Negotiated Binding Agreements[*]

Malachy James Gavan[†]
University of Liverpool

March, 2025
Current Version: [Here]

Abstract

I study binding agreements over play in a game and the possibility of inefficient agreements when mutual confirmation is required. I propose a negotiation protocol where, in each round, agents propose actions from the underlying game that they could agree to, signalling the agreement they would like. The protocol terminates when proposals are mutually confirmed. The model is solved using Subgame Perfect Equilibrium. Since agents only propose actions they could agree to, the agreement outcomes exhibit a self-generating structure. A full characterisation is provided for two-player games, relying on appropriate individual punishments. These individual punishments are used for sufficiency in n-player games and a necessary iterative rationality constraint is introduced. I extend the solution concept to allow cooperative agreements within the negotiation game where generalisations of the main results hold.

Keywords: Agreements, Negotiation, Cooperation
JEL Codes: C70, C71, C72

1 Introduction

Binding agreements over actions in strategic environments are central to economic life. From international trade deals to labour contracts, negotiated agreements shape outcomes across a wide range of contexts. Economic theory, following the logic of the Coase Theorem, predicts that when mutual gains exist and there are no frictions, fully rational agents should bargain until efficiency is achieved.[1] Even when agreement amongst all agents is not achievable,

[*] I thank my PhD supervisor Antonio Penta for his ongoing support. I have also had innumerable useful discussions with Larbi Alaoui, Alexander Frug and Pia Ennuschat, for which I am greatly indebted. I am also particularly grateful to Marco Mariotti and Olivier Bochet. I also thank (in alphabetical order) Nemanja Antic, Josefina Cenzon, Francesco Cerigioni, Vincent Crawford, Faruk Gul, Gianmarco Leon, Gilat Levy, Raquel Lorenzo, Zoel Martín Vilató, Rosemarie Nagel, Maria Ptashkina, Debraj Ray, Danila Smirnov and participants in a number of seminars and conferences. All faults are my own.

[†] *email:* malachy.gavan@liverpool.ac.uk.

[1] Although the statement is usually in reference to environments with transfers the idea is more broadly applied. Nonetheless, in this paper I will remain agnostic on whether transfers are within the underlying game or not, but in either case the main results will apply and inefficiencies will be possible.

existing models often imply that the most efficient feasible outcome will be reached among the agreeing parties (e.g., Ray and Vohra (1997); Ellingsen and Paltseva (2016)).

Yet in practice, inefficient agreements are widespread. Trade agreements often maintain positive tariffs despite mutual gains from liberalization. Climate treaties fall short of optimal emissions targets, and labour negotiations frequently result in rigid and suboptimal working conditions. Concretely, post-Brexit negotiations between the UK and EU yielded a trade deal that imposed costly frictions, where it was clear from the outset both sides would have preferred to avoid. These outcomes raise a central question: what drives inefficient agreements even when agents are fully rational and information is complete? Additionally, what kind of inefficiency can arise?

This paper presents a model of negotiation over play in a game, where inefficiencies emerge endogenously from the structure of agreement. Agents make binding proposals about which action profile to implement in the underlying game. Agreements are formed when proposals coincide and are mutually confirmed, that is, both parties need to sign off on the proposal. The negotiation is dynamic and potentially unbounded in time, but agents are not impatient. Importantly, all proposals must be agreement actions – those that could in principle be agreed to – which leads to a self-generation structure in equilibrium. Nonetheless, the need for mutual sign off, which is prevalent in economic agreements, can lead to inefficiency. Simply put, a proposal that is inefficient may arise due to the mutual belief that the other player(s) would not agree to a more efficient proposal. This provides no incentive a) to put a more efficient proposal on the table and b) even confirm a more efficient agreement, as lack of confirmation from the other party means it would not be implemented regardless. However, there is the natural question of what kind of agreements can persist in this environment.

Two main results to answer this question follow. First, in two-player games, I fully characterize the set of negotiated agreement outcomes. Each outcome must provide players with payoffs above a player-specific punishment threshold – the worst credible agreement each player can enforce. Simply, the player specific punishment must provide a) the punished player with a best response payoff in the underlying game and b) the punishing player a payoff higher than the one pinned down by their punishment. These thresholds are constructed from the underlying game and can be seen as similar to self-enforcing punishments as in repeated games (Fudenberg and Maskin, 1986; Abreu et al., 1994) and commitment folk theorems of contract theory (Tennenholtz, 2004; Kalai et al., 2010; Peters and Szentes, 2012), but are constrained to be agreement outcomes themselves. The result bears resemblance to folk theorems but is generally more restrictive, as the need to only propose what could actually be agreed to prevents threats that are *contemporaneously* not credible.[2] In comparison to predictions of folk theorems, negotiated binding agreement outcomes must satisfy some lower bound of efficiency, dictated by these player-specific punishment bounds.

Second, in n-player games, the two-player characterization provides sufficient conditions for agreement. To derive necessary conditions, I introduce the concept of "iterated elimination of individually irrational actions" in the underlying game. An action is eliminated if, even under the most optimistic beliefs about others' actions, it yields a lower payoff than some guaranteed response. This process identifies actions that cannot credibly be proposed or agreed upon. Any equilibrium agreement must survive this iterative process. In many games, the necessary and sufficient conditions lead to a tight characterisation on outcomes.

To illustrate these results, I analyse two benchmark games. First, a Cournot duopoly with asymmetric costs shows how bargaining power emerges endogenously and shapes the set of sustainable agreement outcomes. Unlike the full folk theorem predictions, the set of agreement outcomes may exclude highly inefficient profiles and favor the lower-cost firm. Second, a first-price auction with heterogeneous bidders demonstrates how minimal payoff guarantees constrain the allocation: bidders with low valuations cannot be awarded the good with probability one under any equilibrium agreement.

Finally, I extend the framework to allow for coalitional deviations. That is, agents may make binding agreements not only over final actions but over the negotiation strategies themselves. I define a generalized equilibrium concept, the $\mathcal{C}$-Subgame Perfect Equilibrium, where no coalition in a specified collection can jointly deviate profitably. This leads to a refinement of outcomes consistent with the β-core (Aumann, 1961), subject to strategic constraints. While allowing coalitional deviations can sometimes restore efficiency, it may also result in non-existence, underscoring the fragility of efficient agreement when negotiating.

This approach complements existing work on bargaining, repeated games, and contract theory by offering a tractable model of agreement formation that allows for full strategic reasoning without relying on a mediator to ensure cooperative outcomes. It identifies an important source of inefficiency that arise solely from the requirement of mutual agreement in a dynamic setting, even under complete information. In doing so, it sheds light on the structure of real-world agreements and their frequent departures from efficiency and provides a model of what could occur under these settings.

[2] The contracting literature has not imposed such a restriction and the flow of payoffs to reward punishments today with prizes tomorrow ensures enforcement for infinitely repeated games.

2 Model

Let the underlying game being negotiated over be $G = \langle N, (u_i, A_i)_{i \in N} \rangle$ where $N = \{1, 2, 3, ..., n\}$ is a finite set of players, A_i is a set of actions for each player with typical element $a_i \in A_i$. $A = \times_{i \in N} A_i$ is the set of action profiles with typical element $a \in A$. u_i is utility function such that $u_i : A \to \mathbb{R}$ and u_i is bounded for all $i \in N$. Let $A_{-i} = \times_{j \neq i} A_j$.

I now define the *negotiation game* over G. There will be potentially infinitely many periods to reach an agreement and the process will take the following form. In each period, agents make a proposal of their own agreement action, a_i, they will take within the underlying game G. Agents then observe the proposal made by all others. After doing so, they may simultaneously decide whether to "confirm" their choice by proposing the same action again, or alternatively propose a new action. If all agents confirm the proposal, an agreement is made, and that action profile is implemented in a binding way. If not, they continue to the next round and the same process occurs until confirmation is made by all agents, leading to an agreement. As the proposals made are required to be agreement actions a self-generation argument is used. If there are infinitely many periods without agreement, I refer to this as *perpetual disagreement.*[3]

Formally, let the set of partial histories consist of all $h = (a^1, a^2, ..., a^k)$ such that $a^t \neq a^{t-1}$ for any $t \leq k$ where $a^t = (a_i^t)_{i \in N}$ denotes the profile of proposals made in period t. I will denote the set of all partial histories by H. Proposals are assumed to be made simultaneously within a period, and therefore no history is such that only some agents have made proposals.[4]

A history is terminal if, either:

1. the same action profile is proposed in consecutive periods, and no earlier occurrence of consecutive repetition is present. That is, $z = (a^1, ..., a^{k-1}, a^k)$ is terminal if $a^k = a^{k-1}$ and $a^m \neq a^{m-1}$ for all $m < k$. Let the set of such histories be denoted by Z' and refer to such histories as with *agreement.*

2. or there is an infinite sequence of proposed action profiles where the same action profile is never proposed consecutively. Let the set of such histories be denoted by Z''. I will refer to these as histories with *perpetual disagreement.*

Let the set of all terminal histories be given by $Z = Z' \cup Z''$.

Let $U_i : Z \to \mathbb{R}$ denote the payoff for player $i \in N$ of the negotiation game.

[3] Formally, this game is similar to that used in the farsighted stable set for games, which is discussed at length in the literature review in section 6.

[4] A previous working paper version considered the extension of non-simultaneous proposals where the main insights are upheld.

Whenever there is an agreement, it is assumed that the payoff is that of the agreed-upon action profile. Formally, whenever $z = (a^1, ..., a^k) \in Z'$, that is a history that ends in agreement, let $U_i(z) = u_i(a^k)$ for all $i \in N$.

Whenever there is perpetual disagreement, the payoff is defined to be between the lim inf and lim sup of the utility in the underlying game of the proposals made.[5] Formally, whenever $z = (a^1, a^2, ..., a^k, ...) \in Z''$, that is a terminal history with perpetual disagreement, I assume that $U_i(z) \in [\liminf_{t\to\infty} u_i(a^t), \limsup_{t\to\infty} u_i(a^t)]$. This assumption, which is primarily used to prevent discounting and have payoffs pinned down by agreement outcomes, will be discussed more at length shortly.

At each round of the negotiation game, before agreements have been made, agents consider all previous proposals, both of themselves and others, and decide on a new proposal to make. With this, strategies map each partial history to a new proposal of what they will play in an underlying game. Formally, at each partial history, $h \in H$ the strategy of $i \in N$ dictates the proposal i would make in the next round: $s_i : H \to A_i$. Let S_i be the space of all such mappings. Let $s : H \to A$ be the joint strategy, such that $s(h) = (s_i(h))_{i \in N}$.

For a partial history $h \in H$ and a joint strategy s let $(s|h)$ denote the continuation history of h given by s. That is, $(s|h) = z \in Z$ such that $z = (h, a'^{,1}, a'^{,2}, ..., a'^{,k}, ...)$ where $a'^{,1} = s(h)$, $a'^{,2} = s((h, a'^{,1}))$, $a'^{,k} = s((h, a'^{,1}, a'^{,2}, ..., a'^{,k-1}))$. With some abuse of notation, let $U_i(s|h) = U_i(z')$ when $z' \in Z'$. Additionally, let $U_i(s|h) = U_i(z'')$, where $(s|h) = (h, z'') \in Z''$, that is, only take the continuation of the history h for perpetual disagreement. When $z = (a^1, a^2, ..., a^k) \in Z'$, i.e. an agreement is made, let $a(z) = a^k$ and $a_i(z) = a_i^k$.

The structure of the negotiation game has some similarities to the structure of repeated games, due to the structure of the partial histories and payoff of perpetual disagreement. There are a few important differences. Firstly, repeated games only have one type of terminal history, where the underlying game has been repeated the specified number of times, be that some finite number or infinitely. This negotiation game allows for two distinct types of terminal histories, those with agreement (finite) and those without (infinite). Secondly, repeated games use flow payoffs, receiving a payoff in each period of play to guide strategic behaviour. This negotiation game only allows for payoffs to be realised upon termination. Identical disparities between negotiation games and repeated games are common in the literature (see Kalai 1981; Bhaskar 1989; Kimya 2020; Nishihara 2022, etc.).

[5] A previous working paper version of the paper considered alternative specifications, such as an exogenous outside option, where the main insight of inefficiency is upheld.

2.1. Discussion of Perpetual Disagreement Payoffs

Recall that the perpetual disagreement payoffs are such that

$$U_i(z) \in [\liminf_{t\to\infty} u_i(a^t), \limsup_{t\to\infty} u_i(a^t)]$$

for $z \in Z''$. This restriction is consistent with a standard probabilistic termination model, where the proposal today is implemented with probability $(1-\delta)$ for each period, while the process continues with probability δ, taking δ to 1. Therefore, this can also be interpreted as a limiting version of the condition used within Kimya (2020), where there is a probability that the negotiation will end at the currently proposed actions.[6] This is formalised by the following lemma, and the proof is provided in the appendix.

Lemma 1. *For* $z = (a^1, a^2, ..., a^t, ...) \in Z''$

$$\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(a^t) \in \left[\liminf_{k\to\infty} u_i(a^k), \limsup_{k\to\infty} u_i(a^k)\right]$$

By taking the view that the payoff of perpetual disagreement can take on *any* value from this set it weakens the reliance on the specific method of confirmation for agreement. So long as this confirmation is simultaneously made by all agents, the results would remain the same. To see this, notice that *any* payoff in the underlying game that is proposed countably infinitely many times can be used for the payoff of perpetual disagreement. Equally, if more than one profile of proposals is made a countably infinite number of times, one can easily be ignored. With this, it is possible to use a proposal to specifically avoid agreement, without it being used within the payoff of perpetual disagreement. Therefore, a proposal could be used to avoid a consecutive repetition without impacting payoffs. With this, one may consider a single action of the underlying game being used as an "object" button, while confirmation of the previous choice is seen as an "accept" button, and unanimity of acceptance is needed for agreement. Given lemma 1, one interpretation of the payoff of perpetual disagreement is that there is an ϵ probability of each player mistakenly pressing accept with ϵ taken to 0. This would be independent of the payoff that would be received if no agreement is made, conditional on such "trembles".

This specification may also embed the approach of infinitely repeated games with no discounting: i.e. using the limit of means criteria when well defined (Rubinstein, 1994; Aumann and Shapley, 1994) where joint commitment to continue to play an agreement action profile is modelled.

[6] Similar notions also exist in the context of Rubinstein (1982) bargaining, where Busch and Wen (1995) take a game to be played in each rejection phase, which is implemented with probability $1-\delta$ and continuation occurs to a new proposal happens with probability δ, allowing for an endogenous outside option.

2.2. Solution Concept

As the negotiation game is used in order to reach an agreement, in each period each agent will *signal* the agreement action they would like to take.[7] As this is the case, we must define the set of agreement actions for each player. Let $A_i^*(s)$ be the set of agreement outcomes for player i under the strategy profile s. With this, in equilibrium $s_i^*(h) \in A_i^*(s^*)$ for all histories, $h \in H$. This will be referred to as *respecting signalling of agreement* for s^*. This prevents threatening to use unagreeable outcomes in the event of deviations.[8] The idea here is that if an agent were to propose an action it must be because they see some possibility of agreeing to it, and therefore do not propose actions they would never agree to. This ensures that payoffs are defined only with respect to agreement outcomes rather than any possible outcome. This negotiation protocol defines a dynamic game with complete information therefore Subgame Perfect Equilibrium (SPE) is well defined and appropriate. This does not restrict deviations to respect $s_i'(h) \in A_i^*(s^*)$, and therefore agents are effectively permitted to *change* the agreement actions within the negotiation. The following defines a Negotiated Binding Agreement formally:

Definition. *a^* is a Negotiated Binding Agreement if there exits an equilibrium s^* of the negotiation game such that:*

1. *$s^*(\emptyset) = a^*$.*

2. *s^* is a Subgame Perfect Equilibrium.*

3. *s^* respects signalling of agreement, i.e. for all $i \in N, h \in H$, $s_i^*(h) \in A_i^*(s^*)$ where*

$$A_i^*(s^*) = \{a_i \in A_i | a_i = a_i(s^*|h) \text{ for some } h \in H \text{ such that } (s^*|h) \in Z'\}$$

is the set of all possible agreement outcomes for player $i \in N$.

I will occasionally refer to s^* satisfying these conditions as an *equilibrium.*

A few important points are worth noting. Firstly, as the respect for agreement signalling is defined with respect to the possible agreement outcomes induced by the strategy profile itself, this definition requires an internal consistency notion to apply. Therefore, $A^*(s^*)$

[7] A previous working paper version considered when proposals of action *profiles* are instead considered where the key takeaways are upheld.

[8] It is clear that without this restriction, and when taking the payoff to be the lim inf for perpetual disagreement, a folk theorem applies. This is as perpetual disagreement can always make up part of equilibrium and players can alternate between the lowest individually rational outcomes for each player, ensuring anything is possible. However, this relies on a very particular specification of the perpetual disagreement payoff, and would not be robust outside of that, while the results provided in the remainder of the paper will be.

is self-generating. Further, by this definition, the set of agreement outcomes is always non-empty.

One important feature is the definition does not imply an immediate agreement must be made and there may be mis-coordination in the signalling in early periods. However, as there is no discounting, there is no cost associated with this. Note that any agreement outcome could be a Negotiated Binding Agreement as one can simply "shift" the strategies starting from any history h to start from the initial history. Further, if any two strategies lead to distinct agreement outcomes, one could perform the same reasoning to realise the union of the set of agreement outcomes would also be consistent with agreement outcomes for some other SPE. With this, studying the set of agreement outcomes $A^*(s^*)$ and the set of possible Negotiated Binding Agreements is one and the same.

3 Negotiated Binding Agreement Action Profiles for Two-Player Games

In this section, I provide a full characterisation of the set of all possible Negotiated Binding Agreements for two-player games where the underlying action space is compact and the utility function is continuous. As outlined in the introduction, the logic of the characterisation is as follows. Each player will be willing to agree to an outcome if it is better than the worst possible agreement from said players perspective. Given this, there is a "punishment" agreement for each player which gives the worst possible agreement payoff. Call them $\underline{a}^1$ and $\underline{a}^2$ respectively. By definition, $u_i(\underline{a}^i) \leq u_i(a)$ for any agreement outcome a, including $\underline{a}^{-i}$. Further, it must be that player i is willing to agree to their worst agreement. Ensuring that there is no unilateral deviation to this action profile in the underlying game will make such punishment profile agreeable. Therefore, the Negotiated Binding Agreements can be completely characterised with easy-to-check conditions using purely information from the underlying game. To further demonstrate the logic of this result, I now turn to a Cournot Duopoly with Linear Demand and Heterogeneous costs. I will also discuss the distinction between Negotiated Binding Agreement outcomes, player specific punishment (Fudenberg and Maskin, 1986; Abreu et al., 1994) and commitment folk theorems (Peters and Szentes, 2012).

3.1. Leading Example and Preview of Results for two-player games

Consider a simple Cournot Duopoly model as the underlying game, G. Let $q_1, q_2 \in [0, b] = A_i$ be the quantities produced and the inverse demand be given by $p(q_1, q_2) = \max\{b - q_1 - q_2, 0\}$. Let firms have potentially heterogeneous costs, c_1 and c_2. Without loss of generality let $c_1 \geq c_2 \geq 0$. Assume that firm 1 is a viable competitor: $\frac{b+c_2}{2} \geq c_1$. Profits are given by

$\pi_i(q_1, q_2) = q_i(p(q_1, q_2) - c_i)$.

Notice that the best responses in the underlying game are given by:

$$q_i^*(q_{-i}) = \begin{cases} 0 & \text{if } q_{-i} \geq b - c_i \\ \frac{b - c_i - q_{-i}}{2} & \text{if } q_{-i} < b - c_i \end{cases}$$

The Nash equilibrium of this underlying game is given by quantities of $\left(\frac{b+c_2-2c_1}{3}, \frac{b+c_1-2c_2}{3}\right)$, leading to payoffs of $\left(\left(\frac{b+c_2-2c_1}{3}\right)^2, \left(\frac{b+c_1-2c_2}{3}\right)^2\right)$.

Consider a profile of (q_1^*, q_2^*) such that $\pi_1(q_1^*, q_2^*) \geq 0$ and $\pi_2(q_1^*, q_2^*) \geq (c_1 - c_2)^2$. It will be shown that the profile (q_1^*, q_2^*) is a Negotiated Binding Agreement action profile. Note given the assumption that $\frac{b+c_2}{2} \geq c_1$ it follows that $(\frac{b-c_2}{2})^2 \geq (c_1 - c_2)^2$ and therefore such a profile exists.

Consider the following strategies. Take $\underline{q}_2^1 = b - c_1$ and $\underline{q}^2 = (b - 2c_1 + c_2, c_1 - c_2)$.

1. [Firm 1's punishment for deviating] Let $s^*(h') = (0, \underline{q}_2^1)$ whenever $h' = (q^1, q^2, ..., (q_1', q_2^*))$, $q_1' \neq q_1^*$, $h' = (q^1, q^2, ..., (q_1'', \underline{q}_2^1))$, or $(q^1, q^2, ..., (q_1', \underline{q}_2^2))$, $q_1' \neq \underline{q}_1^2$.
2. [Firm 2's punishment for deviating] Let $s^*(h'') = \underline{q}^2$ whenever $h'' = (q_1, q_2, ..., (q_1^*, q_2'))$, $q_2' \neq q_2^*$, $h'' = (q_1, q_2, ..., (\underline{q}_1^2, q_2''))$, or $h'' = (q^1, q^2, ..., (0, q_2''))$, $q_2'' \neq \underline{q}_2^1$.
3. [No / multilateral deviations] Otherwise, $s^*(\emptyset) = s^*(h) = (q_1^*, q_2^*)$ for all other h.

The intuition of this strategy is as follows. In case a firm does not act as expected the other proposes to *partially* flood the market. The partial element comes from the fact that they can only choose a quantity that allows them to maintain positive profits, where they understand the other firm will propose their best response in the underlying game. Notice that if $c_1 > c_2$ firm 1 cannot entirely flood firm 2 out of the market while maintaining positive profits. Note that this equilibrium can also be constructed as Markov perfect and only depends on the proposal made in the previous period.

Now it will be shown that this strategy is an SPE satisfying consistency of agreement signalling, that leads to the Negotiated Binding Agreement action profile (q_1^*, q_2^*).

To see this is satisfies consistency of agreement signalling, notice that signalling of agreement applies as all three rules are absorbing, therefore the proposal itself must be part of an agreement. Therefore all that is left is to check that s^* is a Subgame Perfect Equilibrium of the negotiation game.

Let us consider firm 1 and take case 1 (where firm 1 faces punishment for deviation). In this case, regardless of what they propose, firm 2 will continue to propose $b - c_1$ for all periods. Given this, firm 1 cannot profitably produce a positive quantity in any period, and therefore any deviation leads to a payoff of at most 0 profits. This is not profitable as their current strategy would provide them with a profit of 0 under the continuation. Now let us consider case 2, the punishment of firm 2. Under the current strategy and rule, no deviation will lead to an agreement of $\underline{q}^2$. Firm 1 receives a profit of 0 by construction. A deviation from the prescribed strategies for firm 1 can only lead to firm 2 proposing $b - c_1$ in every period. With this, firm 1's payoff would be pinned down by $\pi_1(q_1', b - c_1) \leq 0$. With this, it cannot be profitable to deviate. Finally, by the same logic, deviating from case 3, which yields a weakly positive profit, cannot be profitable.

Now instead consider firm 2. Firstly, consider case 1 (where firm 1 is facing their punishment). If firm 2 does not deviate, this will lead to a profit of $\pi_2(0, \underline{q}_2^1) = (b - c_1)(c_1 - c_2)$. However, a deviation will lead to firm 1 proposing $\underline{q}_1^2$ in all subsequent periods. With this, a deviation will lead to a payoff at most the static best response to $\underline{q}_1^2$, $\pi_2(\underline{q}^2) = (c_1 - c_2)^2$. However, this can not be profitable due to the viable competitor assumption. In a similar vein as firm 1, it cannot be that it is profitable for 2 to deviate from their punishment due to statically best responding to their own punishment. They can also not profitably deviate from case 3 as this would lead to an agreement for q^*. No deviation would lead to a payoff of $\pi_2(q^*) \geq (c_1 - c_2)^2$, while a deviation would lead to a payoff of $(c_1 - c_2)^2$.

Now we will study why the Negotiated Binding Agreement must be necessarily better than that of the outlined punishment $\underline{q}^i$. Notice that due to both firms being restricted to only making proposals that they can agree to at any history $s^{*,\prime}(h) \in Q^*$, where Q^* is some set of agreement outcome quantities. Now notice that for any $s^{*,\prime}$ it is not possible that some firm i receives a lower payoff than the one prescribed by their minimal best response payoff in Q^*, as they could elect to best respond statically in each period.[9] As it is the case that a) the $\underline{q}_{-i}^i \in Q_{-i}^*$ which gives the minimal best response payoff is agreeable, and b) the payoff received from it must be higher than the minimal best response, we conclude that such an outcome of $\underline{q}^i$ such that $\underline{q}_i^i \in \arg\max_{q_i \in Q_i} \pi_i(q_i, \underline{q}_{-i}^i)$ must be an agreement outcome. This $\underline{q}^i$ therefore pins down the lowest agreement payoff for i. Notice that if a profile is included in Q^*, then any profile that provides both players with a higher payoff must also be included in Q^*, as the same punishment could be used to incentivise the more efficient agreement as the less efficient. in this case, as $\pi_1(\underline{q}^1) = 0$, it is clear that this will prescribe the lowest best response payoff. Further, as any agreement outcome must guarantee firm 1 a payoff of at least 0, even with firm 2's quantity taken into account, it must be that in any agreement outcome $q^{*,\prime} \in Q^*$ $\pi_1(q^{*,\prime}) \geq 0$. With this, taking into account firm 2's best response firm 1 can produce at most $b - 2c_1 + c_2$. This leads to a minimum payoff for firm 2 of $(c_1 - c_2)^2$. Showing that above strategy fully characterises the q^*s that are Negotiated

[9] I leave the argument that Q_{-i}^* is compact for the formal proof of theorem 1.

Binding Agreements action profiles.

Note that this analysis provides us with some natural comparative statics and comparison to player specific punishments of Fudenberg and Maskin (1986); Abreu et al. (1994) and the commitment folk theorems of Peters and Szentes (2012). If $c_1 = c_2$, then both players may agree to an outcome that provides any payoff above their individually rational payoff of 0. In this case, the resulting set of agreement payoffs is identical to that of the commitment folk theorem and the payoff space of individual punishments. However, if it is not the case, and $c_1 > c_2$, then the agreement outcome of the commitment folk theorems and the payoff space of individual punishments remains unchanged. However, under Negotiated Binding Agreements, the feasible set of outcomes is restricted to account for the additional bargaining power of firm 2, given that firm 1 will not propose or accept outcomes that are clearly detrimental to its own interests. In the most extreme case, when firm 1 is no longer a viable competitor, $c_1 = \frac{b+c_2}{2}$, we conclude that firm 2's profit is $\pi_2(\underline{q}^2) = \left(\frac{b-c_2}{2}\right)^2$, their monopoly profit.[10] In this sense, the bargaining power of each firm is dictated by the difference in marginal costs, i.e. the parameters of the *underlying game* they play, rather than the bargaining protocol. Nonetheless, inefficient agreements can arise.

3.2. Results

The logic of the previous example is formalised in general by the following theorem.

Theorem 1 (Full Characterisation for Two-Player Games)**.** *For any game G such that $N = \{1, 2\}$, A_i is a compact subset of a metric space for $i = 1, 2$ and u_i is continuous, then a^* is a Negotiated Binding Agreement action profile if and only if $\exists\{\underline{a}^1, \underline{a}^2\} \subseteq A$ such that:*

1. $\underline{a}^i_i \in \arg\max_{a_i \in A_i} u_i(a_i, \underline{a}^i_{-i})$.

2. $u_i(\underline{a}^j) \geq u_i(\underline{a}^i)$ *for all* $i \neq j$.

3. $u_i(a^*) \geq u_i(\underline{a}^i)$.

It is worth noting that any pure Nash equilibrium of the game G is supported by a Negotiated Binding Agreement. Further, any action profile that Pareto dominates a pure Nash equilibrium in the underlying game can be sustained by this reasoning. Indeed, a natural question is why the so called Coase Theorem does not apply - that in this environment bargaining may not lead to an efficient outcome.[11] To understand this, let us consider

[10] Note that if $c_1 > \frac{b+c_2}{2}$ then the only outcome that can be supported by a Negotiated Binding Agreement is $q_1^* = 0$, $q_2^* = \frac{b-c_2}{2}$, while under commitment folk theorems and individual punishments all individually rational payoffs would still be supported.

[11] Note that there are not explicitly transfers, so in this respect the environment does not meet the requirements set out by Coase. Nonetheless, even if the underlying game did have transfers the statement may fail in this environment which is discussed within the next subsection.

the Nash equilibrium of the game as a Negotiated Binding Agreement action profile. One may assume that if a player signals an action that can lead to a more efficient outcome the other would follow suit, inferring this change as an attempt to reach efficiency, to the benefit of both players. If this would be the case, an inefficient Nash equilibrium outcome would not be consistent with an SPE of the negotiation game. However, if both agents had correct beliefs that neither agent would change their proposal away from the Nash outcome then this would not be profitable as confirmation of the more efficient outcome would not occur–effectively, agents are free to infer *nothing* from the signals of others. This is purely driven by the need for multilateral confirmation and the strategic nature of negotiation.

Indeed, many negotiation and bargaining protocols do not lead to efficiency for various reasons including the one stated above. The 2-player bargaining model of Rubinstein (1979) does necessarily lead to efficiency when the cost of time is constant, rather than hyperbolic as there may be equilibria with costly delay. In the hyperbolic discounting case, when the outside option of Rubinstein's model is taken to be endogenous, a la Busch and Wen (1995), a folk theorem is obtained, although for a different reason from this paper, as the endogenous outside option there need not be (and in fact is not permitted to be) consistent with agreement. When there are more than two-players, additional restrictions on the equilibrium notion are needed to regain efficiency (Chatterjee et al., 1993) due to potential strategic co-ordination and multiple agents needing to confirm, where in their model inefficiency is driven by delay as opposed to inefficient agreements per se. The work of Harstad (2022) shows that a pledge-and-review bargaining game for contributions to a public good may also lead to inefficient outcomes due to the mutual confirmation needed. However this is shown for only the class of public goods games and considers a slightly different bargaining protocol. Additionally, inefficiencies are common in the contracting literature, for instance in contractable contracts (Tennenholtz, 2004; Kalai et al., 2010; Peters and Szentes, 2012) and strategic contract settings (Jackson and Wilkie, 2005; Yamada, 2003; Ellingsen and Paltseva, 2016).

3.3. Discussion of Transferable Utility Games

One important class of games are those with transferable utility. That is, agents' may receive utility from actions with an underlying game, while they may also make (potentially contingent) transfers between them. A natural question in light of the Coase Theorem is whether the inclusion of transfers can overcome the inefficiencies in the setting of mutual confirmation. It turns out this is not the case.

To study this more concretely, we can represent a game that allows for transferable utility (with two players) in the following way: $G = \langle \{1,2\}, (v_i)_{i\in\{1,2\}}, (A_i \times \mathbb{R}_+^{|A_1\times A_2|})_{i\in\{1,2\}}\rangle$ where

$u_i(a, t_i, t_j) = v_i(a) + t_j(a) - t_i(a)$. That is, agents' receive their utility from actions, v_i and (contingent) utility from net transfers $t_j(a) - t_i(a)$. Here, transfers are non-negative - i.e. player i can decide how much utility to transfer to player $j \neq i$, but cannot decide how much they receive. In this transferable utility game it is possible to find best response payoff that are represented by $t_i(a) = 0$ for all $a \in A$ and $a_i^* \in \text{argmax}_{a_i} v_i(a_i, a_j)$, i.e when no transfers occur and player i best responds based on v_i. In fact, as transfers are non-negative, these will provide the lowest best response payoffs.

Therefore the characterisation and associated payoffs the lower bound of utility remains in respect to the non-transferable component, and therefore the inclusion of transfers makes no substantial difference to the possible outcomes. For example, in the Cournot Duopoly example, the ability to transfer utility would only convexify the space, but not change the message of inefficiencies nor the characterisation. The same is true for the results regarding n-player games for identical reasons.

4 Negotiated Binding Agreement Action Profiles for n-Player Games

In this section, I will explore the necessary and sufficient conditions for n-player games. First, I show that the idea of the characterisation for two-player games is still sufficient for n-player games. Therefore the key takeaway of potentially inefficiencies and their "flavour" persists. However, it is no longer necessary. This is as the signalling of agreement condition does not impose strong coordination on the action *profile* proposed by players. Therefore there may be *profile* proposed by other players, while consistent with the signalling of agreement, itself may not be mutually agreeable. However, when strong conditions on coordination of agreement outcome *profiles*, i.e. not only does a player have to propose an action they would agree to, but the profile of actions proposed by any set of agents is such that they would jointly agree, then individual punishment condition returns to being necessary. Outside of imposing this condition, I show that an iterative individual rationality constraint on the underlying game, which I call *iterated elimination of individually irrational actions*, is necessary for any Negotiated Binding Agreement or agreement outcome.

4.1. Sufficiency

The logic of the sufficient condition for agreement outcomes of two-player games can be generalised to n-player games. Specifically, an outcome can be a Negotiated Binding Agreement action profile if each player has a punishment profile that they best respond to in the underlying game, they prefer to punish than being punished, and the candidate outcome is preferred to their punishment.

Theorem 2. *Take any underlying game, G, such that $\exists\{a^*, \underline{a}^1, ..., \underline{a}^n\} \subseteq A$ such that:*

1. $\underline{a}_i^i \in \arg\max_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
2. $u_i(a^*) \geq u_i(\underline{a}^i)$
3. $u_i(\underline{a}^j) \geq u_i(\underline{a}^i)$ *for all* $i, j \in N$

Then a^* is a Negotiated Binding Agreement action profile.

The proof follows identical logic to that of the sufficiency of 1 and therefore is omitted.

4.1.1. An Equilibrium Refinement where Outcomes are Fully Characterised

Further justification for the general sufficient conditions can be found. For a refinement of the solution concept, where the focus is upon SPE that end in immediate agreement following from each history, the sufficient conditions for agreement outcomes are also necessary. This *No Delay* condition applies for all possible histories, and therefore applies both on and off the path. I refer to this solution as *No Delay SPE* and is similar to the no delay equilibrium proposed by Chatterjee et al. (1993), who impose this requirement to regain efficiency in their setting. Therefore, for the class of No Delay SPE, I fully characterise the set of outcomes that can be supported. This refines the solution concept to ensure that all proposals of $n-1$ players are mutually agreeable (for some action of the remaining player), and therefore strengthens the requirement of signalling of agreement.

Definition 1 (No Delay Subgame Perfect Equilibrium)**.** *s^* is a No Delay Subgame Perfect Equilibrium supporting $a^* = a(s^*|\emptyset)$ if:*

1. *s^* is a Subgame Perfect Equilibrium of the negotiation game.*
2. *No Delay: For all partial histories $h \in H$, $s^*(h) = s^*(h, s^*(h)) = a^*(s^*|h)$.*

Proposition 1. *For any underlying game G such that A_i is a compact subset of a metric space and u_i is continuous for all $i \in N$, a^* is supported by a No Delay Subgame Perfect Equilibrium, s^*, if and only if $\exists\{\underline{a}^1, ..., \underline{a}^n\} \subseteq A$ such that:*

1. $\underline{a}_i^i \in \arg\max_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
2. $u_i(a^*) \geq u_i(\underline{a}^i)$
3. $u_i(\underline{a}^j) \geq u_i(\underline{a}^i)$ *for all* $i, j \in N$

The proof is identical to that of 1 and therefore is omitted.

Finally, note that within the literature on agreements it is common to use the notion of Perfect Equilibrium of Selten (1988), for instance in Kalai (1981) and Bhaskar (1989). Notice that this does not have a significant change in the results, and to ensure the sufficient conditions for agreement outcomes remain true for this refinement, as well as the signalling of agreement and agreement for all histories condition, the only check is to ensure that the action $\underline{a}_i^i$ is not weakly dominated in the underlying game G.

Before moving to the necessary conditions for the Negotiated Binding Agreement action profiles, I turn to the following example to preview the logic.

4.2. Preview of Necessary Conditions n-player games

Let the underlying game G be a 3 player single unit First Price Auction with heterogeneous valuations. Specifically, there are three bidders, $N = \{1, 2, 3\}$. Each bidder has a value for the good, v_i. It is assumed that $v_1 = 6$, while $v_2 = 5$ and $v_3 = 2$. Each bidder may bid an integer from 0 to 7, $b_i \in \{0, 1, .., 7\}$.[12] The highest bidders wins the good with uniform probability and pay their bid. Bidders who do not win the good receive a utility of 0. Therefore utility is given by their probability of winning, multiplied by their surplus value. Formally,

$$u_i(b) = \begin{cases} \frac{v_i - b_i}{|\arg\min_{j\in\{1,2,3\}} b_j|} & \text{if } i \in \arg\min_{j\in\{1,2,3\}} b_j \\ 0 & \text{if } i \notin \arg\min_{j\in\{1,2,3\}} b_j \end{cases}$$

Firstly, can it be that any bidder agrees to the maximal bid, $b_i = 7$, in a Negotiated Binding Agreement action profile? If this were the case, bidder i would receive a strictly negative utility, as they would certainly win the auction with positive probability and at a price above their valuation. However, they could avoid such an outcome by deciding to propose their own valuation in every round of negotiation, $s_i(h) = v_i$ for all $h \in H$. If they did so, regardless of whether the negotiation game ended in agreement or perpetual disagreement, they would receive a payoff of 0. More concretely, bidding $b_i = 7$ is *individually irrational* in the underlying game, which will be formalised in the next section, as they can guarantee themselves a higher payoff. With this, it cannot be that agreeing to bid $b_i = 7$ be part of a Negotiated Binding Agreement, as such a strategy cannot be a Subgame Perfect Equilibrium of the negotiation game. Given this, as signalling of agreement actions holds, this action cannot be proposed by any agent.

Now consider whether it is the case that bidders 2 or 3 could agree to bid 6. By the

[12] The maximal bid being 7 is not important for the analysis, we only need to ensure payoffs are bounded by including some maximum bid.

previous argument, we conclude that agreeing to bid 6 will result in winning the good with positive probability, as we know no bidder will ever bid 7. With this, as the valuations of bidders 2 and 3 are below 6, it must be they receive a strictly negative payoff from such an agreement. However, we can again consider a deviation of these firms in the negotiation game to always propose their valuation, ensuring a payoff of 0. More concretely, bidding 6 is *individually irrational* for bidders 2 and 3 in the underlying game, again formalised in the next section, as they can guarantee themselves a higher payoff, given that 7 cannot be bid, and therefore cannot be agreed to. With this, we conclude that such an agreement cannot be part of a Negotiated Binding Agreement. Again, this implies that a bid of 6 by players 2 and 3 cannot be proposed by the signalling of agreement action.

We can continue this induction, concluding that bidder 1 would also never bid 6 once bidders 2 and 3 will not. Bidder 3 would never bid 5, due to this bid being *iteratively individually irrational* in the underlying game.

By the same argument as ruling out such bids, we conclude that any Negotiated Binding Agreement action profile must provide bidders 2 and 3 with a payoff of at least 0. Notice in any Negotiated Binding Agreement it must be that bidder 1 receives a payoff of at least 1/2. In the worst possible stream of proposals for bidder 1 is that bidders 2 and 3 bid their highest possible bid in every round of the negotiation game, 5 and 4 respectively. Given this, bidder 1 can simply respond by bidding 5 in every round of the negotiation game, guaranteeing a payoff of 1/2.

Given the sufficient conditions provided lead to the same conclusion, this completely characterises the set of Negotiated Binding Agreement outcomes in this context. However, the logic of the necessary and sufficient conditions are distinct. The sufficient conditions were based on *player-specific punishments*, while the necessary conditions are based on iteratively ruling out outcomes that could not be agreed to.

With this, I move on to provide general necessary conditions, which this example has already pointed to.

4.3. Necessary Conditions

Now I will show that any action proposed at some history must survive a procedure of *iterated deletion of individually irrational actions* in the underlying game, which to my knowledge is a novel definition. This procedure works inductively as follows. If an individual's action, regardless of the action profile of other agents, always provides a payoff that is not individually rational, in the sense of inf-sup utility, then it is individually irrational. In the iterated elimination we can therefore remove said actions from consideration. Now, upon deleting such actions, we proceed inductively. If an individual's action, regardless of the action profile of other agents chosen *within* the set that has survived iterated deletion of individually irrational actions, always provides a payoff that is not individually rational (on the remaining actions), then it does not survive iterated deletion of individually irrational actions. The formal definition of individual irrational actions and the iterated deletion notion associated are given below.

Definition 2 (Individually Irrational actions given $C_{-i} \subseteq A_{-i}$)**.** *For a game G, $a_i \in A_i$ is individually irrational given $C_{-i} \subseteq A_{-i}$ if:*

$$\inf_{a'_{-i} \in C_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > \sup_{a_{-i} \in C_{-i}} u_i(a_i, a_{-i})$$

Denote the set of actions that are individually irrational given C_{-i} by $D_i(C_{-i})$.

This notion is similar to the notion of absolute dominance by Salcedo (2017), simultaneously developed in Halpern and Pass (2018), who instead compare the best case of one action and the worst case of another, whereas I compare based on the best case of an action compared to the inf-sup.[13] Therefore the set that survives elimination of individually irrational actions is smaller. If in a normal form game there is a single action that is not absolutely dominated given A_{-i}, then this action is an obviously dominant strategy as defined by Li (2017). Therefore if a single action is not individually irrational it is also obviously dominant.

Definition 3 (Iterated Deletion of Individually Irrational Actions)**.** *For a game G, let $\tilde{A}^0_i = A_i$ for all $i \in N$. Let $\tilde{A}^0_{-i} = A_{-i}$. Then for all $m > 0$ let $\tilde{A}^m_i = \tilde{A}^{m-1}_i \backslash D_i(\tilde{A}^{m-1}_{-i})$ where $\tilde{A}^{m-1}_{-i} = \times_{j \neq i} \tilde{A}^{m-1}_j$.*

The set of actions that survive iterated deletion of individually irrational actions, or those that are iteratively individually rational, for i is given by $IIR_i = \bigcap_{m \geq 0} \tilde{A}^m_i$. Let $IIR = \times_{i \in N} IIR_i$.

Given these definitions, we can present the first necessary condition of Negotiated Binding Agreement action profile profiles and equilibrium strategies in n-player games, which states that any proposal, at any history – on and off the path of play, must survive iterated elimination of individually irrational actions in the underlying game. This exact process was used in order to find the possible proposals in the first price auction with heterogeneous values.

Theorem 3. *If s^* is an equilibrium then for all $h \in H$, $s^*_i(h) \in IIR_i$.*

[13] The notion of absolute dominance was more recently used by Doval and Ely (2020), who extend this concept to incomplete information.

To better understand the set of actions that survives iterated elimination of individually irrational actions, note the following. In a large class of games, non-emptiness of the set of actions that are iteratively individually rational is implied by the fact that the set of actions that survive iterated elimination of never best responses to pure actions, a refinement of rationalizable strategies as defined by Bernheim (1984); Pearce (1984), also survive iterated elimination of individually irrational actions.[14] This is formalised in the following definition and lemma.

Definition 4. *Let $a_i \in A_i$ be a never best response to a pure action in $C_{-i} \subseteq A_{-i}$ if, for all $a_{-i} \in C_{-i}$ there is some $a_i' \in A_i$ for which $u_i(a_i', a_{-i}) > u_i(a_i, a_{-i})$. Denote the set of actions that are never best responses to pure actions in C_{-i} by $NBR_i(C_{-i})$.*

Let $B_i^0 = A_i$. Let $B_i^k = B_i^{k-1} \backslash NBR_i(A_{-i}^{k-1})$. Let $B^k = \times_{i \in N} B_i^k$ and $B_{-i}^k = \times_{j \neq i} B_j^k$. Let the set of actions that survive iterated elimination of never best responses to pure actions be given by $IENBR = \bigcap_{k \geq 1} B^k$.

Lemma 2. *The set of actions that survive iterated elimination of never best responses to pure actions also survives iterated elimination of iterated deletion of individually irrational actions: $IENBR \subseteq IIR$.*

Note that the set of actions that survives iterated elimination of never best responses is necessarily non-empty in finite games. Typically even more profiles may survive iterated elimination of individually irrationally actions than never best responses to pure actions. To see this, consider the following underlying game.

Example 1. Let the underlying game, G, be the following prisoners' dilemma.

1\2	C	D
C	3,3	0,4
D	4,0	1,1

D is strictly dominant for both players, hence (D, D) is the only profile that survives iterated elimination of never best responses to pure actions. Yet, in IIR, all action profiles survive. This is as the maximum payoff for playing C given by 3. The individually rational payoff is given by 1. Therefore C is not individually irrational. ▼

Any action profile satisfying the conditions of the sufficient conditions will be held in IIR, and therefore all pure Nash equilibria must be included.

The next result provides further necessary conditions, providing a relationship between to Negotiated Binding Agreement payoffs with iterative individual rationality considerations in the underlying game.

[14] As all proposals are pure the notion is defined with respect to pure actions. It is a simple extension to show that when mixed proposals are permitted similar results hold in relation to a version of rationalizability.

Theorem 4. *if s^* is an equilibrium then:*

$$U_i(s^*|h) \geq \inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i})$$

for all $h \in H$ and $i \in N$.

Therefore if a^* is a Negotiated Binding Agreement action profile then:

$$u_i(a^*) \geq \inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i})$$

for all $i \in N$.

I illustrate the use of this result with the same underlying prisoners' dilemma game as in example 1.

Example 1. revisited Again consider the underlying game, G, to be that of example 1.

No actions are individually irrational for any player, as previously argued. However, notice that the min-max payoff for each player is 1. The min-max is given by 1, as the worst outcome is the other player selecting D. Therefore we conclude that (D, C) and (C, D) cannot be a Negotiated Binding Agreement. However, the necessary conditions do not rule out the possibility of (C, C). ▼

Note that for any underlying game the inf-sup restricted to the set of actions that survives iterated elimination of individually irrational actions is always weakly higher than the inf-sup without this restriction.

Remark 1. *For any underlying game, G, such that $\underline{u}_i$ is well defined then*

$\inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) \geq \inf_{a_{-i} \in A_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$.

Notice this inequality holds strictly within the leading example: the min-max payoff for bidder 1 is 0, via other firms setting bids of 7, however the min-max payoff when we restrict ourselves to IIR is 1/2.

The results of this section bear resemblance to the analysis of infinitely repeated games, where individual rationality constraints must be satisfied. However, this iterated version can be substantially more restrictive. For instance, in the First Price auction it would only rule out bidders having a net negative valuation, and would not provide a lower bound on the surplus of bidder 1.

Before moving forward, I point to the following corollary, which provides a class of game for which the Negotiated Binding Agreements are fully characterised.

Corollary 1. *If a^{NE} is a pure Nash equilibrium of the underlying game G such that:*

$$u_i(a^{NE}) = \min_{a_{-i} \in IIR_{-i}} \max_{a_i \in A_i} u_i(a_i, a_{-i})$$

i.e. the IIR min-max profiles are mutual, then a^* is a Negotiated Binding Agreement if and only if $u_i(a^*) \geq u_i(a^{NE})$.

This is a direct implication of theorems 2 and 4. This provides a class of games for which the Negotiated Binding Agreements are fully characterised by action profiles that Pareto Dominate a Nash equilibrium in the underlying game. Specifically, if a Nash equilibrium provides agents with their individually rational payoffs over the set of actions that survives iterated deletion of individually irrational actions in the underlying game, then an action profile is a Negotiated Binding Agreement if and only if said action profile Pareto Dominates this Nash equilibrium of the underlying game. This is the case in the three bidder first price auction used as a leading example for this section.

5 Coalitional Deviations

In principle, allowing for jointly beneficial (multilateral) deviations can resolve inefficiencies. However, as will be discussed shortly, this is not always the case since allowing for such deviations can lead to inexistence.

To study this, I allow for a collection of permissible coalitions, where a coalition may jointly deviate. The richest of all such possibilities is the power set of N, which allows *any* possible subset of players to jointly deviate. I will look for the most robust form of equilibrium, that prevents any permissible coalition from deviating, where coalitions are permitted to agree to any deviation.

Allowing for any possible deviation may be seen as overly permissive, as it increases the possibility of equilibrium nonexistence. In principle, one might prefer to restrict the set of permissible deviations. For example, we could require deviations to be the result of some form of agreement among coalition members. However, this would require assumptions about the internal negotiation procedures within coalitions. Do agents have veto power? Can they jointly pre-commit to exclude certain outcomes from any agreement? Different assumptions about these processes can yield markedly different predictions about the resulting equilibria.

If we aim for predictions that are robust to the specifics of how coalitional negotiations unfold, it is reasonable to allow any deviation that a coalition can agree upon. In such cases, the existence of an equilibrium under unrestricted deviations guarantees equilibrium existence under any more limited set of deviations.

The analysis below makes two key points. First, if the stated conditions are satisfied, the resulting equilibrium is robust to group deviations. Second, if these conditions are not met, equilibrium may fail to exist, and predictions become sensitive to the precise structure of permitted deviations. In such cases, no general claims about efficiency gains from coalitional deviations can be made.

5.1. Definitions

I first introduce the notation of a coalition and coalition configuration. A coalition configuration defines the set of coalitions that may make a binding agreement within the negotiation. I let a coalition configuration be denoted by $\mathcal{C}$, and only restrict $\mathcal{C}$ to be a cover of N. That is, for all $i \in N$, there is some coalition $C \in \mathcal{C}$ such that $i \in C$. For a coalition configuration $\mathcal{C}$, if $C \in \mathcal{C}$ I will refer to C as permissible.

Further to this, for a non-empty coalition $C \in \mathcal{C}$, let $a_C = (a_i)_{i \in C}$, $A_C = \times_{i \in C} A_i$, $s_C = (s_i)_{i \in C}$ and $S_C = \times_{i \in C} S_i$. Let $a_{-C} = (a_i)_{i \notin C}$, $A_{-C} = \times_{i \notin C} A_i$, $s_{-C} = (s_i)_{i \notin C}$ and $S_{-C} = \times_{i \notin C} S_i$. For a set $B \subset A$, which may or may not have a product structure, let $B_C = \{a_C \in A_C | \exists a'_{-C} \in A_{-C} \text{ s.t. } (a_C, a'_{-C}) \in B\}$ and $B_{-C} = \{a_{-C} \in A_{-C} | \exists a_C \in A_C \text{ s.t. } (a_C, a_{-C}) \in B\}$.

With this, I go on to define the natural extension of Subgame Perfect Equilibrium when coalitions are permitted to jointly deviate. This will be referred to as $\mathcal{C}$-Subgame Perfect Equilibrium and will require that strategies are such that, at no history of the negotiation game, is there a way for *any* permissible coalition of players, $C \in \mathcal{C}$, to jointly deviate and improve the utility of all players within that coalition.[15]

Definition ($\mathcal{C}$-Subgame Perfect Equilibrium)**.** *s^* is a $\mathcal{C}$-Subgame Perfect Equilibrium if, for all partial histories $h \in H$, there does not exist a non-empty coalition $C \in \mathcal{C}$ and a joint strategy $s_C \in \times_{i \in C} S_i$, such that $u_i(s_C, s^*_{-C}|h) > U_i(s^*|h)$ for all $i \in C$.*

This concept generalises a number of solution concepts. Firstly, whenever $\mathcal{C} = \{\{i\}_{i \in N}\}$, $\mathcal{C}$-Subgame Perfect Equilibrium and Subgame Perfect Equilibrium of Selten (1965) coincide. Further to this, whenever $\{\{i\}_{i \in N}\} \subset \mathcal{C}$, $\mathcal{C}$-Subgame Perfect Equilibrium is a refinement of Subgame Perfect Equilibrium. Whenever $\mathcal{C} = 2^N \backslash \{\emptyset\}$, $\mathcal{C}$-Subgame Perfect Equilibrium coincides with the concept of strong perfect equilibrium of Rubinstein (1980), in this case I will refer to this concept as strong in its place. Note that any strong Subgame Perfect Equilibrium would also be a $\mathcal{C}$-Subgame Perfect Equilibrium for any $\mathcal{C}$. Finally, when $\mathcal{C}$ is a

[15] In essence, this is assuming that, at any history, any permissible coalition may write a private binding agreement that dictates the behaviour they will take going forward. If the agreements were public, the concept would be closer to a coalitional version of Tennenholtz (2004)'s program equilibrium.

partition of N, $\mathcal{C}$-Subgame Perfect Equilibrium can be seen as the extension of coalitional equilibrium of Ray and Vohra (1997) to extensive form games.

To find the set of *$\mathcal{C}$-Negotiated Binding Agreement*, $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game and require a signalling of agreement condition, as previously done.

Definition 5 ($\mathcal{C}$-Negotiated Binding Agreement)**.** *a^* is a $\mathcal{C}$-Negotiated Binding Agreement action profile if there is some strategy profile of the negotiation game where*

1. *$s^*(\emptyset) = a^*$.*

2. *s^* is a $\mathcal{C}$-Subgame Perfect Equilibrium.*

3. *s^* respects coalitional signalling of agreements, i.e. for all $C \in \mathcal{C}, h \in H$ $s^*_C(h) \in A^*_C(s^*)$ where*

$$A^*_C(s^*) = \{a_C \in A_C | a_C = a_C(s^*|h) \text{ for some } h \in H \text{ such that } (s^*|h) \in Z'\}$$

is the set of all possible agreement outcomes for coalition $C \in \mathcal{C}$.

When $\mathcal{C} = 2^N \backslash \{\emptyset\}$ I refer to this as a strong Negotiated Binding Agreement action profile. Whenever $\{i\}_{i \in N} \subset \mathcal{C}$, $\mathcal{C}$-Negotiated Binding Agreements are a subset of Negotiated Binding Agreements and therefore necessary conditions still hold. However, we can strengthen these conditions, and provide conditions that hold for a general coalition configuration $\mathcal{C}$. I show that natural extensions of the necessary and sufficient conditions used for Negotiated Binding Agreement hold for $\mathcal{C}$-Negotiated Binding Agreement.

5.2. $\mathcal{C}$-Negotiated Binding Agreement Outcomes

5.2.1. Necessary Conditions

First, I will show that in any $\mathcal{C}$-Negotiated Binding Agreement the action profile must survive a procedure of *iterated deletion of coalitionally irrational actions* on the underlying game. This procedure generalises the notion of Iterated Elimination of Individually Irrational actions to allow coalitions of players in $\mathcal{C}$ to be the unit of decision making. This provides a recursive version of Aumann (1961)'s β-core, where the "punishments" themselves must be justified. This, therefore, provides one answer to the question posed by Scarf (1971), providing a notion of the core for normal form games that is fully justified.[16]

[16] Chakrabarti (1988) offers a different solution to this question by taking the punishments to be such that they cannot be coalitionally dominated for any action.

Definition 6. *For any underlying game G, for a coalition C, a joint action $a_C \in A_C$ is coalitionally irrational with respect to $B_{-C} \subseteq A_{-C}$ if, for some $a'_C : B_{-C} \to A_C$:*

$$\inf_{a_{-C} \in B_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a_{-C} \in B_{-C}} u_i(a_C, a_{-C}) \qquad \forall i \in C$$

Denote the set of joint actions that are coalitionally irrational with respect to B_{-C} by $D_C(B_{-C})$.

Definition 7 (Iterated Elimination of Coalitionally Irrationality Actions with Respect to $\mathcal{C}$)**.** *For any game G, let $\tilde{A}^0(\mathcal{C}) = A$. For $m > 0$ let:*

$$\tilde{A}^m(\mathcal{C}) = \tilde{A}^{m-1}(\mathcal{C}) \backslash \left[\bigcup_{C \in \mathcal{C}} \left[[D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})] \times A_{-C} \right] \right]$$

Let the set of action profiles that survive iterated elimination of coalitionally irrational actions, or those that are iteratively coalitionally rational, with respect to $\mathcal{C}$ be denoted by $ICIR(\mathcal{C})$ where $ICIR(\mathcal{C}) = \bigcap_{m>0} \tilde{A}^m(\mathcal{C})$.

Note, unlike iterated elimination of individually irrational actions, iterated elimination of coalitionally irrational actions may be empty, even in finite games. To see this, consider the following example.

Example 2. Consider the following two-player game. Let $\mathcal{C} = \{\{1,2\},\{1\},\{2\}\}$.

1\2	L	C	R
T	20,0	20,0	20,0
M	0,7.5	0,7.5	30,5
D	10,10	0,0	0,0

Notice that only (M, R) and (D, L) survive iterated elimination of coalitionally irrational actions for the coalition $C = \{1, 2\}$. However, D cannot survive elimination of individually irrational actions for player 1, as the maximum payoff of D is 10 while the min-max utility for player 1 is 20. Therefore we conclude that within the first round of iterated elimination of coalitionally irrational actions only (M, R) survives. This implies that R is individually irrational with respect to M for player 2, as the profile (M, R) gives a payoff of 5 while the min-max utility, when restricting attention to player 1 playing R is 7.5. Therefore $ICIR(\mathcal{C}) = \emptyset$. ▼

$ICIR(\mathcal{C})$ of course may be non-empty, even when a rich set of coalitions are permitted. Before doing so, notice the following. If $\mathcal{C}' \subset \mathcal{C}$, then $ICIR(\mathcal{C}) \subseteq ICIR(\mathcal{C}')$. Given this, if some action profile survives $ICIR(2^N \backslash \{\emptyset\})$ then it survives any other $\mathcal{C}$.

Example 3. Consider the following two-player game. Let $\mathcal{C} = \{\{1,2\},\{1\},\{2\}\}$.

1\2	L	C	R
T	2,7	2,8	0,6
M	1,4	0,8	2,3
D	1,9	0,8	20,7.5

Notice that (D, R), and (D, L) and (T, C) are the set of Pareto efficient outcomes, therefore, as $\{1, 2\} \in \mathcal{C}$, it must be all other action profiles are ruled out in $\tilde{A}^1(\mathcal{C})$. Further, R is individually irrational for 2 as it provides a payoff of at most 7.5, while the min-max payoff is 8. We conclude that $\tilde{A}^1(\mathcal{C}) = \{(D, L), (T, C)\}$. Now notice that D is individually irrational for 1 with respect to $\tilde{A}^1_{-1}$, where $\tilde{A}^1_{-1} = \{L, C\}$, as the highest payoff that D can provide is 1 while the min-max payoff over this set is 2. We conclude that $\tilde{A}^2(\mathcal{C}) = \{(T, C)\}$. Finally, note that neither T or C are individually irrational given $B_{-1} = \{C\}$ and $B_{-2} = \{T\}$ respectively. Therefore $ICIR(\mathcal{C}) = \{(T, C)\}$. ▼

One condition that ensures non-emptiness of $ICIR(\mathcal{C})$, regardless of the coalition configuration, is the existence of a strong Nash equilibrium.

Lemma 3. *For any Strong Nash equilibrium a^{SNE} of G, $a^{SNE} \in ICIR(\mathcal{C})$ regardless of $\mathcal{C}$.*

A similar necessary condition to theorem 3 holds, linking $ICIR(\mathcal{C})$ of the underlying game to the $\mathcal{C}$-Negotiated Binding Agreements.

Theorem 5. *For any $\mathcal{C}$-Subgame Perfect Equilibrium satisfying coalitional signalling of agreement, s^*, and any $h \in H$, $s^*(h) \in ICIR(\mathcal{C})$.*

Notice once again that this holds for all histories. Further to this, by the definition of $ICIR(\mathcal{C})$, whenever $N \in \mathcal{C}$, it follows that no proposal is coalitionally irrational for the coalition N. This implies that only proposals that are weakly Pareto optimal in the underlying game may be used.

The following corollary links the observation surrounding the potential emptiness of $ICIR(\mathcal{C})$ of the underlying game to the emptiness of $\mathcal{C}$-Negotiated Binding Agreement.

Corollary 2. *If $ICIR(\mathcal{C}) = \emptyset$ then no $\mathcal{C}$-Negotiated Binding Agreement can exist.*

This is an immediate implication of theorem 5. Note that this is possible, i.e. in example 2, and may imply that there is no Negotiated Binding Agreement that is robust to the concerns of coalitions for a specific coalition structure $\mathcal{C}$.

A result analogous to theorem 4 also holds. This result will state that at any history h, a $\mathcal{C}$-Negotiated Binding Agreement must give a payoff that is coalitionally rational for any coalition C in the underlying game, with respect to $[ICIR(\mathcal{C})]_{-C}$. A payoff is not coalitional rational, with respect to $[ICIR(\mathcal{C})]_{-C}$, if, for any punishment a coalition can find some joint action $a_C \in A_C$ such that the utility is higher for all agents. To understand the implications of this result more fully, I define a notion of the β-core Aumann (1961), which I refer to as the β-core with respect to $ICIR(\mathcal{C})$.

Definition 8. *$a^* \in A$ is in the β-core with respect to $ICIR(\mathcal{C})$ if, there is no $C \in \mathcal{C}$ and $a_C : [ICIR(\mathcal{C})]_{-C} \to A_C$ such that $\inf_{a_{-C} \in [ICIR(\mathcal{C})]_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > u_i(a^*)$ for all $i \in C$.*

For an action profile to be in the β-core the payoff of this profile must be higher than the coalitional rational with respect to A_{-i}, in the sense that a coalition understands that they can only be punished for a deviation with a specific profile of actions. However, the actions used to prevent deviations are not necessarily justifiable. The β-core with respect to $ICIR(\mathcal{C})$ partially resolves this problem, as upon deviating the actions of others are restricted to a set of actions that is consistent with respect to itself and is defined in a similar way to the β-core restriction itself.

With this, I formalise the result connecting $\mathcal{C}$-Negotiated Binding Agreement to the β-core with respect to $ICIR(\mathcal{C})$.

Theorem 6. *For any $\mathcal{C}$-Negotiated Binding Agreement s^* must be such that, for any history h, and for any coalition $C \in \mathcal{C}$, there is no $a'_C : [ICIR(\mathcal{C})]_{-C} \to A_C$ such that:*

$$\inf_{a_{-C} \in [ICIR(\mathcal{C})]_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > U_i(s^*|h)$$

for all $i \in C$.

In other words, $a(s^*|h)$ must be in the β-core with respect to $ICIR(\mathcal{C})$ for all histories.

Note that it may be that an outcome is both Pareto efficient and individually rational in the underlying game, yet it is not possible to sustain such an outcome via a $\mathcal{C}$-Negotiated Binding Agreement for $\{N, \{i\}_{i \in N}\} \subseteq \mathcal{C}$.

Example 4. Let the following two-player game be the underlying game G. Consider the richest set of coalitions $\mathcal{C} = \{\{1\}, \{2\}, \{1,2\}\} = 2^N \backslash \{\emptyset\}$.

1\2	LL	L	R	RR
TT	$\mathbf{6,6}$	0,4	$\mathbf{1,\underline{12}}$	0,0
T	4,0	0,0	$\mathbf{7,\underline{2}}$	$\underline{1}$,1
D	$\mathbf{\underline{12},1}$	$\mathbf{\underline{2},7}$	4,4	0,$\underline{8}$
DD	0,0	1,$\underline{1}$	$\underline{8}$,0	0,0

I have labelled the weakly Pareto efficient outcomes of G in bold blue font, and therefore must be the only actions in $\tilde{A}^1$ are $\{(TT, LL), (TT, R), (T, R), (D, LL), (D, L)\}$. No further deletion can take place therefore:

$$ICIR(2^N\backslash\{\emptyset\}) = \{(TT, LL), (TT, R), (T, R), (D, LL), (D, L)\}$$

(TT, R) necessarily cannot be sustained in a strong Negotiated Binding Agreement, as it provides a payoff of 1, while the min-max payoff, given that player 2 must choose from $[ICIR(2^N\backslash\{\emptyset\})]_2 = \{LL, L, R\}$, is given by 2. Therefore we conclude that despite the fact that (TT, R) is Pareto efficient in G, and provides a higher payoff than the min-max over all possible profiles it cannot be sustained in a strong Negotiated Binding Agreement. ▼

With these results, I now turn to providing sufficient conditions for $\mathcal{C}$-Negotiated Binding Agreement.

5.2.2. Sufficient Conditions

To provide sufficient conditions for the outcomes of a $\mathcal{C}$-Negotiated Binding Agreement, as with theorem 2, I will rely on conditions of the underlying game G. To provide these conditions, I again rely on a structure that does not focus on the deviation that a coalition takes, but only on the deviating coalition. In this case, a coalition must prefer the punishment of others to their own and a coalition must not be able to improve all members' utility by changing their action profile in G, holding the punishment used against them constant. Note, due to the rich deletion that can take place, the inclusion of such profiles in $ICIR(\mathcal{C})$ is now required and not implied.

Theorem 7. *Take any underlying game such that there is some $a^* = \underline{a}^N \in ICIR(\mathcal{C})$ and for all $C \in \mathcal{C}\backslash N$ $\exists \underline{a}^C \in ICIR(\mathcal{C})$ such that:*

1. *$\nexists a'_C \in A_C$ such that $u_i(a'_C, \underline{a}^C_{-C}) > u_i(\underline{a}^C)$ for all $i \in C$*
2. *for all $C \in \mathcal{C}$ there is some $i \in C$ such that $u_i(a^*) \geq u_i(\underline{a}^C)$*
3. *For all $C, C' \in \mathcal{C}$ there is some $i \in C$ such that $u_i(\underline{a}^{C'}) \geq u_i(\underline{a}^C)$*

Then a^* can be supported in a $\mathcal{C}$-Negotiated Binding Agreement.

Combining this result with the result of lemma 3, which states that if a strong Nash equilibrium of G exists it is within $ICIR(\mathcal{C})$, implies that any strong Nash equilibrium of G can be supported in a $\mathcal{C}$-Negotiated Binding Agreement. However, these conditions can apply in underlying games with no strong Nash equilibrium, and therefore are a more general set of conditions.[17] To see this, consider the following example.

Example 4. revisited Consider again the following two-player game as the underlying game, G, given in example 4. All possible coalitions are permitted, $\mathcal{C} = 2^N \backslash \{\emptyset\}$.

Here there is no strong Nash equilibrium of G. In fact, as there is no pure Nash equilibrium in G, there is no pure coalition proof Nash equilibrium. However, the conditions of theorem 7 apply.

Given the previous analysis we may take $\underline{a}^N = a^* = (TT, LL)$, $\underline{a}^1 = (D, L)$ and $\underline{a}^2 = (T, R)$. Concluding that (TT, LL) can be sustained in $2^N \backslash \{\emptyset\}$-Negotiated Binding Agreement. ▼

The sufficient conditions for outcomes of $\mathcal{C}$-Negotiated Binding Agreements presented in theorem 7 can be seen as a further refinement of the β-core of Aumann (1961), where within the β-core any constant action profile in G of those outside of a coalition may be used in order to prevent deviations, whereas in this paper we must satisfy additional conditions to ensure such a profile in G can be mutually justified by all coalitions. Note that this is not necessarily true in the notion of the β-core with respect to $ICIR(\mathcal{C})$, as some profiles within $ICIR(\mathcal{C})$ do not satisfy this notion of mutual coalitional rationality.

6 Literature Review

A number of papers have approached the question of which binding agreements can be made for normal form games using an approach close to or inspired by the farsighted stable set of Harsanyi (1974). I instead take a more non-cooperative game theoretic approach, exploring the SPE in a fully specified negotiation game where agents signal the agreement they would like to take. Within this strand of literature, Mariotti (1997) has the closest model and also considers an explicit negotiation protocol. The extensive form of the negotiation protocol is similar, but the payoff of perpetual disagreement is set to $-\infty$. In this work, Mariotti (1997) takes an approach close to the strong Subgame Perfect Equilibrium of Rubinstein (1980). He also imposes a refinement on this subgame perfect type concept based on the farsighted stable set. Mariotti (1997) does not provide general conditions for his solution concept, due to the complexity that the history-dependent negotiation entails. He instead proposes a history-independent version of his solution concept, in line with Harsanyi (1974), where agents strategies only map from the current proposal to the next proposal, rather than all possible previous proposals being considered. In this history independent version,

[17] Shubik (2012) examines the 78 2x2 games which can be induced by strict ordinal preferences, of these 78, 67 allow for the sufficient conditions for outcomes of a $\mathcal{C}$-Negotiated Binding Agreement to be applied. Note that is only 2 less than the existence of Nash equilibrium in pure strategies. In this sense, these sufficient conditions apply to more scenarios than initial inspection may suggest.

Mariotti (1997) provides some necessary conditions for agreement outcomes similar to those provided in this paper for both Negotiated Binding Agreements and $\mathcal{C}$-Negotiated Binding Agreements. He also provides sufficient conditions for agreement outcomes for a class of two-player games with conditions on the Pareto Frontier, similarly using a notion of individual punishments.

Chwe (1994); Xue (1998); Ray and Vohra (2015, 2019) also consider versions of the farsighted stable set. The closest with respect to my paper is Ray and Vohra (2019), which games with transferable utility, and defines the notion of the maximal farsighted stable set, which additionally requires a subgame perfect-like condition, imposing optimality given others' strategies at all histories of the negotiation. They provide general conditions linking the farsighted stable set as defined in Ray and Vohra (2015) to this concept. I instead take an approach that looks at general games, rather than a game with transferable utility, and instead link the concept of $\mathcal{C}$-Negotiated Binding Agreements to an alternative cooperative game theoretic notion of the β-core of Aumann (1959, 1961). Finding the farsighted stable set is challenging and some papers have looked at finding the farsighted stable set for a specific underlying game (Suzuki and Muto, 2005; Nakanishi, 2009).

Other papers have also proposed fully non-cooperative models of negotiation over binding agreements for normal form games, based on a dynamic game of negotiation. Kalai (1981) looks at a fully specified model of negotiation by proposing a non-cooperative extensive form game. In that model, agents propose an individual action in the underlying game. If an agent changes their proposal within a period then they are no longer permitted to change their proposal again. The process ends at time t with the proposal profile proposed in that period. Kalai (1981) looks at the perfect equilibria of Selten (1988) and shows that only cooperation can be sustained in the 2-player prisoners' dilemma game. Nishihara (2022) has extended this to an n-player prisoners' dilemma, maintaining Kalai's negotiation protocol. The philosophy of Kalai's approach is similar to that of this paper, where agents negotiate over the agreement and can do so by proposing their own action. Bhaskar (1989) examines a model of pre-play agreement over a symmetric two-player Bertrand game. In a similar sense to this model, agents make proposals of the prices they will take, and have the opportunity to revise their proposals sequentially. Confirmation requires one agent not to change their proposal after seeing the others. Bhaskar (1989) looks at the perfect equilibria of such an agreement game and concludes that only the monopoly price can be sustained. The closest model in the non-cooperative literature is that of Harstad (2022), who proposes a "pledge-and-review" bargaining protocol, similar to the one in this paper, for public goods games. In his model, Harstad (2022) shows that when agents confirm by default, and discounting is hyperbolic, a folk theorem remains for the subgame perfect equilibria outcomes of this game. When considering stationary subgame perfect equilibria, each agent's pledge must be the result of maximising *some* weighted Nash product. However, the weights used by each agent may differ, and therefore many inefficient equilibria arise despite this. In my work, I instead consider a more general class of games and consider and alternative refinement of SPE. Note that by the construction of the equilibria, imposing stationarity does not substantially change sufficiency.

A number of papers have provided a more cooperative game theoretic approach for the agreements that can be made for games, for instance Strong Nash equilibrium (Aumann, 1959) and the β-core (Aumann, 1959, 1961). In my paper, $\mathcal{C}$-Negotiated Binding Agreement outcomes lie somewhere between the β-core and Strong Nash equilibrium and is fully backed by a negotiation procedure. Given this, my paper can also be seen in the light of the Nash program pointed to in Nash (1953), as the necessary and sufficient conditions $\mathcal{C}$-Negotiated Binding Agreement outcomes can be seen as a perturbed version of the β-core.

There are a number of other related papers that take the cooperative game theoretic approach. Notably, the γ-core (Chander and Tulkens, 1997). Chander (2007) provides further justification for the γ-core by showing it is *an* equilibrium to an infinitely repeated game where agents decide whether to cooperate or not in each round. Chander and Wooders (2020) define a notion of coalitional Subgame Perfect Equilibrium for underlying games with transferable utility, where a coalition's deviation payoff is with respect to the best Subgame Perfect Equilibrium assuming all other players act without cooperation. A number of papers have also proposed notions of rationalizability for coalitions in a cooperative sense, for instance Herings et al. (2004); Ambrus (2006, 2009); Grandjean et al. (2017), which iterative elimination of coalitionally irrational actions can be seen as, but are all distinct. A strand of literature abstracts from the negotiation process *within* a group and takes a cooperative perspective, focusing on Pareto undominated actions that prevent new groups from breaking and forming (Ray and Vohra, 1997; Diamantoudi and Xue, 2007).

A number of papers consider a form of communication for equilibrium selection (Bernheim et al. (1987); Farrell and Maskin (1989); Bernheim and Ray (1989); Rabin (1994), etc.). My paper is related in the sense that agents can communicate via the negotiation procedure, proposing the agreement action they would like, to select the outcome of the underlying game that will be played. However, the perspective is different, as these concepts are about refining a given set of non-binding agreements represented by the (potential mix over) SPE or Nash Equilibria of an underlying game, whereas I allow agents to make a binding agreement of potentially any outcome.

In the contracting literature, the closest work is of Jackson and Wilkie (2005); Yamada (2003); Ellingsen and Paltseva (2016) who all propose model allowing agents all have a strategic input on the *structure* of the contract over an underlying strategic environment.

In a similar way, Negotiated Binding Agreements allows for all agents to have a strategic input on the action they will agree to in the underlying game. On the other hand, Kalai et al. (2010), Peters and Szentes (2012) and Tennenholtz (2004) all consider the possibility of all agents proposing contracts surrounding their own play in an underlying game, where these contracts can be a function of the contracts of others. This allows agents to specify reactions to deviations in full, and can allow for these to be fully specified at a higher level also. In contrast to these, my paper is requires that agents are required to only propose actions they could agree to, whereas these papers allow contracts to specify actions in the contract that would never be the result of equilibrium. Inefficiency persists, but with a novel lower bound.

The way payoffs are defined for perpetual disagreement can be seen as similar to the literature of infinitely repeated games with no discounting (Aumann and Shapley, 1994; Rubinstein, 1994). The individual punishment results within the paper are also similar to player-specific punishment is used in infinitely repeated games (Fudenberg and Maskin, 1986; Abreu et al., 1994). The sufficient conditions I use are more restrictive as player-specific punishment only requires that their punishments' provide them an individually rational payoff and they prefer to punish rather than be punished. In contrast, I also require that individuals are best responding to their punishment in the underlying game. These are used as there are no further rewards from following their punishments, which are held in the continuation of an infinitely repeated game. Therefore it must be the case that agents cannot improve the utility they would get facing the constant punishment of others, requiring that they best respond.

7 Conclusion

This paper has developed a new theoretical framework for understanding how agents negotiate over actions in strategic settings, and how inefficiencies can arise even when agreements are binding, information is complete, and there is no cost of delay. The model introduces a simple yet powerful negotiation protocol in which agents must confirm proposals to reach a binding agreement, and the analysis is conducted through the lens of subgame perfect equilibrium.

The central insight is that inefficiencies emerge endogenously due to two key structural features of negotiation: (i) each agent retains control over their own action, and (ii) any agreement must be mutually confirmed. These features mean that even in a setting with full strategic rationality, inefficiencies can persist—not due to external frictions, but as a consequence of the strategic interdependence built into the negotiation process itself.

In two-player games, I provide a full characterization of the agreement outcomes, showing that a strategy profile can be sustained in equilibrium if and only if it guarantees each player at least their individual punishment payoff. For n-player games, a similar logic yields a set of sufficient conditions, while a novel iterative rationality constraint offers a necessary condition, significantly narrowing the space of sustainable outcomes. These conditions help identify when a particular agreement is strategically feasible, and when it is not, based solely on the structure of the underlying game.

The model is also extended to allow coalitional deviations, introducing the concept of $\mathcal{C}$-Negotiated Binding Agreements, which generalizes the β-core of cooperative game theory in a fully strategic setting. While such agreements may eliminate some inefficiencies, they can also result in non-existence, highlighting the inherent tension between robustness and feasibility in negotiated outcomes.

Together, these results offer a tractable and robust approach to understanding strategic agreement formation, and explain why suboptimal agreements frequently arise in practice – even in an idealised settings. The model speaks directly to real-world inefficiencies observed in numerous agreements such as trade agreements, climate negotiations, labour bargaining, and industrial collusion. The formal analysis of these questions is left for future research to do them justice. This paper shows that, apart from informational or institutional frictions alone, the strategic structure of negotiation alone limits the scope for efficient cooperation.

A Proofs

Proof of lemma 1: Notice that $\liminf_{k\to\infty} u_i(a^k) = (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\liminf_{k\to\infty} u_i(a^k)$. Therefore by continuity of subtraction we have that

$$\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(a^t) - \liminf_{k\to\infty} u_i(a^k) = \lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\left(u_i(a^t) - \liminf_{k\to\infty} u_i(a^k)\right)$$

Note by definition of the $\liminf$, for all $\epsilon > 0$ $\exists T \in \mathbb{N}$ such that $\forall t > T$ we have that $u_i(a^t) - \liminf_{k\to\infty} u_i(a^k) > -\epsilon$. Therefore, for any such T, we may decompose the expression as follows.

$$\begin{aligned}\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right) &= \lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{T}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right)+\cdots\\ &\cdots+\lim_{\delta\to 1}(1-\delta)\sum_{t=T+1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right)\\ &= \lim_{\delta\to 1}(1-\delta)\sum_{t=T+1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right)\\ &> \lim_{\delta\to 1}(1-\delta)\sum_{t=T+1}^{\infty}\delta^{t-1}(-\epsilon)=\lim_{\delta\to 1}-\delta^{T+1}\epsilon=-\epsilon\end{aligned}$$

Therefore $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right) > -\epsilon \ \forall\epsilon > 0$, concluding that $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right)\geq 0$ and therefore

By analogy $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(a^t)\leq \limsup_{k\to\infty}u_i(a^k)$. ■

Proof of Theorem 1:

Sufficiency: Note within this proof I maintain the notation a^k to refer to the k^{th} period proposal in a history h, while I use $\underline{a}^j$ to denote the action profile used in equilibrium as a punishment for j. Let s^* be as follows:

1. if $h=(a^1,...,a^k)$ is such that there is some $j\in N$, such that $a_{-j}^{k-1}=s_{-j}^*((a^1,...,a^{k-2}))$ and either:
   (a) $a_l^k=s_l^*(h\backslash a^{k-1}) \quad \forall l\neq j$ while $a_j^k\neq s_j^*(h\backslash a^{k-1})$.
   (b) or $a_{-j}^k=\underline{a}_{-j}^j$.

   then $s_i^*(h)=\underline{a}_i^j$.
2. $s_i^*(h)=a_i^*$ otherwise.

First note that from any history the continuation is terminal within two periods and therefore signalling of agreement is satisfied. Now to show that s^* is a Subgame Perfect Equilibrium of the negotiation game. Suppose that a profitable deviation exists at a history $h\in H$ for $i\in N$. If the deviation does not include some different proposal within two periods of h it cannot be profitable, as the outcome remains the same. Therefore any deviation must occur within two periods. Any such deviation, denoted by s_i', if it does not lead to the same terminal history and therefore cannot be profitable, of $i\in N$ must lead to $\underline{a}_{-i}^i$ for all periods following. Let the terminal history following the deviation be denoted by $(s_{-i}^*,s_i'|h)=(h,a^k,a^{k+1},....,a^t,...)$. When $(s_{-i}^*,s_i'|h)\in Z'$ let

$(s^*_{-i}, s'_i|h) = (h, a'^{,1}, a'^{,2}, ..., a((s^*_{-i}, s'_i|h)), a((s^*_{-i}, s'_i|h)), a((s^*_{-i}, s'_i|h)), ...)$, i.e let the agreement that $(s^*_{-i}, s'_i|h)$ concludes in be infinitely repeated at the end of the sequence, with some abuse of notation. However, by construction, it must be that $\limsup_{t\to\infty} u(a^t) \leq u_i(\underline{a}^i)$ and therefore it must be at least weakly worse than any terminal history of the strategy s^*. Therefore no profitable deviation exists.

Necessity: To see that only such a^* can be sustained, take any a^* that is a Negotiated Binding Agreement outcome in the two-player case given by the SPE s^*. Denote $\tilde{A} = \{a \in A | \exists h \in H \text{ s.t. } s^*(h) = a\}$. Note by signalling of agreement these are the only actions that can be proposed in equilibrium. Therefore $s^*_{-i}(h) \in \tilde{A}_{-i}$ for all $h \in H$. As s^* is an SPE it must be that there is no profitable deviation. Notice that $U_i(s^*|h) \geq \inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i})$. Suppose not $U_i(s^*|h) < \inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i})$. It follows that $\inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i}) - U_i(s^*|h) > 0$. Consider a deviation to s'_i such that $s'_i(h') = s^*_i(h')$ for all h' such that $h = (h', h'')$ while $s'_i(h')$ is such that $u_i((s'_i, s^*_{-i})(h')) = \max_{a_i\in A_i} u_i(a_i, s^*_{-i}(h'))$ for all other histories. Suppose such a deviation leads to perpetual disagreement. Denote the sequence induced by such a strategy by $z' = (a^1, a^2, ..., a^t, ...)$. Notice that $u_i(a^t_i, a^t_{-i}) = \max_{a_i\in A_i} u_i(a_i, a^t_{-i})$. Note that therefore

$$u_i(a^t_i, a^t_{-i}) \geq \inf_{a_{-i}\in\{a'_{-i}\in A_{-i}|a'_{-i}=a^k_{-i}\}} \max_{a_i\in A_i} u_i(a_i, a_{-i})$$

By definition:

$$\begin{aligned}
U_i(s_i, s^*_{-i}|h) &\geq \liminf_{t\to\infty} u_i(a^t) \\
&\geq \liminf_{t\to\infty} \inf_{a_{-i}\in\{a'_{-i}\in A_{-i}|a'_{-i}=a^k_{-i}\}} \max_{a_i\in A_i} u_i(a_i, a_{-i}) \\
&= \inf_{a_{-i}\in\{a'_{-i}\in A_{-i}|a'_{-i}=a^k_{-i}\}} \max_{a_i\in A_i} u_i(a_i, a_{-i}) \\
&\geq \inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i}) \Rightarrow U_i(s_i, s^*_{-i}|h) > U_i(s^*|h)
\end{aligned}$$

therefore it cannot be that s^* is an SPE if the deviation ends in perpetual disagreement. The argument for agreement is direct from the definition.

Therefore it must be that $U_i(s^*|h) \geq \inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i})$. As all profiles in $\tilde{A}$ are agreed upon, therefore $\forall\tilde{a} \in \tilde{A}$ $u_i(\tilde{a}) \geq \inf_{a_{-i}\in\tilde{A}_{-i}} \max_{a_i\in A_i} u_i(a_i, a_{-i})$. Therefore $\exists a'_{-i} \in \bar{\tilde{A}}_{-i}$, where $\bar{\tilde{A}}_{-i}$ is the limit points of $\tilde{A}_{-i}$ such that $u_i(\tilde{a}) \geq \max_{a_i\in A_i} u_i(a_i, a'_{-i})$. As this holds for all $\tilde{a} \in \tilde{A}$ it follows that $u_i(a') \geq \max_{a_i\in A_i} u_i(a_i, a'_{-i})$ therefore $u_i(a') = \max_{a_i\in A_i} u_i(a_i, a'_{-i})$. therefore $\exists a^i \in \bar{\tilde{A}}$ such that $u_i(\tilde{a}) \geq u_i(a^i) = \max_{a_i\in A_i} u_i(a_i, a^i_{-i})$. Notice that: $u_i(\tilde{a}) \geq u_i(a^i)$ for all $\bar{\tilde{A}}$ and therefore $u_i(\underline{a}^j) \geq u_i(\underline{a}^i)$ and $u_i(a^*) \geq u_i(a^i)$. Therefore such a profile of action profiles must exist for a^* to be supported. ■

Proof of theorem 3: Suppose not, for some history $h' \in H$ we have that $s_i(h') = a_i$.

By consistency of agreement signalling it follows that there exists some $h \in H$ such that $a_i(s|h) = a_i$. Therefore it must be that $U_i(s|h) = u_i(a(s|h)) \leq \sup_{a'_{-i} \in A_{-i}} u_i(a_i, a'_{-i})$. Take

$$\epsilon = \inf_{a'_{-i} \in A_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - u_i(a(s|h)) > 0$$

Take a function $\tilde{a}_i : A_{-i} \to A_i$ such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$. Consider a deviation $s'_i(h'') = \tilde{a}_i(s_{-i}(h''))$ for all $h'' \in H$. It follows that: $U_i(s'_i, s_{-i}|h) \geq \inf_{a_{-i} \in A_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \inf_{a_{-i} \in A_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$. Therefore it follows that $U_i(s'_i, s_{-i}|h) > u_i(a(s|h)) = U_i(s|h)$, concluding that a profitable deviation exists and therefore it cannot be that s is a Subgame Perfect Equilibrium of the negotiation game. By consistency of agreement signalling, we conclude that $s_i(h) \notin D_i(A_{-i})$ for any $h \in H$.

Now suppose by contradiction that, for all $j \in N$ $s_j(h') \in \tilde{A}^k_j$ $\forall k < m$ and $h' \in H$ but for some $i \in N$ $s_j(h') = a_i \notin \tilde{A}^{m+1}_j$ for some $h' \in H$. By consistency of agreement signalling it must be that a) $s_{-i}(h') \in \tilde{A}^m_i$ for all h' and b) by consistency of agreement signalling there is some $h \in H$ for which $a_i(s|h) = a_i$. Therefore it must be that $U_i(s|h) = u_i(a(s|h)) \leq \sup_{a'_{-i} \in \tilde{A}^m_{-i}} u_i(a_i, a'_{-i})$. Take $\epsilon = \inf_{a'_{-i} \in \tilde{A}^m_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - u_i(a(s|h)) > 0$. Take a function $\tilde{a}_i : \tilde{A}^m_{-i} \to A_i$ such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$ for all $a_{-i} \in \tilde{A}^m_{-i}$. Consider a deviation $s'_i(h'') = \tilde{a}_i(s_{-i}(h''))$ for all $h'' \in H$. It follows that: $U_i(s'_i, s_{-i}|h) \geq \inf_{a_{-i} \in \tilde{A}^m_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \inf_{a_{-i} \in \tilde{A}^m_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$. Therefore it follows that $U_i(s'_i, s_{-i}|h) > u_i(a(s|h)) = U_i(s|h)$, concluding that a profitable deviation exists and therefore it cannot be that s is a Subgame Perfect Equilibrium of the negotiation game. By consistency of agreement signalling, we conclude that $s_i(h) \notin D_i(\tilde{A}^m_{-i})$ for any $h \in H$ and therefore $s_i(h) \in \tilde{A}^{k+1}_i$, a contradiction.■

Proof of lemma 2: Note that $B^0 = \tilde{A}^0$. Now we will show that $B^k \subseteq \tilde{A}^k$ for all $k \geq 0$. By the inductive hypothesis suppose that $B^m \subseteq \tilde{A}^m$ for all $m < k$. Now notice that for any $a_i \in B^k_i$ we have that there is some $a_{-i} \in B^{k-1}_{-i} \subseteq \tilde{A}^{k-1}_{-i}$ such $u_i(a_i, a_{-i}) \geq u_i(a'_i, a_{-i})$ for all $a'_i \in A_i$. It follows that $u_i(a_i, a_{-i}) \geq \inf_{a'_{-i} \in B^{k-1}_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) \geq \inf_{a'_{-i} \in \tilde{A}^{k-1}_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i})$. Further,

$$u_i(a_i, a_{-i}) \leq \sup_{a''_{-i} \in B^k_{-i}} u_i(a_i, a''_{-i}) \leq \sup_{a''_{-i} \in \tilde{A}^k_{-i}} u_i(a_i, a''_{-i})$$

and therefore we conclude that if $a_i \in B^k_i$ then $a_i \in \tilde{A}^k_i$, concluding the proof. ■

Proof of theorem 4: Suppose not, then there is some $i \in N$ and $h \in H$ such that that $\inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > U_i(s^*|h)$. It must be that a) s^* is a Subgame Perfect Equilibrium of the negotiation game and b) by theorem 3 it must be that $s^*_{-i}(h) \in IIR_{-i}$ for all $h \in H$. Let $\epsilon = \inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - U_i(s^*|h) > 0$. Construct $\tilde{a}_i : IIR_{-i} \to A_i$ such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) \geq \sup_{a_i \in A_i} u_i(a_i, a_{-i}) - \frac{\epsilon}{2}$ for all $a_{-i} \in IIR_{-i}$. Consider a deviation to $s'_i(h')$ such that $s'_i(h') = \tilde{a}(s^*_{-i}(h'))$ for all $h' \in H$ at the history h. It follows that:

$$U_i(s'_i, s^*_{-i}|h) \geq \inf_{a_{-i} \in IIR_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) = \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i}) - \frac{\epsilon}{2}$$
$$= \frac{\inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i}) + U_i(s^*|h)}{2} > U_i(s^*|h)$$

A contradiction, as therefore s^* is not a Subgame Perfect Equilibrium of the negotiation game. ■

Proof of lemma 3: As a^* is a strong Nash equilibrium, it follows that $\nexists C \in 2^N \backslash \{\emptyset\}, a'_C \in A_C$ such that $u_i(a'_C, a^*_{-C}) > u_i(a^*)$ for all $i \in C$. Therefore a^* is not coalitionally irrational. Now suppose that $a^* \in \tilde{A}^m(\mathcal{C})$ for all $m < k$. Notice that by the same statement this implies that $a^* \in \tilde{A}^{m+1}(\mathcal{C})$. This implies that $a^* \in ICIR(\mathcal{C})$ for all $\mathcal{C}$. ■

Proof of theorem 5: Suppose not, for some history $h' \in H$ we have that $s_C(h') = a_C$. By coalitional signalling of agreement it follows that there exists some $h \in H$ such that $a_C(s|h) = a_C$. Therefore it must be that $U_i(s^*|h) = u_i(a(s^*|h)) \leq \sup_{a'_{-C} \in A_{-C}} u_i(a_C, a'_{-C})$ for all $i \in C$. By definition of a_C being not coalitionally rational, there exists a function $a'_C : A_{-C} \to A_C$ such that $\inf_{a_{-C} \in A_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in A_{-C}} u_i(a_C, a'_{-C})$. Consider a deviation of C at history h such that $s_C(h') = a'_C(s_{-C}(h'))$ for all $h' \in H$. It follows that $U_i(s'_C, s^*_{-C}|h) \geq \inf_{a_{-C} \in A_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in A_{-C}} u_i(a_C, a'_{-C}) \geq U_i(s^*|h)$ for all $i \in C$. Concluding that s^* is not a $\mathcal{C}$-Subgame Perfect Equilibrium.

Now suppose by contradiction that $s(h') \in \tilde{A}^k(\mathcal{C})\ \forall k < m$ and $h' \in H$ but $s(h') = a \notin \tilde{A}^{m+1}(\mathcal{C})$ for some $h' \in H$. By definition, it must be that $a \in \bigcup_{C \in \mathcal{C}} [D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C}) \times A_{-C}]$. Therefore it must be that $a_C \in D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})$ for some $C \in \mathcal{C}$. By coalition agreement signalling we have that $\exists h \in H$ such that $a_C = a^*_C(s^*|h)$. By definition of coalition rationality given $\tilde{A}^{m-1}(\mathcal{C})_{-C}$, as $a_C \in D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})$ there must be some that there is some $a'_C : \tilde{A}^{m-1}(\mathcal{C})_{-C}$ such that $\inf_{a_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a_C, a'_{-C})$. Consider a deviation of C at history h such that $s_C(h') = a'_C(s_{-C}(h'))$ for all $h' \in H$. It follows that:

$$U_i(s'_C, s^*_{-C}|h) \geq \inf_{a_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a_C, a'_{-C})$$

Therefore $U_i(s'_C, s^*_{-C}|h) > U_i(s^*|h)$ for all $i \in C$. Concluding that s^* is not a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game. A contradiction. ■

Proof of theorem 6: Suppose this is not the case. There is some $C \in \mathcal{C}$ $a'_C : [ICIR(\mathcal{C})]_{-C} \to A_C$ such that $\inf_{a_{-C} \in [ICIR(\mathcal{C})]_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > U_i(s^*|h)$ for all $i \in C$. It must be that s^* is a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game, and therefore there cannot exist a profitable deviation for C. Notice that $s_i^*(h) \in [ICIR(\mathcal{C})]_i$ for all $i \in N$.

Consider a joint deviation from coalition C such that $s'_C(h) = a'_C(s^*_{-C}(h))$ for all $h \in H$. By the definition of the utilities that this can induce, it is clear that: $U_i(s'_C, s^*_{-C}|h) \geq \inf_{a_{-C}\in[ICIR(\mathcal{C})]_{-C}} u_i(a'_C(a_{-C}), a_{-C})$ for all $i \in C$, and therefore $u_i(s'_C, s^*_{-C}|h) > U_i(s^*|h)$ for all $i \in C$. In conclusion, s^* cannot be a $\mathcal{C}$-Subgame Perfect Equilibrium.■

Proof of theorem 7: Consider the following strategy:

1. if $h = (a^1, ..., a^k)$ is such that there is some $C \in \mathcal{C}$, such that $a^{k-1}_{-C} = s^*_{-C}((a^1, ..., a^{k-2}))$ and either $a^k_l = s^*_l(h \backslash a^{k-1}) \quad \forall l \notin C$ while $a^k_j \neq s^*_j(h \backslash a^{k-1})$ for all $j \in C$ or $a^k_{-C} = \underline{a}^C_{-C}$ then $s^*_i(h) = \underline{a}^C_i$.
2. $s^*_i(h) = a^*_i$ otherwise.

By definite, at no history can N deviate as a coalition to improve all their utilities if $N \in \mathcal{C}$. Now assume that some other coalition $C \in \mathcal{C}$ has a profitable deviation. If $a_j \neq s^*_j(h)$ for all $j \in C$, then it cannot be profitable as it leads to a history that induces the $\underline{a}^C_{-C}$ for all periods. If $a_j \neq s^*_j(h)$ for all $j \in B$, where $B \subset C$, while $a^*_j = s^*_j(h)$. Then it must induce a path such that either a member of B is worse off, or further deviations within C take place. Either way, it cannot be that this is a profitable deviation.

As all histories end within 2 periods we satisfy consistency of coalition agreement signalling and therefore we have a $\mathcal{C}$-SPE leading to a $\mathcal{C}$-Negotiated Binding Agreement outcome a^*.■

References

Abreu, D., Dutta, P. K., and Smith, L. (1994). The Folk Theorem for Repeated Games: A Neu Condition. *Econometrica*, 62(4):939–948.

Ambrus, A. (2006). Coalitional Rationalizability. *The Quarterly Journal of Economics*, 121(3):903–929.

Ambrus, A. (2009). Theories of Coalitional Rationality. *Journal of Economic Theory*, 144(2):676–695.

Aumann, R. J. (1959). Acceptable Points in General Cooperative n-person Games. *Contributions to the Theory of Games (AM-40)*, 4:287–324.

Aumann, R. J. (1961). The Core of a Cooperative Game without Side Payments. *Transactions of the American Mathematical Society*, 98(3):539–552.

Aumann, R. J. and Shapley, L. S. (1994). Long-Term Competition—a game-theoretic analysis. In *Essays in game theory*, pages 1–15. Springer.

Bernheim, B. D. (1984). Rationalizable Strategic Behavior. *Econometrica: Journal of the Econometric Society*, pages 1007–1028.

Bernheim, B. D., Peleg, B., and Whinston, M. D. (1987). Coalition-Proof Nash Equilibria i. Concepts. *Journal of Economic Theory*, 42(1):1–12.

Bernheim, B. D. and Ray, D. (1989). Collective Dynamic Consistency in Repeated Games. *Games and Economic Behavior*, 1(4):295–326.

Bhaskar, V. (1989). Quick Responses in Duopoly Ensure Monopoly Pricing. *Economics Letters*, 29(2):103–107.

Busch, L.-A. and Wen, Q. (1995). Perfect equilibria in a negotiation model. *Econometrica: Journal of the Econometric Society*, pages 545–565.

Chakrabarti, S. K. (1988). Refinements of the β-core and the strong equilibrium and the aumann proposition. *International Journal of Game Theory*, 17:205–224.

Chander, P. (2007). The gamma-Core and Coalition Formation. *International Journal of Game Theory*, 35(4):539–556.

Chander, P. and Tulkens, H. (1997). The Core of an Economy with Multilateral Environmental Externalities. *International Journal of Game Theory*, 26(3):379–401.

Chander, P. and Wooders, M. (2020). Subgame-Perfect Cooperation in an Extensive Game. *Journal of Economic Theory*, page 105017.

Chatterjee, K., Dutta, B., Ray, D., and Sengupta, K. (1993). A Noncooperative Theory of Coalitional Bargaining. *The Review of Economic Studies*, 60(2):463–477.

Chwe, M. S.-Y. (1994). Farsighted Coalitional Stability. *Journal of Economic Theory*, 63(2):299–325.

Diamantoudi, E. and Xue, L. (2007). Coalitions, Agreements and Efficiency. *Journal of Economic Theory*, 136(1):105–125.

Doval, L. and Ely, J. C. (2020). Sequential information design. *Econometrica*, 88(6):2575–2608.

Ellingsen, T. and Paltseva, E. (2016). Confining the Coase Theorem: contracting, ownership, and free-riding. *The Review of Economic Studies*, 83(2):547–586.

Farrell, J. and Maskin, E. (1989). Renegotiation in Reeated Games. *Games and Economic Behavior*, 1(4):327–360.

Fudenberg, D. and Maskin, E. (1986). The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica*, 54(3):533–554.

Grandjean, G., Mauleon, A., and Vannetelbosch, V. (2017). Strongly Rational Sets for normal-form Games. *Economic Theory Bulletin*, 5(1):35–46.

Halpern, J. Y. and Pass, R. (2018). Game Theory with Translucent Players. *International Journal of Game Theory*, 47(3):949–976.

Harsanyi, J. C. (1974). An Equilibrium-Point Interpretation of Stable Sets and a Proposed Alternative Definition. *Management Science*, 20(11):1472–1495.

Harstad, B. (2022). A theory of pledge-and-review bargaining. *Journal of Economic Theory*, page 105574.

Herings, P. J.-J., Mauleon, A., and Vannetelbosch, V. J. (2004). Rationalizability for Social Environments. *Games and Economic Behavior*, 49(1):135–156.

Jackson, M. O. and Wilkie, S. (2005). Endogenous games and mechanisms: Side payments among players. *The Review of Economic Studies*, 72(2):543–566.

Kalai, A. T., Kalai, E., Lehrer, E., and Samet, D. (2010). A Commitment Folk Theorem. *Games and Economic Behavior*, 69(1):127–137. Special Issue In Honor of Robert Aumann.

Kalai, E. (1981). Preplay Negotiations and the Prisoner's Dilemma. *Mathematical Social Sciences*, 1(4):375–379.

Kimya, M. (2020). Equilibrium coalitional behavior. *Theoretical Economics*, 15(2):669–714.

Li, S. (2017). Obviously Strategy-Proof Mechanisms. *American Economic Review*, 107(11):3257–87.

Mariotti, M. (1997). A Model of Agreements in Strategic Form Games. *Journal of Economic Theory*, 74(1):196–217.

Nakanishi, N. (2009). Noncooperative Farsighted Stable Set in an n-player Prisoners' Dilemma. *International Journal of Game Theory*, 38(2):249–261.

Nash, J. (1953). Two-person Cooperative Games. *Econometrica: Journal of the Econometric Society*, pages 128–140.

Nishihara, K. (2022). Resolution of the N-Person Prisoners' Dilemma by Kalai's Preplay Negotiation Procedure. *Available at SSRN 4112007*.

Pearce, D. G. (1984). Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica: Journal of the Econometric Society*, pages 1029–1050.

Peters, M. and Szentes, B. (2012). Definable and Contractible Contracts. *Econometrica*, 80(1):363–411.

Rabin, M. (1994). A Model of pre-game Communication. *Journal of Economic Theory*, 63(2):370–391.

Ray, D. and Vohra, R. (1997). Equilibrium Binding Agreements. *Journal of Economic Theory*, 73(1):30–78.

Ray, D. and Vohra, R. (2015). The Farsighted Stable Set. *Econometrica*, 83(3):977–1011.

Ray, D. and Vohra, R. (2019). Maximality in the Farsighted Stable Set. *Econometrica*, 87(5):1763–1779.

Rubinstein, A. (1979). Equilibrium in Supergames with the Overtaking Criterion. *Journal of Economic Theory*, 21(1):1–9.

Rubinstein, A. (1980). Strong perfect equilibrium in supergames. *International Journal of Game Theory*, 9(1):1–12.

Rubinstein, A. (1982). Perfect Equilibrium in a Bargaining Model. *Econometrica: Journal of the Econometric Society*, pages 97–109.

Rubinstein, A. (1994). Equilibrium in Supergames. In *Essays in Game Theory*, pages 17–27. Springer.

Salcedo, B. (2017). Interdependent Choices. Technical report, University of Western Ontario.

Scarf, H. E. (1971). On the Existence of a Coopertive Solution for a General Class of N-person Games. *Journal of Economic Theory*, 3(2):169–181.

Selten, R. (1965). Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit: Teil i: Bestimmung des dynamischen preisgleichgewichts. *Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics*, (H. 2):301–324.

Selten, R. (1988). *Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games*, pages 1–31. Springer Netherlands, Dordrecht.

Shubik, M. (2012). What is a Solution to a Matrix Game. *Cowles Foundation Discussion Paper N. 1866, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2220772.*

Suzuki, A. and Muto, S. (2005). Farsighted Stability in an n-Person Prisoner's Dilemma. *International Journal of Game Theory*, 33(3):431–445.

Tennenholtz, M. (2004). Program Equilibrium. *Games and Economic Behavior*, 49(2):363–373.

Xue, L. (1998). Coalitional Stability under Perfect Foresight. *Economic Theory*, 11(3):603–627.

Yamada, A. (2003). Efficient Equilibrium Side Contracts. *Economics Bulletin*, 3(6):1–7.