Negotiated Binding Agreements*

Malachy James Gavan[†]
University of Liverpool

January, 2024
Current Version: [Here]

Abstract

I study binding agreements over play in a game. I propose a negotiation protocol where, in each round, agents propose actions from the underlying game. The protocol terminates when proposals are confirmed. I study the outcomes of Negotiated Binding Agreements of the negotiation protocol, a refinement of Subgame Perfect Equilibrium. 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. Generalisations of the main results hold.

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

1 Introduction

Negotiations and their resulting binding agreements play an important role in the economy.[1] Many such negotiations are over the behaviour that will be taken, rather the division of some abstract surplus. For instance, when prospective employees and employers negotiate, they may do so over pay but also the opportunity for flexible working, parental leave, vacation time, or other work benefits and conditions. In many cases, the non-monetary, or

[*] This paper was previously circulated under the title of “Negotiated Equilibrium”. This paper was chapter 1 of my PhD thesis, supervised by Antonio Penta, and has benefited greatly from 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] To provide just one example, in 2022 alone, the Office of United States Trade Representatives estimated that trade between U.S. and Mexico was valued at over 800 billion USD, all of which was facilitated under the binding agreement of NAFTA. The latest round of this agreement was negotiated over many years.

non-transferable, components may be the only aspect of negotiation. This would be true, for instance, if pay is fixed within a range, a common occurrence in public institutions or sectors with a strong union presence.[2] In this sense, the negotiation can be seen as over an agreement of the behaviour in game representing the provision of benefits by the employer and the acceptance of an offer by the employee, rather than a division of a surplus. Similarly, countries negotiate tariffs, quotas and regulation while committees may negotiate contributions to a public good, but may not have the ability to make direct monetary transfers due to political or legal reasons.[3] All such situations are well described as a negotiation over the behaviour that will be taken within an underlying strategic environment.

Negotiations that are not over the "split of the pie" but rather by an agreement over what to play in a game and are clearly prevalent. Despite this, providing theoretical predictions for negotiated binding agreements over such environments has proven difficult. On one hand, some works provide fully specified models of negotiation and agreements, that ensure credibility of behaviour at all stages of the negotiation (Kalai, 1981; Bhaskar, 1989; Chwe, 1994; Mariotti, 1997). However, due to the complexity that these models entail, they do not provide results that can be applied to a broad range of environments. On the other hand, there are models that provide easy-to-use conditions for what can be agreed upon in the underlying environment (Aumann, 1961; Chander and Tulkens, 1997; Currarini and Marini, 2003). However, they abstract from credibility while negotiating, allowing for behaviour that may never be agreed upon to be taken when agents do not negotiate as expected. Presently, the tension between tractability of agreement outcomes and credibility of negotiation behaviour has been difficult to resolve. To bridge this gap, I propose a negotiation protocol for agreements over what to play in the underlying game and a refinement of subgame perfect equilibrium, ensuring full credibility, while also showing that easy-to-use conditions for agreement outcomes result.

The negotiation protocol I consider regarding the behaviour players take in the underlying game takes the following form. In each period, each agent makes a proposal of the action they will take in the underlying game.[4] Agents then observe the proposals made by all others and can decide whether to "confirm" their choice or propose a new action. Confirmation is modelled by proposing the same action again. If all agents confirm their choice of action, a binding agreement is made. This binding agreement is to take the confirmed

[2] Negotiating over wages is rare in the US school system. An exception is in Wisconsin due to Act 10, which allowed individual negotiation over the employment contracts for teachers, leading to a better understanding of negotiation empirically (Baron 2018; Biasi 2021; Biasi and Sarsons 2021, 2022).

[3] Limao (2016) points to the fact that direct transfers may not be possible for international trade agreements and an increasing number of trade agreements include agreements surrounding policy, such as IP rights and foreign direct investment. For further discussion and modelling of these non-tariff barriers see Grossman et al. (2021).

[4] This is similar to the approach of Kalai (1981); Bhaskar (1989); Kimya (2020); Nishihara (2022), etc.

action profile and therefore each agent receives the payoff of said outcome. If any agent does not confirm their choice, the new proposed actions are observed and all agents make the same confirmation or new proposal choice. Agents repeat this process until confirmation, or agreement, is made by all agents. When agents never agree, or there is *perpetual disagreement*, I make a weak assumption on the payoffs that result, consistent with many interpretations, which is discussed in full in section 2.[5]

As the aforementioned negotiation protocol defines a dynamic game with complete information, to ensure agents always credibly negotiate in their own best interest, I explore a refinement of Subgame Perfect Equilibrium. I impose two main refinement criteria: firstly, I only consider Subgame Perfect Equilibria of the negotiation game that result in agreement, as the agreement outcomes are the key objects of interest of this paper. Secondly, I impose a *no babbling* condition where agents only make proposals of actions they could agree to.[6] I refer to this solution concept as *Negotiated Binding Agreements*. The agreement outcomes of the Negotiated Binding Agreements will be referred to as *supported* by a Negotiated Binding Agreement. As the negotiation game has infinitely many histories, with different types of terminal histories, this is a complex object to consider. However, I show that Negotiated Binding Agreements allow for a tractable solution, which I outline next.

Firstly, in section 3, I provide a full characterisation for the agreement outcome of Negotiated Binding Agreements in two-player games. I show that any outcome of the underlying game can be supported by a Negotiated Binding Agreement if and only if it gives each players a payoff weakly higher than an "individual punishment" profile, also defined within the underlying game. These punishments are used as "threat" agreements, where the punishment of an agent will be agreed to in the case that they do not act as expected when negotiating. The individual punishment profiles are such that (i) the payoff for any other players' punishment is weakly better than the payoff of their own punishment and (ii) when being punished, each player is prescribed to play their best response within the underlying game to their punishment profile. The logic of the result is simple. A player makes an agreement if they believe it is better than the worst agreement that could be made for them. However, the worst agreement must in it self be agreeable. Therefore there must have no reason to deviate from such an agreement, in this case by having no incentive to deviate in the underlying game.[7] This result reflects, although is more restrictive than, the

[5] For instance, it is consistent with probabilistic termination, taking this probability of termination to 0 or taking the weighted average of all proposals made. Additionally, I show that the results of the paper are consistent with a number of variations in this *baseline procedure*, including in the timing of proposals, proposing action profiles, and in the payoff of perpetual disagreement, studied in the online appendix.

[6] The no babbling assumption can embed a form of no delay equilibrium used within bargaining games with a large number of players (Chatterjee et al., 1993), imposing that all proposed divisions of surplus can be agreed to.

[7] Note that this logic is distinct from that of the solution to the Rubinstein (1982) bargaining game, where

player-specific punishments used in the literature of infinitely repeated games, for example in Fudenberg and Maskin (1986) and Abreu et al. (1994) - which in it self can be seen as an agreement to play a strategy based on the threat of future punishment. Additionally, this provides a link to the Commitment Folk Theorems in the literature on contractable contracts (Peters and Szentes, 2012). Nonetheless, the characterisation of agreement outcomes can be substantially more restrictive than both in a number of games. As one of the most canonical examples in economics, I explore a leading example of a Cournot Duopoly with linear demand and heterogeneous marginal costs to illustrate the key ideas of the proof, as well as to demonstrate the key difference to the aforementioned models in a well understood environment.

As for games of n-players, in section 4, I show that the characterisation of agreement outcomes for two-player games can be used as a sufficient condition for n-player games. I show that a necessary condition for any action profile supported or proposed in a Negotiated Binding Agreement must survive *iterated elimination of individually irrational actions* of the underlying game, which I introduce in this paper. Specifically, an action a is *individually irrational* if, given the most optimistic beliefs the agent can have when evaluating it, the payoff it induces is still strictly worse than the minimum payoff that an agent can receive from best responding to some action profile of others. Performing this process iteratively, deleting *all* individually irrational actions within a round before moving to the next, results in actions that survive iterated elimination of individually irrational actions. I show that the minimum payoff that an agent can receive from a Negotiated Binding Agreement outcome is always weakly higher than the worst best response payoff in the underlying game, taken over the set of actions that survives iterated elimination of individually irrational actions. The conditions for deletion and calculating the minimum payoff are simple to implement in any underlying game that is finite or with smooth utility functions. Further, I show that in an important class of n-player games the characterisation is tight.

To illustrate the logic of this necessary result, I use an underlying game of a simple three-bidder First Price Auction with heterogeneous valuations. In this case, in any profile of bids supported by a Negotiated Binding Agreement the bidder with the highest valuation must receive the good with positive probability. However, in comparison to the Nash equilibria of this underlying game, it is possible that *any* bidder receives the good with positive probability, at many different prices. Nonetheless, these possibilities are restricted by the minimal payoffs that bidders must receive, and therefore it is not the case that any outcome is possible.

Negotiated Binding Agreements as a solution concept only contemplates unilateral devi-

instead agents fear that lack of acceptance will lead to a discounted future agreement or no agreement at all. In this case, there is no cost of delay, and therefore no possibility of discounted future agreement.

ations, but we may also be interested in the possibility of agents making binding agreements over *how* they will negotiate. To allow for this, I extend the solution concept to allow for cooperative agreements within the very negotiation game. I do so by introducing coalitions of agents to jointly choose a new strategy and will do so if it is profitable for all agents within the coalition. This may include permissible coalitions that overlap.

To capture the possibility of agents acting in such a way within the negotiation procedure, in section 5, I define the concept of $\mathcal{C}$-Negotiated Binding Agreement. Here, no coalition in a predefined set $\mathcal{C}$ can profitably deviate at any history and a no babbling condition is imposed. I show that the natural extension of the baseline necessary and sufficient conditions for agreement outcomes in n-player games hold. Within $\mathcal{C}$-Negotiated Binding Agreement, players only make proposals from the set of actions that survives iterated elimination of *coalitionally* irrational actions in the underlying game, defined similarly to individually irrational actions taking coalition-wide preferences into account. Further, for all permissible coalition the outcome of negotiation must satisfy a notion of *coalitional rationality* in the underlying game. These conditions can be viewed as a perturbed version of the cooperative game theoretic notion of the β-core (Aumann, 1961).[8] I provide sufficient conditions of the outcomes of the underlying game that can be supported using coalition-specific punishments; a further refined version of the β-core.

Finally, in section 6, I expand on the relations to these and other works within the literature review. I conclude the paper in section 7, pointing to a number of directions for future work.

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 the action 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

[8] The β-core allows any outcome that is better than the worse-case scenario for any group to be agreed upon, even if these worst-case scenarios make use of non-credible behaviour that could never be agreed upon.

round and the same process occurs until confirmation is made by all agents, leading to an agreement. If there are infinitely many periods without agreement, I refer to this as *perpetual disagreement*.[9]

Formally, let the set of partial histories consists 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.[10]

A history is terminal if, either:

1. the same action profile is proposed twice 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 $U_i : Z \rightarrow \mathbb{R}$ denote the payoff for player $i \in N$ of the negotiation game.

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.[11] 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 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, if the probability of continuation is taken 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.[12] This is formalised by the following lemma, and the proof is provided in the appendix.

[9] 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.

[10] See the online appendix for the extension of non-simultaneous proposals.

[11] See the online appendix for alternative specifications.

[12] Similar notions also exist in the context of Rubinstein (1982) bargaining, where Busch and Wen (1995)

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, leading to confirmation, 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. Therefore, one interpretation of the payoff of perpetual disagreement is that an there is an ϵ probability of each player mistakenly pressing accept, and such ϵ is taken to 0. This specification may also embed, for example, 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 is modelled.

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 and those without. 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.).

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 $i \in N$

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.

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 $U_i(s|h) = U_i(z')$ where $z' \in Z'$ is defined as before and $U_i(s|h) = U_i(z'')$, 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 $a_i(z) = a_i^k$.

2.1. Solution Concept

This negotiation protocol defines a dynamic game with complete information therefore Subgame Perfect Equilibrium (SPE) is well defined.

Definition (Subgame Perfect Equilibrium)**.** *s^* is Subgame Perfect Equilibrium, if for all partial histories $h \in H$, for all $i \in N$, $U_i(s^*|h) \geq U_i(s_i, s^*_{-i}|h)$, for all $s_i \in S_i$.*

Due to the structure of the negotiation protocol, in any SPE agents must receive a payoff weakly higher than their inf-sup payoff in the underlying game. This is true for any history. This is formalised by the following lemma.

Lemma 2. *For any Subgame Perfect Equilibrium s^*, for any partial history $h \in H$*

$$U_i(s^*|h) \geq \inf_{a_{-i} \in A_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$$

As I do not require that the utility functions are continuous and actions are from a compact set, the minimum or maximum need not exist. However, whenever the underlying game being considered has well defined maxima and minima I will refer to them as such.

Note that the set of SPE of the negotiation game trivially includes many perpetual disagreement outcomes. As this work is primarily focused on the agreements that can be supported by some equilibrium, I focus on the SPE of the negotiation game that reaches an agreement from the initial history. I will also restrict attention to SPE where proposals only involve actions that can be part of an agreement. This rules out agents proposing actions that they would never agree to on or off the path of play. I will refer to such property as agreements having *no babbling*. Similar concepts have been used within the literature on bargaining. For instance, no delay equilibrium of Chatterjee et al. (1993), where the proposals can only be made, at any history, if they could be accepted. I will refer to this concept as *Negotiated Binding Agreement*. Formally:

Definition 1 (Negotiated Binding Agreement)**.** *s^* is a Negotiated Binding Agreement if:*

1. *s^* is a Subgame Perfect Equilibrium of the negotiation game.*
2. *No babbling: $\forall h \in H$, $\exists h' \in H$ such that $s_i^*(h) = a_i(s^*|h')$.*

a^* is supported by s^* if $a^* = a(s^*|\emptyset)$.

Note that the set of actions that may be supported by a Negotiated Binding Agreement does not change if the no babbling condition were to be defined as only making proposals that could be agreed to in *some* Negotiated Binding Agreement, rather than no babbling requiring proposed actions must be agreed to in the Negotiated Binding Agreement being considered.

3 Negotiated Binding Agreement Outcomes for Two-Player Games

In this section, I provide a full characterisation of the Negotiated Binding Agreement outcomes 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 that player 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 agreement outcomes for the Negotiated Binding Agreements can be completely characterised purely with information of the underlying game, in a simple way. 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 supporting (q_1^*, q_2^*) such that $\pi_1(q_1^*, q_2^*) \geq 0$ and $\pi_2(q_1^*, q_2^*) \geq (c_1 - c_2)^2$ in a Negotiated Binding Agreement. 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 to do so. 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 Negotiated Binding Agreement is to have each firm propose to flood the market as much as possible whenever the other firm is not acting as expected when negotiating, while maintaining a positive profit, understanding the other agent 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, when firm 2 has not negotiated as expected.

To see this is a Negotiated Binding Agreement, notice that no babbling applies as all three rules are absorbing. Therefore all that is left is to check that s^* is a Subgame Perfect Equilibrium of the negotiation game.

First, let us consider firm 1. Firstly, consider the strategy in 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 statically, and therefore they cannot improve upon the current strategy. Now let us consider a deviation of firm 1 from the punishment of firm 2. Under the current strategy and rule, no deviation will lead to an agreement of $\underline{q}^2$. In which case, firm 1 receives a profit of 0. However, a deviation 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, consider a deviation from any other history. No deviation will lead to an agreement that yields weakly positive profit, either via rule 2 or 3. Again, any deviation can only lead to firm 2 proposing $\underline{q}_2^1$ for all subsequent periods. Given this, it must be that the payoff of said deviation is again at most 0, and therefore cannot be profitable.

Now instead consider firm 2. Firstly, consider the first case above, and consider whether there could be a profitable deviation from punishing firm 1. 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, nor case 3, as this would lead to an agreement for q^*, which provides them with a higher payoff than $(c_1 - c_2)^2$.

Now we will study why the Negotiated Binding Agreement outcome must be necessarily better than that of the outlined punishment $\underline{q}^i$. Firstly, notice that due to both firms being restricted to only making proposals that they can agree to, in any Negotiated Binding Agreement $s^{*,\prime}$, at any history $s^{*,\prime}(h) \in Q^*$, where Q^* is some set of agreement outcome quantities. Now notice that in any Negotiated Binding Agreement $s^{*,\prime}$ 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.[13] As it is the case that a) the $q_{-i}^* \in Q_{-i}^*$ 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 q^* such that $q_i^* \in \arg\max_{q_i \in Q_i} \pi_i(q_i, q_{-i}^*)$ must be an agreement outcome. Further, it pins down the lowest agreement payoff for i. Further, 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 later agreement as the former. 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 may be supported under Negotiated Binding Agreements.

Note that the construction provides us with some natural comparative statics and comparison to player specific punishments of Fudenberg and Maskin (1986); Abreu et al. (1994)

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

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 outcomes 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 space is restricted to reflect the additional bargaining power firm 2 has due to firm 1 not proposing or agreeing to outcomes known to be bad for themself. In the most extreme case, when firm 1 is no longer a viable competitor, when $c_1 = \frac{b+c_2}{2}$, we conclude that firm 2's profit is $\pi_2(\underline{q}^2) = \left(\frac{b-c_2}{2}\right)^2$, their monopoly profit.[14]

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 supported by a Negotiated Binding Agreement 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.

4 Negotiated Binding Agreement Outcomes 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. However, it is no longer necessary, as the no babbling condition does not impose strong coordination on the action *profile* proposed by players after a deviation. Therefore it is not possible to look for a punishment that is best responded to with strong conditions on coordination. However, when strong conditions on coordination of agreement

[14] 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.

outcome *profiles*, i.e. not only does a player have to propose an action thy would agree to, but the profile of actions proposed by any set of agents is such that they would jointly agree, then this 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 agreement outcomes to be supported by a Negotiated Binding Agreement.

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 supported by a Negotiated Binding Agreement 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^* can be supported in a Negotiated Binding Agreement.

4.1.1. A Refinement of Negotiated Binding Agreements where outcomes are Fully Characterised

Further justification for the general sufficient conditions can be found. For a refinement of Negotiated Binding Agreements, where the focus is upon SPE that end in immediate agreement following from each history, the sufficient conditions for agreement outcomes are also necessary for underlying games where the action space is a compact subset of a metric space and utility is continuous. 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 Negotiated Agreements and is similar to the no delay equilibrium proposed by Chatterjee et al. (1993). Therefore, for the class of No Delay Negotiated Binding Agreements, I fully characterise the set of outcomes that can be supported.

Definition 2 (No Delay Negotiated Binding Agreement)**.** *s^* is a No Delay Negotiated Binding Agreement 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 Negotiated Binding Agreement, 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 $i, j \in N$*

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 no babbling 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 outcomes, I turn to the following example to preview the logic.

4.2. Preview of Necessary Conditions n-player games

Here the underlying game G is 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, $b_i \in \{0, 1, .., 7\}$. 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 value minus their bid. 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? 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, there 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. This is because the payoff can only be pinned down by losing the auction, or by winning at their valuation, leading to 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 $b_i = 7$ is supported by a Negotiated Binding Agreement, as such a strategy cannot be a Subgame Perfect Equilibrium of the negotiation game. Further, by no babbling, it cannot be that proposing to bid 7 occurs in *any* Negotiated Binding Agreement.

Now consider whether it is the case that bidders 2 or 3 could agree to bid 6 in a Negotiated Binding Agreement. By the 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 a Negotiated Binding Agreement, as it would not be a Subgame Perfect Equilibrium of the negotiation game. Further, by the no babbling condition, we conclude that in no Negotiated Binding Agreement can bidding 6 *ever* be proposed by bidders 2 and 3.

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 must provide bidders 2 and 3 with a payoff of at least 0. Also notice in any Negotiated Binding Agreement it must be that bidder 1 receives a payoff of at least 1/2. To see this, notice that 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.

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

4.3. Necessary Conditions

Within this section, I characterise a number of necessary conditions for a Negotiated Binding Agreement outcomes and strategies for n-player games. First, I will show that any action proposed at some history of Negotiated Binding Agreement 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 chosen, 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, in the sense of inf-sup utility, where the inf is taken *over the set of actions that survives iterated individual rationality*, then it does not survive iterated deletion of individually irrational actions. The formal definition of individual irrational actions and iterated deletion of individually irrational actions are formally defined below.

Definition 3 (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.[15] Therefore the set that survives elimination of individually irrational actions is smaller. Note that, 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 4 (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 $\tilde{A}^{m-1}_{-i} = \times_{j \neq i} \tilde{A}^{m-1}_j$.*

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

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

Given these definitions, we can present the first necessary condition of Negotiated Binding Agreement 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 a Negotiated Binding Agreement, then for all $h \in H$, $s_i^*(h) \in IIR_i$.*

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.[16] This is formalised in the following definition and lemma.

Definition 5. *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 3. *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.

D is strictly dominant for both players, hence (D, D) is the only profile that survives survive iterated elimination of never best responses to pure actions. Yet, in IIR, all action

[16] 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.

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

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, shows the relation to Negotiated Binding Agreement payoffs with individual rationality considerations in the underlying game, when taken over the set of actions that survive iterated elimination of individually irrational actions.

Theorem 4. *if s^* is a Negotiated Binding Agreement 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 $i \in N$.

I illustrate the use of this result with the same underlying prisoner's 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 no Negotiated Binding Agreement can support the action profile (D, C) or (C, D). 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^* can be supported by 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 can be supported by 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, a negotiation may be susceptible to a collection of agents making binding agreements over how they will act *within* the negotiation process itself. This is particularly important given that the negotiation protocol does not generically lead to a unique and efficient outcome.[17] To address the concern of susceptibility to groups agreeing to deviate, I now extend the analysis to allow for this possibility. To do so, I include 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.

[17] Many negotiation and bargaining protocols do not lead to efficiency. The 2-player Rubinstein (1979) model does not when cost of time is constant, rather than hyperbolic. In the hyperbolic discounting case, when the outside option of this model is taken to be endogenous a la Busch and Wen (1995), a folk theorem is obtained. When there are more than two-players, additional restrictions on the equilibrium notion are needed to regain efficiency (Chatterjee et al., 1993). 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. 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 setting (Jackson and Wilkie, 2005; Yamada, 2003; Ellingsen and Paltseva, 2016).

In this analysis, 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. This can be seen as stronger than necessary, as we may wish for the deviations to face the same criticism of stability, where these deviations must be the result of some agreement.[18] However, if it were possible to make a binding agreement to not make new binding agreements, agents may take this option upon deviating. Therefore, in the context of robust binding agreements, if we do not wish to make assumptions surrounding the game that is induced to negotiate over when a deviation occurs then this approach ensures no misspecification. That is, do we allow for agents within a coalition to have veto power? Do we allow agents to make agreements over what can be within the agreement in the sense that they pre-commit to rule out some options? This can potentially allow for different conclusions in the outcome of the negotiation game. Nonetheless, if all deviations of a coalition are permitted, this includes the outcomes of processes, and therefore if we have an equilibrium that allows for all possible deviations we certainly have an equilibrium when all such deviations are not allowed. In this sense, the aim of this analysis differs from the previous sections in that I will provide easy-to-check conditions for a robust Negotiated Binding Agreement, where no specified coalition could deviate, rather than searching for all possible Negotiated Binding Agreements that could occur when coalitions can deviate in a specific and specified way.

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

[18] This would renegotiation proofness a la Farrell and Maskin (1989); Bernheim and Ray (1989).

improve the utility of all players within that coalition.[19]

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). Whenever $\mathcal{C} = 2^N \backslash \{\emptyset\}$ 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 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.

Although specified for any coalition configuration, I will take $\{i\}_{i \in N} \subseteq \mathcal{C}$ as implicit within the discussion, although it is not necessary for the formal results. I will also pay particular attention to the grand coalition being permitted; $N \in \mathcal{C}$.

We can now introduce $\mathcal{C}$-Negotiated Binding Agreements, where we will require that we have a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game and require a no babbling condition. Note that the use of $\mathcal{C}$-Subgame Perfect Equilibria of the negotiation game when $N \in \mathcal{C}$, gives further justification for no babbling agreements, and indeed no delay agreements. To see this, suppose that there was some $\epsilon > 0$ cost for delay for all agents. If this were the case, then there would be no $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game that ends in more than two periods as a joint deviation could reduce the cost of delay.

Definition 6 ($\mathcal{C}$-Negotiated Binding Agreement)**.** *s^* is a $\mathcal{C}$-Negotiated Binding Agreement if:*

1. *s^* is a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game*
2. *$\mathcal{C}$-no babbling: $\forall h \in H$, $\exists h' \in H$ such that $s^*_C(h) = a_C(s^*|h')$.*

a^* is supporting by s^* if $a^* = a(s^*|\emptyset)$.

When $\mathcal{C} = 2^N \backslash \{\emptyset\}$ I refer to this as a strong Negotiated Binding Agreement. Whenever $\{i\}_{i \in N} \subset \mathcal{C}$, $\mathcal{C}$-Negotiated Binding Agreement are a subset of Negotiated Binding Agreement

[19] 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.

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 any action proposed in the negotiation game 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.[20]

Definition 7. *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 8 (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

[20] 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.

the following example.

Example 2. Consider the following 2 player game to be the underlying game G. 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. However, 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 $ICIR(\mathcal{C}) = \emptyset$. ▼

However, it 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 2 player game as the underlying game, G. 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 rules 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 4. *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 proposals made in $\mathcal{C}$-Negotiated Binding Agreement of the negotiation game.

Theorem 5. *For any $\mathcal{C}$-Negotiated Binding 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 9. *$a^* \in A$ 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	**6,6**	0,4	**1,$\underline{12}$**	0,0
T	4,0	0,0	**7,$\underline{2}$**	$\underline{1}$,1
D	**$\underline{12}$,1**	**$\underline{2}$,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 4, 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.[21] 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.

[21] 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.

6 Literature Review

A number of papers have approached the question of binding agreements that 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 a refinement of SPE in a fully specified negotiation game. 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, 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 typical refinements and variations of subgame perfect equilibrium (stationary subgame perfection and local perfection / trembling hand), 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.

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 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.

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

I propose a model of negotiated binding agreements over agents' play in an underlying normal form game. I study the outcomes of the underlying game that be supported using a refinement of Subgame Perfect Equilibrium, where agents only propose actions that they could agree to. I refer to this concept as Negotiated Binding Agreements. I show that the outcomes of the underlying game that can be agreed upon must satisfy a condition of *iterative* individual rationality. Further, any outcome in the underlying game, where appropriate individual punishments can be found, can be agreed to. These individual punishments are defined on the underlying game, where agents must be prescribed the action that best responds to their punishment in the baseline game. The sufficient condition for agreement outcomes is also shown to be necessary for two-player games, leading to a full characterisation within this class. By providing conditions for outcomes that can be agreed upon that are solely based on characteristics of the underlying game, I reconcile the rigour of the solution of a fully specified model of negotiation with easy-to-use conditions for agreement outcomes for the underlying game.

To display the ease of use of these conditions, I explore two key applications. In a Cournot Duopoly, I show that when marginal costs are the same, any profile of payoffs such that each player receives positive profits is sustainable. In contrast, when marginal costs are very different only the firm with the lowest marginal cost receiving their monopoly profit is supported. In a simple First Price Auction, I show that these conditions lead to intuitive results about what can be agreed upon, where a minimal bound is put on the payoff the highest valuation bidder. In these applications, I fully characterise the Negotiated Binding Agreement outcomes.

I show how the necessary and sufficient conditions for the outcomes of the Negotiated Binding Agreements within this negotiation game naturally generalise to the case where agents may agree upon *how* to negotiate. I show these generalised conditions are linked to a perturbed version of the cooperative game theoretic notion of the β-core of Aumann (1961), while having the full backing of a fully specified negotiation procedure.

A number of questions remain open. Firstly, there are a number of applied theory questions that can use the results of this paper. A number of applied theory papers have made use of cooperative solutions, for example in environmental agreements (Chander and Tulkens, 1997; Carraro, 1998; Carraro et al., 2006) and trade agreements (Aghion et al., 2007; Conconi and Perroni, 2002). Due to the easy-to-use conditions, my results may also provide some interesting insights in some applied theoretical settings, while having the backing of a fully specified negotiation protocol.

Additionally, the results of this paper may shed light on which environments should be negotiated jointly, that is, when is bundling issues or games in negotiation beneficial.[22] This is particularly interesting from the applied theory perspective. For instance, international trade agreements involve simultaneously negotiating tariffs for multiple markets and, for instance, environmental policy.[23] However this is not always the case and therefore understanding when it is theoretically beneficial is an interesting line to follow.[24] Further, the results of this paper may provide an understanding of when there is a benefit from unilaterally giving up some actions in the underlying game, essentially allowing agents to "take chips off the table". The results of this paper show that unilaterally giving up an action *can* be beneficial. Nonetheless, understanding the removal of *which* actions leads to this improvement is an open question. I leave these questions for future work.

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

[22] Bloch and De Clippel (2010) studies the core of cooperative games and characterise for which cooperative games is the core of the sum of those games the same as the sum of the cores.

[23] When these issues are independent, it is trivial to show that the set of payoffs sustainable is weakly larger, however, when there is interdependence between these games the relation is unclear.

[24] Conconi and Perroni (2002) considers this question for international trade, but do so via a coalition formation procedure, a la Ray and Vohra (1997). Such procedures are fragile to changes in the games and definitions (see, example 1 of Gavan (2022)) and therefore it is difficult to make broad conclusions, whereas the results of my paper may allow for a better understanding within classes of games.

Proof of lemma 2: Suppose not, $U_i(s^*|h) < \inf_{a_{-i}\in A_{-i}} \sup_{a_i\in A_i} u_i(a_i, a_{-i})$. For any $\epsilon > 0$, let $\tilde{a}_i : A_{-i} \to A_i$ be such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \sup_{a_i\in A_i} u_i(a_i, a_{-i}) - \epsilon$. Note such a function exists for any $\epsilon > 0$. Let $s'_i(h) = (\tilde{a}_i(s^*_{-i}(h')), s^*_{-i}(h'))$ for all $h' \in H$. It follows that $U_i(s'_i, s^*_{-i}|h)$ is either such that it ends in agreement, in which case $U_i(s'_i, s^*_{-i}|h) > \inf_{a_{-i}\in A_{-i}} \sup_{a_i\in A_i} u_i(a_i, a_{-i}) - \epsilon$ and therefore, as we can construct such a function for any $\epsilon > 0$, we conclude that $U_i(s'_i, s^*_{-i}|h) \geq \inf_{a_{-i}\in A_{-i}} \sup_{a_i\in A_i} u_i(a_i, a_{-i})$. On the other hand, it may be that $U_i(s'_i, s^*_{-i}|h)$ ends in perpetual disagreement. In which case $(s'_i, s^*_{-i}|h) = (a^1, a^2, ..., a^T, ...)$, where $a^t_i = \tilde{a}_i(a^t_{-i})$. Therefore:

$$U_i(s'_i, s^*_{-i}|h) \geq \liminf_{t\to\infty} u_i(\tilde{a}_i(a^t_{-i}), a^t_{-i}) \geq \inf_{a_{-i}\in A_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \inf_{a_{-i}\in A_{-i}} \sup_{a_i\in A_{-i}} u_i(a_i, a_{-i}) - \epsilon$$

$\Rightarrow U_i(s'_i, s^*_{-i}|h) \geq \inf_{a_{-i}\in A_{-i}} \sup_{a_i\in A_{-i}} u_i(a_i, a_{-i})$. A contradiction. ■

Proof of theorem 1 and proposition 1: By theorem 2 such a^* can be supported and therefore that section of the proof is omitted.

To see that only such a^* can be sustained, take any a^* such that it is supported by a no delay Negotiated Binding Agreement in the n-player case and a Negotiated Binding Agreement 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 no delay these completely define the set of actions that can be agreed upon in the n-player case by no delay. Further to this, not that $s^*_{-i}(h) \in \tilde{A}_{-i}$ for all $h \in H$ by no delay and in the two-player case by no babbling. 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 2: 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 $a^{k-1}_{-j} = s^*_{-j}((a^1, ..., a^{k-2}))$ and either:

   (a) $a^k_l = s^*_l(h \backslash a^{k-1}) \quad \forall l \neq j$ while $a^k_j \neq s^*_j(h \backslash a^{k-1})$.

   (b) or $a^k_{-j} = \underline{a}^j_{-j}$.

   then $s^*_i(h) = \underline{a}^j_i$.

2. $s^*_i(h) = a^*_i$ otherwise.

First note that from any history the continuation is terminal within two periods and therefore no babbling 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. ■

Proof of theorem 5: Suppose not, for some history $h' \in H$ we have that $s_i(h') = a_i$. By no babbling 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 no babbling, we conclude that $s_i(h) \notin D_i(A_{-i})$ for any $h \in H$.

Now suppose by contradiction that, for all $j \in N$ $s_j(h') \in \tilde{A}_j^k$ $\forall k < m$ and $h' \in H$ but for some $i \in N$ $s_j(h') = a_i \notin \tilde{A}_j^{m+1}$ for some $h' \in H$. By no babbling it must be that a) $s_{-i}(h') \in \tilde{A}_i^m$ for all h' and b) by no babbling there is some $h \in H$ for which $a_i(s|h) = a_i$. Therefore it must be that $U_i(s|h) = u_i(a(s|h)) \leq \sup_{a'_{-i} \in \tilde{A}_{-i}^m} u_i(a_i, a'_{-i})$. Take $\epsilon = \inf_{a'_{-i} \in \tilde{A}_{-i}^m} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - u_i(a(s|h)) > 0$. Take a function $\tilde{a}_i : \tilde{A}_{-i}^m \rightarrow A_i$ 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 $a_{-i} \in \tilde{A}_{-i}^m$. Consider a deviation $s'_i(h'') = \tilde{a}_i(s_{-i}(h''))$ for all $h'' \in H$. It follows that: $U_i(s'_i, s_{-i}|h) \geq \inf_{a_{-i} \in \tilde{A}_{-i}^m} u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \inf_{a_{-i} \in \tilde{A}_{-i}^m} \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$. 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 no babbling, we conclude that $s_i(h) \notin D_i(\tilde{A}_{-i}^m)$ for any $h \in H$ and therefore $s_i(h) \in \tilde{A}_i^{k+1}$, a contradiction.■

Proof of lemma 3: 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 $a_i \in B_i^k$ we have that there is some $a_{-i} \in B_{-i}^{k-1} \subseteq \tilde{A}_{-i}^{k-1}$ such $u_i(a_i, a_{-i}) \geq u_i(a'_i, a_{-i})$ for all $a'_i \in A_i$. It follows that $u_i(a_i, a_{-i}) \geq \inf_{a'_{-i} \in B_{-i}^{k-1}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) \geq \inf_{a'_{-i} \in \tilde{A}_{-i}^{k-1}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i})$. Further,

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

and therefore we conclude that if $a_i \in B_i^k$ then $a_i \in \tilde{A}_i^k$, 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 $\tilde{a}_i : IIR_{-i} \rightarrow A_i$ 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:

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

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

Proof of lemma 4: 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 C no babbling 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 $\mathcal{C}$-no babbling 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 the condition of no babbling agreements and therefore we have a $\mathcal{C}$-Negotiated Binding Agreement.■

References

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

Aghion, P., Antràs, P., and Helpman, E. (2007). Negotiating Free Trade. *Journal of International Economics*, 73(1):1–30.

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.

Baron, E. J. (2018). The Effect of Teachers' Unions on Student Achievement in the Short Run: Evidence from Wisconsin's Act 10. *Economics of Education Review*, 67:40–57.

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.

Biasi, B. (2021). The Labor Market for Teachers under Different Pay Schemes. *American Economic Journal: Economic Policy*, 13(3):63–102.

Biasi, B. and Sarsons, H. (2021). Information, Confidence, and the Gender Gap in Bargaining. *AEA Papers and Proceedings*, 111:174–78.

Biasi, B. and Sarsons, H. (2022). Flexible Wages, Bargaining, and the Gender Gap. *The Quarterly Journal of Economics*, 137(1):215–266.

Bloch, F. and De Clippel, G. (2010). Cores of combined games. *Journal of Economic Theory*, 145(6):2424–2434.

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

Carraro, C. (1998). Beyond Kyoto: A Game-Theoretic Perspective. In *the Proceedings of the OECD Workshop on "Climate Change and Economic Modelling. Background Analysis for the Kyoto Protocol", Paris*, pages 17–18. Citeseer.

Carraro, C., Eyckmans, J., and Finus, M. (2006). Optimal Transfers and Participation Decisions in International Environmental Agreements. *The Review of International Organizations*, 1(4):379–396.

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.

Conconi, P. and Perroni, C. (2002). Issue linkage and issue tie-in in multilateral negotiations. *Journal of international Economics*, 57(2):423–447.

Currarini, S. and Marini, M. (2003). A Sequential Approach to the Characteristic Function and the Core in Games with Externalities. In *Advances in Economic Design*, pages 233–249. Springer.

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.

Gavan, M. J. (2022). Weak Coalitional Equilibrium: Existence and Overlapping Coaitions. *Working Paper*.

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

Grossman, G. M., McCalman, P., and Staiger, R. W. (2021). The "new" economics of trade agreements: From trade liberalization to regulatory convergence? *Econometrica*, 89(1):215–249.

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.

Limao, N. (2016). Preferential Trade Agreements. In *Handbook of Commercial Policy*, volume 2, pages 281 – 360. Elsevier.

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.

B Online Appendix

B.1. Appendix: Robustness

In this section, I outline how the results of this paper are robust to changes in how the negotiation game is defined. I do so as follows. In subsection B.1.1. I show that necessarily proposals can only be made from actions that survive iterated elimination of absolutely dominated actions, which are tightly related to those that survive iterated elimination of individually irrational actions, and the sufficient conditions for agreement outcomes hold if agents make proposals sequentially rather than simultaneously in each period. In subsection B.1.2. I show that, if the payoffs of the infinite histories are appropriately defined, both the necessary and sufficient conditions for agreement outcomes hold if agents may make proposals of the joint action, rather than just their own, in each period. In subsection B.1.3., I show that the sufficient conditions for agreement outcomes remain to be true in a model where the payoff of the infinitely terminal histories are taken to be worse than the payoff of any finite terminal history.

In essence, these robustness checks show how the drivers of the results. Specifically, that agents cannot use a non-agreement outcome as a threat of deviating, whereas timing and the proposals used are not an important for driving the results.

B.1.1. Robustness to Order of Proposals

As in section 2, let G be an underlying game with bounded payoffs.

Define the negotiation game with order as follows.

Let $\mathcal{O} : N \to |N|$ be the order in which agents make proposals within a period. Note that this function may not be one-to-one, and therefore it may be that many agents make the proposals at the same time. Assume that if $\mathcal{O}(i) = k > 1$ then $\exists j \in N$ such that $\mathcal{O}(j) = k - 1$. That is, $\mathcal{O}$ naturally defines an order: if I am not first, then there must be someone who proposes before me. I also assume that $\mathcal{O}(i) = 1$ for some $i \in N$ to ensure the first proposer is labelled as such. Let $\mathcal{O}^{-1}(k) = \{i \in N | \mathcal{O}(i) = k\}$, that is, define $\mathcal{O}^{-1}(k)$ is the set of agents who make the k^{th} proposal.

A history will be the empty set or a sequence of proposals for all agents followed by the first k proposals within the last period. That is,

$$h = (a^1, a^2, ..., a^{k-1}, (a^k_{\mathcal{O}^{-1}(2)}, a^k_{\mathcal{O}^{-1}(1)}, ..., a^k_{\mathcal{O}^{-1}(l)}))$$

, with $l \leq n$, i.e. there may be agents who are yet to make a proposal within the current period.

A history is terminal if, either:

1. Where the same action profile is proposed twice in consecutive periods, all agents have made a proposal within the last period, 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 $\tilde{Z}'$ and refer to these histories as ones where an *agreement* is made.

2. an infinite sequence where the same action profile is never proposed consecutively, and all agents have made a proposal within each period. Let the set of such histories be denoted by $\tilde{Z}''$. I will again refer to these as perpetual disagreement histories.

Let the set of terminal histories be given by $\tilde{Z} = \tilde{Z}' \cup \tilde{Z}''$. The set of all possible histories is all terminal histories and all finite histories where there are no consecutive proposals that are the same action for all agents. Let the set of partial histories be denoted by $\tilde{H}$.

As before, whenever $z = (a^1, ..., a^k) \in \tilde{Z}'$ let $U_i(z) = u_i(a^k)$.

Whenever $z \in \tilde{Z}''$ let $U_i(z) \in [\liminf_{t\to\infty} u_i(a^t), \limsup_{t\to\infty} u_i(a^t)]$. Only take these definitions over well-defined action profiles.

Let $\tilde{H}_i$ be the set of partial histories where $i \in N$ is active. That is $h \in \tilde{H}_i$ is such that $h = (a^1, a^2, ..., a^{k-1}, (a^k_{\mathcal{O}^{-1}(1)}, ..., a^k_{\mathcal{O}(i)-1}))$ when $\mathcal{O}(i) \neq 1$ and $h = (a^1, a^2, ..., a^{k-1}, a^k)$. the strategy of $i \in N$ dictates the proposal i would make at any history for which they are active: $s_i : \tilde{H}_i \to A_i$. Let S_i be the space of all such mappings.

For a partial history $h \in \tilde{H}$, let $U_i(s|h)$ denote the payoff that would be received from the terminal history that the strategy s would induce, starting from the history $h \in \tilde{H}$. I will refer to such a history as $(s|h)$. When $z \in \tilde{Z}'$, i.e. an agreement is made, let $a(h)$ as the action profile that terminates z.

I define Subgame Perfect Equilibrium for this model here:

Definition (Subgame Perfect Equilibrium)**.** *s^* is Subgame Perfect Equilibrium, if for all $i \in N$, for all partial histories where $i \in N$ is active $h \in \tilde{H}_i$, $U_i(s^*|h) \geq U_i(s_i, s^*_{-i}|h)$, for all $s_i \in S_i$.*

This leads to the natural definition of Negotiated Binding Agreement in this setting. To make the distinction clear, I refer to this as Negotiated Binding Agreement with order.

Definition 10 (Negotiated Binding Agreement with Order)**.** *s^* is a Negotiated Binding Agreement with order $\mathcal{O}$ supporting $a^* = a * (s^*|\emptyset)$ if:*

1. s^* is a Subgame Perfect Equilibrium.

2. For all $h \in \tilde{H}_i$ $\exists h' \in \tilde{H}_i$ such that $s_i(h) = a_i(s^*|h')$.

Now I show that some necessary conditions related in section 4 remain to be true for this specification of the model. First, I will show that any action proposed at some history of Negotiated Binding Agreement must survive a procedure of iterated deletion of absolutely dominated actions, also known as interdependent choice rationalizability (Salcedo, 2017) and min-max rationalizability (Halpern and Pass, 2018).

Definition 11 (Absolute Domination given $C_{-i} \subseteq A_{-i}$)**.** *$a_i \in A_i$ is absolutely dominated given $C_{-i} \subseteq A_{-i}$ if $\exists a_i' \in A_i$ such that*

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

Denote the set of absolutely dominated actions given C_{-i} by $D_i(C_{-i})$.

As I do not require that the utility functions are continuous and defined over a compact set, the minimum or maximum may not exist. With this, I take the supremum and infimum, which by the assumption that the utility function is bounded are always well-defined. Bar this change, the above definition is equivalent to that of Salcedo (2017). Note that, 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).

Definition 12 (Iterated Elimination of Absolutely Dominated Actions)**.** *Let $\tilde{A}_i^0 = A_i$ for all $i \in N$. Let $\tilde{A}_{-i}^0 = A_{-i}$. Then for all $m > 0$ let $\tilde{A}_i^m = \tilde{A}_i^{m-1} \backslash D_i(\tilde{A}_{-i}^{m-1})$ where $\tilde{A}_{-i}^{m-1} = \times_{j \neq i} \tilde{A}_j^{m-1}$.*

The set of actions that survives Iterated Elimination of Absolutely Dominated Actions (IAD) for i is given by $IAD_i = \bigcap_{m \geq 0} \tilde{A}_i^m$. Let $IAD = \times_{i \in N} IAD_i$.

Note that if at each level of iteration if the min-max and max-min payoffs are the same then IAD coincides with IIR. Note that typically, the concept of iterated elimination of individually irrational actions and iterated elimination of absolutely dominated actions are different, for instance, consider the following example.

Example 5. Consider the following underlying two-player game.

$1\backslash 2$	L	R
U	$1, 2$	$-1, 0.5$
M	$-1, 1$	$1, 0.5$
D	$-0.7, 3$	$-0.7, 3$

Here, in iterated elimination of absolutely dominated actions, all profiles survive. However, if we consider iterated elimination of individually irrational actions, we may remove

D, as the min-max payoff for player 1 is 1. Given this, we may also eliminate R for player 2, as her min-max payoff is 0.5. Finally, we remove M, therefore we conclude that iterated elimination of individually rational actions leads to the unique prediction of (U, L), while iterated elimination of absolutely dominated actions allows for any action profile. ▼

These definitions lead to the following proposition; any proposal made on or off the path in a Negotiated Binding Agreement with order must come from the set of actions that survives iterated elimination of absolutely dominated actions.

Proposition 2. *For any order $\mathcal{O}$, if s^* is a Negotiated Binding Agreement with order then, for all histories where i is active $h \in \tilde{H}_i$, $s_i(h) \in IAD_i$.*

I reserve this proof, and all other proofs within this appendix, for the appendix B.2..

Further to this, the following proposition shows that the sufficient conditions for agreement outcomes are relevant within this specification of the model. Indeed, further to this, any outcome that can be sustained with a Negotiated Binding Agreement can be sustained within a model of negotiation with order, no matter the order. This is highlighted by the following proposition.

Proposition 3. *Take any order $\mathcal{O}$. If a^* is supported in a Negotiated Binding Agreement then it is supported in Negotiated Binding Agreement with order $\mathcal{O}$.*

In essence, this shows that the qualitative results of having agreements be based on player specific punishments and all agreement outcomes having to satisfy some iterative individual rationality constraint are robust to having sequential proposals. Rather, the structure of the terminal histories, and the associated payoffs, as well as the ability of all agents to make some proposal, are the key features of the model. Within the next sub-appendix, I go on to show that when the payoffs of infinite histories are correctly specified, the robustness of these results also holds when agents propose the action profile, rather than only their action. This further highlights this point.

B.1.2. Robustness to Joint Proposals

As in section 2, let G be an underlying game with bounded payoffs.

Define the negotiation game with all proposals as follows.

A history will be the empty set or a sequence of proposals for all agents, where each agent may propose a joint action profile. That is,

$$h = ((a^{1,1}, a^{2,1}, ..., a^{n,1}), (a^{1,2}, a^{2,2}, ..., a^{n,2}), ..., (a^{1,k}, a^{2,k}, ..., a^{n,k}))$$

, where $a^{i,t} \in A$. With some abuse of notation, let $a^t = (a^{1,t}, a^{2,t}, ..., a^{n,t})$.

A history is terminal if, either:

1. Where the same action profile is proposed twice in consecutive periods by all agents and no earlier occurrence of consecutive repetition is present. That is, $h = (a^1, ..., a^{k-1}, a^k)$ is terminal if $a^k = a^{k-1}$, $a^{i,k} = a^{j,k}$ for all $i, j \in N$, and either $a^m \neq a^{m-1}$ for all $m < k$ or $a^{i,m} \neq a^{j,m}$ for some $i, j \in N$. Let the set of such histories be denoted by $\tilde{\tilde{Z}}'$ and refer to these histories as ones where an *agreement* is made.

2. an infinite sequence where the same action profile for all agents is never proposed consecutively. Let the set of such histories be denoted by $\tilde{\tilde{Z}}''$. Refer to these as perpetual disagreement histories.

Let the set of terminal histories be given by $\tilde{\tilde{Z}} = \tilde{\tilde{Z}}' \cup \tilde{\tilde{Z}}''$. The set of all possible histories is all terminal histories and all finite histories where there are no consecutive proposals that are the same action profile for all agents. Let the set of all partial histories given by $\tilde{\tilde{H}}$.

Whenever $z = (a^1, ..., a^k, ...) \in \tilde{\tilde{Z}}'$, let

$$\tilde{h} = ((a_i^{i,1})_{i \in N}, (a_i^{i,2})_{i \in N}, ..., (a_i^{i,k})_{i \in N}, ...)$$

, i.e., take the proposals that each agent makes for themselves. Let this sequence be denoted by $\tilde{z} = (\tilde{a}^1, \tilde{a}^2, ..., \tilde{a}^k, ...)$ Let the lower bound of the utility of $z = (a^1, ..., a^k, ...) \in Z'$ be given by $\liminf_{t \to \infty} u_i(\tilde{a}^t)$ and an upper bound of the $\limsup_{t \to \infty} u_i(\tilde{a}^t)$. This implies that if no agreement is made, then only your own proposals matter, you cannot impact what others do in this case.

the strategy of $i \in N$ dictates the proposal i would make when they are active: $s_i : \tilde{\tilde{H}} \to A$. Let S_i be the space of all such mappings.

With some abuse of notation, for a partial history $h \in \tilde{\tilde{H}}$, let $U_i(s|h)$ denote the payoff that would be received from the terminal history that the strategy s would induce, starting from the history $h \in \tilde{\tilde{H}}$. I will again refer to such a history as $(s|h)$. As before, when $z \in Z'$, i.e. an agreement is made, let $a(h)$ as the action profile that terminates z.

Definition (Subgame Perfect Equilibrium)**.** *s^* is Subgame Perfect Equilibrium, if for all $i \in N$, for all partial histories $h \in \tilde{\tilde{H}}$, for all $i \in N$, $U_i(s^*|h) \geq U_i(s_i, s^*_{-i}|h)$, for all $s_i \in S_i$.*

This leads to the definition of Negotiated Binding Agreement in this setting. To make the distinction clear, I refer to this as Negotiated Binding Agreement with all proposals.

Definition 13 (Negotiated Binding Agreement with all Proposals)**.** *s^* is a Negotiated Binding Agreement with all proposals supporting $a^* = a(s|\emptyset)$ if:*

1. s^* is a Subgame Perfect Equilibrium.

2. $\forall h \in \tilde{\tilde{H}}\ \exists h' \in \tilde{\tilde{H}}$ *such that* $s_i(h) = a(s^*|h)$.

As before, the following proposition shows that the necessary conditions previously shown for Negotiated Binding Agreement hold for this specification of the model.

Proposition 4. *If* s^* *is a Negotiated Binding Agreement with all proposals, for all histories* $h \in \tilde{\tilde{H}}$, $s_i(h) \in IIR_i$.

Further, for any negotiated with order s^* *be be such that, for any history* $h \in \tilde{\tilde{H}}$, $U_i(s^*|h) \geq \underline{u}_i$ *where*

$$\underline{u}_i = \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$$

Further to this, the sufficient conditions for Negotiated Binding Agreement outcomes are also sufficient for Negotiated Binding Agreements with all Proposals. This is captured by the following proposition, which shows us that any Negotiated Binding Agreement can be replicated by a Negotiated Binding Agreement with all proposals.

Proposition 5. a^* *is supported by a Negotiated Binding Agreement with all proposals if* a^* *is supported by a Negotiated Binding Agreement.*

This again highlights the important features and drivers of the results of the model. In essence, it is the ability of agents to make a meaningful impact on their payoff via their proposals, while ensuring they do not force other agents to take some action. This is highlighted by the idea that the payoffs of infinite terminal histories, i.e. when there is no agreement, take the actions for individuals that they propose for themselves.

B.1.3. Robustness to Outside Options

Within this sub-appendix, I take the model to be exactly as in section 2. That is, agents simultaneously propose the action that they will take. The only caveat is that whenever a terminal history is infinite they receive a payoff that is worse that the payoff within the underlying game. That is, when $z \in Z''$ let $U_i(z) = \inf_{a \in A} u_i(a)$. Negotiated Binding Agreement can be defined as before. To distinguish between these cases I will refer to Negotiated Binding Agreement for the model in this sub-appendix as *constant outside option Negotiated Binding Agreement*. In this setting, it is no longer true that the necessary conditions for agreement outcomes hold. However, the sufficient conditions for agreement outcomes remain to be valid. This is highlighted by the following proposition.

Proposition 6. *If* s^* *is a Negotiated Binding Agreement then* s^* *is a constant outside option Negotiated Binding Agreement.*

As Negotiated Binding Agreements do need not make use of the infinitely long terminal histories as part of equilibrium, this result shows us that they are important only for restricting deviations. That is, if we were to make such an option worse for each player, they have less incentive to deviate than before. Therefore Negotiated Binding Agreement captures a set of strategies and outcomes that work regardless of whether the outside option is specified as within this chapter or normalised to be worse than any outcome as typically assumed in bargaining games.

B.2. Proofs for Appendix B.1.

Proof of proposition 2: By induction. Firstly, note that $s_i(h) = a_i \notin D_i(A_{-i})$ for all $h \in \tilde{H}_i$. To see this suppose by contradiction it is not the case. Then $s^*_i(h) = a_i \in D_i(A_{-i})$ for some $i \in N$ and some history $h \in \tilde{H}_i$. It must be that $a_i(s^*|h) = a_i$ for some $h' \in H$. Given this, $U_i(s^*|h) \leq \sup_{a_{-i}\in A_{-i}} u_i(a_i, a_{-i})$. Further, s^* is a Subgame Perfect Equilibrium of the negotiation game, and therefore there is no profitable deviation for s^*_i at any history for which i is active, including h'. Notice that as $a_i \in D_i(A_{-i})$ then $\exists a'_i \in A_i$ such that $\inf_{a'_{-i}\in A_{-i}} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}\in A_{-i}} u_i(a_i, a_{-i})$. Now consider a strategy s'_i such that $s'_i(h'') = a'_i$ for all h'' for which i is active. Notice that, by construction of s'_i, the history $(s'_i, s^*_{-i}|h')$ must either terminate in a'_i or be such that only action profiles with a'_i appear after h. In either case, we can conclude that $U_i(s'_i, s^*_{-i}|h') \geq \inf_{a'_{-i}\in A_{-i}} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}\in A_{-i}} u_i(a_i, a_{-i}) \geq U_i(s^*|h')$. A contradiction to s^* be a Subgame Perfect Equilibrium of the negotiation game.

By the inductive hypothesis, suppose that $s^*_i(h) \in \tilde{A}^m_i$ for all $h \in \tilde{H}_i$ and $i \in N$. Now suppose by contradiction that $s^*_i(h) = a_i \in D_i(\tilde{A}^m_{-i})$. It must be that $a_i(s^*|h') = a_i$. Given this, $U_i(s^*|h') \leq \sup_{a_{-i}\in A^m_{-i}} u_i(a_i, a_{-i})$, as $s^*_{-i}(h'') \in \tilde{A}^m_{-i}$ for all $h'' \in \tilde{H}_i$. Further, s^* is a Subgame Perfect Equilibrium of the negotiation game, and therefore there is no profitable deviation for s^*_i at any history, including h'. Notice that as $a_i \in D_i(\tilde{A}^m_{-i})$ then $\exists a'_i \in A_i$ such that $\inf_{a'_{-i}\in \tilde{A}^m_{-i}} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}\in \tilde{A}^m_{-i}} u_i(a_i, a_{-i})$. Now consider a strategy s'_i such that $s'_i(h'') = a'_i$ for all $h'' \in \tilde{H}_i$. Notice that, by definition and construction of s'_i $U_i(s'_i, s_{-i}|h')$ must only be constructed using the utility of $u_i(a'_i, \cdot)$, as either $(s'_i, s_{-i}|h') \in Z'$, in which case it must terminate in a'_i by definition, or $(s'_i, s_{-i}|h') \in Z''$, in which case all histories following h' use only a'_i. Further, as $s^*_{-i}(h'') \in \tilde{A}^m_{-i}$ that from this history on the only action profiles proposed are a'_i, a'_{-i} such that $a'_{-i} \in \tilde{A}^m_{-i}$. Given this, we can conclude that $U_i(s'_i, s_{-i}|h') \geq \inf_{a'_{-i}\in \tilde{A}^m_{-i}} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}\in \tilde{A}^m_{-i}} u_i(a_i, a_{-i}) \geq U_i(s^*|h')$. A contradiction to s^* being a Subgame Perfect Equilibrium of the negotiation game. ■

Proof of proposition 3: We will show that if a^* is sustained in a Negotiated Binding Agreement then it can be sustained in a Negotiated Binding Agreement with order $\mathcal{O}$ for any order. Take any order $\mathcal{O}$. Take s^* that sustains a^* in a Negotiated Binding Agreement. Let $s'_i : \tilde{H}_i \to A_i$ such that, for all $h \in \tilde{H}_i$ such that $h = (h', (a_{\mathcal{O}^{-1}(1)}, ..., a_{\mathcal{O}^{-1}(i)-1}))$ we have that $s'_i(h) = s^*_i(h')$. First note that $a(s'|\emptyset) = a^*$ and $a(s'|h') = a(s^*|h)$ whenever $h' = h$ while $h' \in \tilde{H}$ and $h \in H$. Next we will show that s' is subgame perfect of the negotiation game. Suppose not, there is some $i \in N$ for which there exists some $h \in H'_i$ and some $s''_i \in S_i$ such that $U_i(s''_i, s'_{-i}|h) > U_i(s'|h)$. However, given agents are rational and the structure of s', they can replicate any deviation from s'_i with a deviation from s^*_i. With this, we must conclude that s^*_i is not subgame perfect of the negotiation game. A contradiction. Concluding that s' is a Negotiated Binding Agreement with order $\mathcal{O}$, leading to the outcome a^*. ■

Proof of proposition 4: By induction. Firstly, note that $s_i(h) = a \notin D_i(A)$ for all $h \in \tilde{\tilde{H}}$. Suppose by contradiction it is the case. Then $s^*_i(h) = a \in D(A)$ for some $i \in N$ and some history $h \in \tilde{\tilde{H}}$. It must be that $a(s^*|h') = a$ for some history $h' \in H$. This implies that $[s^*_i(h')]_j = a_j \in D_j(A_{-j})$ for some j. Given this, $U_j(s^*|h) \leq \sup_{a_{-j} \in A_{-j}} u_j(a_j, a_{-j})$. Further, s^* is a Subgame Perfect Equilibrium of the negotiation game, and therefore there is no profitable deviation for s^*_j at any history, including h. Notice that as $a_j \in D_j(A_{-j})$ then, for all $\epsilon > 0$ $\exists a'_j : A_{-i} \to A_j$ such that $u_i(a'_j(a_{-i}), a'_{-j}) > \sup_{a_{-j} \in A_{-j}} u_i(a_j, a_{-j}) - \epsilon$. Now consider a strategy s'_j such that $s'_j(h'') = (a'_j(s_{-j}(h'')), a''_{-j})$, for some $a''_{-j} \in A_{-j}$ for all h''. Notice that, by construction of s'_i, the history $(s'_j, s^*_{-j}|h')$ must either terminate in a'_j or be such that only action profiles with a'_j appear after h'. In either case, we can conclude that $U_j(s'_j, s^*_{-j}|h') \geq \inf_{a'_{-j} \in A_{-j}} u_i(a'_j(a'_{-j}), a'_{-j}) > \sup_{a_{-j} \in A_{-j}} u_i(a_j, a_{-j}) \geq U_j(s^*|h')$. A contradiction to s^* be a Subgame Perfect Equilibrium of the negotiation game.

By the inductive hypothesis, suppose that $s^*_i(h) \in \tilde{A}^m$ for all $h \in \tilde{H}_i$ and $i \in N$. Now suppose by contradiction that $s^*_i(h) = a \in D(\tilde{A}^m)$. It must be that $a(s^*|h') = a$ for some history $h' \in H$. Further, for some $j \in N$ $a_j \in D(\tilde{A}^m)$. Without loss of generality let $j = i$. Given this, $U_i(s^*|h') \leq \sup_{a_{-i} \in A^m_{-i}} u_i(a_i, a_{-i})$, as $s^*_{-i}(h'') \in \times_{j \neq i} \tilde{A}^m$ for all $h'' \in \tilde{\tilde{H}}$. Further, s^* is a Subgame Perfect Equilibrium of the negotiation game, and therefore there is no profitable deviation for s^*_i at any history, including h. Notice that as $a_j \in D_j(A_{-j})$ then, for all $\epsilon > 0$ $\exists a'_j : \tilde{A}^m_{-i} \to A_j$ such that $u_i(a'_j(a_{-i}), a'_{-j}) > \sup_{a_{-j} \in \tilde{A}^m_{-j}} u_i(a_j, a_{-j}) - \epsilon$. Now consider a strategy s'_i such that $s'_i(h'') = (a'_i(s^*_{-i}(h'')), a''_{-i})$, with $a''_{-i} \notin A_{-i}$ for all $h'' \in \tilde{H}_i$. Notice that, by definition and construction of s'_i $U_i(s'_i, s^*_{-i}|h')$ must only be constructed using the utility of $u_i(a'_i, \cdot)$, as with the before logic, we can only terminate in histories that have a'_i infinitely repeated or an agreement is reached with a'_i. Given this, we can conclude that $U_i(s'_i, s^*_{-i}|h') \geq \inf_{a'_{-i} \in \tilde{A}^m_{-i}} u_i(a'_i(a'_{-i}), a'_{-i}) > \sup_{a_{-i} \in \tilde{A}^m_{-i}} u_i(a_i, a_{-i}) \geq U_i(s^*|h')$. A contradiction to s^* be a Subgame Perfect Equilibrium of the negotiation game.

As proposals are simultaneous, the logic of showing that for any negotiated with order

s^* be be such that, for any history $h \in \tilde{H}$, $U_i(s^*|h) \geq \underline{u}_i$ where

$$\underline{u}_i = \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$$

is identical to theorem 4, where s'_i is selected to intentionally cause perpetual disagreement. ■

Proof of proposition 5: If a^* is supported by a Negotiated Binding Agreement then a^* is supported by a all proposal Negotiated Binding Agreement. Take s^* that supports a^* in a Negotiated Binding Agreement. Construct $s'_i : \tilde{\tilde{H}} \to A$ as follows. Let $s'_i(h'') = s^*(\tilde{h''})$, where $\tilde{h''}$ is as defined to define payoffs of infinite histories. Clearly if s^*_i is optimal so is s'_i as a deviation to a partial infinite history leads to the same payoff that could be achieved under s^*_{-i}. A deviation to another terminal history must be such that it could not be achieved under a deviation from s^*_i. However, by definition of s'_i, this cannot be the case. ■

Proof of proposition 6: As s^* is a Negotiated Binding Agreement it must be that s^* is a Subgame Perfect Equilibrium of the negotiation game with the terminal infinite histories giving a payment as defined in section 2. As s^* never dictates that a history should be infinite and terminal, it follows that there is no profitable deviation where the outcome leads to a deterministic outcome. It follows that the payoff on the path remains the same when the model of a constant outside option is taken. Finally, as there is no profitable deviation when the deviation would induce a terminal infinite history when the payoff is defined as in section 2, there cannot be a profitable deviation when the constant outside option is taken. Therefore s^* is a constant outside option Negotiated Binding Agreement. ■