Commitment Design*

Siyang Xiong†

October 12, 2018

Abstract

A simple model of commitment design is proposed, or equivalently, we rigorously introduce a new paradigm (i.e., *commitment design*) for economic design problems, besides the two current paradigms (i.e., the classical mechanism design and information design). Furthermore, we prove that Pareto efficiency can always be achieved via a commitment protocol. This is in sharp contrast to previous results in the literature, e.g., the impossibility result in Arieli, Babichenko, and Tennenholtz (2017) and the folk theorem in Kalai, Kalai, Lehrer, and Samet (2010).

*Preliminary and incomplete.

†Department of Economics, University of Bristol, siyang.xiong@bristol.ac.uk

...where do commitment devices come from? Is there an outside entity (other than the players of the game) able to construct commitment spaces for the players, or are commitment devices something the players generate on their own? Under the former case, the study of commitment may be viewed as a subarea of the implementation literature,.... The implementation literature raises another issue...., it may be desirable to generate only the Pareto efficient ones or even subsets of these, like ones consisting of "fair" outcomes.

— (Kalai, Kalai, Lehrer, and Samet, 2010, pp.134)

1 Introduction

Suppose that a group of agents is about to play a game denoted by G, which results in inefficient outcomes (e.g., G is the Prisoner's Dilemma game). As social planners, we do not know what G is (i.e., who moves first, who moves second, how the game evolves... etc.). Rather, we know *only* that G comes from a set of games, denoted by $\mathcal{G}$, i.e., $G \in \mathcal{G}$.

Suppose our sole goal is efficiency. Clearly, if we knew G, we could simply enforce the efficient outcome in G. Without knowing G, we do not even know what is the efficient outcome, because it may vary for different games in $\mathcal{G}$. Given this, is there anything we can do so as to *always* achieve efficiency (regardless of G being any game in $\mathcal{G}$)?

It has been well known that players' pre-game commitments (on not playing some strategies) may substantially change the outcome of a game. For instance, consider the "battle of sex" game described as follows.

	boxing	concert
boxing	$3,1$	$0,0$
concert	$0,0$	$1,3$

Clearly, if the row player is the first one to make a pre-game commitment, he would commit to choose "boxing", which would result in his desired equilibrium (boxing, boxing). Similarly, if the column player is the first one to make a pre-game commitment, she would commit to choose "concert", which would result in her desired equilibrium (concert, concert). This example further shows that the order of players making (pre-game) commitments may also substantially change the resulting outcome.

A *commitment protocol* describes *completely* how players make pre-game commit-

ments (i.e., who is the first one to make a commitment, and how we proceed given any commitment made by the first player....). As social planners, we may enforce a commitment protocol, so that players follow the protocol to make commitments, and given the agreed commitments, players follow the commitments to play the true game $G \in \mathcal{G}$. The question we study in this paper is: with knowledge of the set $\mathcal{G}$ only, can we find a commitment protocol, which would *always* induce efficient outcomes for any true game $G \in \mathcal{G}$? We provide several positive answers for this question.

To rigorously study this question, a simple model of commitment design is proposed in this paper. More importantly, we rigorously introduce a new paradigm for economic design problems, i.e., *commitment design*, besides the two current paradigms (i.e., the classical mechanism design and information design). To see this, note that there are three dimensions in an economic design problem: (1) games, (2) information and (3) (pre-game) commitments. In the classical mechanism design, we fix (2) and (3), and construct the optimal game to achieve a social goal. In information design, we fix (1) and (3), and construct the optimal information structure to achieve a social goal. In commitment design, we fix (1) and (2), and construct the optimal commitment protocol to achieve a social goal.

Specifically, we are facing a set of games $\mathcal{G}$, a solution concept $\mathcal{S} : \mathcal{G} \longrightarrow \mathcal{O}$ and a social goal $\mathcal{E} : \mathcal{G} \longrightarrow \mathcal{O}$, where $\mathcal{O}$ is a set of possible social outcomes. The problem that we face is

$$\exists G \in \mathcal{G} \text{ such that } \neg \left[\mathcal{S}(G) \approx \mathcal{E}(G)\right],$$

i.e., $[\mathcal{S}(G) \approx \mathcal{E}(G)]$ does not hold for some game $G \in \mathcal{G}$, or equivalently, we cannot achieve our goal $\mathcal{E}$ for some $G \in \mathcal{G}$. Note that we do not define "$\mathcal{S}(G) \approx \mathcal{E}(G)$" rigorously here, and different definitions of it correspond to different notions of implementation, e.g., partial implementation and full implementation (see definitions in Section 2).

Given a commitment protocol, denoted by CP, a mega-game $CP \circ G$ is defined for each $G \in \mathcal{G}$. That is, in the mega-game $CP \circ G$, players first follow CP to make commitments, and then follow the agreed commitments to play $G \in \mathcal{G}$. Suppose that the solution concept $\mathcal{S} : \mathcal{G} \longrightarrow \mathcal{O}$ extends to $\mathcal{S} : CP \circ \mathcal{G} \longrightarrow \mathcal{O}$, where $CP \circ \mathcal{G} \equiv \{CP \circ G : G \in \mathcal{G}\}$. Then, we say that we achieve a social goal $\mathcal{E}$ via a commitment protocol CP if and only if

$$\mathcal{S}(CP \circ G) \approx \mathcal{E}(G), \forall G \in \mathcal{G}.$$

In this paper, we consider

$$\mathcal{E}: \text{ Pareto efficiency,}$$

$$\mathcal{S}: \text{ subgame perfect Nash equilibria;}$$

$$\mathcal{G}: \text{ an arbitrary set of finite and bounded games;}$$

$$\approx: \text{ either full implementation or partial implementation,}$$

and we find commitment protocols which *always* induce efficient outcomes for any game $G \in \mathcal{G}$.

The commitment protocols proposed in this paper are actually inspired by real-life phenomena. To motivate them, let us draw a couple of observations. First, eighteen rounds of negotiation between China and World Trade Organization (WTO) had been conducted, before China finally joined WTO in 2001.[1]

Why eighteen rounds (as opposed to one round) of negotiation?

In fact, this is not a unique phenomenon. We always observe multiple rounds of negotiation, when two or more international parties try to resolve economic and/or political conflicts (e.g., trade negotiation, truce negotiation). A recent prominent example is the multi-round Brexit negotiation between United Kingdom and European Union.

A simple explanation is that the issue involved in negotiation may be too complicated to resolve in one round. Besides this naive reason, is there any fundamental reason (regarding strategic concerns of the parties involved) that could explain this phenomenon of multi-round negotiation? In particular, we ask this question from the angle of a mechanism designer (or a social planner): is this mechanism of multi-round negotiation superior to other mechanisms (e.g., one-round negotiation)? If yes, in what sense?

Second, at the end of each round of negotiation, a pact (i.e., a contract) is usually signed by all involving parties if an agreement is reached. — Interestingly, this is another common feature shared by almost all international negotiation practices. Does such pact signing serve any other (strategic) purposes besides being a procedure of formality?

We regard the process of multi-round negotiation as a "commitment device," and propose a notion of *K-round negotiation protocol*, which summarizes the features contained in the two observations. To see the connection, consider a typical (and hypothetical) con-

[1] See https://www.wto.org/english/thewto_e/acc_e/a1_chine_e.htm .

versation between China and WTO in the eighteen-round negotiation:

$$\left[\begin{array}{c} \text{China: we promise to reduce tariffs for automobiles by 80\% in 3 years;} \\ \text{WTO: we promise to grant free access to Chinese beef in all of our member countries in 2 years.} \end{array}\right]$$

That is, both parties make commitments (for a game that will be played later) during the negotiation, or equivalently, the negotiation serves as a commitment device which shapes the restricted game that will be played later.

Specifically, under the K-round negotiation protocol, there are K rounds of negotiation in total, like the first observation described above. Furthermore, in each round, players sequentially announce their commitment, followed by a voting by the players regarding whether they accept the proposed commitment profile.—Like pact signing in the second observation, the proposed commitment profile becomes effective if and only if all players vote yes (or equivalently, sign the pact). Whenever an agreement is reached in any round, they proceed to play the true game, following the agreed commitment. If no agreement is reached in all of the K rounds, they proceed to play the true game without commitment.[2]

We prove that, without knowing fine details of the true game, we can always achieve Pareto efficiency via the K-round negotiation protocol described above. This is a quite surprising result, especially when it is compared to that in Arieli, Babichenko, and Tennenholtz (2017),[3] which propose a particular class of commitment protocols, called DFS mechanisms. Arieli, Babichenko, and Tennenholtz (2017) prove a positive result and a negative result: when there are two agents only, Pareto efficiency can always be implemented by an DFS mechanism (for any generic games); and when there are four or more agents, Pareto efficiency cannot be implemented by any DFS meconiums.[4] In contrast, we show that Pareto efficiency can always be implemented, and clearly, the K-round negotiation protocol does not belong to the class of DFS mechanisms.

[2]Our K-round negotiation protocol may not 100% match the WTO-China negotiation process.—The point is not to describe 100% of the real-life negotiation. If we consider a different commitment protocol that summarizes more features of the WTO-China negotiation (than the K-round negotiation protocol does), it may achieve more social goals than that considered in this paper.

[3]The idea of "commitment design" (i.e., an implemenation approach of commitment) is first discussed in Kalai, Kalai, Lehrer, and Samet (2010), and Arieli, Babichenko, and Tennenholtz (2017) is the first paper which takes an implementation approach on commitment, even though a formal model of "commitment design" is not provided in that paper.

[4]For DFS mechanisms, the case of three agents remains an open question.

Nevertheless, the intuition of our result is quite simple, and it consists of two parts, an easy part and a difficult part. Let CP^K denote the K-round negotiation protocol, which induces the mega-game $CP^K \circ G$. If any agent vetoes in the first round under CP^K, we proceed to the subgame $CP^{K-1} \circ G$. This immediate implies that the equilibrium utility for $CP^K \circ G$ is weakly increasing in K, because all agents can always deviate to veto in the first round and get the equilibrium utility of $CP^{K-1} \circ G$. — This is the easy part. The more difficult part is: when Pareto efficiency is not reached yet, the equilibrium utility for $CP^K \circ G$ is, in fact, *strictly* increasing in K, due to backward induction (and with or without a technical genericity condition).[5] Finally, since we consider bounded and finite games, these two parts immediately imply that CP^K always induces efficient outcomes for sufficiently large K, because equilibrium utility cannot increase unboundedly.

The remainder of the paper proceeds as follows: Section 2 defines the model; Section 3 presents the main results; Section 4 extends the results to more general setups; Section 5 discusses related literature and concludes.

2 A simple model of commitment design

Throughout the paper, we consider a finite set of agents, denoted by $\mathcal{I}$, and in particular, we fix an order of the agents, $1, ..., I$, i.e., $\mathcal{I} = \{1, ..., I\}$. Assume $|\mathcal{I}| = I \geq 2$. For notational ease, we focus on extensive games with complete and perfect information, and for simplicity, we just call them "games." The model and the results can be easily extended to imperfect-information games, which is described in Section 4.

Let $\mathcal{O}$ denote a finite set of social outcomes, and $q \notin \mathcal{O}$ denote a holocaust outcome.[6] Furthermore, each agent $i \in \mathcal{I}$ is endowed with a utility function $u_i : \mathcal{O} \cup \{q\} \longrightarrow \mathbb{R}$ such that

$$u_i(q) < u_i(o), \forall (o, i) \in \mathcal{O} \times \mathcal{I}.$$

The interpretation is that if agents do not follow their commitments, we punish them with q.—This is how we model commitment enforcement.

Following Arieli, Babichenko, and Tennenholtz (2017), we define *generic games* as

[5] For different setups, we prove serval lemmas to summarize this intuition. In some of them, we impose a gnericity condition in order to get strong results. In the others, we get rid of the genericity condition, which leads to weaker results.

[6] For example, q may be a sufficiently large monetary penalty imposed to all agents.

follows.

Definition 1 *The genericity condition holds if*

$$u_i(o) = u_i(o') \Longrightarrow \begin{bmatrix} u_j(o) = u_j(o') \\ \forall j \in \mathcal{I} \end{bmatrix}, \forall (o, o', i) \in \mathcal{O} \times \mathcal{O} \times \mathcal{I}.$$

When the genericity condition holds, games are called *generic games*. In a generic game, it is straightforward to see that the subgame perfect Nash equilibrium (SPNE) utility is unique, which makes our prediction on games (regarding equilibrium utility) sharp. Nevertheless, we study both generic games and non-generic games, and provide results that cover all of them.

2.1 A simple definition of games

Usually, extensive games are represented by game trees (see e.g., Mas-Colell, Whinston, and Green (1995)), which is notationally complicated. For notational ease, we propose a new definition of games, which maps the space of game trees to a much more well-shaped space. However, our new notion is quite different from the traditional game trees, and for heuristic purposes, we describe an intermediate notion in this subsection, which is equivalent to our new notion, and meanwhile closer to a game tree. We will further simplify the intermediate notion, and rigorously defined the new notion in Section 2.2.

For simplicity, we focus on finite-action-finite-round games, and throughout the paper, we fix a pair of sufficiently large positive integers, (M, N), and a finite set of actions, A, such that $|A| = M$. The space where a game sits is $A^{I \times N}$.

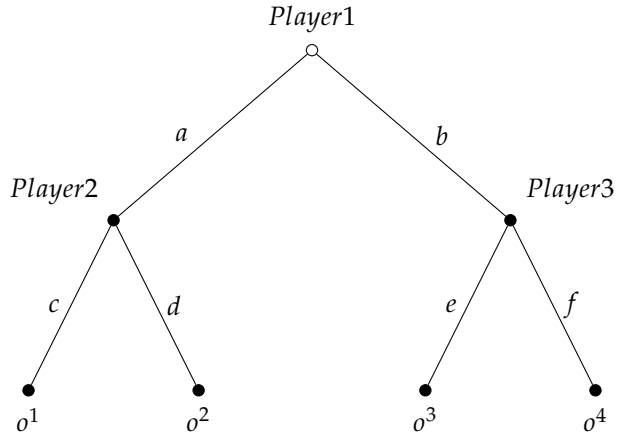
Let G denote a game, and it is fully described by a tuple,

$$G = \left\langle E \in 2^{A^{I \times N}} \setminus \{\varnothing\}, g : E \longrightarrow \mathcal{O} \right\rangle.$$

The interpretation is that agents move in N rounds, and in each round, agents follow the fixed order, $(1, ..., I)$, to move sequentially. For each move, an agent chooses an action in A. The set E denote the set of all possible paths in this game G, and g describes the outcome of the game, i.e., it maps each path in E to a social outcome in $\mathcal{O}$.

It is easy to see that each game $G = \left\langle E \in 2^{A^{I \times N}} \setminus \{\varnothing\}, g : E \longrightarrow \mathcal{O} \right\rangle$ can be translated to a standard game tree, and vice versa. To see the latter, consider the following

game tree.



Clearly, this game tree can be represented as follows.

$$\mathcal{I} = \{1,2,3\}, \mathcal{O} = \left\{o^1, o^2, o^3, o^4\right\},$$
$$M = 6, N = 2, A = \{a,b,c,d,e,f\},$$
$$G = \left\langle E \in 2^{A^{I\times N}} \diagdown \{\varnothing\}, g : E \longrightarrow \mathcal{O} \right\rangle,$$
$$E = \left\{ \begin{array}{l} e^1 \equiv (a,c,e,a,c,e), \\ e^2 \equiv (a,d,e,a,c,e), \\ e^3 \equiv (b,c,e,b,c,e), \\ e^4 \equiv (b,c,f,b,c,e). \end{array} \right\} \text{ and } \begin{array}{l} g\left(e^1\right) = o^1, \\ g\left(e^2\right) = o^2, \\ g\left(e^3\right) = o^3, \\ g\left(e^4\right) = o^4. \end{array}$$

For each game $G = \left\langle E \in 2^{A^{I\times N}} \diagdown \{\varnothing\}, g : E \longrightarrow \mathcal{O} \right\rangle$, define histories of G as follows.

$$H^{(G,0)} = \{\varnothing\}, \text{ and for } n \in \{1,...,I\times N - 1\},$$
$$H^{(G,n)} = \left\{ (x_1,...,x_n) \in A^n : \begin{array}{c} (x_1,...,x_n,x_{n+1},...,x_{I\times N}) \in E \\ \text{for some } (x_{n+1},...,x_{I\times N}) \in A^{I\times N-n} \end{array} \right\},$$
$$H^G = \bigcup_{n=0}^{I\times N-1} H^{(G,n)}.$$

That is, each $(x_1,...,x_n) \in H^{(G,n)}$ denote a history of length n, i.e., n moves have been made by the agents.

A strategy of agent i is function $\sigma_i : H^G \to A$. Clearly, such a strategy carries redundant information. For instance, the first move is taken by agent 1, and $\sigma_1(\varnothing)$ describes

agent 1's move. For any agent $j \neq 1$, $\sigma_j(\varnothing)$ does not play a role in this game. We adopt such a redundant definition of strategy due to notational ease, i.e., we do not need to define the individualized history sets.

Given a profile of strategies, $(\sigma_i)_{i\in\mathcal{I}}$, a unique path is determined. Specifically, agent 1 first moves to $\sigma_1(\varnothing)$; agent 2 moves second to $\sigma_2[\sigma_1(\varnothing)]$; agent 3 moves third to $\sigma_3[\ (\sigma_1(\varnothing), \sigma_2[\sigma_1(\varnothing)])\ ]$;.... Let ϕ denote this process, which maps strategy profiles to resulted paths.

Furthermore, for a valid strategy, agents must play legitimate options in game g. For instance, we should impose

$$\sigma_1(\varnothing) \in H^{(G,1)},$$

and for every positive integer n with $i = [n \text{ modular } I]$

$$\sigma_i(x) \in \left\{a \in A : (x,a) \in H^{(G,n+1)}\right\}, \forall x \in H^{(G,n)}.$$

Clearly, to keep track of all of these is very demanding, and because of this, we propose an equivalent and much simpler model in Section 2.2.

2.2 An even simpler definition of games

A game G is fully defined by a function $g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}$ such that $g^{-1}(\mathcal{O}) \neq \varnothing$, and it is straightforward to see that such a game is equivalent to the game $\left\langle E = g^{-1}(\mathcal{O}), g|_{g^{-1}(\mathcal{O})}\right\rangle$ defined in Section 2.1. Let $\mathcal{G}$ denote the space of all such games, i.e.,

$$\mathcal{G} \equiv \left\{\left\langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\right\rangle \ : \ g^{-1}(\mathcal{O}) \neq \varnothing\right\}.$$

Define histories of G as follows.

$$H^0 = \{\varnothing\},$$

$$H^n = A^n, \forall n \in \{1, ..., I\times N - 1\},$$

$$H = \bigcup_{n=0}^{I\times N-1} H^n.$$

A strategy of agent i is function $\sigma_i : H \to A$. Define $\Sigma \equiv A^H$. Let Σ_i denote the strategy space of agent i, i.e.,

$$\Sigma_i \equiv \Sigma \equiv A^H.$$

Recall that $\phi : \times_{i\in\mathcal{I}}\Sigma_i \longrightarrow A^{I\times N}$ describe the resulted paths for strategy profiles, i.e.,

$$\phi\left[(\sigma_i)_{i\in\mathcal{I}}\right] = \left[\sigma_1(\varnothing),\ \sigma_2\left[\sigma_1(\varnothing)\right],\ \sigma_3\left[\ (\sigma_1(\varnothing),\sigma_2\left[\sigma_1(\varnothing)\right])\ \right],...\right] \in A^{I\times N}.$$

Given a game $G = \langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\rangle$, each history $x = \left(x^1,...,x^n\right) \in H$ defines a subgame $G^x = \langle g^x : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\rangle$ as follows.

$$g^x\left[\left(a^1,...,a^{I\times N}\right)\right] = \begin{cases} g\left[\left(a^1,...,a^{I\times N}\right)\right], & \text{if } \left(a^1,...,a^n\right) = \left(x^1,...,x^n\right), \\ q, & \text{otherwise,} \end{cases}$$
$$\forall\left(a^1,...,a^{I\times N}\right) \in A^{I\times N}.$$

Definition 2 *A strategy profile $\sigma = (\sigma_i)_{i\in I} \in [\times_{i\in\mathcal{I}}\Sigma_i]$ is a Nash equilibrium (hereafter NE) in a game $G = \langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\rangle$ if*

$$u_i\left[g\left(\phi\left[\sigma_i,\sigma_{-i}\right]\right)\right] \geq u_i\left[g\left(\phi\left[\sigma_i',\sigma_{-i}\right]\right)\right],\ \forall\left(i,\sigma'\right) \in \mathcal{I}\times\left[\times_{j\in\mathcal{I}}\Sigma_j\right].$$

Furthermore, σ is a subgame perfect Nash equilibrium (hereafter SPNE) in G if σ is a NE in every subgame of G. Let $SPNE\left[G\right]$ denote the set of SPNEs in G.

2.3 Commitment protocols

We now define a commitment protocol.

Definition 3 *A commitment protocol, denoted by CP, is a tuple*

$$CP = \left\langle B \in 2^{\mathbb{N}}\diagdown\{\varnothing\},\ T\in\mathbb{N},\ l : B^{I\times T} \longrightarrow 2^{A^{I\times N}}\diagdown\{\varnothing\}\right\rangle.$$

Let $\mathcal{C}$ denote the set of all commitment protocols.

That is, a commitment protocol is almost a (finite-action-finite-move) game, and the only difference is that the induced outcome is not an element in $\mathcal{O}\cup\{q\}$. Rather, it is a non-empty subset of $A^{I\times N}$. The interpretation is that agents make commitment sequentially, which results in a path $x \in B^{I\times T}$ and a non-empty subset $l(x) \subset A^{I\times N}$, and agents commit to take strategies which would induce only paths in $l(x)$.

Given a commitment protocol $CP = \left\langle B,\ T, l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\} \right\rangle \in \mathcal{C}$ and a game $G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O} \cup \{q\} \right\rangle \in \mathcal{G}$, agents play a mega-game, denoted by $CP \circ G$, which is defined as follows.

$$CP \circ G = \left\langle g^{CP} : \left[B^{I\times T}\right] \times \left[A^{I\times N}\right] \longrightarrow \mathcal{O} \cup \{q\} \right\rangle \text{ such that}$$

$$g^{CP}[x,y] = \begin{cases} g(y), & \text{if } y \in l(x), \\ q, & \text{otherwise.} \end{cases}$$

That is, agents first follow CP to make commitments, which results in the commitment subset $l(x)$; agents then proceed to play G, and if they do not keep their commitment (i.e., $y \notin l(x)$), the penalty q is imposed.

The definitions of NE, SPNE and ϕ in Section 2.2 extend to mega-games $CP \circ G$.

2.4 Implementation

A goal of an implementation problem is described by a function $\mathcal{E} : \mathcal{G} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\}$. That is, for each game $G \in \mathcal{G}$, the set $\mathcal{E}(G)$ contains all of the paths that are deemed desirable.

We say that we achieve a goal $\mathcal{E}$, if for every game $G \in \mathcal{G}$, one or all of the equilibrium paths fall in the set $\mathcal{E}(G)$. Clearly, this is very demanding. For instance, consider the goal of Pareto efficiency, and it cannot be achieved (e.g., in the Prisoner's Dilemma game).—This leads to an implementation problem: can we find some $CP \in \mathcal{C}$ such that we always achieve efficiency in $CP \circ G$ for every game $G \in \mathcal{G}$? Based on this, we define two notions of implementation, i.e., full implementation and partial implementation.

To simplify the notation, define

$$\mathcal{E}(CP \circ G) \equiv B^{I\times T} \times \mathcal{E}(G),$$

i.e., $\mathcal{E}(CP \circ G)$ contains all the paths in the mega-game $CP \circ G$ which induce the desired paths (dictated by $\mathcal{E}$) in the original game G.

Definition 4 *A commitment protocol* $CP = \left\langle B,\ T, l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\} \right\rangle \in \mathcal{C}$ *fully implements a goal* $\mathcal{E}$ *if*

$$SPNE[CP \circ G] \subset \mathcal{E}(CP \circ G), \forall G \in \mathcal{G}. \tag{1}$$

Furthermore, a goal $\mathcal{E}$ *is fully implementable if it is fully implemented by some* $CP \in \mathcal{C}$.

Definition 5 *A commitment protocol* $CP = \left\langle B,\ T\ , l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\} \right\rangle \in \mathcal{C}$ *partially implements a goal* $\mathcal{E}$ *if*

$$SPNE\left[CP \circ G\right] \bigcap \mathcal{E}\left(CP \circ G\right) \neq \varnothing, \forall G \in \mathcal{G}. \tag{2}$$

Furthermore, a goal $\mathcal{E}$ *is partially implementable if it is partially implemented by some* $CP \in \mathcal{C}$.

3 Main Result

3.1 A summary

For $a = (a_1, ..., a_n), b = (b_1, ..., b_n) \in \mathbb{R}^n$, define

$$a \geq b \Longleftrightarrow a_h \geq b_h, \forall h \in \{1, ..., n\},$$
$$a > b \Longleftrightarrow a \geq b \text{ and } a \neq b,$$
$$a \gg b \Longleftrightarrow a_h > b_h, \forall h \in \{1, ..., n\}.$$

Throughout the paper, we focus on the social goal of Pareto efficiency defined as follows.[7]

$$\mathcal{E}^* : \mathcal{G} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\},$$
$$\mathcal{E}^*(G) = \left\{x \in A^{I\times N} : \nexists x' \in A^{I\times N} \text{ such that } \left(u_i\left[g\left(x'\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g\left(x\right)\right]\right)_{i\in\mathcal{I}}\right\},$$
$$\forall G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O} \cup \{q\} \right\rangle \in \mathcal{G}.$$

I.e., $\mathcal{E}^*(G)$ is the set of paths in G that induce Pareto efficient outcomes in $g\left(A^{I\times N}\right)$.

We provide two positive results, one for full implementation and the other for partial implementation.

Theorem 1 *$\mathcal{E}^*$ is fully implementable, if the genericity condition holds.*

Theorem 2 *$\mathcal{E}^*$ is partially implementable.*

[7] When we define Pareto efficiency, we adopt a weak notion of Pareto dominance. Alternatively, we may adopt a strong notion of Pareto dominance. However, our definition suffers no loss of generality, because we prove positive results.

Theorem 1 says that we can design a commitment protocol under which all SPNEs in *any generic game* always achieve Pareto efficiency, even if we do not know fine details of the game played by the agents. Relaxing the genericity condition, Theorem 2 says that we can design a commitment protocol under which some SPNE in *any game* always achieves Pareto efficiency.

3.2 The K-round negotiation protocol

We define a particular commitment protocol, called the K-round negotiation protocol, which will be used to prove both Theorem 1 and Theorem 2.

In the K-round negotiation protocol, at most K rounds of commitment proposing occur. Each round consists of two sub-stages: the proposing stage and the endorsement stage. Rigorously, it is described as follows.

The K-round negotiation protocol:

the proposing stage at the beginning of round $k \in \{1, 2, ..., K\}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce her commitment σ_i $(\in \Sigma_i \equiv \Sigma)$;[8]

the endorsement stage given the announced commitment profile $(\sigma_i)_{i \in I}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce whether she accepts or rejects $(\sigma_i)_{i \in \mathcal{I}}$:

- **if all players accept** $(\sigma_i)_{i \in \mathcal{I}}$**:** $(\sigma_i)_{i \in \mathcal{I}}$ becomes effective, i.e., each agent i commits to play σ_i in the true game $G \in \mathcal{G}$;
- **otherwise:** $(\sigma_i)_{i \in \mathcal{I}}$ is revoked, and they proceed to round $k + 1$, if $k < K$; and proceed to play the true game G without commitment if $k = K$.

[8] Here, a commitment is defined as "committing to a particular strategy." Alternatively, we may define commitment as "committing to a non-empty subset of strategies." All of our results remain true if the latter is adopted, and for notational ease, we adopt the former.

Let CP^K denote the K-round negotiation protocol. Following the definitions in Section 2.3, we define CP^K rigorously as follows.

$$CP^K = \left\langle B = \Sigma \cup \{yes, no\},\ T = 2 \times K \in \mathbb{N},\ l^{CP^K} : B^{I \times T} \longrightarrow 2^{A^{I \times N}} \diagdown \{\varnothing\} \right\rangle,$$
$$\text{with } \Sigma \equiv A^H \text{ and } \Sigma \cap \{yes, no\} = \varnothing.$$

We need the following notation, before we can define l^{CP^K}. For each $x \in B^{I \times 2 \times K}$, we write it as

$$x = \left(\left[\left(c_i^1\right)_{i \in \mathcal{I}},\ \left(d_i^1\right)_{i \in \mathcal{I}} \right], \left[\left(c_i^2\right)_{i \in \mathcal{I}},\ \left(d_i^2\right)_{i \in \mathcal{I}} \right], ..., \left[\left(c_i^K\right)_{i \in \mathcal{I}},\ \left(d_i^K\right)_{i \in \mathcal{I}} \right] \right) \in B^{I \times 2 \times K},$$
$$\text{with } \left(c_i^k, d_i^k\right) \in B \times B, \forall (i,k) \in \mathcal{I} \times \{1, 2, ..., K\},$$

and define

$$\Omega(x) = \left\{ k \in \{1, 2, ..., K\} : \begin{array}{c} \left(c_i^k\right)_{i \in \mathcal{I}} \in \times_{i \in \mathcal{I}} \Sigma_i \\ \text{and } d_i^k = yes, \forall i \in \mathcal{I} \end{array} \right\},$$
$$k^*(x) = \min \Omega(x) \text{ if } \Omega(x) \neq \varnothing.$$

I.e., $\Omega(x)$ is the set of rounds in which all agents commit to some particular strategies, and they all endorse the commitment, while $k^*(x)$ denotes the earliest round in which such an agreement is reached. Then, define

$$l^{CP^K}(x) = \begin{cases} A^{I \times N}, & \text{if } \Omega(x) = \varnothing, \\ \left\{ \phi \left[\left(c_i^{k^*(x)}\right)_{i \in \mathcal{I}} \right] \right\} \subset A^{I \times N}, & \text{otherwise.} \end{cases}$$

That is, if no agreement is reached in all of the K rounds (i.e., $\Omega(x) = \varnothing$), no restriction is imposed on the true game which will be played later (i.e., $l^{CP^K}(x) = A^{I \times N}$); otherwise, $k^*(x)$ is the earliest round in which the agents reach an agreement, and they commit to play the strategy profile $\left(c_i^{k^*(x)}\right)_{i \in \mathcal{I}} \in \times_{i \in \mathcal{I}} \Sigma_i$, and as a result, only the path $\phi \left[\left(c_i^{k^*(x)}\right)_{i \in \mathcal{I}} \right]$ is induced in G (i.e., $l^{CP^K}(x) = \left\{ \phi \left[\left(c_i^{k^*(x)}\right)_{i \in \mathcal{I}} \right] \right\}$).

For notational ease, define

$$CP^0 \circ G \equiv G, \forall G \in \mathcal{G},$$
$$\text{i.e., } CP^0 \circ G = \left\langle g^{CP^0} \equiv g \right\rangle \in \mathcal{G},$$
$$\forall G = \left\langle g : A^{I \times N} \longrightarrow \mathcal{O} \cup \{q\} \right\rangle \in \mathcal{G}.$$

That is, the game with 0-round negotiation is simply the original game.

3.3 Proof of Theorem 1

The following result is immediately implied by the genericity condition (Definition 1), and we omit the proof.

Lemma 1 *If the genericity condition holds, we have*

$$\begin{gathered}\forall G=\left\langle g: A^{I\times N}\longrightarrow \mathcal{O}\cup\{q\}\right\rangle\in\mathcal{G},\\ \forall\left(\sigma,\sigma'\right)\in SPNE\left(G\right)\times SPNE\left(G\right),\\ \left(u_i\left[g\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}}=\left(u_i\left[g\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}},\end{gathered}$$

and furthermore, we have

$$\begin{gathered}\forall G=\left\langle g: A^{I\times N}\longrightarrow \mathcal{O}\cup\{q\}\right\rangle\in\mathcal{G},\\ \forall CP\in\mathcal{C},\\ \forall\left(\sigma,\sigma'\right)\in SPNE\left[CP\circ G\right]\times SPNE\left[CP\circ G\right],\\ \left(u_i\left[g^{CP}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}}=\left(u_i\left[g^{CP}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.\end{gathered}$$

Lemma 1 says that SPNE utility is unique in any generic game. Throughout this subsection, we impose the genericity condition, and as a result, we have a unique prediction (regarding equilibrium utility) on each generic game.

To prove Theorem 1, we first draw a simple observation on the K-round negotiation protocol: for any true game G, the mega-game $CP^{K-1}\circ G$ is a subgame of $CP^{K}\circ G$. In the mega-game $CP^{K}\circ G$, if any player vote against the proposed commitment in the first round, we proceed to the second round, and have $K-1$ more rounds of negotiation, and as a result, the subgame starting at the second round is equivalent to $CP^{K-1}\circ G$. This simple observation immediately implies the following lemma.

Lemma 2 *Suppose that the genericity condition holds. For any positive integer K, we have*

$$\begin{gathered}\forall G=\left\langle g: A^{I\times N}\longrightarrow \mathcal{O}\cup\{q\}\right\rangle\in\mathcal{G},\\ \forall\left(\sigma,\sigma'\right)\in SPNE\left[CP^{K}\circ G\right]\times SPNE\left[CP^{K-1}\circ G\right],\\ \left(u_i\left[g^{CP^{K}}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}}\geq\left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.\end{gathered}$$

Lemma 2 says that equilibrium utility *weakly increases* as the number of rounds of negotiation increases. The proof is straightforward: in game $\left[CP^K \circ G\right]$, every player can always veto in the first round, and proceed to the second round to play the subgame $\left[CP^{K-1} \circ G\right]$. We omit the proof of Lemma 2.

In fact, equilibrium utility is *strictly increasing*, if Pareto efficiency is not reached yet, which is summarized by the following result.

Lemma 3 *Suppose that the genericity condition holds. For any positive integer K,*

$$\forall G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\right\rangle \in \mathcal{G},$$
$$\forall \left(\sigma,\sigma'\right) \in SPNE\left[CP^K \circ G\right] \times SPNE\left[CP^{K-1} \circ G\right],$$
$$\phi\left[\sigma'\right] \notin \mathcal{E}^*\left(CP^{K-1} \circ G\right) \Longrightarrow \left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.$$

For expositional ease, we relegate the proof of Lemma 3 to Section 3.5.

Proof of Theorem 1: Recall $|\mathcal{O}| < \infty$. Pick $K^* = |\mathcal{O}| + 1$. We show that the commitment protocol CP^{K^*} fully implements $\mathcal{E}^*$. We prove this by contradiction. Fix any $G \in \mathcal{G}$. For each $k \in \{0,1,2,...,K^*\}$, pick any $\sigma^k \in SPNE\left[CP^k \circ G\right]$. Suppose $\phi\left[\sigma^{K^*}\right] \notin \mathcal{E}^*\left(CP^{K^*} \circ G\right)$. Then, by Lemma 2, we have

$$\phi\left[\sigma^k\right] \notin \mathcal{E}^*\left(CP^k \circ G\right), \forall k \in \{0,1,2,...,K^*\}.$$

By lemma 3, we have

$$\left(u_i\left[g^{CP^{K^*}}\left(\phi\left[\sigma^{K^*}\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K^*-1}}\left(\phi\left[\sigma^{K^*-1}\right]\right)\right]\right)_{i\in\mathcal{I}} \gg ... \gg \left(u_i\left[g^{CP^0}\left(\phi\left[\sigma^0\right]\right)\right]\right)_{i\in\mathcal{I}},$$

which implies

$$\left\{\left[g^{CP^k}\left(\phi\left[\sigma^k\right]\right)\right] \in \mathcal{O}\cup\{q\} : k \in \{0,1,2,...,K^*\}\right\} \subset \mathcal{O}\cup\{q\},$$
$$\left|\left\{\left[g^{CP^k}\left(\phi\left[\sigma^k\right]\right)\right] \in \mathcal{O}\cup\{q\} : k \in \{0,1,2,...,K^*\}\right\}\right| = K^*+1 = |\mathcal{O}|+2,$$

and as a result,

$$\begin{aligned} |\mathcal{O}|+2 &= \left|\left\{\left[g^{CP^k}\left(\phi\left[\sigma\right]\right)\right] \in \mathcal{O}\cup\{q\} : k \in \{0,1,2,...,K^*\}\right\}\right| \\ &\leq |\mathcal{O}\cup\{q\}| \\ &= |\mathcal{O}|+1, \end{aligned}$$

i.e., we reach a contradiction.■

3.4 Proof of Theorem 2

To some extent, the proof of Theorem 2 is similar to that of Theorem 1. The following two lemmas are analogous to Lemmas 2 and 3. However, there are subtle differences in quantifiers due to non-uniqueness of equilibrium utility.

Lemma 4 *For any positive integer K, we have*

$$\begin{gathered}\forall G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\right\rangle \in \mathcal{G},\\ \forall \sigma' \in SPNE\left[CP^{K-1}\circ G\right],\\ \exists \sigma \in SPNE\left[CP^{K}\circ G\right],\\ \left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} \geq \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.\end{gathered}$$

Lemma 5 *For any positive integer K,*

$$\begin{gathered}\forall G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\right\rangle \in \mathcal{G},\\ \forall \sigma' \in SPNE\left[CP^{K-1}\circ G\right],\end{gathered}$$

$$\phi\left[\sigma'\right] \notin \mathcal{E}^*\left(CP^{K-1}\circ G\right) \Longrightarrow \left[\begin{array}{c}\exists \sigma \in SPNE\left[CP^{K}\circ G\right],\\ \left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}\end{array}\right].$$

Proof of Lemma 4: Fix any $G \in \mathcal{G}$ and any $\sigma' \in SPNE\left[CP^{K-1}\circ G\right]$. We solve the mega-game $CP^K \circ G$ by backward induction. Consider all subgames in which agents do not reach an agreement in the first round. Rigorously, such a subgame is represented by $\left[CP^K \circ G\right]^h$ with

$$h \in \widehat{H} = B^{I\times 2}\setminus\left\{\left[s,(t_i)_{i\in\mathcal{I}}\right] \in B^I\times B^I : \begin{array}{l} s \in \times_{i\in\mathcal{I}}\Sigma_i\\ \text{and } t_i = yes, \forall i\in\mathcal{I}\end{array}\right\},$$

$$\text{and } B = \Sigma\cup\{yes, no\}.$$

Each subgame $\left[CP^K \circ G\right]^h$ is equivalent to $CP^{K-1}\circ G$, i.e., $(K-1)$ rounds of negotiation are left.

Furthermore, let all agents play the SPNE $\sigma' \in SPNE\left[CP^{K-1} \circ G\right]$ in each subgame $\left[CP^K \circ G\right]^h$ with $h \in \widehat{H}$. Given this, we can solve a SPNE in $CP^K \circ G$ by backward induction, and denote it by $\sigma \in SPNE\left[CP^K \circ G\right]$. Clearly, we have

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} \geq \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}},$$

because every agent i can always veto in the first round of negotiation, and get utility $u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]$.■

For expositional ease, we relegate the proof of Lemma 5 to Section 3.6.

Proof of Theorem 2: Define $K^* = |\mathcal{O}| + 1$. We show that the commitment protocol CP^{K^*} partially implements $\mathcal{E}^*$.

Fix any $G \in \mathcal{G}$. First, we show

$$\begin{aligned} &\exists k \in \{0,1,2,...,K^*\}, \exists \sigma \in SPNE\left[CP^k \circ G\right], \\ &\text{such that } \phi\left[\sigma\right] \in \mathcal{E}^*\left(CP^k \circ G\right). \end{aligned} \tag{3}$$

We prove this by contradiction. Suppose (3) does not hold, i.e.,

$$\begin{aligned} &\forall k \in \{0,1,2,...,K^*\}, \forall \sigma \in SPNE\left[CP^k \circ G\right], \\ &\phi\left[\sigma\right] \notin \mathcal{E}^*\left(CP^k \circ G\right). \end{aligned} \tag{4}$$

Pick any $\sigma^0 \in SPNE\left[CP^0 \circ G\right]$, and $\phi\left[\sigma^0\right] \notin \mathcal{E}^*(G)$. Then, by Lemma 5, we can find $\sigma^1 \in SPNE\left[CP^1 \circ G\right]$ such that

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.$$

Applying (4) and Lemma 5 again, and inductively, we can find

$$\begin{aligned} \left(u_i\left[g^{CP^{K^*}}\left(\phi\left[\sigma^{K^*}\right]\right)\right]\right)_{i\in\mathcal{I}} &> \left(u_i\left[g^{CP^{K^*-1}}\left(\phi\left[\sigma^{K^*-1}\right]\right)\right]\right)_{i\in\mathcal{I}} > ... > \left(u_i\left[g^{CP^0}\left(\phi\left[\sigma^0\right]\right)\right]\right)_{i\in\mathcal{I}}, \\ \text{with } \sigma^k &\in SPNE\left[CP^k \circ G\right] \text{ for every } k \in \{0,1,2,...,K^*\}. \end{aligned}$$

which implies

$$\begin{aligned} |\mathcal{O}| + 2 &= \left|\left\{\left[g^{CP^k}\left(\phi\left[\sigma^k\right]\right)\right] \in \mathcal{O} \cup \{q\} : k \in \{0,1,2,...,K^*\}\right\}\right| \\ &\leq |\mathcal{O} \cup \{q\}| \\ &= |\mathcal{O}| + 1, \end{aligned}$$

i.e., we reach a contradiction. Therefore, (3) holds.

Second, given (3), we get the following by inductively applying Lemma 4.

$$\left(u_i\left[g^{CP^{K^*}}\left(\phi\left[\sigma^{K^*}\right]\right)\right]\right)_{i\in\mathcal{I}} \geq ... \geq \left(u_i\left[g^{CP^{k+1}}\left(\phi\left[\sigma^{k+1}\right]\right)\right]\right) \geq \left(u_i\left[g^{CP^{k}}\left(\phi\left[\sigma^{k}\right]\right)\right]\right)_{i\in\mathcal{I}},$$
$$\text{and } \sigma^h \in SPNE\left[CP^h \circ G\right] \text{ for every } h \in \{k, k+1, ..., K^*\},$$

which, together with $\phi[\sigma] \in \mathcal{E}^*\left(CP^k \circ G\right)$, implies $\sigma^{K^*} \in SPNE\left[CP^{K^*} \circ G\right] \bigcap \mathcal{E}^*\left(CP^{K^*} \circ G\right)$.■

3.5 Proof of Lemma 3

Suppose that the genericity condition holds, which immediately implies

$$\left(u_i[o]\right)_{i\in\mathcal{I}} > \left(u_i\left[o'\right]\right)_{i\in\mathcal{I}} \Longleftrightarrow \left(u_i[o]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[o'\right]\right)_{i\in\mathcal{I}}, \forall o, o' \in \mathcal{O}\cup\{q\}.$$

As a result, we have

$$\begin{aligned}\mathcal{E}^*(G) &\equiv \left\{x \in A^{I\times N} : \nexists x' \in A^{I\times N} \text{ such that } \left(u_i\left[g\left(x'\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g\left(x\right)\right]\right)_{i\in\mathcal{I}}\right\} \\ &= \left\{x \in A^{I\times N} : \nexists x' \in A^{I\times N} \text{ such that } \left(u_i\left[g\left(x'\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g\left(x\right)\right]\right)_{i\in\mathcal{I}}\right\}.\end{aligned}$$

Fix any positive integer K and any game $G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O}\cup\{q\}\right\rangle \in \mathcal{G}$. By Lemma 1, we have

$$\begin{aligned}\forall\left[(\sigma,\widehat{\sigma}),\ (\sigma',\widehat{\sigma}')\right] &\in \left[SPNE\left(CP^K \circ G\right)\right]^2 \times \left[SPNE\left(CP^{K-1} \circ G\right)\right]^2, \\ \left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} &= \left(u_i\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}\right]\right)\right]\right)_{i\in\mathcal{I}}, \\ \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}} &= \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\widehat{\sigma}'\right]\right)\right]\right)_{i\in\mathcal{I}},\end{aligned}$$

i.e., the prediction regarding equilibrium utility is unique. Given an inefficient SPNE σ' in $CP^{K-1} \circ G$, i.e., $\phi[\sigma'] \notin \mathcal{E}^*\left(CP^{K-1} \circ G\right)$, we aim to show

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}, \forall\sigma \in SPNE\left(CP^K \circ G\right). \quad (5)$$

Inefficiency of σ' implies existence of $(s_i)_{i\in\mathcal{I}} \in \times_{i\in\mathcal{I}}\Sigma_i$ such that

$$\left(u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}. \quad (6)$$

That is, $(s_i)_{i\in\mathcal{I}}$ is a strategy profile in G such that the induced outcome $g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)$ strictly dominates that of σ'. Consider a series of histories in the mega-game $CP^K \circ G$, which is listed as follows.

$$\begin{aligned}
h^0 &= \varnothing, \\
h^1 &= (s_1), \\
h^2 &= (s_1, s_2), \\
&\ldots \\
h^I &= (s_1, s_2, ..., s_I), \\
h^{I+1} &= (s_1, s_2, ..., s_I, yes), \\
&\ldots \\
h^{2\times I} &= (s_1, s_2, ..., s_I, yes, ..., yes).
\end{aligned}$$

I.e., in the first round of negotiation, all agents follow $(s_i)_{i\in\mathcal{I}}$ to make commitments and endorse the commitments sequentially, and $h^0, h^1, h^2, ..., h^{2\times I}$ record such histories.

For each history h^k, recall that $\left[CP^K \circ G\right]^{h^k}$ denotes the subgame starting from the history h^k. We now solve $CP^K \circ G$ by backward induction, which will prove (5). First, consider the subgame $\left[CP^K \circ G\right]^{h^{2\times I}}$, and clearly, all agents must honor their commitments, and as a result, the unique SPNE utility is $\left(u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}}$.

Second, consider the subgame $\left[CP^K \circ G\right]^{h^{2\times I-1}}$, i.e., agents have committed to $(s_i)_{i\in\mathcal{I}}$ in G, and agents $1, ..., I-1$ have agreed to this commitment profile. Then, agent I must choose between "yes" and "no" regarding $(s_i)_{i\in\mathcal{I}}$. If agent I chooses yes, she gets utility $u_I\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]$, while agent I gets utility $u_I\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]$ if she chooses "no."[9] By (6), agent I finds it strictly better to choose "yes," and hence, the unique SPNE utility in $\left[CP^K \circ G\right]^{h^{2\times I-1}}$ is $\left(u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}}$.

Third, by backward induction and applying the same argument as above to subgames $\left[CP^K \circ G\right]^{h^{2\times I-2}}$, $\left[CP^K \circ G\right]^{h^{2\times I-3}}$, ..., $\left[CP^K \circ G\right]^{h^I}$, it is easy to see that the unique SPNE utility in all of these subgames is $\left(u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}}$.

Fourth, consider the subgame $\left[CP^K \circ G\right]^{h^{I-1}}$, i.e., agents $1, ..., I-1$ have committed to $(s_1, s_2, ..., s_{I-1})$. Consider any SPNE $\widetilde{\sigma}$ in the subgame $\left[CP^K \circ G\right]^{h^{I-1}}$, and we aim to

[9] If agent I chooses "no," the agents proceed to the subgame $CP^{K-1} \circ G$, i.e., $(K-1)$ rounds of negotiation are left.

show

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}. \tag{7}$$

In the subgame $\left[CP^K \circ G\right]^{h^{I-1}}$, agent I is the first one to make a move immediately after the history h^{I-1}, and she may commit to any strategy in Σ_I. However, regardless of what she chooses, all agents can always veto the commitment and proceed to the next round of negotiation. As a result, we have

$$u_i\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] \geq u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right], \forall i \in \mathcal{I}. \tag{8}$$

Furthermore, agent I may also commit to s_I, and proceeds to the subgame $\left[CP^K \circ G\right]^{h^I}$. As showed above, the (unique) SPNE utility in $\left[CP^K \circ G\right]^{h^I}$ is $\left(u_i\left[g\left(\phi\left[\left(s_i\right)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}}$. Hence, we have

$$u_I\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] \geq u_I\left[g\left(\phi\left[\left(s_i\right)_{i\in\mathcal{I}}\right]\right)\right], \tag{9}$$

which, together with (6), further implies

$$u_I\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] > u_I\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]. \tag{10}$$

We now prove (7) by contradiction. Suppose it does not hold. Then, by (8), we have

$$\exists j \in \mathcal{I}, u_j\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] = u_j\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right],$$

which, together with the genericity condition, implies

$$u_i\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] = u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right], \forall i \in \mathcal{I},$$
$$\text{and in particular, } u_I\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right] = u_I\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right],$$

contradicting (10). Therefore, (7) holds.

Fifth, by backward induction and applying the same argument as above to subgames $\left[CP^K \circ G\right]^{h^{I-2}}$, $\left[CP^K \circ G\right]^{h^{I-3}}$, ..., $\left[CP^K \circ G\right]^{h^1}$ and $\left[CP^K \circ G\right]^{h^0}$, it is easy to see that

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}, \tag{11}$$
$$\forall \widetilde{\sigma} \in \bigcup_{k=0}^{I-2} SPNE\left(\left[CP^K \circ G\right]^{h^k}\right).$$

Finally, note that $h^0 = \varnothing$ and $\left[CP^K \circ G\right]^{h^0} = CP^K \circ G$, which, together with (11), implies

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} \gg \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}},$$
$$\forall \sigma \in SPNE\left(CP^K \circ G\right).$$

This completes the proof of Lemma 3.■

3.6 Proof of Lemma 5

Fix any $G \in \mathcal{G}$ and any inefficient $\sigma' \in SPNE\left[CP^{K-1} \circ G\right]$, i.e., $\phi\left[\sigma'\right] \notin \mathcal{E}^*\left(CP^{K-1} \circ G\right)$, we aim to show existence of a SPNE $\sigma \in SPNE\left[CP^K \circ G\right]$ such that

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.$$

We solve the mega-game $\left[CP^K \circ G\right]$ by backward induction. Consider all histories which describe fully what have occurred in the first round. In particular, the set of all such histories can be partition into two parts.

$$\begin{aligned} & \left\{\left[s,(t_i)_{i\in\mathcal{I}}\right] \in B^I \times B^I : x \in \times_{i\in\mathcal{I}}\Sigma \text{ and } t_i = yes, \forall i \in \mathcal{I}\right\} \text{ and} \\ \widehat{H} \quad = \quad & B^{I\times 2} \diagdown \left\{\left[s,(t_i)_{i\in\mathcal{I}}\right] \in B^I \times B^I : x \in \times_{i\in\mathcal{I}}\Sigma \text{ and } t_i = yes, \forall i \in \mathcal{I}\right\}. \end{aligned}$$

The first set (i.e., $B^{I\times 2} \diagdown \widehat{H}$) contains all of the histories in which an agreement is reached in the first round, and the second set (i.e., $\widehat{H}$) contains all of those in which an agreement is not reached.

In each subgame $\left[CP^K \circ G\right]^h$ with $h \in B^{I\times 2} \diagdown \widehat{H}$, all agents must follow their commitments.

Note that each subgame $\left[CP^K \circ G\right]^h$ with $h \in \widehat{H}$ is equivalent to $\left[CP^{K-1} \circ G\right]$. From now on, let all agents play the SPNE $\sigma' \in SPNE\left[CP^{K-1} \circ G\right]$ in each subgame $\left[CP^K \circ G\right]^h$ with $h \in \widehat{H}$.

Inefficiency of σ' implies existence of $(s_i)_{i\in\mathcal{I}} \in \times_{i\in\mathcal{I}}\Sigma_i$ such that

$$\left(u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}. \tag{12}$$

Consider a series of histories in the mega-game $CP^K \circ G$, which is listed as follows.

$$\begin{aligned} h^0 &= \varnothing, \\ h^1 &= (s_1), \\ h^2 &= (s_1, s_2), \\ &\ldots \\ h^I &= (s_1, s_2, ..., s_I), \\ h^{I+1} &= (s_1, s_2, ..., s_I, yes), \\ &\ldots \\ h^{2\times I} &= (s_1, s_2, ..., s_I, yes, ..., yes). \end{aligned}$$

I.e., in the first round of negotiation, all agents follow $(s_i)_{i\in\mathcal{I}}$ to make commitments and endorse the commitments sequentially, and $h^0, h^1, h^2, ..., h^{2\times I}$ record such histories.

First, we show existence of $\sigma^I \in SPNE\left(\left[CP^K \circ G\right]^{h^I}\right)$ such that

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^I\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}. \tag{13}$$

Let σ^I denote a strategy profile in which all agents vote "yes" immediately after $(s_1, s_2, ..., s_I)$ are proposed by the agents in the first round; then, all agent play $(s_1, s_2, ..., s_I)$ in the true game G. By backward induction, for each agent $i = I, I-1, ..., 1$, if she vetoes, she gets $u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]$, while she gets $u_i\left[g\left(\phi\left[(s_i)_{i\in\mathcal{I}}\right]\right)\right]$ if she vote yes. And by (12), σ^I is a SPNE and (13) holds.

From now on, let all agents play the SPNE $\sigma^I \in SPNE\left(\left[CP^K \circ G\right]^{h^I}\right)$ in the subgame $\left[CP^K \circ G\right]^{h^I}$.

Second, consider the subgame $\left[CP^K \circ G\right]^{h^{I-1}}$, and we show existence of $\sigma^{I-1} \in SPNE\left(\left[CP^K \circ G\right]^{h^{I-1}}\right)$ such that

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^{I-1}\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}. \tag{14}$$

Given the history h^{I-1}, agent I is the player to make the next move, and she may commit to any strategy $\widetilde{\sigma}_I \in \Sigma_I$. If she commits to s_I, we proceed to subgame $\left[CP^K \circ G\right]^{h^I}$, and agents play the SPNE σ^I described above.

If agent I commits to $\widetilde{\sigma}_I \in \Sigma_I \setminus \{s_I\}$, we proceed to subgame $\left[CP^K \circ G\right]^{\left(h^{I-1}, \widetilde{\sigma}_I\right)}$, and fix any SPNE $\widehat{\sigma}^{\left(h^{I-1}, \widetilde{\sigma}_I\right)} \in SPNE\left(\left[CP^K \circ G\right]^{\left(h^{I-1}, \widetilde{\sigma}_I\right)}\right)$ such that the agents play the SPNE σ' in each subgame $\left[CP^K \circ G\right]^h$ with $h \in \widehat{H}$. This immediately implies

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1}, \widetilde{\sigma}_i\right)}\right]\right)\right]\right)_{i\in\mathcal{I}} \geq \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}, \forall \widetilde{\sigma}_I \in \Sigma_I \setminus \{s_I\}. \tag{15}$$

because every agent can always veto and proceed to the subgame $\left[CP^K \circ G\right]^h$. Pick any $\widetilde{\sigma}_I^* \in \Sigma_I \setminus \{s_I\}$ such that

$$u_I\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1}, \widetilde{\sigma}_I^*\right)}\right]\right)\right] \geq u_I\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1}, \widetilde{\sigma}_I\right)}\right]\right)\right], \forall \widetilde{\sigma}_I \in \Sigma_I \setminus \{s_I\},$$

i.e., $\widetilde{\sigma}_I^*$ is a best option for agent I in the set $\Sigma_I \setminus \{s_I\}$, which induces the highest SPNE utility for agent I. We now consider two cases. In Case (1), we have

$$u_I\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1}, \widetilde{\sigma}_I^*\right)}\right]\right)\right] > u_I\left[g^{CP^K}\left(\phi\left[\sigma^I\right]\right)\right], \tag{16}$$

i.e., $\widetilde{\sigma}_i^*$ is actually a best move in Σ_I. Note that (13) and (16) imply

$$u_I\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1},\widetilde{\sigma}_I^*\right)}\right]\right)\right] > u_I\left[g^{CP^K}\left(\phi\left[\sigma'\right]\right)\right]. \tag{17}$$

Define σ^{I-1} in subgame $\left[CP^K \circ G\right]^{h^{I-1}}$ as follows.

$$\sigma^{I-1}: \left[\begin{array}{c} \text{agent } I \text{ first commits to } \widetilde{\sigma}_I^*; \\ \text{for the subgame } \left[CP^K \circ G\right]^{h^I}, \text{agents play } \sigma^I \in SPNE\left(\left[CP^K \circ G\right]^{h^I}\right); \\ \text{for subgame } \left[CP^K \circ G\right]^{\left(h^{I-1},\widetilde{\sigma}_i\right)} \text{ with } \widetilde{\sigma}_i \in \Sigma_i \diagdown \{s_I\}, \\ \text{agents play } \widehat{\sigma}^{\left(h^{I-1},\widetilde{\sigma}_i\right)} \in SPNE\left(\left[CP^K \circ G\right]^{\left(h^{I-1},\widetilde{\sigma}_i\right)}\right). \end{array}\right]$$

Clearly, σ^{I-1} is a SPNE in $\left[CP^K \circ G\right]^{h^{I-1}}$, and

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^{I-1}\right]\right)\right]\right)_{i\in\mathcal{I}} = \left(u_i\left[g^{CP^K}\left(\phi\left[\widehat{\sigma}^{\left(h^{I-1},\widetilde{\sigma}_i^*\right)}\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}, \tag{18}$$

where the strict inequality follows from (15) and (17).

In Case (2), we have

$$u_I\left[g^{CP^K}\left(\phi\left[\widetilde{\sigma}^{\left(h^{I-1},\widetilde{\sigma}_i^*\right)}\right]\right)\right] \leq u_I\left[g^{CP^K}\left(\phi\left[\sigma^I\right]\right)\right],$$

i.e., s_I is actually a best move in Σ_i. Define σ^{I-1} in subgame $\left[CP^K \circ G\right]^{h^{I-1}}$ as follows.

$$\sigma^{I-1}: \left[\begin{array}{c} \text{agent } I \text{ first commits to } s_I; \\ \text{for the subgame } \left[CP^K \circ G\right]^{h^I}, \text{agents play } \sigma^I \in SPNE\left(\left[CP^K \circ G\right]^{h^I}\right); \\ \text{for subgame } \left[CP^K \circ G\right]^{\left(h^{I-1},\widetilde{\sigma}_i\right)} \text{ with } \widetilde{\sigma}_i \in \Sigma_i \diagdown \{s_I\}, \\ \text{agents play } \widehat{\sigma}^{\left(h^{I-1},\widetilde{\sigma}_i\right)} \in SPNE\left(\left[CP^K \circ G\right]^{\left(h^{I-1},\widetilde{\sigma}_i\right)}\right). \end{array}\right]$$

Clearly, σ^{I-1} is a SPNE in $\left[CP^K \circ G\right]^{h^{I-1}}$, and

$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^{I-1}\right]\right)\right]\right)_{i\in\mathcal{I}} = \left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^I\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}, \tag{19}$$

where the strict inequality follows from (13). Therefore, (18) and (19) in the two cases imply (14).

Third, by backward induction and applying the same argument as above to subgames $\left[CP^K \circ G\right]^{h^{I-2}}$, $\left[CP^K \circ G\right]^{h^{I-3}}$, ..., $\left[CP^K \circ G\right]^{h^1}$ and $\left[CP^K \circ G\right]^{h^0}$, we can show

$$\forall k = I-2, I-3, ..., 1, 0, \tag{20}$$
$$\exists \sigma^k \in SPNE\left(\left[CP^K \circ G\right]^{h^k}\right)$$
$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma^k\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.$$

In particular, $CP^K \circ G = \left[CP^K \circ G\right]^{h^0}$, and (20) implies

$$\exists \sigma \in SPNE\left(CP^K \circ G\right)$$
$$\left(u_i\left[g^{CP^K}\left(\phi\left[\sigma\right]\right)\right]\right)_{i\in\mathcal{I}} > \left(u_i\left[g^{CP^{K-1}}\left(\phi\left[\sigma'\right]\right)\right]\right)_{i\in\mathcal{I}}.$$

This completes the proof of Lemma 5.■

4 Extension

Till now, we have focused our study on complete-information and perfect-information games. In fact, our results can be generalized beyond this class. In this section, we extend them to imperfect-information games. By Harsanyi (1967), this suffers no loss of generality, because an incomplete-information game can be represented by an imperfect-information game (with nature being an additional player), and for notational ease, we focus on the latter.[10]

4.1 A simple definition of imperfect-information games

First, we slightly modify the definition of games. A game is a tuple,

$$G = \left\langle g : A^{I\times N} \longrightarrow \mathcal{O} \cup \{q\},\ (f_i : H \longrightarrow \mathbb{R})_{i\in\mathcal{I}} \right\rangle,$$
$$\text{such that } g^{-1}(\mathcal{O}) \neq \varnothing.$$

[10]For incomplete-information games, all of our results can be extended, if we modify the notion of Pareto efficiency appropriately, i.e., efficiency should be defined as efficient outcomes at the time when commitments are made. For instance, if commitments are made at the ex-ante stage, we should adopt ex-ante Pareto efficiency, and if commitments are made at the interim stage, we should adopt interim Pareto efficiency.

That is, the only new ingredient is $f_i : H \longrightarrow \mathbb{R}$ for each $i \in \mathcal{I}$, which describes agent i's information structure. For instance, after agent 1 makes a first move $a \in A = H^1$. Agent 2's information regarding 1's move can be describes by the following partition of H^1.

$$\left\{ H^1 \cap f_2^{-1}\left(\{t\}\right) : t \in f_2\left(H^1\right) \right\}.$$

That is, for two histories, $x, x' \in H^1$ such that $f_2(x) = f_2(x')$, agent 2 cannot distinguish the two. Furthermore, the definition of strategy should be modified accordingly. A strategy of agent i in game $G = \left\langle g : A^{I \times N} \longrightarrow \mathcal{O} \cup \{q\}, \left(f_i : H \longrightarrow \mathbb{R}\right)_{i \in \mathcal{I}} \right\rangle$ is a function $\sigma_i : H \to A$ such that

$$f_i(x) = f_i\left(x'\right) \Longrightarrow \sigma_i(x) = \sigma_i\left(x'\right), \forall k \in \{1, ..., I \times N - 1\}, \forall x, x' \in H^k,$$

i..e., agents' strategies are measurable with respect to their information structure. Finally, the rest of the definitions remain the same.

Let $\mathcal{G}^*$ denote the space of all such games. It is worthy of noting:

$$\mathcal{G} \subset \mathcal{G}^*,$$

i.e., the games defined in Section 2 are specific cases, with an additional requirement that f_i is injective for every $i \in \mathcal{I}$.

To implement $\mathcal{E}^*$ in $\mathcal{G}^*$, we need to slightly modify the definition of commitment protocols. Specifically, we consider two alternative ways.

4.2 Commitment semi-protocols

We first define commitment semi-protocols.

Definition 6 *A commitment semi-protocol, denoted by CSP is a tuple*

$$CSP = \left\langle B \in 2^{\mathbb{N}} \diagdown \{\varnothing\}, T \in \mathbb{N}, l : B^{I \times T} \longrightarrow 2^{A^{I \times N}} \right\rangle.$$

Let $\mathcal{C}^{CSP}$ denote the set of all commitment semi-protocols.

The sole difference between CP (Definition 3) and CSP is $l(x) \in 2^{A^{I \times N}} \diagdown \{\varnothing\}$ for the former, and we may have $l(x) = \varnothing$ for the latter. That is, in a CP, we allow for *voluntary*

punishment only, i.e., we can punish players if and only if they break their commitment. However, in a CSP, we allow for involuntary punishment, i.e., given $l(x) = \varnothing$, we punish players regardless of whether they break their commitment.

Like above, the definitions of $CSP \circ G$ and $\mathcal{E}(CSP \circ G)$ are defined similarly, and we omit the details. We also modify the definitions of implementation accordingly.

Definition 7 *A commitment semi-protocol* $CSP = \left\langle B \in 2^{\mathbb{N}} \setminus \{\varnothing\}, T \in \mathbb{N}, l : B^{I \times T} \longrightarrow 2^{A^{I \times N}} \right\rangle \in \mathcal{C}^{CSP}$ *fully implements a goal* $\mathcal{E}$ *in* $\mathcal{G}^*$ *if*

$$SPNE[CSP \circ G] \subset \mathcal{E}(CSP \circ G), \forall G \in \mathcal{G}^*.$$

Furthermore, a goal $\mathcal{E}$ *is* $\left(\mathcal{C}^{CSP}, \mathcal{G}^*\right)$*-fully-implementable if it is fully implemented by some* $CSP \in \mathcal{C}^{CSP}$ *in* $\mathcal{G}^*$.

Definition 8 *A commitment semi-protocol* $CSP = \left\langle B \in 2^{\mathbb{N}} \setminus \{\varnothing\}, T \in \mathbb{N}, l : B^{I \times T} \longrightarrow 2^{A^{I \times N}} \right\rangle \in \mathcal{C}^{CSP}$ *partially implements a goal* $\mathcal{E}$ *in* $\mathcal{G}^*$ *if*

$$SPNE[CSP \circ G] \bigcap \mathcal{E}(CSP \circ G) \neq \varnothing, \forall G \in \mathcal{G}^*.$$

Furthermore, a goal $\mathcal{E}$ *is* $\left(\mathcal{C}^{CSP}, \mathcal{G}^*\right)$*-partially-implementable if it is partially implemented by some* $CSP \in \mathcal{C}^{CSP}$ *in* $\mathcal{G}^*$.

The following theorems generalize Theorems 1 and 2.

Theorem 3 $\mathcal{E}^*$ *is* $\left(\mathcal{C}^{CSP}, \mathcal{G}^*\right)$*-fully-implementable, if the genericity condition holds.*

Theorem 4 $\mathcal{E}^*$ *is* $\left(\mathcal{C}^{CSP}, \mathcal{G}^*\right)$*-partially-implementable.*

To prove these theorems, we modify the K-round negotiation protocol (i.e., CP^K) slightly to a CSP, and call it *"the K-round negotiation-with-penalty protocol,"* denoted by CSP^K. We define CSP^K rigorously as follows.

The K-round negotiation-with-penalty protocol:

the proposing stage at the beginning of round $k \in \{1, 2, ..., K\}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce her commitment, σ_i $(\in \Sigma_i \equiv \Sigma)$;

the endorsement stage given the announced commitment profile $(\sigma_i)_{i \in I}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce whether she accepts or rejects $(\sigma_i)_{i \in \mathcal{I}}$:

if all players accept $(\sigma_i)_{i \in \mathcal{I}}$: $(\sigma_i)_{i \in \mathcal{I}}$ becomes effective, i.e., each agent i commits to play σ_i in the true game $G \in \mathcal{G}$;

otherwise: $(\sigma_i)_{i \in \mathcal{I}}$ is revoked. Furthermore, they proceed to round $k+1$, if $k < K$, and **if $k = K$, the punishment q is always implemented in the true game G.**

The sole difference between CP^K and CSP^K is that, when an agreement is not reached in all of the K rounds, the agents proceed to play the true game $G \in \mathcal{G}$ without commitment in CP^K, while the punishment q is always implemented in CSP^K.

Consider the mega-game $CSP^1 \circ G$ (i.e., $K = 1$). Because of the penalty for no agreement, it is always a best reply for every agent to endorse any commitments proposed. Then, by backward induction, we effectively transform an imperfect-information game into a perfect-information game.[11] As a result, in every SPNE in $CSP^1 \circ G$, agents agree on a commitment profile. The rest of the proofs of Theorems 3 and 4 are the same as those of Theorems 1 and 2.

4.3 Commitment quasi-protocols

Clearly, involuntary punishment plays a critical role in Theorems 3 and 4. If involuntary punishment is not allowed, we still can generalize Theorem 2 by using a commitment quasi-protocol defined below, which forbids involuntary punishment.

[11] Consider any history of the first round of negotiation, denoted by h. First, if an agreement is not reached under h, the only SPNE outcome in subgame G^h is q, and we can replace G^h by q in the backward induction. Second, if a commitment $\sigma = (\sigma_i)_{i \in \mathcal{I}}$ is agreed under h, the only SPNE outcome in subgame G^h is $g[\phi(\sigma)]$, and we can replace G^h by $g[\phi(\sigma)]$ in the backward induction. Given these, the reduced game resulted from backward induction is a perfect-information game.

Definition 9 *A commitment quasi-protocol, denoted by CQP, is a tuple*

$$CQP = \left\langle B \in 2^{\mathbb{N}} \diagdown \{\varnothing\}, T \in \mathbb{N}, l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\}, \left(f_i : \bigcup_{n=0}^{I\times T} B^n \longrightarrow \mathbb{R}\right)_{i\in\mathcal{I}} \right\rangle,$$
$$\text{with } B^0 \equiv \{\varnothing\}.$$

Let $\mathcal{C}^{CQP}$ denote the set of all commitment quasi-protocols.

In CQP, we allow for voluntary punishment only, i.e., $l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\}$, and the sole difference between CP (Definition 3) and CQP is that the phase of K-round negotiation is modeled as a perfect-information game for the former, and as an imperfect-information game for the latter via $\left(f_i : \bigcup_{n=0}^{I\times T} B^n \longrightarrow \mathbb{R}\right)_{i\in\mathcal{I}}$.

Like above, the definitions of $CQP \circ G$ and $\mathcal{E}(CQP \circ G)$ are defined similarly, and we omit the details. We also modify the definition of partial implementation accordingly.

Definition 10 *A commitment quasi-protocol*

$$CQP = \left\langle B \in 2^{\mathbb{N}} \diagdown \{\varnothing\}, T \in \mathbb{N}, l : B^{I\times T} \longrightarrow 2^{A^{I\times N}} \diagdown \{\varnothing\}, \left(f_i : \bigcup_{n=0}^{I\times T} B^n \longrightarrow \mathbb{R}\right)_{i\in\mathcal{I}} \right\rangle \in \mathcal{C}^{CQP}$$

partially implements a goal $\mathcal{E}$ in $\mathcal{G}^*$ if

$$SPNE[CQP \circ G] \bigcap \mathcal{E}(CQP \circ G) \neq \varnothing, \forall G \in \mathcal{G}^*.$$

Furthermore, a goal $\mathcal{E}$ is $(\mathcal{C}^{CQP}, \mathcal{G}^*)$-partially-implementable if it is partially implemented by some $CQP \in \mathcal{C}^{CQP}$ in $\mathcal{G}^*$.

The following theorem generalize Theorem 2.

Theorem 5 *$\mathcal{E}^*$ is $(\mathcal{C}^{CQP}, \mathcal{G}^*)$-partially-implementable.*

To prove this theorems, we modify the K-round negotiation protocol (i.e., CP^K) slightly to a CQP, and call it "*the K-round negotiation-with-simultaneous-voting protocol*," denoted by CQP^K. We define CQP^K rigorously as follows.

The K-round negotiation-with-simultaneous-voting protocol:

For Round $k \in \{1, 2, ..., K-1\}$:

the proposing stage at the beginning of round $k \in \{1, 2, ..., K-1\}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce her commitment, $\sigma_i \, (\in \Sigma_i \equiv \Sigma)$;

the endorsement stage given the announced commitment profile $(\sigma_i)_{i \in I}$, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce whether she accepts or rejects $(\sigma_i)_{i \in \mathcal{I}}$:

if all players accept $(\sigma_i)_{i \in \mathcal{I}}$: $(\sigma_i)_{i \in \mathcal{I}}$ becomes effective, i.e., each agent i commits to play σ_i in the true game $G \in \mathcal{G}$;

otherwise: $(\sigma_i)_{i \in \mathcal{I}}$ is revoked, and they proceed to round $k+1$,.

and furthermore, for Round $k = K$:

the proposing stage at the beginning of round K, each player follows the fixed order, $(1, 2, ..., I)$, to sequentially and publicly announce her commitment, $\sigma_i \, (\in \Sigma_i \equiv \Sigma)$;

the endorsement stage given the announced commitment profile $(\sigma_i)_{i \in I}$, all player player simultaneously announce whether she accepts or rejects $(\sigma_i)_{i \in \mathcal{I}}$:

if all agents reject $(\sigma_i)_{i \in \mathcal{I}}$: $(\sigma_i)_{i \in \mathcal{I}}$ is revoked and they proceed to play the true game $G \in \mathcal{G}$;

otherwise: $(\sigma_i)_{i \in \mathcal{I}}$ becomes effective, i.e., each agent i commits to play σ_i in the true game $G \in \mathcal{G}$.

That is, the sole difference between CP^K and CQP^K is the "endorsement stage" at the *last round* of negotiation. Specifically, there are two subtle differences in the last round: (1) a sequential voting is adopted in the former, and a simultaneous voting for the latter; (2) in order for a commitment profile become effective, all agents must vote yes for the former, and one agent voting yes suffices for the latter.

Consider the mega-game $CQP^1 \circ G$ (i.e., $K = 1$). Due to the simultaneous voting (and the unanimity rule for "rejection"), there is a SPNE such that all agents vote yes for any commitment profile proposed and it becomes effective. Thus, by backward induction, we effectively transform an imperfect-information game into a perfect-information game. As a result, there is a SPNE in $CSP^1 \circ G$ such that the agents agree on a commitment profile. I.e., we achieve the same for both CSP^1 and CQP^1, though we achieve it by different tools: involuntary punishment for CSP^1 and simultaneous voting (and partial implementation) for CQP^1. The rest of the proof of Theorem 5 is the same as those of Theorems 2 and 4.

5 Conclusion

We take an implementation approach on commitment, and provide protocols that always induce Pareto efficiency outcomes.

References

ARIELI, I., Y. BABICHENKO, AND M. TENNENHOLTZ (2017): "Sequential commitment games," *Games and Economic Behavior*, 105, 297–315.

HARSANYI, J. C. (1967): "Games with Incomplete Information Played by Bayesian Players, I-III Part I. The Basic Model," *Management Science*, 14, 159–182.

KALAI, A. T., E. KALAI, E. LEHRER, AND D. SAMET (2010): "A commitment folk theorem," *Games and Economic Behavior*, 69, 127–137.

MAS-COLELL, A., M. D. WHINSTON, AND J. R. GREEN (1995): *Microeconomic Theory*. Oxford University Press.