Negotiated Binding Agreements*

Malachy James Gavan†
Universitat Pompeu Fabra and Barcelona School of Economics

December 3, 2022

[Job Market Paper]
Current Version: [Here]

Abstract

I study the binding agreements that may result from players negotiating over their behaviour in an underlying strategic environment. To do so, I propose a negotiation protocol where, in each round of negotiation, agents make public proposals of the action they will take in the underlying game. The protocol terminates when these proposals are confirmed. Confirmation results in a binding agreement over the action profile and payoffs are the corresponding ones in the underlying game. I study the outcomes of Negotiated Binding Agreements of the negotiation protocol, which is a refinement of Subgame Perfect Equilibrium that I introduce in this context to obtain both credibility and tractability. The main results show that any outcome of the underlying game that is agreed upon must satisfy an *iterative* individual rationality constraint. Additionally, an outcome of the underlying game can be agreed upon if appropriate individual punishments in the underlying game can be found. A full characterisation is provided for two-player games. Finally, to allow for the possibility that agents make binding agreements over how they will negotiate, I extend the solution concept to allow for cooperative agreements within the negotiation game. Generalisations of the main results hold and refine the set of agreement outcomes.

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

*This paper was previously circulated under the title of "Negotiated Equilibrium". This paper has benefited greatly from numerous suggestions and comments from colleagues. I pay particular thanks to Antonio Penta for his ongoing supervision and support throughout my Ph.D.. I have also had innumerable useful discussions with Larbi Alaoui, Alexander Frug and Pia Ennuschat, for which I am greatly indebted. 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 seminar participants at UPF, the BSE PhD jamboree, the 12th Conference on Economic Design, the International Conference on Game Theory and Applications, the 2022 Conference on Mechanism and Institution Design, the 33rd Stony Brook International Conference on Game Theory, and the Asian School in Economic Theory for a number of useful comments and suggestions. All faults are my own.

†*email:* malachy.gavan@upf.edu.

1 Introduction

Negotiations and their resulting binding agreements play an important role within the economy.[1] Many such negotiations take place over multiple dimensions of a problem. 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 some cases, the non-monetary 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] Given this, the negotiation is over agents' behaviour in an the underlying strategic environment, such as the provision of benefits by the employer and the acceptance of an offer by the employee. Similarly, negotiation also occurs when buying a house, where the timing of the move and renovations needed may be key. Countries negotiate tariffs and quotas and committees may negotiate contributions to a public good, both of which are well represented by a negotiation over underlying strategic behaviour.

Despite their empirical relevance, understanding the theoretical predictions of outcomes of negotiated binding agreements over strategic behaviour has proven challenging. 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 and tractable 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 punishments that may never be agreed upon 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.

In this paper, I bridge this gap between credibility of negotiation and tractability of agreement outcomes. I do so by proposing a negotiation protocol for binding agreements over behaviour in an arbitrary underlying strategic environment, or game. I study the outcomes of *Negotiated Binding Agreements* of the negotiation protocol, a refinement of Subgame Perfect Equilibrium to ensure that agreement outcomes are the result of agents negotiating credibly. I show that for a general class of settings, the outcome that may be sustained by Negotiated Binding Agreements of the negotiation must satisfy an *iterative* individual rationality constraint in the underlying game. Additionally, an outcome can be agreed to if it provides each agent with a payoff higher than received by an "individual punishment" profile in the underlying game. These individual punishments are such that, relative to the target agreement, the cost of punishing others is smaller than the cost of being punished. Further, each agent must best respond to their punishment profile in the underlying game.[3] In this sense, I reconcile the rigour of Subgame Perfection in the underlying non-cooperative negotiation game with easy-to-check conditions for agreement outcomes in the underlying game.

[1] In 2021 alone the National Association of Realtors estimated 6.12 million existing homes were sold at a median price over $350,000 in the US. All such sales involve a binding agreement and are negotiated in some respect.

[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] This condition is similar, but distinct, from the conditions of player-specific punishment in infinitely repeated games (Fudenberg and Maskin, 1986; Abreu et al., 1994) which is discussed at greater length within the paper.

A number of applications are used to show how the results lead to intuitive agreement outcomes. For instance, taking the underlying game to be a public goods game, I show that an agent can agree to contribute if and only if they are sufficiently "compensated" by others' contributions. When the underlying game is a Cournot Duopoly and quantities are negotiated over, the possible agreement outcomes depend on the difference between firms' marginal costs.[4] When marginal costs are the same all outcomes that provide both firms with non-negative profits can be agreed to. This is because a firm can credibly agree to flood the other firm out of the market, given the other could do the same, leading to 0 profits in either case. With this, a firm flooding the market can be used as the punishment profile for the other, allowing any target that provide a weakly higher profit than this to be agreed upon. However, outside of the case when marginal costs are the same the inefficient firm cannot credibly agree to flooding the efficient firm out of the market, as this would lead to a loss for the inefficient firm. Therefore there is a minimal bound on the profits that the efficient firm can receive from any agreement, as only limited punishments can be used against them, restricting the outcomes that can be agreed to. When marginal costs are extremely different only the efficient firm receiving their monopoly profits can be agreed to.

The negotiation protocol I consider regarding the behaviour players should take in the underlying game takes the following form. In each period, agents make a public proposal of the action they will take in the underlying game. They continue making proposals in this form until there is "confirmation" of their choices, modelled by the same proposal profile being made for two consecutive periods.[5] At this point, a binding agreement to play this action profile, or outcome, is made and each agent receives the payoff of said outcome. When agents do not agree, or there is *perpetual disagreement*, I make a weak assumption on the payoffs that result. consistent with many interpretations, which are discussed in full in section 2.[6]

As the aforementioned negotiation protocol defines a dynamic game with complete information, I explore a refinement of Subgame Perfect Equilibrium to ensure agents always credibly negotiate in their own best interest. 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.[7] 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, 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*

[4]Beyond an agreement between two firms, this environment could represent two countries deciding on quotas in a trade agreement or co-owned firms making an agreement.

[5]Other methods of locking in their proposals, if simultaneous and all respected, would lead to the same results.

[6]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 Appendix A.

[7]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.

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.

Secondly, I provide sufficient conditions for the outcomes supported by a Negotiated Binding Agreement. I show that any outcome in the underlying game that gives all players a payoff weakly higher than their "individual punishment" profile can be supported by a Negotiated Binding Agreement. 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. This is similar, although more restrictive, to *player-specific punishments* used in the literature of infinitely repeated games, for example in Fudenberg and Maskin (1986) and Abreu et al. (1994), which will be discussed further within the paper. These sufficient conditions imply that any action profile that Pareto dominates a pure Nash equilibrium in the underlying game can always be supported in a Negotiated Binding Agreement. In two-player games, this sufficient condition is also necessary, leading to a full characterisation in this class.

I explore three key applications. The first, where the underlying game is a simple three-bidder First Price Auction with heterogeneous valuations, is used as a leading example to display the key results of the paper and provide the intuition behind them. 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. The second application, studied in section 4, takes the underlying game to be a public goods game. Here I show that an agent can contribute if and only if a minimal level of aggregate contribution is reached, permitting both full contribution or no contribution at all. With this, an agent would only agree to contribute if they are sufficiently compensated by the contributions of others. The third application, also in section 4, considers a Cournot Duopoly with linear demand and heterogeneous marginal costs. I show that when marginal costs are the same, any profile of payoffs that gives both players positive profits is supportable. In contrast, when marginal costs are extremely different, only the efficient firm receiving their monopoly profit can be supported. These results are intuitive and display the ease of use of the conditions.

Negotiated Binding Agreements as a solution concept only contemplates unilateral deviations, 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. The novelty of permitting coalitions to overlap can offer new insights into environments when agents may be members of multiple groups simultaneously. For example, within international trade and politics, groups with international agreements regularly overlap. This occurs when country A is in a coalition with country B, country B is in a coalition with country C, but there is no coalition which contains countries A and C.[8] Such arrangements of groups frequently occur in economic environments but are not typically considered in the literature.[9]

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. In section 6, I show that the natural extension of the baseline necessary and sufficient conditions for agreement outcomes 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).[10] 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. In a simple Cournot model with fixed costs, I show that all the β-core outcomes can be sustained in $\mathcal{C}$-Negotiated Binding Agreement, while having the backing of a fully specified negotiation procedure, which is not the case for the β-core itself.

By providing tractable results for outcomes of a given underlying game that are based on a refinement of Subgame Perfect Equilibrium of a fully specified model of negotiation, I contribute to a number of important strands of literature on the theory of binding agreements over strategic environments. The first strand that this paper contributes is where papers that provide easy-to-check conditions on the underlying game for agreements. Existing work in this strand focuses on what can be sustained by taking a cooperative game theory approach. In that literature, reactions to a deviation from a candidate agreement are not credible in the sense they do not have to satisfy the same conditions as an agreement itself (Aumann, 1959, 1961; Chander and Tulkens, 1997; Currarini and Marini, 2003). For instance, these concepts may allow for any punishment to be used to prevent deviations, even if such a punishment may *never* be agreed to. Alternatively, they may take away the ability for groups to act jointly in reaction to a deviation, while they may act jointly in a deviation. This in essence requires that after a deviation the reaction may not be agreed upon in the same sense as the original notion. These concepts take a cooperative view that abstracts from why or how agents would act in such a way. Making such

[8] For instance, Australia is in The Regional Comprehensive Economic Partnership (RCEP), which can be seen as a coalition in multinational negotiations. RCEP also includes members of the Association of Southeast Asian Nations, which Australia is not a member of.

[9] As far as I am aware, all results within the literature are based on the assumption of no overlapping coalitions. One exception to this, providing results when allowing for overlapping coalitions, is Gavan (2022), where I provide an existence result for an equilibrium concept that allows for overlapping coalitions.

[10] 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.

assumptions allows for a much larger degree of tractability, making it simple for an analyst to evaluate what can be sustained in the underlying game with such notions. Nonetheless, they can make unreasonable predictions for agreements in some applications, due to allowing potentially non-credible threats.[11] I contribute to this literature by ensuring that the conditions I provide are backed by a fully specified model of negotiation where the solution concept ensures credibility at all stages. In a second strand, there are approaches that provide a negotiation procedure and companion solution concept where agents believe that others will act credibly within their own best interest, with consistent tools at all stages of the negotiation (Kalai, 1981; Bhaskar, 1989; Mariotti, 1997; Ray and Vohra, 2019). These concepts are fully specified and have well defined behaviour to back them, but often intractable due to the richness of behaviour that these procedures allow for. As a result, they fail to provide easy-to-check conditions for a general underlying game. I contribute to this strand of literature by providing easy-to-check conditions for agreement outcomes for any underlying game. In section 7, I expand on the relations to these and other works within the literature review. I conclude the paper in section 8, 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 make no further restrictions on the underlying game. In particular, A need not be finite nor compact, nor u_i be continuous.

I now define the *negotiation game* over G. There will be potentially infinitely many periods to reach an agreement, 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. If all agents make the same proposal that they made in the previous period, that action profile is implemented in a binding way. If not, they continue to the next round and continue the same process until an agreement is made. If there are infinitely many periods without agreement, I refer to this as *perpetual disagreement*.

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.[12]

A history is terminal if, either:

a) 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*.

[11] Scarf (1971) provides an early observation of this issue in reference to Aumann (1961)'s α-core, pointing to the potential unreasonable use of any punishment to prevent deviations.

[12] See appendix A for this extension.

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

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

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

Whenever there is 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, it is assumed that the payoff is defined to be between the lim inf and lim sup of the utility in the underlying game of the proposals made.[13] 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 ensures the following properties hold:

1. For any $z' \in Z'$ and $z = (a^1, a^2, ..., a^k, ...) \in Z''$ if $\exists T \in \mathbb{N}$ such that $\inf_{a^t, t>T} u_i(a^t) > U_i(z')$ then $U_i(z) > U_i(z')$.
2. For any $z' \in Z'$ and $z = (a^1, a^2, ..., a^k, ...) \in Z''$ if $\exists T \in \mathbb{N}$ such that $\sup_{a^t, t>T} u_i(a^t) \leq U_i(z')$ then $U_i(z) \leq U_i(z')$.
3. $U_i(z)$ is bounded for any history, and is bounded by the same bounds as u_i in G.

This restriction is consistent a the standard 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 current proposed actions.[14] This is formalised by the following lemma.

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

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

Proof: (see appendix B).

More generally, the assumed payoffs of agents are based on proposals that are sufficiently far along the history. This is consistent with the interpretation that an outside party decides the proposals, which must be taken from sufficiently far along the path of proposals, that will be

[13] See Appendix A for alternative specifications.

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

implemented, in a way that is known to the agents. Within finite underlying games G it can also be interpreted as taking any weighted average of the proposals made infinitely often along the entire path. 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).

By taking a 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.

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 changes. 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. 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 $i \in N$ 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, rather than using the infimum and supremum.

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:*

a) s^* is a Subgame Perfect Equilibrium of the negotiation game.

b) 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.

Further motivation can be found behind ensuring that the strategy to reach agreement has no babbling, as it rules out the proposals that are payoff irrelevant *both on and off* the path of play. These would be understood to be payoff irrelevant by agents due to complete information and correct beliefs of equilibrium. To see the use for this, it may be that an SPE may induce proposals that are not used for the purpose of agreement at *any* history, even in the event an agreement is eventually reached. For instance, consider the following example.

Example 1. Let the underlying game, G, be the following:

1\2	L	C	R
U	2, 3	0, 1	1, 1
D	3, 2	1, 1	1, 0

The only Nash equilibrium of the underlying game is (D, L). Notice that the individual min-max for player 1 in the underlying game is given by 1, while the individual min-max for player 2 in the underlying game is given by 2. By lemma 2, players cannot receive below their min-max payoff in the negotiation game. Consider the payoff of perpetual disagreement to be taken to be as defined in lemma 1.

Now, consider the following SPE of the negotiation game.

$$s_2^*(h) = \begin{cases} R & \text{if } h = (a^1, ..., a^k), a^k = (D, C) \text{ or } a^k = (D, L) \\ C & \text{if } h = (a^1, ..., a^k), a^k = (U, R) \\ L & \text{otherwise} \end{cases} \qquad s_1^*(h) = U, \quad \forall h \in H$$

On the path of play, starting from the initial history, (U, L) is proposed within the first two periods and the negotiation game terminates. On the other hand, if the history is such that (D, C) has been proposed in the previous period, $t-1$, then (U, R) is proposed in period t, (U, C) in period $t+1$, (U, L) in period $t+2$ and the negotiation game terminates in period $t+3$ as (U, L) is proposed again. We can see that this strategy always leads to terminal histories that end in (U, L). Note that there is no profitable deviation for any history, and therefore this is a SPE of the negotiation game. Clearly player 2 cannot improve their utility, as L is the best response to U in the baseline game, and player 1 only makes proposals of U. On the other hand, player 1 cannot profitably deviate as they cannot receive a payoff higher than 2 given the strategy of player 2. To see this, notice that due to the strategy of 2, it is only possible for the negotiation game to terminate in an agreement with (U, L). Now consider any strategy of 1. In order to be profitable, it must be that player 2 plays L sufficiently often, as this (D, L) is the only outcome that provides a higher payoff. However, by the strategy of player 2 the play cannot terminate in (D, L), therefore the only possibility of a profitable deviation is to induce perpetual disagreement where (D, L) is proposed frequently enough. In order for it to be perpetual disagreement, and given the strategy of 2 it must be C and R are played at least as frequently as (D, L). Therefore this leads to a payoff of at most 2. Therefore it cannot be that utility is improved.

However, within this negotiation game, from a history h such that $s^*(h) = (U, C)$, it is clear the strategy is not directly used for any agreement itself, both on and off the path, and therefore we do not have a Negotiated Binding Agreement. ▼

2.2. Leading Example and Preview of Results

Here I provide a leading example to illustrate the key ideas in the paper, where 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, $b_i \in \{0, 1, .., 7\}$.[15] 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,

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

multiplied by their value minus their bid. Formally,

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

The Nash equilibria of this underlying game are such that either a) $b^* = (5, 5, b_3^*)$ with $b_3^* \leq 4$ leading to payoffs of $(1/2, 0, 0)$, b) $b^* = (5, b_2^*, b_3^*)$ with $\max\{b_2^*, b_3^*\} = 4$ with payoffs of $(1, 0, 0)$ or c) $b^* = (4, 4, b_3^*)$ with $b_3^* \leq 3$, leading to payoffs of $(1, 1/2, 0)$. Notice those within a) and b) make use of weakly dominated actions. The lowest payoffs in any Nash equilibria are $(1/2, 0, 0)$, with $b^* = (5, 5, b_3^*)$.

First, we will understand what cannot be supported by a Negotiated Binding Agreement. 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.

I will now argue that *any* such profile b^* with $u_1(b^*) \geq 1/2$ and $u_2(b^*), u_3(b^*) \geq 0$ is supported in a Negotiated Binding Agreement. Suppose that the strategies are such that b_i^* is proposed in the first period of the negotiation game. In the second round, if all players propose b_i^* in the first round, they propose b_i^* once again, leading to a binding agreement to play b^*. That is, $s_i^*(\emptyset) = b_i^*$ and $s_i^*(b^*) = b_i^*$. In all other cases, let each player propose their part of the worst Nash equilibrium of the game, $(5, 5, 4)$, for all possible periods. That is, for all histories $h \in H$ such that $h \neq \emptyset$ and $h \neq (b^*)$, let $s_1^*(h) = s_2^*(h) = 5$ and $s_3^*(h) = 4$. To see there is no incentive to deviate in the negotiation game, notice that any deviation at any point leads to the others playing their part of the worst Nash equilibrium. Given that each bidders' payoff of the negotiation game will be defined with respect to this, the best possible deviation would lead to a payoff no higher than their static best response to this profile. By the definition of Nash equilibrium, it cannot be that a player can improve their payoff from any history, as either it is the first two periods where they receive a payoff weakly higher than this or it is not, and they receive exactly this payoff by not deviating.

In conclusion, we have that any vector of bids b^* can be supported by a Negotiated Binding Agreement if and only if $u_1(b^*) \geq \frac{1}{2}$, $u_2(b^*), u_3^*(b^*) \geq 0$. This can include all bidders receiving the good with positive probability, but necessarily bidder 1 must receive the good with positive probability.

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

3 Negotiated Binding Agreement Outcomes

3.1. Necessary Conditions

Within this section, I characterise a number of necessary conditions for a Negotiated Binding Agreement outcomes and strategies. 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. 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 2 (Individually Irrational actions given $C_{-i} \subseteq A_{-i}$)**.** *For a game G, $a_i \in A_i$ is individually irrational given $C_{-i} \subseteq A_{-i}$ if*

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

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

This notion is similar to the notion of absolute dominance by Salcedo (2017), simultaneously developed in Halpern and Pass (2018), who instead compare the best case of one action and the worst case of another, whereas I compare based on the best case of an action compared to the inf-sup.[16] Therefore the set that survives elimination of individually irrational actions is smaller, as if an action is obviously dominated it is also individually irrational. 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 3 (Iterated Deletion of Individually Irrational Actions)**.** *For a game G, 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 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, which states that any proposal must survive iterated elimination of individually irrational actions in the underlying game.

Theorem 1. *If s^* is a Negotiated Binding Agreement, then for all $h \in H$, $s_i^*(h) \in IIR_i$.*

Notice that this applies for all histories, be that on or off the path of play. Therefore any proposal being made must have survived iterated elimination of individually irrational actions. Notice that this is the exact process and result that was used in order to find the proposals that could occur within the leading example, resulting in no proposal including a bid of 6 or 7 for any bidder and bidder 3 not proposing a bid of 5.

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. This is formalised in the following definition and lemma.

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

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

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. Further, 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 2. Let the underlying game, G, be the following prisoners' dilemma.

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

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

Further, and importantly for the case of Negotiated Binding Agreements, the result of lemma 3 gives rise to the following corollary, that any pure Nash equilibrium of the underlying game is contained in IIR.

Corollary 1. *If* a^{NE} *is a pure Nash equilibrium of* G *then* $a^{NE} \in IIR$.

A further useful set of action profiles, and using a similar logic to example 2, is contained in IIR. If there is a chain of action profiles such that: for each player there is a single action profile that prescribes their best response to the action profile prescribed to others, while the other profiles give them a weakly higher utility, then such a set of action profiles is in IIR. Note, joint with lemma 1, this implies all profiles that Pareto dominate a pure Nash equilibrium in the underlying game remain in IIR. Having such a set of profiles within IIR will be further leveraged for sufficient conditions. This is given formally in the following lemma.

Lemma 4. *For any game* G, *if* $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A$ *satisfy:*

1. $\underline{a}_i^i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
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 $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq IIR$.

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 2. *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 2.

Example 2. revisited Again consider the underlying game, G, being the following prisoners' dilemma game.

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

In this case, 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 the following inequality holds:*

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

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, as is discussed in the literature review (section 7).

Finally, before moving to the sufficient conditions for an action profile to be supported by a Negotiated Binding Agreement, note that in two-player underlying games, where the action space is a compact subset of metric space and utility is continuous, the conditions of lemma 4 are also necessary. That is, a^* can be supported by a Negotiated Binding Agreement only if, in the underlying game, there exists a punishment for each player, where each player is prescribed the best response to their punishment within their punishment profile, they prefer the other players'

punishment to their own, and they prefer a^* to their punishment. This is formalised by the following theorem.

Theorem 3. *For any underlying game, G, such that $N = \{1, 2\}$, A_i is a compact subset of a metric space and u_i is continuous for all $i \in \{1, 2\}$, a^* is supported by a Negotiated Binding Agreement, s^*, only if $\exists\{\underline{a}^1, \underline{a}^2\} \subseteq A$ such that:*

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

The proof is similar to that of theorem 5, and therefore is omitted.

3.2. Sufficient Conditions

The first sufficient condition I provide states that if the outcome of iterated elimination of individually irrational actions is unique for all players then that only profile can be supported by any Negotiated Binding Agreement.

Corollary 2 (Conditions for a Unique Outcome)**.** *For any underlying game. G. such that A_i is a compact subset of a metric space and u_i is continuous for all $i \in N$, if $IIR = \{a^*\}$ then there is a unique Negotiated Binding Agreement and it is such that $s_i^*(h) = a_i^*$ for all histories.*

Of course, this uniqueness result requires strong conditions. Nonetheless, examples of this result do exist. This is true in example 1, where only (D, L) survives iterated elimination of individually irrational actions. This corollary is a joint implication of theorem 1 and theorem 4 that follows.

For the more general conditions, we require that each agent has a specific action profile in the underlying game, which I will denote $\underline{a}^i$. This can be thought of as the punishment of deviation used for i. For this action profile, i will best respond to $\underline{a}_{-i}^i$ in the baseline game G. The action of the underlying game that is sustained in Negotiated Binding Agreement, which I will denote a^*, must, for each player i, give a weakly higher payoff than $\underline{a}^i$. Further, I will require that the punishment of other agents gives a weakly higher payoff than the punishment for i in the underlying game. If such a collection of action profiles exist, then a^* can be supported by a Negotiated Binding Agreement. Notice by lemma 4, such a set of actions will be within IIR.[17] In essence, this relies on player-specific punishment strategies, that have been used for sufficiency for SPE in infinitely repeated games (Fudenberg and Maskin, 1986; Abreu et al., 1994). The requirements in my paper are more stringent, as the profile used to punish i must use i's best response to the punishment in the baseline game. This is because there is no future payoff to compensate for abiding to the punishment, as the agreement to play such an action is binding and the negotiation game terminates. I discuss the reasoning for this disparity further within the literature review. I state this formally in the following theorem.

[17] These conditions can be viewed as similar to *player contingent threats* of Greenberg (1990), as the negotiation game can be viewed as a special case of a social situation.

Theorem 4. *Take any underlying game, G, such that $\exists\{a^*, \underline{a}^1, ..., \underline{a}^n\} \subseteq A$ such that:*

1. $\underline{a}_i^i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
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.

It is worth noting that any pure Nash equilibrium of the game G is indeed supported by a Negotiated Binding Agreement. This is immediately implied by the fact any pure Nash equilibrium of the underlying game, denoted by a^{NE}, are in IIR via lemma 1, and can be used as the punishment for all individuals. That is, $\underline{a}^i = a^{NE}$ for all players. Further, any action profile that Pareto dominates a pure Nash equilibrium in the underlying game can be sustained by this reasoning. However, in underlying games where no pure Nash equilibrium exists there may exist a Negotiated Binding Agreement due to the above sufficient conditions.

Example 3. Consider the underlying game, G, being the following two-player game. For clarity, I have underlined the corresponding best responses in the baseline game.

1\2	L	C	R
T	7,7	<u>4</u>,4	0,<u>12</u>
M	4,<u>4</u>	0,0	<u>2</u>,3
D	<u>12</u>,0	3,<u>2</u>	1,1

Notice that there is no pure Nash equilibrium in this underlying game. However, there exists a Negotiated Binding Agreement. Specifically, applying theorem 4 take $a^* = (T, L)$, while taking $\underline{a}^1 = (M, R)$ and $\underline{a}^2 = (D, C)$, which satisfies the assumptions. Therefore there exists a Negotiated Binding Agreement that supports (T, L), while there is no pure Nash equilibrium in the underlying game. ▼

In two-player underlying games where the action space is a compact subset of a metric space and u_i is continuous for each player such conditions are both necessary and sufficient.

Corollary 3. *For any underlying game G such that $N = \{1, 2\}$, A_i is compact subset of a metric space and u_i is continuous for all $i \in \{1, 2\}$, a^* is supported by a Negotiated Binding Agreement, s^*, if and only if $\exists\{\underline{a}^1, \underline{a}^2\} \subseteq A$ such that:*

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

This is a direct implication of theorems 3 and 4.

Before moving forward, I point to the following corollary.

Corollary 4. *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 that 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.

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. Here I formally define No Delay Negotiated Binding Agreement and state the formal result.

Definition 5 (No Delay Negotiated Binding Agreement)**.** *s^* is a No Delay Negotiated Binding Agreement supporting $a^* = a(s^*|\emptyset)$ if:*

a) s^* is a Subgame Perfect Equilibrium of the negotiation game.

b) No Delay: For all partial histories $h \in H$, $s^*(h) = s^*(h, s^*(h)) = a^*(s^*|h)$.

With this, I turn to formally stating the result.

Theorem 5. *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. $\underline{a}^i_i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}^i_{-i})$
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). This is a Subgame Perfect Equilibrium that does not make use of weakly dominated strategies at any history. 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. One would also need to rule out weakly dominated actions from IIR, keeping the necessary conditions for agreement outcomes the same for this refinement of IIR. This, for instance, would rule out the possibility of using the worst Nash equilibrium as a punishment in the leading example where the underlying game was a First Price Auction. However, the other Nash equilibrium of the game, where $b^* = (4, 4, b_3^*)$ with $b_3^* \leq 1$ would provide the same logical result, albeit changing the lower bar of utility that bidders 1 and 2 must receive.

With these results, I now turn to some applications.

4 Applications

In this section, I explore two key applications of Negotiated Binding Agreements, within which the necessary and sufficient conditions allow for a full characterisation of the profiles supported by Negotiated Binding Agreements. These applications also provide the intuitions surrounding the proofs of the results presented in the paper.

In application 1, I explore a public goods game as the underlying environment and fully characterise the set action profiles that can be supported in Negotiated Binding Agreement. In this setting, contribution of an agent can only be supported if sufficiently many other agents also contribute. With this, full and no contribution can be supported. In application 2, I consider a simple Cournot Duopoly with potentially heterogeneous marginal costs as the underlying game. I fully characterise the set of actions that can be supported by Negotiated Binding Agreements and show that when marginal costs are the same, any profile of payoffs that gives both players positive profits is supportable. In contrast, when marginal costs are extremely different, only the firm with the lowest marginal cost receiving their monopoly profit can be supported.

Application 1. (Public Goods Game) Consider the underlying game, G, to be the following Public Goods Game. $N = \{1, 2, ..., n\}$. Let $A_i = \{c, d\}$ for each i. Let $u_i(a) = 1 + k\left[\sum_{j \in N} \mathbf{1}_{a_j = c}\right] - \mathbf{1}_{a_i = c}$ with $k \in \left(\frac{1}{n}, 1\right)$.

Firstly notice that for any player it is strictly dominant to choose d and hence the only Nash equilibrium payoff is 1.

I will now construct a strategy that allows for any action profile that Pareto dominates the Nash equilibrium to be supported by Negotiated Binding Agreement. Specifically, let a^* denote an action profile such that $u_i(a^*) = 1 + k|\{i \in N : a_i^* = c\}| - \mathbf{1}_{a_i^* = c} \geq 1$. Now construct s^* as follows. Let $s_i^*(\emptyset) = s_i^*(a^*) = a_i^*$. For all other partial histories let $s_i^*(h) = d$. First, notice that for the partial histories $\emptyset$ and (a^*) we have that $s^*(h) = a^*$ while $a(s^*|h) = a^*$. Secondly, notice that for all other partial histories we have $s^*(h) = (d)_{i \in N}$ while $a(s^*|h) = (d)_{i \in N}$. Concluding the condition for a no babbling agreement is always satisfied. All that is left to show is that s^* is a Subgame Perfect Equilibrium of the negotiation game. Suppose not, there is some partial history $h \in H$ such that there is some other strategy $s_i \in S_i$ such that $U_i(s_i, s_{-i}^*|h) > U_i(s^*|h)$. There are two possible cases.

1. The first possibility is that $h = \emptyset$ or (a^*). Notice for a deviation to be profitable it must be such that $a(s_i, s_{-i}^*|h) \neq a^*$, as otherwise a strict inequality cannot hold. Given this, the

strategy s_i, s^*_{-i} must induce a history $h' \neq (a^*, a^*)$. Therefore, it must be that all other players choose d for all periods other than the first. There are three possibilities.

(a) Firstly, it may be that the strategy s_i, s^*_{-i} induces a terminal history, $z \in Z'$, with the agreement $(d)_{i \in N}$. This induces a payoff of 1 for player i, while $U_i(s^*|\emptyset) = u_i(a^*) \geq 1 = U_i(s_i, s^*_{-i}|\emptyset)$, therefore this cannot be profitable.

(b) It may be that the strategy induces a terminal history, $z \in Z'$, with the agreement $((d)_{j \neq i}, c)$. However, this leads to a payoff of 0 for player i, while $U_i(s^*|\emptyset) = u_i(a^*) \geq 1 > 0 = U_i(s_i, s^*_{-i}|\emptyset)$, therefore this cannot be profitable.

(c) It may be that s_i, s^*_{-i} induces a terminal history, z such that d is played by all other players in all but the first period, and no agreement is made, i.e. $z \in Z''$. As no agreement is made, it must be that there are no two consecutive periods where the same action profile is played by all players it must be that s_i alternates between d and c. This implies that the lim sup of utilities induces by the proposals is 1. As $U_i(s^*|\emptyset) = u_i(a^*) \geq 1 = U_i(s_i, s^*_{-i}|\emptyset)$, this cannot be a profitable deviation.

2. Now suppose the history is partial and such that $h \neq \emptyset$ and $h \neq a^*$. No deviation leads to the agreement $(d)_{i \in N}$, with a payoff of 1. A deviation can only lead to the three cases examined above. Given this, the logic of the previous case remains true.

In conclusion, for any a^* such that $u_i(a^*) = 1 + k|\{i \in N : a^*_i = c\}| - \mathbf{1}_{a^*_i = c} \geq 1$ holds, we can provide a Negotiated Binding Agreement that supports such a profile. Further to this, it provides some intuition behind the sufficiency proof of theorem 4 which would imply this result.

To explore this further, notice that this implies that $a^* = (d)_{i \in N}$ may be supported. Further to this, a number of action profiles that maintain contribution can be supported by a Negotiated Binding Agreement. Specifically, for some a^* such that there exists some i such that $a^*_i = c$, we have that $1 + k|\{i \in N : a^*_i = c\}| > k|\{i \in N : a^*_i = c\}|$, i.e. the number of players contributing have a strictly lower utility that those who are not. With this, we can see that any a^* such that $k|\{i \in N : a^*_i = c\}| \geq 1$. More succinctly, when the number of contributors is above a lower bound, $|\{i \in N : a^*_i = c\}| \geq \frac{1}{k}$, the action profile can be supported by a Negotiated Binding Agreement. As $\frac{1}{k} < n$ this implies that full cooperation can be sustained.

Finally, to show that this fully characterises the Negotiated Binding Agreement, suppose that there is some equilibrium s^* that supports some a^* such that $u_i(a^*) < 1$ for some $i \in N$. For this to be the case it must be that $a(s^*|\emptyset) = a^*$. Now consider a deviation of $i \in N$ such that $s_i(h') = d$ for all histories $h' \in H$ at $h = \emptyset$. Such a deviation ensures that in any terminal history the payoff is pinned down by $u_i(d, a_{-i})$, be that if the history ends in agreement or not. If it does not end in agreement, it is pinned down by between some $u_i(d, a_{-i})$ with $a_{-i} \in \{c, d\}^{n-1}$. However, $u_i(d, a_{-i}) \geq 1$ for all possible $a_{-i} \in \{c, d\}^{n-1}$. Therefore $U_i(s_i, s^*_{-i}|\emptyset) \geq 1 > U_i(s^*|\emptyset)$. Therefore it cannot be that s^* is a Subgame Perfect Equilibrium of the negotiation game and therefore cannot be a Negotiated Binding Agreement. ▼

Application 2. (Cournot Duopoly with Heterogeneous Marginal Costs) Consider a simple Cournot Duopoly model as the underlying game, G, where $q_1, q_2 \in [0, b] = A_i$ where inverse demand is given by $p(q_1, q_2) = \min\{b - q_1 - q_2, 0\}$. Let firms have heterogeneous costs, c_1 and c_2 where without loss of generality $c_1 \geq c_2 \geq 0$. Assume that $\frac{b+c_2}{2} \geq c_1$. 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

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

This leads to profits of

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

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.

1. $s^*(\emptyset) = s^*(h) = (q_1^*, q_2^*)$ whenever $h = (q^1, q^2, ..., (q_1^*, q_2^*))$.
2. Otherwise, let $s^*(h') = (0, \underline{q}_2^1)$, where $\underline{q}_2^1 = b - c_1$ if $h' = (q^1, q^2, .., (q_1', q_2^*))$ for $q_1' \neq q_1^*$, $h' = (q^1, q^2, .., (q_1, \underline{q}_2^1))$, and $s^*(h'') = \underline{q}^2 = (b - 2c_1 + c_2, c_1 - c_2)$ for all other histories.

The intuition of this agreement is to have each firm to 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 and understand the other agent will propose their best response in the underlying game. Notice firm 1 can not sufficient flood the market when firm 2 has not negotiated as expected, due to their higher marginal costs.

Such a strategy profile satisfies agreement for all histories and no babbling. Therefore all that is left is to check that s^* is a Subgame Perfect Equilibrium of the negotiation game. Suppose that firm 1 has a profitable deviation at any history. It cannot be that it is profitable to deviate from $h = \emptyset$ or $h = (q^1, q^2, ..., (q_1^*, q_2^*))$ as this leads to player 2 playing $b - c_1$ for all periods. Therefore firm 1 can receive a utility of at most 0 via any deviation, as the static best response to $b - c_1$ is to set a quantity of 0. The same logic holds for all other cases, as, by construction, the static utility at every period of any other history is weakly less than 0, no matter s_i'. Suppose that firm 2 has a profitable deviation. It cannot be that a profitable deviation exists from $h = \emptyset$ or $h = (q^1, q^2, ..., (q_1^*, q_2^*))$ as this will lead to firm 1 proposing $b - 2c_1 + c_2$ in all periods. Therefore the highest possible utility is given by the static best response utility to such a quantity, given by $(c_1 - c_2)^2$. By construction, $U_i(s^*|\emptyset) = U_i(s^*|(q^1, q^2, ..., q^*)) \geq (c_1 - c_2)^2$, therefore it cannot be profitable. Further, *any* deviation leads to $s_1(h) = b - 2c_1 + c_2$, for which the static best response in each period would be $c_1 - c_2$, and therefore leading to a payoff no higher than $(c_1 - c_2)^2$. Finally, note that $U_i(s^*|q^1, q^2, .., (q_1', q_2^*)) = U_i(s^*|q^1, q^2, .., (q_1, \underline{q}_2^1)) = (b - c_1)(c_1 - c_2) \geq (c_1 - c_2)^2$ and therefore it cannot be that it is profitable to deviate from such a history either. Note that by corollary 3, this fully characterises the set of payoffs and strategies that can be supported by

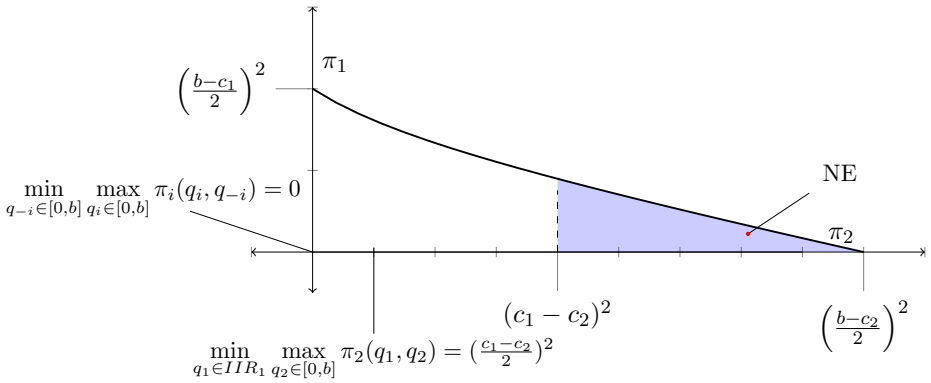

Figure 1: The payoff space of the Cournot Duopoly in example 2, where costs are such that $0 \leq c_2 < c_1 < \frac{c_2+b}{2}$. The black curve represents the payoff frontier. The shaded blue area represents that payoffs that can be sustained in Negotiated Binding Agreements.

a Negotiated Binding Agreement, as it gives both firms 1 and 2 the lowest possible payoffs they could best respond to, while maintaining that they prefer the punishment of the other firm to their own.

The payoff space is represented by figure 1.

Note that the construction provides us with some natural comparative statics. If $c_1 = c_2$, then both players may receive any payoff above 0. If $c_1 = \frac{b+c_2}{2}$ then we conclude that $\pi_2(\underline{q}^2) = \left(\frac{b-c_2}{2}\right)^2$, their monopoly profits.[18] ▼

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. To address these concerns, 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.

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.[19] 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 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

[18] Note that if $c_1 > \frac{b+c_2}{2}$ then the only outcome that can be supported by a Negotiated Binding Agreement is $q_1^* = 0$, $q_2^* = \frac{b-c_2}{2}$.

[19] This would be in line with a concept of renegotiation proofness a la Farrell and Maskin (1989) and Bernheim and Ray (1989).

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.

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 $B_C = \{a_C \in A_C | \exists a'_{-C} \in A_{-C}$ s.t. $(a_C, a'_{-C}) \in B\}$ and $B_{-C} = \{a_{-C} \in A_{-C} | \exists a_C \in A_C$ s.t. $(a_C, a_{-C}) \in B\}$.

With this, I go on to define the natural extension of Subgame Perfect Equilibrium when coalitions are permitted to jointly deviate. This will be referred to as $\mathcal{C}$-Subgame Perfect Equilibrium and will require that strategies are such that, at no history of the negotiation game, is there a way for *any* permissible coalition of players, $C \in \mathcal{C}$, to jointly deviate and improve the utility of all players within that coalition. 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. Note that the assumption that these agreements are private is important within this setting to ensure that the strategy of those outside are not dependent on the agreement itself. If the agreements were public, the concept would be closer to a coalitional version of Tennenholtz (2004)'s program equilibrium. I now define $\mathcal{C}$-Subgame Perfect Equilibrium formally.

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, which I outline here:

1. 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.

2. 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 $\mathcal{C}$.

3. 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.

Before defining the notion of Negotiated Binding Agreement with respect to this concept, it is worth noting that some coalition configurations can be seen as more reasonable than others in this case. Firstly, it seems reasonable to include all singletons within the coalition configuration, as allowing individuals to make unilateral deviations is in the essence of individual rationality. With this, I will concentrate the remainder of the analysis taking $\{i\}_{i\in N} \subseteq \mathcal{C}$ 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; $N \in \mathcal{C}$.

With this, I turn to defining $\mathcal{C}$-Negotiated Binding Agreement. This simply extends the notion of Negotiated Binding Agreement, instead of requiring a Negotiated Binding Agreement is a Subgame Perfect Equilibrium of the negotiation game, that has no babbling, I will require instead that it is a $\mathcal{C}$-Subgame Perfect Equilibrium of the underlying, that has a form of no babbling. 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 concluded in more than two periods. To see this, suppose that the equilibrium concludes in a and did so in more than 2 periods from the current one. This is as if this were the case, the grand coalition containing all agents would be able to profitably deviate to a joint strategy that ends in two periods and concludes in a. With this, I turn to formally define $\mathcal{C}$-Negotiated Binding Agreement.

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 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.

6 $\mathcal{C}$-Negotiated Binding Agreement Outcomes

6.1. 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 works inductively as follows. Consider some joint action of those within a coalition $C \in \mathcal{C}$ in the underlying game, a_C. If, for a coalition $C \in \mathcal{C}$ there is some function, that maps the joint action of those outside of the coalition to a joint action of the coalition, which, even in the worst case said function can provide a higher payoff than the joint

action a_C, then a_C is a coalitionally irrational joint action. This generalises the notion of individual rationality.[20] Notice this is exactly the notion of Aumann (1961)'s β-core. We may proceed inductively. Remove all coalitionally irrational actions for all coalitions $C \in \mathcal{C}$ in the underlying game. Consider some joint action of those within a coalition $C \in \mathcal{C}$, a_C, which survives iterated elimination of coalitionally irrational actions up to some iteration k. If, for a coalition $C \in \mathcal{C}$ there is some function, that maps the joint actions that have so far *survived iterated coalitionally irrational actions* of those outside of the coalition to a joint action of the coalition, which, even in the worst case of the joint actions outside of C that survives iterated elimination of coalitionally irrational actions is taken, said function provides a higher payoff than the joint action a_C, then a_C is a coalitionally irrational joint action at the iteration at hand. 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. The formal definition of coalitionally irrational actions and iterated elimination of coalitionally irrational joint actions is formally defined below.

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

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

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

Example 4. 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

[20] For underling games with compact action spaces and continuous utility they are identical when $\mathcal{C} = \{\{i\}_{i \in N}\}$.

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. Here I provide an example that shows how to find $ICIR(\mathcal{C})$. 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 5. 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.[21]

Lemma 5. *For any Strong Nash equilibrium a^{SNE} of G, $a^{SNE} \in ICIR(\mathcal{C})$ regardless of $\mathcal{C}$.*

With this definition, a similar necessary condition to theorem 1, linking $ICIR(\mathcal{C})$ of the underlying game to the proposals made in $\mathcal{C}$-Negotiated Binding Agreement of the negotiation game, exists.

Theorem 6. *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 5. *If $ICIR(\mathcal{C}) = \emptyset$ then no $\mathcal{C}$-Negotiated Binding Agreement can exist.*

[21] Recall a strong Nash equilibrium is an action profile a^{SNE} such that for all $C \in 2^N \backslash \{\emptyset\}$ $\nexists a_C \in A_C$ such that $u_i(a_C, a^{SNE}_{-C}) > u_i(a^{SNE})$ for all $i \in C$.

This is an immediate implication of theorem 6. Note that this is possible, i.e. in example 4, 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 2 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 for deviation 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 $i \in C$.*

This includes a notion of making sure the action profile at hand is jointly coalitionally rational in the underlying game. Note the similarity to the β-core of Aumann (1961). 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 $ICIR(\mathcal{C})$, defined over the underlying game, by the following theorem.

Theorem 7. *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.

This result can provide us with some insight into the types of agreements that may not be sustained. For instance, 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 $\{N, \{i\}_{i \in N}\} \subseteq \mathcal{C}$. This is illustrated by the following example.

Example 6. 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	**1,12**	0,0
T	4,0	0,0	**7,2**	1,1
D	**12,1**	**2,7**	4,4	0,8
DD	0,0	1,1	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 $\tilde{A}^1 = \{(TT, LL), (TT, R), (T, R), (D, LL), (D, L)\}$. No further deletion can take place, as the individually rational payoffs over this set are given by 2, the lowest payoff given by a profile in this set, therefore:

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

Now notice that the outcome (TT, R) necessarily cannot be sustained in equilibrium, 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, which is 1 from best responding to RR, but it cannot be sustained in a $2^N \backslash \{\emptyset\}$-Negotiated Binding Agreement. ▼

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

6.2. Sufficient Conditions

Similarly to theorem 4, I provide sufficient conditions for $\mathcal{C}$-Negotiated Binding Agreement that prevents deviations from equilibrium based on the identities of the deviators, rather than the deviation they perform. Firstly, similarly to corollary 2, if there is only a single action profile consistent with $ICIR(\mathcal{C})$ then this must be supported in equilibrium, and further to this is the only profile that can be the outcome of $\mathcal{C}$-Negotiated Binding Agreement. I state this formally here.

Corollary 6. *If G is such that u_i is continuous and A_i is compact for all agents, if $ICIR(\mathcal{C}) = \{a^*\}$, then a^*, then s^* is a $\mathcal{C}$-Negotiated Binding Agreement if and only if $s^*(h) = a^*$ for all $h \in H$.*

Note that this condition may occur in more environments than corollary 2 when $\{i\}_{i \in N} \subset \mathcal{C}$, as $ICIR(\mathcal{C})$ may involve more deletion in the underlying game G. However, as $ICIR(\mathcal{C})$ may be empty and leave us with no $\mathcal{C}$-Negotiated Binding Agreement.

Nonetheless, a more general set of sufficient conditions apply, as with theorem 4 and will again 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. These are as before: 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, the inclusion of such profiles in $ICIR(\mathcal{C})$ is now required and not implied due to the rich deletion that can take place.

Theorem 8. *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 5, 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.[22] To see this, consider the following example.

Example 6. revisited Consider again the following two-player game as the underlying game, G. All possible coalitions are permitted, $\mathcal{C} = 2^N \backslash \{\emptyset\}$.

1\2	LL	L	R	RR
TT	**6,6**	0,4	**1,12**	0,0
T	4,0	0,0	**7,2**	1,1
D	**12,1**	**2,7**	4,4	0,8
DD	0,0	1,1	8,0	0,0

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 8 apply. From the previous analysis we know that:

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

Therefore we may take, for example, $\underline{a}^N = a^* = (TT, LL)$, $\underline{a}^1 = (D, L)$ 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 8 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.

I now turn to an application.

6.3. An Application of $\mathcal{C}$-Negotiated Binding Agreements

As with strong Nash equilibrium, conditions for existence of a $\mathcal{C}$-Negotiated Binding Agreement are not generically satisfied. Nonetheless, there exist interesting applications for which $\mathcal{C}$-Negotiated Binding Agreements exist. Consider the following Cournot game.

Application 3. (Symmetric Cournot with Fixed Cost) Consider a simple model of Cournot with fixed costs as the underlying game G. These fixed costs depend on the total number of firms

[22] 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.

that enter the market. This captures a situation where the fixed cost is due to the purchase of equipment. The cost of equipment itself is dictated by the law of supply and demand and therefore this cost increases with the number of firms purchasing this.

I model this in the following way. Let there be $n = 4$ firms. Let each firm choose the quantity that they will sell, $q_i \geq 0$. Let total demand, as a function of the total quantity, be given by $\max\{b - \sum_{j=1}^{4} q_j, 0\}$, where $b > 0$. I assume that the marginal cost is constant and symmetric, therefore it is without loss to set it to 0. Therefore gross profits for player $i \in \{1, 2, 3, 4\}$ are given by $\max\{(b - \sum_{j=1}^{n} q_j), 0\}q_i$. Let fixed costs take the following form: $\left(\frac{3}{32}b\sum_{j\neq i}\mathbf{1}_{q_j>0}\right)^2\mathbf{1}_{q_i>0}$. Notice that this does indeed increase with the number of firms entering, and the first firm to enter the market may do so for free. Therefore utility takes the following form:

$$u_i(q) = \max\left\{\left(b - \sum_{j=1}^{4} q_j\right), 0\right\}q_i - \left(\frac{3}{32}b\sum_{j\neq i}\mathbf{1}_{q_j>0}\right)^2\mathbf{1}_{q_i>0}$$

Notice that in this model the individual best responses are given by the following expressions

$$q_i^*(q_{-i}) = \begin{cases} \left\{\frac{b-\sum_{j\neq i} q_j}{2}\right\} & \text{if } \sum_{j\neq i} q_j < b - \frac{3}{16}m\sum_{j\neq i}\mathbf{1}_{q_j>0} \\ \left\{0, \frac{b-\sum_{j\neq i} q_j}{2}\right\} & \text{if } \sum_{j\neq i} q_j = b - \frac{3}{16}b\sum_{j\neq i}\mathbf{1}_{q_j>0} \\ \{0\} & \text{if } \sum_{j\neq i} q_j > b - \frac{3}{16}b\sum_{j\neq i}\mathbf{1}_{q_j>0} \end{cases}$$

Notice that this implies the following:

1. There is a Nash equilibrium when 2 firms enters the market, and both choose quantities of $q_i^* = \frac{b}{3}$. This leads to a payoff of $\frac{943}{9216}b^2$ for the firm who enters and 0 for those who do not. There are no other Nash equilibria in this game. Further, note that this is not a Strong Nash equilibrium, as the two producing firms could split the monopoly profits equally. Such quantities are a coalition-proof Nash equilibrium.

2. There are many Pareto efficient outcomes. For example, any outcome such that the firms who sell, in aggregate sell the monopoly quantity, $\sum_{i=1}^{4} q_i = \frac{b}{2}$, while profits are strictly positive for all those who produce strictly positive quantities, are Pareto efficient. Note that there exists such a profile for any number of firms entering. For instance, all firms producing $\frac{b}{8}$ leads to profits of $u_i(q^{p,all}) = \frac{55}{1024}b^2 > 0$. Further note that, due to the fixed costs, the Pareto efficient outcomes do not require the monopoly quantity unless all firms enter, but all involve each firm receiving a weakly positive payoff.

3. Any Pareto efficient profile is in the α-core and the β-core.

Consider $\mathcal{C} = 2^N\backslash\{\emptyset\}$. Let $q^* = (q_1^*, q_2^*, q_3^*, q_4^*)$ be the quantity that is trying to be sustained. I will argue that it is possible to sustain an efficient outcome where all agents produce in strong Negotiated Binding Agreement. That is, an agreement such that $q_i^* = \frac{b}{8}$ can be sustained. Consider the following strategies.

1. If $h = (q^1, q^2, ..., q^k)$ is such that $q_{-C}^{k-1} = s_{-C}^*((q^1, q^2, ..., q^{k-2}))$ and either

(a) $q_l^k = s_l^*(q^1, q^2, ..., q^{k-1})$ for all $l \notin C$ and $q_j^k \neq s_j^*(q^1, q^2, ..., q^{k-1})$ for all $j \in C$

(b) or $q_{-C}^k = \frac{16+\sqrt{137}}{64}b$ if $|C| = 3$, $q_{-C}^k = \left(\frac{16+\sqrt{137}}{64}b, \frac{16+\sqrt{137}}{64}b\right)$ if $|C| = 2$ and $q_{-C}^k = \left(\frac{16+\sqrt{137}}{64}b, \frac{16+\sqrt{137}}{64}b, 0\right)$ if $|C| = 1$

Then:

- $s_i^*(h) = \frac{16+\sqrt{137}}{64}b$ for $i = \min_{j \notin C} j$ if $|C| \leq 3$ or $i = 1$ if $|C| = 4$.
- $s_i^*(h) = \frac{16+\sqrt{137}}{64}b$ for $i = \min_{j \notin C \setminus \{\min_{j \notin C} j\}} j$ if $|C| \leq 2$ or $i =$ mod $(j+2, 4)$, $j \in C$, $j \geq k$, $k \in C$ otherwise.
- $s_i^*(h) = 0$ for all other $i \in N$.

2. For all other histories, let $s_i^*(h) = \frac{b}{8}$

The logic of this strategy is as follows. Suppose that we are at a history only one coalition has deviated in the penultimate period of the history, while in the previous period either all players have played the assigned strategy or only those within the deviating coalition have deviated. Note that this may involve a smaller coalition deviating in the penultimate period, while in the next a larger coalition deviates. If this is the case, assign two players to play a strategy that gives them exactly the payoff in aggregate of all players entering and producing the efficient quantity, while all other players do not produce. At least one of these players is not within the deviating coalition if the cardinality of that coalition is 3 or less. At all other histories all agents propose their share of the equal division of the monopoly quantity.

Now I will show that this does indeed constitute a strong Negotiated Binding Agreement.

First consider a coalition deviating from a history that does not fall into case 2, where no deviation leads to the agreement that all firms enter and divide the monopoly quantity. It cannot be that the grand coalition deviates to improve the utility of all members. Therefore it must be that deviation does not involve one firm. By the structure of case 1, which any deviation must lead to, it is then the case that those outside the coalition are proposing, in aggregate, at least $\frac{16+\sqrt{137}}{64}b$ in every period. As $\frac{16+\sqrt{137}}{64}b > \frac{1}{8}b$ it follows that it cannot be that all firms who deviate are producing and improving the utility of all members. Therefore it must be that the deviation only involves a coalition of at most two firms. It cannot be that they are both assigned to not produce in all periods, as this implies that the profits are bounded above by $\frac{249-32\sqrt{137}}{4096}b < 0$ if producing and 0 if not. This bound is the same if only one firm deviates. Therefore no profitable deviation can exist from case 2. Now suppose that a profitable deviation exists from case 1. By a similar logic, it cannot be that two firms deviate and improve their utility. This is because a "punishing" firm does not wish to deviate, as they would be punished for this. A "punished" firm also does not wish to deviate, as the punishment is sufficiently high to ensure that they do not wish to enter the market. Further, it cannot be that all agents jointly deviate, as there are two punishing firms, who, in aggregate, receive the utility that is the maximum that can be achieved for all firms entering. Therefore they have no incentive to do so.

Further to this point, notice that the punishments are such that 1 punishing firm does not produce, leading to a profit of 0. Given for any deviating coalition such a punishment could be used, we conclude that any Pareto efficient profile could be sustained, leading to an equivalence to the β-core, while being fully justified via the use of $\mathcal{C}$-Negotiated Binding Agreements. ▼

7 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). The philosophy of the farsighted stable set is that an outcome is in the farsighted stable set if no group of agents can change the current outcome, inducing a chain reaction of groups who do the same, where at each possible point in the chain the group who changes the outcome prefers the final outcome induced to the outcome they move from.[23] 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), where at no history can any coalition improve their utility by jointly deviating to a new strategy. He also imposes a refinement on this subgame perfect type concept based on the farsighted stable set, where a group of agents only make a new proposal at period t if it is in their benefit (given the outcome it would induce) *compared to the current proposal in period* t. This is similar to this work in the sense that ($\mathcal{C}$-)Negotiated Binding Agreements takes a refinement of the ($\mathcal{C}$-)Subgame Perfect Equilibria of the negotiation game, but I do so in a way that requires no babbling. 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 based on Pareto optimality in the case of a unique outcome and individual rationality, similar to those provided in this paper for both Negotiated Binding Agreements and $\mathcal{C}$-Negotiated Binding Agreements, while I do not need to impose history independence in the strategies for negotiation. 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. My sufficient conditions are similar in flavour but apply to a more general class of games.

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. For instance, Suzuki and Muto (2005) find the farsighted stable set for an underlying prisoner's dilemma game, allowing all possible groups to make new proposals at each round, concluding cooperation can be sustained.

[23] This was posed as a criticism of Von Neumann and Morgenstern (1994)'s vNM stable set, which only considers a notion of a group moving to a different stable outcome for a group if it improves their utility, but does not consider making a new proposal purely to induce a chain that eventually improves their utility.

Nakanishi (2009) looks at a version of the farsighted stable set of a prisoner's dilemma game where only individuals can make a new proposal, rather than groups of individuals. He again concludes that cooperation is sustainable. My paper instead provides necessary and sufficient conditions for agreement outcomes in a general underlying game, both in the case groups can jointly deviate from a candidate strategy while negotiating (sections 5 and 6) and when only individuals can (sections 2 to 4), which are based fully on a refinement of the SPE in the negotiation game, rather than a stable-set-like concept.

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 may make proposals of what they will individually play in the underlying game and if agents change their proposal within a period then they are no longer permitted to change their proposal again. Kalai (1981) looks at the perfect equilibria of Selten (1988)[24] and shows that only cooperation can be sustained in the 2-player prisoners' dilemma game. More recently, 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. Kalai (1981) also relies on a refinement of SPE as a solution concept to his negotiation game. The model of my paper does not impose that changing the proposal from the last leads to no longer being able to change the proposal again. My approach gains a large degree of tractability and allows for general conditions on a given underlying game for agreement outcomes of the negotiation, rather than focusing on prisoners' dilemma. 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. If there is a sequence of three proposals (two for a player and one for the other) where the player who has proposed first and last does not revise their latest proposal, then the prices are implemented. Bhaskar (1989) looks at the perfect equilibria of such an agreement game and concludes that only the monopoly price can be sustained. The key driving force behind the disparity of the results my paper and in Bhaskar (1989), where in his work only a unique efficient outcome can be sustained, is the use of perfect equilibrium rather than Subgame Perfect Equilibrium. In his case, this rules out the possibility of using the Nash equilibrium of the underlying Bertrand game as a threat agreement, as the Nash equilibrium is in weakly dominated strategies.

A number of papers have provided a more cooperative game theoretic approach for the agreements that can be made for games. The closest to this paper are a) the Strong Nash equilibria (Aumann, 1959) of a game, where, holding the actions of all those outside of a group constant, no group can deviate and induce an improvement for all those within the group and b) the β-core Aumann (1959, 1961), where no group can improve the utility of all members within their group, understanding that those outside of the group will use a coalition specific punishment in response, which need not satisfy any credibility constraint. In my paper, $\mathcal{C}$-Negotiated Binding Agreement outcomes lie somewhere between the β-core and strong Nash equilibrium, as agents are permitted to change their proposals when they observe a proposal of others change, but can only do so in a way pinned down by an optimal strategy in the sense of equilibrium. Given this, my paper can also be seen in the light of the Nash program pointed to in Nash (1953), as the necessary

[24]These are the Subgame Perfect Equilibria that do not permit the use of weakly dominated strategies at any history.

and sufficient conditions $\mathcal{C}$-Negotiated Binding Agreement outcomes can be seen as a perturbed version of the β-core. This provides the result of an equilibrium of a fully specified negotiation game while pointing to the solution having a flavour of cooperative game theory. Notable contributions to this literature include Rubinstein (1982); Chatterjee et al. (1993) on bargaining over the division of surplus.

There are a number of other related papers that take the cooperative game theoretic approach. Chander and Tulkens (1997) define the γ-core, where upon a coalition deviating they consider all other agents acting individually, and do not permit the remainder to form any coalitions. My solution of $\mathcal{C}$-Negotiated Binding Agreements instead takes an equilibrium approach, where deviations are ruled out due to the response of others that would entail, which are defined in the same way as the equilibrium itself, rather than using alternative punishments. 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. Their solution relies on an assumption that agents know a coalition has deviated and is shown to be related to a perturbed version of the α-core of Aumann (1961).[25] This provides a similar link to core-like solutions for agreements over underlying games, but does so from a different perspective. Ismail (2021) considers cooperation in extensive form games, but does so by adding new elements, such as a coalition utility function which, defines the (common) utility all members would receive based on the outcomes, and including it as a strategic choice to respect the coalition's choice via a coalition strategy game. In a similar sense, $\mathcal{C}$-Negotiated Binding Agreements must require optimality of all coalitions and individuals at all points, but do not require that all members of a coalition receive the same utility. 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). I instead focus on the outcomes achievable via negotiation when all agents must negotiate, taking a more non-cooperative approach.

A number of papers consider a form of communication for equilibrium selection. 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 SPE or Nash Equilibria of an underlying game, whereas my concept allows agents to make a binding agreement of potentially any outcome. The closest in this literature is Rabin (1994). Rabin (1994) explicitly models a negotiation over the (potentially mixed) choice of Nash equilibrium of the underlying game.[26] This is similar to the notion of Negotiated Binding Agreements, as they represent the agreements that can occur via the specific negotiation process within my paper. However, my

[25] The α-core is defined similarly to the β-core, however in response to a deviation of a coalition can be responded to in a way specific to the deviation itself, rather than the coalition.

[26] In the model of Rabin (1994), only a mix of Nash outcomes can be proposed in the negotiation, and therefore only a mix of Nash outcomes can be agreed upon, inducing public correlation via negotiation.

paper does not require that a mix of Nash equilibria of the underlying game is selected in the case of agreement, as binding agreements can be made. Another related concept in this strain is Bernheim et al. (1987)'s coalition proof equilibrium, where coalitions are permitted to deviate, but can only do so in a private non-binding way and therefore deviations must be self-enforcing a la Nash. $\mathcal{C}$-Negotiated Binding Agreements also allows coalitions to make private agreements, but does so in a binding way, and therefore need not be self-enforcing. Similarly, Farrell and Maskin (1989) and Bernheim and Ray (1989) develop a concept of renegotiation proof equilibrium, where the grand coalition may deviate to a preferred SPE at a point in a repeated game. The same disparity between my paper and theirs occurs as within coalition proof equilibrium.

Negotiated Binding Agreements is also related to another literature on binding agreements, contract theory. Most closely related are the works 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, rather than allowing one agent or a mediator to completely define the structure and then make a take-it-or-leave-it offer on the contract. 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. This strain of literature does not consider the possibility of joint deviations, however conceptually this is similar to $\mathcal{C}$-Negotiated Binding Agreement, as this can be viewed as the agreements that result when agreements over how to negotiate can be made.

The way payoffs are defined for perpetual disagreement can be seen as similar to the literature of infinitely repeated games with no discounting. Notably, when well defined, the limit of means criteria of Aumann and Shapley (1994); Rubinstein (1994) can be used. The sufficient conditions within the paper are also similar to the sufficient conditions of player-specific punishment is used in infinitely repeated games, i.e 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.

8 Conclusion

In this paper, 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 three key applications. 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 must receive. I show that when the underlying game is a public goods game, agents can agree to contribute only if the aggregate level of contribution sufficiently compensates them for their contribution. 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 all 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. I apply this to a Cournot model, with a fixed cost that depends on the number of entrants. Within this setting, the outcomes coincide with the β-core, but are fully justified by a fully specified negotiation protocol.

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. 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. However this is not always the case and therefore understanding when it is theoretically beneficial is an interesting line to follow.[27] 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.[28] Nonetheless, understanding the removal

[27] 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, for instance, 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.

[28] To see this, consider the two-player case. Notice that ruling out an action for player 1 in the underlying game makes the harshest punishment for player 2 weakly better (from player 2's perspective). However, as player 2 must prefer punishing player 1 than being punished, this can limit the punishments that player 2 can use against player

of *which* actions leads to this improvement is an open question. The well-pinned-down nature of the results makes exploring such questions achievable and I leave these for future work.

Finally, allowing coalitions to overlap provides a direction for interesting insights. Such arrangements of groups frequently occur in economic environments, such as international relations and trade,[29] but are not typically considered in the literature.[30] In this work, I allow for groups to overlap, but take the set of permissible coalitions to be exogenously set. An interesting question is *which* coalitions would form when they are permitted to overlap, allowing for an endogenous formation of coalitions. This would build on the literature of endogenous coalition formation, for instance Ray and Vohra (1997); Diamantoudi and Xue (2007), but add the novelty of allowing agents to be a member of multiple groups at once. Exploring overlapping coalition formation is especially interesting when transfers are not permitted, as then it is not possible to treat those coalitions that overlap equally as one larger coalition, by moving transfers between those agents to align incentives.

1. With this, it may lead to an indirect improvement in the agreements that can occur for player 1.

[29] For instance, Australia is in The Regional Comprehensive Economic Partnership (RCEP), which can be seen as a coalition in multinational negotiations. RCEP also includes members of the Association of Southeast Asian Nations, which Australia is not a member of.

[30] As far as I am aware, all results within the literature are based on the assumption of no overlapping coalitions. One exception to this, providing results when allowing for overlapping coalitions, is Gavan (2022), where I provide an existence result for an equilibrium concept that allows for overlapping coalitions.

A 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 A.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 A.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 A.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. An alternative specification, where, in the case of perpetual disagreement, agents believe that the worst agreeable action is played by all other players, while they may individually deviate, is considered in A.4.. Here both the necessary and sufficient conditions for agreement outcomes hold.

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.

A.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, followed by a sequence of proposals for all agents, and then 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:

a) Where the same action profile is proposed twice in consecutive periods, and 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 this histories as ones where an *agreement* is made.

b) 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 off 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:*

a) s^* is a Subgame Perfect Equilibrium.

b) 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 3 remains 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 minmax 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 need 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 payoff 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 7. 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.

Proposition 1. *For any order $\mathcal{O}$, if s^* is a Negotiated Binding Agreement with order then, for all histories for i is active $h \in \tilde{H}_i$, $s_i(h) \in IAD_i$.*

I reserve this proof, and all other proofs within this section, for the appendix C.

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 2. *Take any order $\mathcal{O}$. 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 order of proposals is not important, nor is important that the proposals are made simultaneously. Rather, the structure of the terminal histories, and the associated payoffs, as well as the ability for all agents to make some proposal, are the key features

of the model. Within the next subsection, 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.

A.2. Robustness to Joint Proposals

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

Define the negotiation game with all proposals.

A history will be the empty set, followed by 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:

a) 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 this histories as ones where an *agreement* is made.

b) 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 for all agents. Let the set of all histories, terminal and partial, be 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 bound of the $\liminf_{t \to \infty} u_i(\tilde{a}^t)$ 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 off 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:*

a) s^* is a Subgame Perfect Equilibrium.

b) $\forall h \in \tilde{\tilde{H}}$ $\exists h' \in \tilde{\tilde{H}}$ 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 3. *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 agreement outcomes also hold. 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 4. *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 for 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.

A.3. Robustness to Outside Options

Within this subsection, 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 subsection as *constant outside option Negotiated Binding Agreement*. In this setting, it is no longer true that the necessary conditions for agreement outcomes remain to be true. However, the sufficient conditions for agreement outcomes remain to be valid. This is highlighted by the following proposition.

Proposition 5. *If s^* is a Negotiated Binding Agreement then s^* is a constant outside option Negotiated Binding Agreement.*

As Negotiated Binding Agreement do need not to 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 paper or normalised to be worse than any outcome as typically assumed in bargaining games.

A.4. Robustness to Worst Agreement of Others for Perpetual Disagreement

In this section I discuss how the results are robust to an alternative specification of the payoffs of perpetual disagreement. Here, it will be assumed that agents believe that the actions of others will be pinned down by the worst outcome of agreement for all other agents, while they will be permitted to unilaterally deviate. Formally, this will be described as follows.

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 a joint action $A = \times_{i \in N} A_i$. u_i is utility function such that $u_i : A \to \mathbb{R}$ and u_i is bounded for all $i \in N$. Let $A_{-i} = \times_{j \neq i} A_j$.

The set of partial histories consists of all $h = (a^1, a^2, ..., a^k)$ 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.

A history is terminal if, either:

a) Where 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 this histories as ones where an *agreement* is made.

b) an infinite sequence 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 terminal histories be given by $Z = Z' \cup Z''$. 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 U_i denote the payoff for player $i \in N$ of the negotiation game.

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 $i \in N$.

Let $s_i : H \to A_i$ be the strategy for each player. Notice that from any history $h \in H$ an agent can choose a strategy such that the continuation lies in Z'', regardless of the strategies of others.

Let A^{agree} be the set of agreements outcomes that can be supported in equilibrium. Let this set be constructed in the following way:

1. $\exists s^*$ be a strategy profile such that:

   (a) $s^*(\emptyset) = s^*(a) = a \in A^{agree}$.

(b) For any $h \in H$, $\nexists s_i' \in S_i$ such that the continuation of $(s_i', s_{-i}^*|h) \in Z'$ and $U_i(s_i', s_{-i}^*|h) > U_i(s^*|h)$.

2. $\forall i \in N \ \forall a \in A^{agree} \ \exists a'_{-i} \in A_{-i}^{agree}$ such that $\sup_{a_i' \in A_i} u_i(a_i', a'_{-i}) \leq u_i(a)$.

1. ensures that agents can not reach a better agreement within the negotiation it self. 2. ensures that it is not the case that an agent can cause perpetual disagreement, and in doing so can induce others playing any of the actions that can be agreed upon, while ensuring that the deviating agent can increase their utility.

Here both the necessary and sufficient conditions for agreement outcomes presented in the main paper still hold, as demonstrated by the following two propositions.

Proposition 6. *For all $a \in A^{agree}$, $a \in IIR$.*

Further, if $a^* \in A^{agree}$, then $u_i(a^*) \geq \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$

Proposition 7. *If $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A$ satisfy:*

1. $\underline{a}_i^i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
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 $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A^{agree}$.

Notice that proposition 7 follows by definition of the agreement set and the fact that, by lemma 4, these actions are in IIR, and therefore I forgo the proof.

B Appendix: Proofs from Main Text

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

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

Proof. 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 we may conclude that $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}\left(u_i(a^t)-\liminf_{k\to\infty}u_i(a^k)\right)>-\epsilon\ \forall\epsilon>0$, 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

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

The analogous proof works for showing $\lim_{\delta\to 1}(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(a^t)\leq\limsup_{k\to\infty}u_i(a^k)$. □

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

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

Proof. 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 $a_i^t=\tilde{a}_i(a^t_{-i})$. Therefore

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

and therefore $U_i(s_i',s^*_{-i}|h)\geq\inf_{a_{-i}\in A_{-i}}\sup_{a_i\in A_{-i}}u_i(a_i,a_{-i})$. A contradiction as there exists a profitable deviation. □

Theorem. *1 If s^* is a Negotiated Binding Agreement, then for all $h\in H$, $s_i^*(h)\in IIR_i$.*

Proof. 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 $j \in N$ $s_j(h') \in \tilde{A}^k_j \ \forall k < m$ and $h' \in H$ but for some $i \in N$ $s_j(h') = a_i \notin \tilde{A}^{m+1}_j$ for some $h' \in H$. By no babbling it must be that a) $s_{-i}(h') \in \tilde{A}^m_i$ 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 $U_i(s|h) = u_i(a(s|h)) \leq \sup_{a'_{-i} \in \tilde{A}^m_{-i}} u_i(a_i, a'_{-i})$. Take $\epsilon = \inf_{a'_{-i} \in \tilde{A}^m_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - u_i(a(s|h)) > 0$. Take a function $\tilde{a}_i : \tilde{A}^m_{-i} \to A_i$ such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$ for all $a_{-i} \in \tilde{A}^m_{-i}$. Consider a deviation $s'_i(h'') = \tilde{a}_i(s_{-i}(h''))$ for all $h'' \in H$. It follows that

$$U_i(s'_i, s_{-i}|h) \geq \inf_{a_{-i} \in \tilde{A}^m_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) > \inf_{a_{-i} \in \tilde{A}^m_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a_{-i}) - \epsilon$$

. Therefore it follows that $U_i(s'_i, s_{-i}|h) > u_i(a(s|h)) = U_i(s|h)$, concluding that a profitable deviation exists and therefore it cannot be that s is a Subgame Perfect Equilibrium of the negotiation game. By no babbling, we conclude that $s_i(h) \notin D_i(\tilde{A}^m_{-i})$ for any $h \in H$ and therefore $s_i(h) \in \tilde{A}^{k+1}_i$, a contradiction. □

Lemma. *3 The set of actions that survive iterated elimination of never best responses to pure actions it also survives iterated elimination of iterated deletion of individually irrational actions: $IENBR \subseteq IIR$.*

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

Lemma. *4 If $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A$ satisfy:*

1. $\underline{a}^i_i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}^i_{-i})$
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 $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq IIR$.

Proof. We proceed inductively. By definition, $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A = \tilde{A}^0$.

Now suppose that $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq \tilde{A}^k$ for all $k \leq m$ for $m \geq 0$. Note that

$$\underline{a}_i^i \in \underset{a_i \in A_i}{\text{argmax}}\, u_i(a_i, \underline{a}_{-i}^i) = \arg \sup_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$$

and therefore

$$u_i(\underline{a}_i, \underline{a}_{-i}) \geq \inf_{a_{-i} \in \tilde{A}_{-i}^m} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$$

. Therefore by definition

$$u_i(\underline{a}^j) \geq \inf_{a_{-i} \in \tilde{A}_{-i}^m} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$$

and $u_i(a^*) \geq \inf_{a_{-i} \in \tilde{A}_{-i}^m} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$. Further, we have that $\sup_{a_{-i} \in \tilde{A}_{-i}^m} u_i(\underline{a}_i^i, a_{-i}) \geq u_i(\underline{a}^i)$, $\sup_{a_{-i} \in \tilde{A}_{-i}^m} u_i(\underline{a}_j^i, a_{-i}) \geq u_i(\underline{a}^j)$ and $\sup_{a_{-i} \in \tilde{A}_{-i}^m} u_i(a_i*, a_{-i}) \geq u_i(a^*)$. Therefore we may conclude that $\underline{a}_i^i, \underline{a}_i^j, a_i^* \in \tilde{A}_i^{m+1}$. □

Theorem. *2 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 $i \in N$.*

Proof. 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 1 it must be that $s^*_{-i}(h) \in IIR_{-i}$ for all $h \in H$. Let $\epsilon = \inf_{a'_{-i} \in IIR_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) - U_i(s^*|h) > 0$. Construct $\tilde{a}_i : IIR_{-i} \to A_i$ such that $u_i(\tilde{a}_i(a_{-i}), a_{-i}) \geq \sup_{a_i \in A_i} u_i(a_i, a_{-i}) - \frac{\epsilon}{2}$ for all $a_{-i} \in IIR_{-i}$. Consider a deviation to $s'_i(h')$ such that $s'_i(h') = \tilde{a}(s^*_{-i}(h'))$ for all $h' \in H$ at the history h. It follows that

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

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

Theorem. *4 Take any underlying game such that $\exists \{a^*, \underline{a}^1, ..., \underline{a}^n\} \subseteq A$ such that:*

1. $\underline{a}_i^i \in \text{argmax}_{a_i \in A_i}\, u_i(a_i, \underline{a}_{-i}^i) = \text{argmax}_{a_i \in A_i}\, u_i(a_i, \underline{a}_{-i}^i)$
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.

Proof. 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. □

Theorem. *5 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. $\underline{a}^i_i \in \text{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}^i_{-i})$
2. $u_i(a^*) \geq u_i(\underline{a}^i)$
3. $u_i(\underline{a}^j) \geq u_i(\underline{a}^i)$ *for all* $i, j \in N$

Proof. By the construction of theorem 4 such a^* can be supported.

To see that only such a^* can be sustained, take any a^* such that it is supported by a no delay Negotiated Binding Agreement given by the SPE s^*. Denote $\tilde{A} = \{a \in A | \exists h \in H \text{ s.t. } s^*(h) = a\}$.

Note by strict no babbling these completely define the set of actions that can be agreed upon. Further to this, not that $s^*_{-i}(h) \in \tilde{A}_{-i}$ for all $h \in H$ by strict no delay. As s^* is an SPE it must be that there is no profitable deviation. Notice that $U_i(s^*|h) \geq \max_{a_i \in A_i} \inf_{a_{-i} \in \tilde{A}_{-i}} u_i(a_i, a_{-i})$. Suppose not $U_i(s^*|h) < \max_{a_i \in A_i} \inf_{a_{-i} \in \tilde{A}_{-i}} u_i(a_i, a_{-i})$. It follows that $\max_{a_i \in A_i} \inf_{a_{-i} \in \tilde{A}_{-i}} u_i(a_i, a_{-i}) - U_i(s^*|h) > 0$. 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

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

By definition,

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

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 $U_i(s^*|h) \geq \max_{a_i \in A_i} \inf_{a_{-i} \in \tilde{A}_{-i}} u_i(a_i, a_{-i})$. As $\tilde{A}$ are agreed upon, it must therefore be that $\forall \tilde{a} \in \tilde{A}$ we have that $u_i(\tilde{a}) \geq \max_{a_i \in A_i} \inf_{a_{-i} \in \tilde{A}_{-i}} u_i(a_i, a_{-i})$. Therefore there must be some $a'_{-i} \in \bar{\tilde{A}}_{-i}$, 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. □

Lemma. *5 For any Strong Nash equilibrium a^{SNE} of G, $a^{SNE} \in ICIR(\mathcal{C})$ regardless of $\mathcal{C}$.*

Proof. 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 $\mathcal{C}$. □

Theorem. *6 For any $\mathcal{C}$-Negotiated Binding Agreement, s^*, and any $h \in H$, $s^*(h) \in ICIR(\mathcal{C})$.*

Proof. 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

$$\begin{aligned} U_i(s'_C, s^*_{-C}|h) &\geq \inf_{a_{-C} \in A_{-C}} u_i(a'_C(a_{-C}), a_{-C}) \\ &> \sup_{a'_{-C} \in A_{-C}} u_i(a_C, a'_{-C}) \\ &\geq U_i(s^*|h) \end{aligned}$$

for all $i \in C$. Concluding that s^* is not a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game.

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. □

Theorem. *7 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.

Proof. 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 to a strategy s'_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 of the negotiation game, and therefore cannot be a $\mathcal{C}$-Negotiated Binding Agreement. □

Theorem. *8 Take any underlying game G 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.

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

   (b) or $a^k_{-C} = \underline{a}^C_{-C}$

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

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

Now I will show that s^* is a $\mathcal{C}$-Negotiated Binding Agreement.

First, I will show that s^* is a $\mathcal{C}$-Subgame Perfect Equilibrium of the negotiation game. First, by assumption, at no history can N deviate as a coalition to improve all their utilities if $N \in \mathcal{C}$, as all $\underline{a}^C$ are weakly Pareto optimal in this case by the definition of $ICIR(\mathcal{C})$. Now assume that some other coalition $C \in \mathcal{C}$ has a profitable deviation. Now,suppose that $a_j \neq s^*_j(h)$ 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. Now suppose that $a_j \neq s^*_j(h)$ 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. □

C Proofs for Appendix A

Proposition. *1 If s^* is a Negotiated Binding Agreement with order then, for all histories for i is active $h \in \tilde{H}_i$, $s_i(h) \in IAD_i$.*

Proof. 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. □

Proposition. *2 Take any order $\mathcal{O}$. a^* is supported in a Negotiated Binding Agreement then it is supported in Negotiated Binding Agreement with order $\mathcal{O}$.*

Proof. 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^*. □

Proposition. *3 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

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

.

Proof. 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 2, where s'_i is selected to intentionally cause perpetual disagreement. □

Proposition. *4 a^* is supported by a Negotiated Binding Agreement with all proposals if a^* is supported by a Negotiated Binding Agreement.*

Proof. 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. □

Proposition. *5 If s^* is a Negotiated Binding Agreement then s^* is a constant outside option Negotiated Binding Agreement.*

Proof. 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. □

Proposition. *6 For all $a \in A^{agree}$, $a \in IIR$.*

Further, if $a^* \in A^{agree}$, then $u_i(a^*) \geq \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i})$.

Proof. We proceed inductively. First by contradiction suppose that $a_i \in D_i(A_{-i})$ while $a_i \in A^{agree}$. As $a_i \in D_i(A_{-i})$ it follows that $\inf_{a'_{-i} \in A_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > \sup_{a_{-i} \in A_{-i}} u_i(a_i, a_{-i})$. Therefore, it follows that $\forall a'_{-i} \in A^{agree}$ for which $\sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}} u_i(a_i, a_{-i})$. Therefore we conclude that for any $a_{-i} \in A^{agree}_{-i}$ we have that $\sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > u_i(a_i, a_{-i})$, violating 2.. Therefore we conclude that $A^{agree} \subseteq \tilde{A}^1$.

Inductively, assume that $A^{agree} \subseteq \tilde{A}^k$ for all $k > 0$, we will show that $A^{agree} \subseteq \tilde{A}^{k+1}$. Suppose not, there is some $a \in \tilde{A}^{k+1}$ such that $a \notin A^{agree}$. It follows that for some $a_i \in A^{agree}_i$, while $a_i \in D_i(\tilde{A}^k_{-i})$. As $a_i \in D_i(\tilde{A}^k_{-i})$ it follows that t $\inf_{a'_{-i} \in \tilde{A}^k_{-i}} \sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > \sup_{a_{-i} \in \tilde{A}^k_{-i}} u_i(a_i, a_{-i})$. Therefore, it follows that $\forall a'_{-i} \in A^{agree}$ for which $\sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > \sup_{a_{-i}} u_i(a_i, a_{-i})$. Therefore we conclude that for any $a_{-i} \in A^{agree}_{-i}$ we have that $\sup_{a'_i \in A_i} u_i(a'_i, a'_{-i}) > u_i(a_i, a_{-i})$, violating 2.. Therefore we conclude that $A^{agree} \subseteq \tilde{A}^{k+1}$. Therefore we can conclude that $A^{agree} \subseteq IIR$.

Finally, $u_i(a^*) \geq \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i}) = \underline{u}_i, \forall a^* \in A^{agree}$ is immediately implied by 2.. □

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.

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.

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 Repeated Games. *Games and Economic Behavior*, 1(4):327–360.

Feenstra, R. C. (2015). *Advanced international trade: theory and evidence*. Princeton university press.

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.

Greenberg, J. (1990). *The theory of social situations: an alternative game-theoretic approach*. Cambridge University Press.

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.

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

Ismail, M. (2021). The Strategy of Conflict and Cooperation. *Available at SSRN 3785149*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rubinstein, A. (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.

Von Neumann, J. and Morgenstern, O. (1994). *Theory of games and economic behavior*. Princeton university press; 2nd edn. 1947; 3rd edn. 1953, commemorative edition, 2007.

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.