to a' and I can continue as in Case 1. □

The UBRD process may be rather slow in some games. Two ideas may speed it up. First, if the actions space A_i is one-dimensional, $A_i \subset \mathbb{R}_+$, then the CCM mechanism can be modified by picking in Step 2 the unique action profile $a \in A^{CCM}(m)$ with the largest sum of contributions $\sum_i a_i$.[18] This should speed up the convergence to a Pareto optimal action profile considerably. Second, if the action profile is multi-dimensional, players could be asked to indicate which of their two conditions they would prefer if implemented. The mechanism then could implement those profiles in $A^{CCM}(m)$ with higher probability that are preferred by a majority of players. This would not change the recurrent class of the

[18] The action profile is unique as shown in Oechssler et al. (2022, equation 3)

CCM but may speed up convergence.

5 Conclusion

This study has introduced and examined the conditional contribution mechanism (CCM) designed to address a wide class of dilemma games. The mechanism relies on binding unilateral commitments that are conditional on others' contributions, thereby eliminating the need for a central authority to enforce multilateral agreements. The main result is that under unexploitable better response dynamics, the CCM converges to recurrent states that are necessarily Pareto optimal.

By extending the framework beyond binary action spaces or linear public good settings to multidimensional action spaces, the analysis substantially broadens the applicability of conditional contribution mechanisms. The results thus provide a theoretical foundation for self-enforcing agreements in contexts as diverse as international climate negotiations, common pool resource management, and coordination problems with multiple Pareto-ranked equilibria. Crucially, the mechanism remains effective under incomplete information and heterogeneous preferences, highlighting its robustness in realistic strategic environments.

While the convergence process may in some cases be slow, the proposed refinements suggest avenues for accelerating adjustment without undermining stability. Future research could further investigate such refinements, explore the empirical performance of the mechanism in larger and more heterogeneous populations, and examine its interaction with alternative institutional arrangements. Overall, the findings underscore the potential of conditional contribution mechanisms as a theoretically rigorous and practically viable approach to overcoming persistent inefficiencies in social dilemma situations.

References

Agranov, M. (2024). Communication in stag hunt games: When does it really help? *Economics Letters*, 244:111991.

Aumann, R. (1990). Nash equilibria are not self-enforcing. In Gabszewicz, J. J., Richard, J.-F., and Wolsey, L., editors, *Economic decision making: Games, econometrics and optimisation: Essays in Honor of Jacques Dreze*, pages 201–206. Amsterdam, Elsevier.

Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. *The American Economic Review*, 84(2):391–395.

Buchholz, W. and Sandler, T. (2021). Global public goods: a survey. *Journal of Economic Literature*, 59(2):488–545.

Cabrales, A. and Serrano, R. (2011). Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms. *Games and Economic Behavior*, 73(2):360 – 374.

Casari, M., Ordaz Cuevas, J., and Tavoni, A. (2025). I will if you will in climate mitigation: Conditional cooperation in the lab. Technical report, University of Bologna.

Charness, G. (2000). Self-serving cheap talk: A test of aumann's conjecture. *Games and Economic Behavior*, 33(2):177–194.

Clark, K., Kay, S., and Sefton, M. (2001). When are nash equilibria self-enforcing? an experimental analysis. *International Journal of Game Theory*, 29(4):495–515.

Clarke, E. (1971). Multipart pricing of public goods. *Public Choice*, 11:17–33.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31:169Ű193.

Duffy, J. and Feltovich, N. (2002). Do actions speak louder than words? an experimental comparison of observation and cheap talk. *Games and Economic Behavior*, 39(1):1–27.

Groves, T. and Ledyard, J. (1977). Optimal allocation of public goods: A solution to the "free rider" problem. *Econometrica*, 45(4):783–809.

Gürdal, M. Y., Gürerk, Ö., Kacamak, Y., and Kart, E. (2024). How to increase and sustain cooperation in public goods games: Conditional commitments via a mediator. *Journal of Economic Behavior & Organization*, 228:106789.

Guttman, J. M. (1978). Understanding collective action: matching behavior. *The American Economic Review*, 68(2):251–255.

Guttman, J. M. (1986). Matching behavior and collective action: Some experimental evidence. *Journal of Economic Behavior & Organization*, 7(2):171–198.

Harstad, B. (2024). The politics of global public goods. Technical report, National Bureau of Economic Research.

Healy, P. J. (2006). Learning dynamics for mechanism design: An experimental comparison of public goods mechanisms. *Journal of Economic Theory*, 129(1):114 – 149.

Healy, P. J. and Mathevet, L. (2012). Designing stable mechanisms for economic environments. *Theoretical Economics*, 7(3):609–661.

Heitzig, J. (2019). Efficient non-cooperative provision of costly positive externalities via conditional commitments. *Available at SSRN*.

Huck, S., Normann, H.-T., and Oechssler, J. (1999). Learning in Cournot oligopoly–an experiment. *The Economic Journal*, 109(454):80–95.

Karlin, S. and Taylor, H. M. (1975). *A first course in stochastic processes*. Academic Press, second edition.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, 24(1):183–214.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In Kagel John, Roth, A., editor, *Handbook of Experimental Economics*. Princeton University Press, Princeton.

Liebrand, W. B. (1983). A classification of social dilemma games. *Simulation & Games*, 14(2):123–138.

MacKay, D. J. C., Cramton, P., Ockenfels, A., and Stoft, S. (2015). Price carbon - I will if you will. *Nature News*, 526(7573):315–316.

Nowak, M. A. (2012). Evolving cooperation. *Journal of theoretical biology*, 299:1–8.

Oechssler, J., Reischmann, A., and Sofianos, A. (2022). The conditional contribution mechanism for repeated public goods–the general case. *Journal of Economic Theory*, 203:105488.

Peña, J. and Nöldeke, G. (2023). Cooperative dilemmas with binary actions and multiple players. *Dynamic Games and Applications*, 13(4):1156–1193.

Reischmann, A. and Oechssler, J. (2018). The binary conditional contribution mechanism for public good provision in dynamic settings - theory and experimental evidence. *Journal of Public Economics*, 159:104–115.

Riedl, A., Rohde, I. M., and Strobel, M. (2016). Efficient coordination in weakest-link games. *The Review of Economic Studies*, 83(2):737–767.

Schmidt, K. M. and Ockenfels, A. (2021). Focusing climate negotiations on a uniform common commitment can promote cooperation. *Proceedings of the National Academy of Sciences (PNAS)*, 118:e2013070118.

Tullock, G. (1980). Efficient rent seeking. In Buchanan, J. M., Tollison, R. D., and Tullock, G., editors, *Toward a theory of the rent-seeking society*, pages 97–112. Texas A&M University Press A&M University Press, College Station.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review*, 80(1):234–248.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37.