Negotiated Binding Agreements

Malachy James Gavan*

Universitat Pompeu Fabra and Barcelona School of Economics†

July 4, 2022

Working Paper Version

Abstract

I study a negotiation protocol with potentially infinitely many periods to reach an agreement over what will be played in a normal form game. In each period, agents publicly make a proposal of the action they will take. The protocol terminates whenever all players make the same proposal twice, at which point an agreement is made to play this proposal, and no further deviations can take place. The payoff is that of the agreed-upon action profile. When there is perpetual disagreement the payoffs are assumed to be anywhere between the $\limsup$ and $\liminf$ utility of the proposals made. I study subgame perfect equilibria where players only propose actions that will be agreed upon and agreement is made from the initial history. I refer to this solution concept as *Negotiated Binding Agreements*. I provide easy-to-check necessary and sufficient conditions for the outcomes that can be supported by Negotiated Binding Agreements and explore a number of applications. I show that these conditions are robust to perturbations in the negotiation procedure including timing of proposals, proposing action profiles, and variation in the payoff of perpetual disagreement. I show that the necessary and sufficient conditions generalise when coalitions may jointly deviate in a cooperative way and show that these are consistent with perturbed versions of the β-core.

Keywords: Agreements, Negotiation, Cooperation

JEL: C70, C71, C72

1 Introduction

Negotiations and their resulting agreements play an important role within the economy. For instance, the purchase of a house is often negotiated in multiple rounds before a sale is made. When agents reach an

**email:* malachy.gavan@upf.edu.

†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.. This has significantly improved both this work as well as my outlook on economics and beyond. I have also had innumerable useful discussions with Alexander Frug and Pia Ennuschat, for which I am greatly indebted. I am also grateful to (in alphabetical order) Nemanja Antic, Josefina Cenzon, Vincent Crawford, Faruk Gül, Gilat Levy, Raquel Lorenzo, Zoel Martín Vilató, Rosemarie Nagel, Debraj Ray, Danila Smirnov and seminar participants at UPF, the BSE PhD jamboree, the 12th Conference on Economic Design, and the International Conference on Game Theory and Applications. All faults are my own.

agreement, the sale goes ahead at the stipulated terms without further deviations. A similar process exists for job offers. Cartels may negotiate price fixing and committees may negotiate contributions to a public good. All such situations involve complex agreements that may depend on many things. For instance, an agreement surrounding the sale of a house may stipulate the price of the house, the renovations the seller must undertake before the sale takes place, and the moving date. A job negotiation may stipulate working hours and conditions rather than just pay. However, due to the difficulty in modelling such negotiations, there has been little progress in using tools that reflect this within theory. Broadly speaking, this is due to two reasons. Firstly, there are a class of negotiation procedure and companion solution concepts that ensure agents always negotiate in an optimal way, regardless of what has previously occurred within a negotiation (Kalai, 1981; Bhaskar, 1989; Chwe, 1994; Mariotti, 1997; Ray and Vohra, 2019). These concepts are reasonable but often intractable, due to the consistency of rationality they require and the richness of behaviour that these procedures allow for. Secondly, a class of concepts instead focus on what can be sustained under the assumption that behaviour that may not be consistent with rationality can occur in the event of a deviation from the candidate outcome (Aumann, 1959, 1961; Currarini and Marini, 2003). These concepts take a cooperative view that abstracts from why such behaviour takes place. Doing so allows for a much larger degree of tractability. However, they do not capture what can reasonably be used to prevent deviations and therefore can be considered unreasonable for many applications.[1] In a similar vein, other concepts abstract from the negotiation process *within* a group and take a cooperative perspective that focuses on Pareto undominated actions that prevent new groups from breaking and forming, understanding that such groups would act with the same behaviour (Ray and Vohra, 1997; Diamantoudi et al., 2007). These may capture what can be sustained by such threats and are often tractable, but they do not speak to what can be achieved purely within a group itself. Due to these polar difficulties, such concepts are often left as purely theoretical and do not allow for a broad use for applications. Within this paper, I define a model of negotiation and corresponding solution concept that provides tractable results, while maintaining consistent and fully rational beliefs, laying down a bridge between several of these approaches.

To do so, I consider the following negotiation protocol over the behaviour players should take in a game.

[1] 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, and such punishments need not even be agreed upon.

Suppose that, within a period, agents can make a proposal of the action they will take. In doing so, they may consider all proposals, both their own and others, from previous periods. Hence, agents' strategies in the negotiation map each history of past joint proposals to a new individual proposal. They continue making proposals in this form, with potentially many rounds, until the same proposal is made for two consecutive periods. At this stage, agents jointly agree to make this final proposed action within the game, no further deviations are permitted, and each agent receives the payoff of the resulting action profile. When there is perpetual disagreement the payoffs are assumed to be anywhere between the $\limsup$ and $\liminf$ utility of the proposals made.[2] This is consistent with the interpretation that an outside party decides what will be implemented in this case. However, she is restricted to respect the agents' proposals, and can only implement a proposal that has been made sufficiently often. It is also consistent with the limiting case of the probability of the proposal profile being implemented in period t being $1-\delta$ while they continue to the next round probability δ, when the probability of continuation is taken to 1. As this negotiation protocol defines a dynamic game with complete information, I explore a refinement of the subgame perfect equilibria where agreement is made in the continuation of the initial history and I restrict players to only propose actions that will be agreed upon on the continuation of *some* history, which I will refer to as a *no babbling* condition.[3] This refinement captures two elements. Firstly, as this paper is concerned with the agreements that can be made, this refinement ensures that the uninteresting equilibria that result in perpetual disagreement are not considered. Secondly, by restricting attention to equilibria that are no babbling, this prevents unreasonable proposals from ever being made, in the sense that agents would not agree to play such an action. I refer to these as *Negotiated Binding Agreements*. I refer to the action profiles of the baseline game that are selected by a Negotiated Binding Agreement as *supported* by a Negotiated Binding Agreement. I will refer to this as the baseline negotiation procedure throughout and show that the results of the paper are consistent with a number of perturbations of this procedure, that I will discuss shortly.

This model allows for full rationality with respect to individual preference, where agents understand the

[2]The game is similar to that of Mariotti (1997), both in terms of histories and strategies. The only difference in the game is that the payoff for perpetual disagreement of Mariotti (1997) is taken to be $-\infty$. The solution concept in this paper is substantially different to that of Mariotti (1997), as will be discussed in the literature review.

[3]This assumption can embed a form of no delay equilibrium used within bargaining games with a large number of players (Chatterjee et al., 1993), where the proposal would have to be the one from the continuation of the history in consideration.

implications of a deviation perfectly, and these implications are justifiable via the use of subgame perfection.[4] As this game has infinitely many histories, with different types of terminal histories, this is a complex object to consider. However, the solution concept proposed, Negotiated Binding Agreements, and corresponding dynamic game of negotiation allows for a tractable solution. I outline those results here. Firstly, I show that a necessary condition for any Negotiated Binding Agreement is that all actions proposed survive iterated elimination of individually irrational actions. An action is a is individually irrational if, given the most optimistic beliefs an agent can have when evaluating a, the payoff is still worse than the minimum payoff they 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. Secondly, it must be that the payoff received is necessarily higher than the min max payoff a player could receive, after iterative deletion of individually irrational actions. The conditions of deletion and payoffs are easy to implement and check for any finite game or game with differentiable utility. I go on to provide sufficient conditions. If only a single action profile survives iterative deletion of individually irrational actions it must be the agreed-upon action profile in *any* Negotiated Binding Agreement. More generally, I show that if for each player there exists an individual "punishment" action profile in the baseline game such that: a) the payoff for any other players' punishment is better than the payoff of the player's own punishment, b) within the baseline game each player best responds to the action profile of their punishment, then such profiles survive iterated elimination of individually irrational actions and any action profile in the baseline game that gives each player a higher payoff than their individual punishment can be supported by a Negotiated Binding Agreement. This is similar to the approach of individual punishments used in infinitely repeated games, for example in Fudenberg and Maskin (1986) and Abreu et al. (1994), which will be discussed further within the paper. This immediately implies that any action profile in the baseline game that Pareto dominates a pure Nash equilibrium can always be supported in Negotiated Binding Agreement. In two player games where the action set of each agent is a compact subset of a metric space and utility is continuous the existence of such a profile of action profiles is also necessary for an action profile to be supported by Negotiated Binding Agreements, therefore in this class of games the described general sufficient conditions fully characterise what can be supported

[4]This can be seen as in contrast with a strain of the cooperative literature that does not consider a justification for why such strategies take place in the event of deviations.

by Negotiated Binding Agreements. I go on to consider a refinement of Negotiated Binding Agreements, where there is no delay in agreements, that is that agents agree within two periods after every partial history. This is similar to the no delay equilibrium of Chatterjee et al. (1993). I refer to these as No Delay Negotiated Binding Agreements. I show that, in games where the action space is a compact subset of a metric space and utility is continuous for each player, an action profile can be supported by No Delay Negotiated Binding Agreement *if and only if* there are action profiles that work as the described player specific punishment.

I use a simple three firm Bertrand model as a leading example, that displays the key results of the paper and the intuition behind them. In this example, in any action profile supported by a Negotiated Binding Agreement, the firm with the lowest marginal cost must at least partially serve the market. I go on to explore applications of a public goods game and a Cournot Duopoly with linear demand. Within the public goods game, I fully characterise the set of action profiles that can be supported in a Negotiated Binding Agreement, which includes both full cooperation and full defection. In a simple Cournot Duopoly with linear demand with potentially heterogeneous marginal costs, 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. In all three settings, I show that the general sufficient conditions fully characterise the set of outcomes that can be supported in a Negotiated Binding Agreement.

Next, I show that the necessary and sufficient conditions are robust in a number of ways. In particular, the necessary and sufficient conditions hold in the following perturbations. (i) if agents make proposals sequentially rather than simultaneously within a period then the sufficient conditions holds, while agents may only propose actions from those that survive iterated elimination of absolutely dominated actions in the sense of Salcedo (2017), which in many cases is the same as those that survive iterated elimination of individually irrational actions. (ii) if agents may make proposals of the action profiles, so as long as the payoff of perpetual disagreement are taken to be between the $\limsup$ and $\liminf$ of the induced sequence that takes the proposals agents make of their own action in each period. (iii) if the result of perpetual disagreement is that each agent believes that the worst agreement for them will be played by all other players, but they may deviate. I also show that if the payoff of an infinite history with no consecutive repetition is instead exogenously set to be worse than any agreement, as is typically taken to be the case in bargaining games, then the Negotiated Binding Agreements in the baseline model remain to be Negotiated Binding Agreements in this perturbed model.

In the second part of this paper, I allow for the possibility of cooperative agreements within the very negotiation process. I do so by allowing coalitions of agents to jointly choose a new strategy, and will do so if it is profitable for all agents within the coalition. I allow for any possible set of coalitions to be permissible, be that a partition of players, or allowing for coalitions to overlap. I allow for externalities across coalitions via the use of a game. To capture the possibility of agents acting in such a way, I define the concept of $\mathcal{C}$-Negotiated Binding Agreement, when no coalition in a predefined set $\mathcal{C}$ can profitably deviate at any history. Further, as in Negotiated Binding Agreement, a $\mathcal{C}$-Negotiated Binding Agreement requires a no babbling condition, where coalitions only propose actions they would agree to on the continuation of some history. I show that the natural extension of the baseline necessary and sufficient conditions hold in this setting, where coalitional versions of both individually irrational and player specific punishment are used. An action a is coalitionally irrational if the coalition could provide a function, that maps the action profiles of those outside the coalition to a joint action of the coalition that, even in the worst case, ensures those within the coalition a higher payoff that even the best case from playing a. This is performed iteratively, ruling out all action profiles that are coalitionally irrational round by round, reducing the set of actions the coalition considers for those outside, for all permissible coalitions $C \in \mathcal{C}$. Within $\mathcal{C}$-Negotiated Binding Agreement, players must only make proposals within the set of actions that survives iterated elimination of coalitionally irrational actions. Further all permissible coalition must play jointly *coalitionally rationally*, where they cannot improve the utility of all members, understanding that agents outside of the coalition will choose some constant action from the set that survives iterated elimination of coalitionally irrational actions, therefore ensuring consistent across coalitions which iterated elimination of coalitionally irrational actions may not. I argue that these conditions can be viewed as a perturbed version of the β-core of Aumann (1961). I provide sufficient conditions that use coalition specific punishment and argue that these can be viewed as a further refined version of β-core, where further joint consistency is required. These conditions are generally demanding, but do provide sharp results in some interesting applications. I apply this concept to a linear best response Cournot model with fixed costs. I show that the Pareto efficient outcome, where agents equally divide the monopoly quantity can be sustained in $\mathcal{C}$-Negotiated Binding Agreement.

The remainder of the paper is organised as follows. In section 2 I outline the model of negotiation and define the concept of Negotiated Binding Agreements and provide a leading example. Section 3 provides necessary and sufficient conditions for Negotiated Binding Agreement outcomes. Section 4 provides applications. Section 5 defines $\mathcal{C}$-Negotiated Binding Agreement. Section 6 provides necessary and sufficient conditions for $\mathcal{C}$-Negotiated Binding Agreement outcomes. Section 7 provides an application of $\mathcal{C}$-Negotiated Binding Agreement. Section 8 discusses the related literature and section 9 concludes the paper. Appendix A provides the results of robustness to perturbations.

2 Model

Let the 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 $a_i \in A_i$, with a joint action $A = \times_{i\in N} A_i$ 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 game. In particular, A need not be finite nor compact, nor u_i be continuous.

I will 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. Within a period, each agent observes the history of all previous proposals of all agents and will make a proposal of the action they will take in G. Hence, agents' strategies in the negotiation game map each history of joint proposals to a new individual proposal. If all agents make the same proposal that they made in the previous period, that action profile is implemented in a binding way. Therefore the payoff of such a history is the payoff of the agreed upon action. If not, they continue to the next round and continue the same process until an agreement is made. When there is perpetual disagreement the payoffs are defined to be the $\limsup$ and $\liminf$ utility of the proposals made.

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 implemented. It is also consistent with there being some probability of the current periods proposal being accepted, taking this probability to 0. I define this formally below.

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.

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*.

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 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 $z = (a^1, ..., a^k) \in Z'$, that is a history that ends in agreement, let $U_i(z) = u_i(a^k)$ for all $i \in N$.

Whenever $z = (a^1, a^2, ..., a^k, ...) \in Z''$, 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 assumes that the payoff of agents is 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 implemented, in a way that is known to the agents. Within finite games G it can also be interpreted as taking any weighted average of the proposals made infinitely often along the entire path. Note that such a utility function always exists in this class of game and may meaningfully use *all* proposals sufficiently far along the path.[5]

This specification for the payoff of perpetual disagreement may 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) and the overtaking criterion (Rubinstein, 1979). This can also be interpreted as a non-discounted version of the condition used within Kimya (2020). I reserve a discussion of the relation of this model to those for the literature review. Further to this, this restriction is consistent with 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. This is formalised by the following lemma.

[5]To see this note the following construction. For any history h let h^T denote the sequence that takes the first T elements of the sequence. For a history h denote the consequence $(v^1, v^2..., v^T, v^{T+1}, ...)$ to be such that $v_i \in \mathbb{R}^n$ such that $v_i^T = \sum_{a \in \{A|a \in h\}} \frac{|\{k \in \mathbb{N}| a^k = a, a^k \in h^T\}|}{T} u_i(a)$. Let $Conv(h) = \{v \in \mathbb{R}^n | \exists (v^{l_1}, v^{l_2}, ...) \subset (v^1, v^2..., v^T, v^{T+1}, ...)$ s.t. $v = \lim_{k \to \infty} v^{l_k}\}$. As $u_i(A)$ is bounded it follows that v_i^T is bounded for all T, and therefore so is v^T. Therefore there exists a convergent subsequence in $\mathbb{R}$. Therefore $Conv(h)$ is non-empty. Further, whenever the limit of the mean utility, $\lim_{T \to \infty} \sum_{a \in A} \frac{|\{k \in \mathbb{N}| a^k = a, a^k \in h^T\}|}{T} u_i(a)$ is well defined, $Conv(h)$ is a singleton containing only this value for each player. For $\sup[Conv(h)]_i$, the utility is only defined on infinite sequences, and therefore if there is a strict relation on all proposals after some finite T we ensure the relation translates to $u_i(h)$. Further, $\sup[Conv(h)]_i \geq \inf_{a^t, t>T} u_i(a^t)$ for any finite T. Therefore if $\inf_{a^t, t>T} u_i(a^t) > u_i(h')$ it follows that $\sup[Conv(h)]_i > u_i(h')$. Note that $\sup_{a^t, t>T} u_i(a^t) \geq \sup[Conv(h)]_i$ and therefore if $u_i(h') \geq \sup_{a^t, t>T} u_i(a^t)$ then $u_i(h') \geq \sup_{a^t, t>T} u_i(a^t)$. Further, as this value is bounded, we have the desired properties.

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

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

I reserve this proof and all other proofs for appendix C.

At each partial history $h \in H$ the strategy of $i \in N$ dictates the proposal i would make in the next round for player i: $s_i : H \to A_i$. 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 equilibria (SPE) and refinements of SPE can be seen as the appropriate solution concept.

Definition (Subgame Perfect Equilibria)**.** 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. 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* $h \in H$

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

Note that SPE trivially include many perpetual disagreement outcomes. As this work is primarily focused on the agreements that can be supported by some equilibria, I focus on SPE that reaches an agreement from the initial history. I will also restrict attention to the case where proposals of individuals can only be used if there is some continuation in which said proposal would make up part of an agreement. This rules out agents proposing actions that they would never agree to. I will refer to such property as agreements having *no babbling*. Note that similar concepts have been used within the literature of bargaining. For instance, the study of no delay equilibrium by Chatterjee et al. (1993), where the proposals can only be made if they could be accepted.[6] I will refer to this concept as *Negotiated Binding Agreement*.

Definition 1 (Negotiated Binding Agreement)**.** *s^* is a Negotiated Binding Agreement supporting $a^* = a(s^*|\emptyset)$ if:*

a) s^* is a subgame perfect equilibrium.

b) No babbling: $\forall h \in H$, $\exists h' \in H$ such that $s_i^*(h) = a_i(s^*|h')$.

Further motivation can be found behind ensuring that the agreement is no babbling, as it rules out the possibility of making proposals that are always payoff irrelevant on the equilibrium path, which 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, even in the event an agreement is eventually reached. For instance, consider the following example.

Example 1. Take the following game

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

Notice that the individual min max for player 1 is given by 1, while the individual min max for player 2 is given by 2. By lemma 2, players cannot receive below their min max payoff.

[6] This is in the view of efficiency due to discounting in their model while in this model discounting is not assumed.

Now, consider the following SPE.

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

On the path of play, starting from the initial history, this induces a path such that (U, L) is proposed within the first two periods and the 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 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. 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 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. However, 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 game, from a history h such that $s^*(h) = (U, C)$, it is clear the strategy only delays the eventual agreement of (U, L), and therefore we do not have a Negotiated Binding Agreement. ▼

I now turn to a leading example.

2.2. Leading Example and Preview of Results

Here I provide a leading example to illustrate the key ideas in the paper. The environment will be a 3 player Bertrand Oligopoly with heterogeneous marginal costs and a unit demand. Specifically, there are three firms, $N = \{1, 2, 3\}$, who may each set an integer price $p_i \in \{0, 1, 2, 3, ..., 10\}$, where 10 is the highest price a firm may set.[7] Each firm faces a marginal cost c_i. It is assumed that: $c_1 = 1$, $c_2 = 3$ and $c_3 = 4$. The firms face a unit demand and the firms who set the lowest price share the demand equally. Therefore utility is given by their individual demand multiplied by the price they set minus their marginal costs. Therefore, when they do not set the lowest price they receive a utility of 0. Formally,

$$u_i(p) = \begin{cases} \frac{p_i - c_i}{|\operatorname{argmin}_{j \in \{1,2,3\}} p_j|} & \text{if } i \in \operatorname{argmin}_{j \in \{1,2,3\}} p_j \\ 0 & \text{if } i \notin \operatorname{argmin}_{j \in \{1,2,3\}} p_j \end{cases}$$

Note that in any Nash equilibrium firms 2 and 3 gain a utility of 0, while firm 1 gains a utility of at least 1. For instance, the profile of prices given by $(p_1^*, p_2^*, p_3^*) = (2, 3, 3)$ is a Nash equilibrium, yielding a payoff of 1 for player 1 and 0 for both other players.

First, we will understand what cannot be supported in a Negotiated Binding Agreement. Firstly, can any firm setting a price of 0 be supported in a Negotiated Binding Agreement s^*? If this were the case, then firm i would receive a strictly negative utility in equilibrium. However, such an agent could deviate to a new strategy that proposes a price of c_i at all partial histories; $s_i'(h) = c_i, \quad \forall h \in H$. If they do so, by the definition of U_i, we have that $U_i(s_i', s_{-i}^*|\emptyset) = 0 > U_i(s^*|\emptyset)$. This is *regardless* of s_{-i}^*, as either a) s_i', s_{-i}^* ends in agreement such that firm i agrees to c_i and improves their utility orb) s_i', s_{-i}^* results in perpetual disagreement, in which case, the payoff is defined with respect to the proposals along the path of s_i', s_{-i}^*, and therefore is defined by proposals involving only c_i for firm i. In either case, the utility is 0. In other words, setting a price of 0 is *individually irrational*, which will be formally introduced and discussed in the next section, by setting a price of c_i. This is as no matter the prices that others set, which may change depending on the price you set, the min max payoff, which is weakly higher than that of setting a price of c_i,

[7] It could equally be assumed that setting a price beyond 10 faces demand 0 without a change in the results.

is higher than the payoff of setting a price of 0. Therefore it cannot be that any firm setting a price of 0 can be supported by a Negotiated Binding Agreement. By no babbling, it cannot be 0 is proposed at any history.

We have concluded that there is no Negotiated Binding Agreement in which 0 is *ever* proposed by any firm. With this, let us now consider whether any Negotiated Binding Agreement can support firm 2 or 3 agreeing to a price of 1. It must be that they would receive a strictly negative utility, as they certainly serve at least part of the market as 1 is the minimum price that can be set now we have eliminated the possibility of others setting a price of 0, and is certainly below their marginal cost. Using the same logic as before, we can see that proposing a price of c_i is certainly better than agreeing to set a price of 1. In other words, for firm 2 or 3 a price of 1 does not survive *iterated elimination of individually irrational actions*, which will be formally defined and discussed in the next section. Therefore it cannot be that either firms 2 or 3 setting a price of 1 can be supported, and it cannot be that firm 2 or 3 proposes 1 at any history. Using the same induction, we may conclude that firm 1 will also never propose a price of 1 and firms 2 and 3 will never propose a price of 2.

In conclusion, it is necessarily the case that in any Negotiated Binding Agreement no proposal of firm 1 is less than 2, and no proposal of firms 2 and 3 is less than 3. Given that we have ruled out what cannot be supported by Negotiated Binding Agreement, we will go on to examine what can be supported. Firstly, notice that for any vector of prices that are supported by a Negotiated Binding Agreement it must be that firm 1 receives a payoff greater or equal to 1. As firms 2 and 3 never propose less than 3, if firm 1 were to deviate to a constant strategy that proposes a price of 2 for all histories firm 1 would guarantee themselves a payoff of at least 1. This is regardless of whether the history ends in perpetual disagreement or an agreement. Given this, we may conclude that any Negotiated Binding Agreement must provide firm 1 with a payoff of at least 1 to ensure such a deviation is not profitable. Similarly, firms 2 and 3 must receive a payoff of at least 0. We will now show that *any* such profile is supported, denoted by p^* such that $u_1(p^*) \geq 1$, $u_2(p^*) \geq 0$, and $u_3(p^*) \geq 0$. Consider the following strategy profile. In the initial history, $s_i^*(\emptyset) = p_i^*$. In the history where p^* has been proposed in the first round, players again propose p_i^* and the game terminates: $s_i^*(p^*) = p_i^*$. For any other history, let $s_1^*(h) = 2$, $s_2^*(h) = s_3^*(h) = 3$. Now let us consider whether a deviation is profitable.

For firms 2 and 3 there is no incentive to deviate from the proposals at the initial history or the history p^*, as this will lead to firm 1 proposing 2 in all subsequent periods, and therefore either the negotiation ends in an agreement for which firm 1 plays 2, which implies that the deviating firm receives a profit of at most 0, which is weakly worse than not deviating. Similarly, if the negotiation results in perpetual disagreement, it must be the payoff is with respect to firm 1 proposing a price of 2, which again cannot give a payoff greater than 0. Similarly, they cannot deviate from any other history. A similar logic holds for firm 1.

In conclusion, we have that any vector of prices p^* such that $u_1(p^*) \geq 1$, $u_2(p^*) \geq 0$, and $u_3(p^*) \geq 0$ can be supported by a Negotiated Binding Agreement.

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. 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 of we can therefore remove said actions from consideration. Now, upon deleting such actions, we proceed inductively, where 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 min is taken *over the set of actions that survives iterated individual rationality*, then it does not survive iterated deletion of individually irrational actions. In this subsection, I show that any action that is proposed must survive iterated deletion of individually irrational actions. The formal definition of individual rationality and iterated individual rationality is formally defined below.

Definition 2 (Individually Irrational actions given $C_{-i} \subseteq A_{-i}$)**.** *$a_i \in A_i$ 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 $D_i(C_{-i})$.

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. Whenever the game being considered has well defined maxima and minima I will refer to them as such, rather than using the infimum and supremum. This notion is similar to the notion of absolute dominance by Salcedo (2017), simultaneously developed in Halpern and Pass (2018), instead of comparing the best case of one action and the worst case of another we compare based on the best case of an action compared to the inf sup. 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), as well as Salcedo (2017)'s absolute dominance, and therefore if single action is not individually irrational it is also obviously dominant.

Definition 3 (Iterated Deletion of Individually Irrational 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 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.

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. 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 price of 0 or 1 for any firm and firms 2 and 3 proposing a price of 2.

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

Let $B_i^0 = A_i$. Let $B_i^k = B_i^{k-1} \backslash NBR_i(A_{-i}^{k-1})$. Let $B^k = \times_{i \in N} B_i^k$ and $B_{-i}^k = \times_{j \neq i} B_j^k$. Let the set of actions that are 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 it also survives iterated elimination of iterated deletion of individually irrational actions: $IENBR \subseteq IIR$.*

Typically, even more profiles may survive iterated elimination of never best responses to pure actions. To see this, consider the following game.

Example 2. Consider the following prisoners' dilemma game.

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

D is strictly dominant for both players, hence 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 that C can receive is 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 is contained in IIR.

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

Using a similar logic to example 2, we can find a chain of action profiles such that: for each player there is an action profile such that it prescribes their best respond to the action prescribed to others in that profile, while the other profiles give them a weakly higher utility. Note, joint with lemma 1, this implies all profiles that Pareto dominate a pure Nash equilibrium 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. *If $\{a^*, \underline{a}^1, .., \underline{a}^n\} \subseteq A$ satisfy:*

1. $\underline{a}_i^i \in \operatorname{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$
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, that combine the necessary conditions in theorem 1 with individual rationality considerations 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 prisoner's dilemma game as in example 2.

Example 2. revisited Again consider the following prisoners' dilemma game

$1\backslash 2$	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 the min max restricted to the set of actions that survives iterated elimination of individually irrational actions is always weakly higher than the min max without this restriction.

Remark 1. *For any 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 firm 1 is 0, via other firms setting prices of 0, however the min max payoff when we restrict ourselves to IIR is 1.

The results of this section bear resemblance to the analysis of infinitely repeated games, where individual rationality constraints must be satisfied. I reserve this discussion for the literature review in section 8.

Finally, before moving to the sufficient conditions, note that in 2 player 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 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 the a^* to their punishment. This is formalised by the following theorem.

Theorem 3. *For any 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^*, 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)**.** *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 would be 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 follow.

For the more general conditions, we require that each agent has a specific action profile, 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 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 higher payoff than the punishment for i. 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. 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 here 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 as there are no future payoff to compensate for abiding to the punishment, as the agreement to play such an action is binding and the game terminates. I discuss the reasoning for this disparity further within the literature review. I state this formally in the following theorem.

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

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

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 are indeed supported by a Negotiated Binding Agreement. This is immediately implied by the fact any pure Nash equilibrium, 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 equilibria can be sustained by this reasoning. However, in games where no pure Nash equilibria exist there may exist a Negotiated Binding Agreement due to the above sufficient conditions.

Example 3. Take 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	$\underline{4}$,4	0,$\underline{12}$
M	4,$\underline{4}$	0,0	$\underline{2}$,3
D	$\underline{12}$,0	3,$\underline{2}$	1,1

Notice that there is no pure Nash equilibrium in this 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 2 player 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 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 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 minmax 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.

Further justification for these 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, which I refer to as No Delay Negotiated Agreements, the sufficient conditions are also necessary in games where the action space is a compact subset of a metric space and utility is continuous. This 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.

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 game G such that A_i is 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 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 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 Bertrand game in the leading example. However the other Nash equilibrium of the game, where $p_1^* = 3$, $\min\{p_2^*, p_3^*\} = 4$ would provide the same logical result, albeit changing the lower bar of utility that firm 1 must receive.

With these results, I now turn to some applications.

4 Applications

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

In application 1, I explore a public good game and fully characterise the set action profiles that can be supported in Negotiated Binding Agreement, and show that both full and no contribution can be supported. In application 2, I consider a simple Cournot Duopoly with potentially heterogeneous marginal costs. 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) $N = \{1, 2, ..., n\}$. 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. 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 $\limsup$ 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 and the result of said theorem 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 and therefore cannot be a Negotiated Binding Agreement. ▼

Application 2. (Cournot Duopoly with Heterogeneous Marginal Costs) Consider a simple Cournot Duopoly model where $q_1, q_2 \in [0, b] = A_i$ where the inverse demand curve 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 $c_1 \geq c_2$. 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 static best responses 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

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

Consider the following strategy for proposals for 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. $s^*(\emptyset) = s^*(h) = (q_1^*, q_2^*)$ such that $h = (q^1, q^2, ..., (q_1^*, q_2^*))$, $s^*(h') = (0, \underline{q}_1^1)$, 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. Notice that 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. 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 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 can be represented by the following graph:

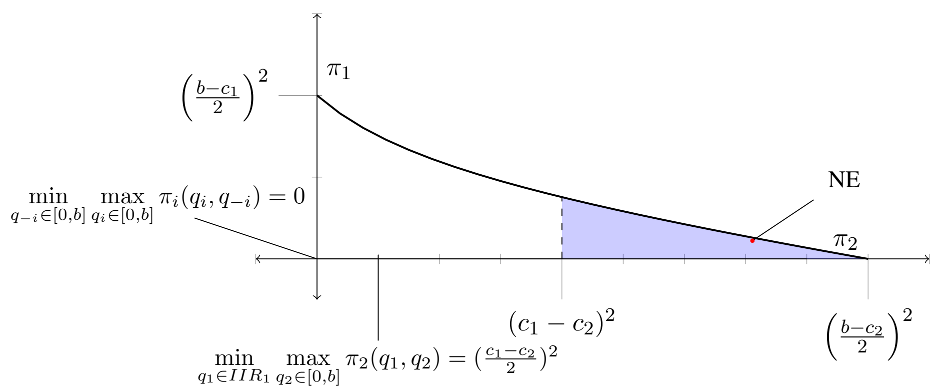

Figure 1: $c_2 < c_1 < \frac{c_2+b}{2}$

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.[8] ▼

[8]Note that if $c_1 > \frac{b+c_2}{2}$ then the only outcome that can be supported by a negotiated binding agreement is $q_1^* = 0, q_2^* = \frac{b-c_2}{2}$

5 Coalitional Deviations

In principle, a negotiation may be susceptible to a collection of agents making binding agreements over how they will play *within* the negotiation process itself. To address these concerns, I now extend the analysis to allow for this possibility. To do so, I allow 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.[9] However, if it were possible to make a binding agreement to not make a 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 when a deviation occurs then this approach ensures no misspecification. That is, do we allow for agents within a coalition to have veto power? Do we allow for 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 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.

To do so, 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

[9] This would be in line with a concept of renegotiation proofness a la Farrell and Maskin (1989) and Bernheim and Ray (1989), I discuss this further in the extended literature review in appendix B.

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

With this, I go on to define the natural extension of subgame perfect equilibria 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, 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 Equilibria). *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 includes 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 equilibria are 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 the coalitional equilibrium concept of Ray and Vohra (1997) to extensive form games.

I reserve a more in-depth discussion of the relation of this concept with these within the literature review.

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, that has no babbling, I will require instead that it is a $\mathcal{C}$-subgame perfect equilibrium, that has a form of no babbling. Before formally defining this concept, note that the use of $\mathcal{C}$-subgame perfect equilibria 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 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 supporting $a^* = a(s^*|\emptyset)$ if:*

1. *s^* is a $\mathcal{C}$-subgame perfect equilibrium*
2. *$\mathcal{C}$-no babbling: $\forall h \in H$, $\exists h' \in H$ such that $s^*_C(h) = a_C(s^*|h')$.*

Again, 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 any action proposed at some history of $\mathcal{C}$-Negotiated Binding Agreement must survive a procedure of iterated deletion of coalitionally irrational actions. This procedure works inductively as follows. Consider some joint action of those within a coalition $C \in \mathcal{C}$, 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.[10] Notice this is essentially the notion of Aumann (1961)'s β-core. We may proceed inductively. Remove all coalitionally irrational actions for all coalitions $C \in \mathcal{C}$. 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 survive 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 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 coalitionally rational. 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 coalitional rationality and iterated elimination of coalitionally irrational joint actions is formally defined below.

Definition 7. *For a coalition C, a joint action $a_C \in A_C$ is coalitionally irrational with respect to $B_{-C} \subseteq$*

[10]For games with compact action spaces and continuous utility they are identical, outside of this case, coalitional rationality is a stronger notion.

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 $\mathcal{C}$)**.** *Let $\tilde{A}^0(\mathcal{C}) = A$. For $m > 0$ let $\tilde{A}^m(\mathcal{C}) = \tilde{A}^{m-1}(\mathcal{C}) \backslash \left[\bigcup_{C \in \mathcal{C}} \left[[D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})] \times A_{-C}\right]\right]$.*

Let the set of action profiles that survives iterated elimination of coalitionally irrational actions, or those that are iteratively coalitionally rational, with respect to $\mathcal{C}$ be denote $ICIR(\mathcal{C})$ 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 in finite games. To see this, consider the following example.

Example 4. Consider the following 2 player game where $\mathcal{C} = \{\{1,2\},\{1\},\{2\}\}$, i.e. both players may make unilateral deviations and may write a binding agreement at any history.

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

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

However, it may be non-empty, even when a rich set of coalitions are permitted. 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 $ICIR(\mathcal{C}) \subseteq ICIR(\mathcal{C}')$. That is, 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 where $\mathcal{C} = \{\{1,2\},\{1\},\{2\}\}$, i.e. both players may make unilateral deviations and may write a binding agreement at any history.

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, regardless of the coalition configuration, is the existence of a strong Nash equilibrium.[11]

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 $ICIR(\mathcal{C})$ and $\mathcal{C}$-Negotiated Binding Agreement, 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 may be used.

[11] 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 $i \in C$.

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

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

This is an immediate implication of theorem 6. Note that this is possible, and may imply that there is no Negotiated Binding Agreement that is robust to the concerns of coalitions for a specific coalition structure $\mathcal{C}$.

Further to this, a result analogous to theorem 2 holds. This result will state that at any history h, s^* must give a payoff that is coalitionally rational for any coalition C, 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, 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. Note the similarity to the β-core of Aumann (1961). Within the β-core the payoff of equilibria 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.

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

I reserve an in-depth discussion of the relation of this concept to other related concepts for the extended literature review in appendix B.

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, but 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 result.

Example 6. Consider the following two player game and consider the richest set of coalitions $\mathcal{C} = \{\{1\}, \{2\}, \{1, 2\}\} = 2^N \setminus \{\emptyset\}$.

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

I have labelled the weakly Pareto efficient outcomes 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 $ICIR(2^N \setminus \{\emptyset\}) = \{(TT, LL), (TT, R), (T, R), (D, LL), (D, L)\}$. 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 $[ICIR(2^N \setminus \{\emptyset\})]_2 = \{LL, L, R\}$, is given by 2. Therefore we conclude that despite the fact that TT, R is Pareto efficient, and provides a higher payoff than the mix max over all possible profiles, it cannot be sustained in a $2^N \setminus \{\emptyset\}$-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 prevent deviations from equilibrium based on the identity 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 sustainable 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 $ICIR(\mathcal{C}) = \{a^*\}$, then a^*, then s^* is a $\mathcal{C}$-Negotiated Binding Agreement if and only if $s_i^*(h) = a_i^*$ 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. 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. 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, holding the punishment used against them constant. Note, inclusion of such profiles in $ICIR(\mathcal{C})$ is now required and not implied due to the rich deletion that can take place. This is formalised by the following theorem.

Theorem 8. *Take any 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 exists, it must be within $ICIR(\mathcal{C})$, implies that any strong Nash equilibrium can be supported in a $\mathcal{C}$-Negotiated Binding Agreement. However, these conditions can apply in games with no strong Nash equilibrium, and therefore are a more general set of conditions. To see this, consider the following example.

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

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

Here there is no strong Nash equilibrium. In fact, as there is no pure Nash equilibrium there is no pure coalition proof Nash equilibrium. However, the conditions of theorem 8 apply and from the previous analysis we know that $ICIR(2^N \backslash \{\emptyset\}) = \{(TT, LL), (TT, R), (T, R), (D, LL), (D, L)\}$. To see this, take, for example, $\underline{a}^N = a^* = (TT, LL)$, $\underline{a}^1 = (D, L)$ and $\underline{a}^2 = (T, R)$. Therefore TT, LL can be sustained in $2^N \backslash \{\emptyset\}$-Negotiated Binding Agreement. ▼

The sufficient conditions 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 of those outside of a coalition may be used in order to prevent deviations, whereas here we must satisfy additional conditions to ensure such a profile 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 individual rationality.[12]

I now turn to an application.

[12]Shubik (2012) examines the 78 2x2 games which can be induced by strict ordinal preferences, of these 78, 67 allow for the sufficient conditions to be applied. Note that is only 2 less than the existence of Nash equilibrium in pure strategies. In this sense, the sufficient conditions apply to more scenarios than initial inspection may suggest.

7 Application of $\mathcal{C}$-Negotiated Binding Agreement

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

Application 3. (Symmetric Cournot with Fixed Cost) Consider a simple model of Cournot with fixed costs. These fixed costs depend on the total number of firms 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 $m - \sum_{j=1}^{4} q_j$, where $m > 0$. I assume that the marginal cost is constant and symmetric, therefore, without loss, I set it to 0. Therefore gross profits for player $i \in \{1, 2, 3, 4\}$ are given by $(m - \sum_{j=1}^{n} q_j)q_i$. Let fixed costs take the following form: $\left(\frac{3}{32} m \sum_{j \neq i} \mathbf{1}_{q_j > 0}\right)^2 \mathbf{1}_{q_i > 0}$. 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:

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

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

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

Notice that this implies the following:

1. There is a Nash equilibrium when 2 firms enters the market, and both choose quantities of $q_i^* = \frac{m}{3}$. This leads to a payoff of $\frac{943}{9216} m^2$ for the firm who enters and 0 for those who do not. There are

no other Nash equilibrium of 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 coalition-proof Nash equilibrium.

2. There are many Pareto efficient outcomes. Any outcome such that the firms who sell, in aggregate sell the monopoly quantity, $\sum_{i=1}^{2} q_i = \frac{m}{2}$, 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 $\frac{m}{8}$ leads to profits of $u_i(q^{p,all}) = \frac{55}{1024}m^2 > 0$.

3. Any profile that is such that at least one player receives a payoff of at least $\frac{55}{1024}m^2$ 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 the efficient outcome where all agents produce in strong Negotiated Binding Agreement. That is, an agreement such that $q_i^* = \frac{m}{8}$ 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 $q_{-C}^k = \frac{16+\sqrt{137}}{64}m$ if $|C| = 3$, $q_{-C}^k = \left(\frac{16+\sqrt{137}}{64}m, \frac{16+\sqrt{137}}{64}m\right)$ if $|C| = 2$ and $q_{-C}^k = \left(\frac{16+\sqrt{137}}{64}m, \frac{16+\sqrt{137}}{64}m, 0\right)$ if $|C| = 1$

   Then:

   - $s_i^*(h) = \frac{16+\sqrt{137}}{64}m$ for $i = \min_{j \notin C} j$ if $|C| \leq 3$ or $i = 1$ if $|C| = 4$.
   - $s_i^*(h) = \frac{16+\sqrt{137}}{64}m$ for $i = \min_{j \notin C \backslash \{\min_{j \notin C} j\}} j$ 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 $s_i^*(h) = \frac{m}{8}$

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 has 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 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 $\frac{16+\sqrt{137}}{64}m$ in every period. As $\frac{16+\sqrt{137}}{64}m > \frac{m}{8}$ 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 $\frac{249-32\sqrt{137}}{4096}m < 0$ 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 as 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. ▼

Before concluding the paper, I turn to some related literature.

8 Literature Review

Here I outline the key areas of literature and the main related papers within them. Within appendix B, I provide a more in detail discussion of the relation to these papers and a wider body of work.

Mariotti (1997) has the closest model to this paper. Histories take a similar form, both terminal and partial, whereas in Mariotti (1997) the payoff of perpetual disagreement is normalised to $-\infty$. Mariotti (1997) similarly allows for a fully general specification of permissible coalitions. However, Mariotti (1997) looks at a different form of equilibrium, where coalitions make new proposals only if it is strictly in their benefit to do so from the current proposal. In this sense, Mariotti (1997) takes a solution concept closer to those used in cooperative theory, whereas this paper takes a more non-cooperative approach. Mariotti (1997) provides some general necessary conditions and some sufficient conditions for two player games. Both Kamada and Kandori (2020) and Caruana and Einav (2008) consider models of revision strategies, where agents may revise their strategies before some final time and the last proposed action is implemented in a binding way. This is a similar sense in which a twice repeated profile is binding within the model of this paper, however, in my model the time taken to have this binding agreement is not bounded. Kalai (1981) also looks at a model of negotiation, where agents may make proposals, where agents have finitely many periods to reach an agreement, 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)[13] and show that only cooperation can be sustained in a prisoners' dilemma game. Bhaskar (1989) examines a model of pre-play agreement for a symmetric 2 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 this proposal sequentially. If there is a sequence of 3 proposals, 2 for 1 player and 1 for the other, where the player who has proposed first and last does not revise on her last proposal, then the prices are implemented. Bhaskar (1989) looks at perfect equilibria of such a game and concludes that only the monopoly price can be sustained.

A number of papers have provided cooperative solutions for normal form games. Chwe (1994); Nakanishi (2009); Ray and Vohra (2019) all consider versions of the farsighted stable set, with Ray and Vohra

[13] These are subgame perfect equilibria that do not permit the use of weakly dominated strategies at any history.

(2019) being the closest to $\mathcal{C}$-Negotiated Binding Agreement as it looks for well defined strategies to pin down the stable set. The concept of Ray and Vohra (2019) and Chwe (1994) are similar to that of Mariotti (1997), while Mariotti (1997) demands full optimality with respect to the strategies of others in comparison to Chwe (1994) and allows for a general utility function in comparison to Ray and Vohra (2019). Both allow for a fully general specification of permissible coalitions. Ray and Vohra (2019) is similar to this paper, however assumes that there are no externalities across coalitions, in this paper I instead allow for externalities via the use of a game. Ray and Vohra (2019) provide general conditions for existence in games with a general coalition structure in games with transferable utility. Aumann (1959, 1961) defines strong Nash equilibrium, the α-core, and the β-core. In this paper, my solution of $\mathcal{C}$-Negotiated Binding Agreement lies somewhere between the β-core and strong Nash equilibria, as agents are permitted to change their proposals when they observe a proposal of others change, but can only do so in a way consistent with rationality, and must be pinned down by an optimal strategy in the sense of equilibrium. In the work of Aumann (1961) the traditional cooperative approach of using any means to ensure the agreement is kept is used. The sufficient conditions provided in this paper are also close in nature to the β-core, where mutual coalitional rationality consistency of punishments is required. Bernheim et al. (1987) look at coalition proof Nash equilibria, which has a similar flavour of private agreements as $\mathcal{C}$-Negotiated Binding Agreements, however in that case they are non-binding, and as there is no observability as with proposals in the negotiation game, all other players outside of the agreement choose a constant strategy. Chander and Wooders (2020) define a notion of coalitional subgame perfect equilibrium for 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, but their solution is shown to be related to a perturbed version of the α-core. Herings et al. (2004); Ambrus (2006, 2009); Grandjean et al. (2017) all consider versions of coalitional rationalizability, all of which are different to the concept of iterated elimination of coalitionally irrational actions.

In contract theory, Jackson and Wilkie (2005); Yamada (2003); Ellingsen and Paltseva (2016) all consider models of contracting where agents may all have an input into the contracts that are proposed and agreed upon. Negotiated Binding Agreement has a similar flavour in this respect, as all agents have a meaningful impact, i.e. rather than a take-it-or-leave-it offer, on the action they take. Kalai et al. (2010) and Peters and Szentes (2012) look at a notion similar to Tennenholtz (2004) program equilibria, where agents may contract over contracts. This allows agents to specify reactions to deviations in full, and can allow for these to be fully specified at a higher level also. They do not consider this is a cooperative way, 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.

This paper can also be seen in the light of the Nash agenda pointed to in Nash (1953), as the sufficient conditions $\mathcal{C}$-Negotiated Binding Agreement can be seen as a perturbed version of the β-core, while it is the result of a fully specified game with fully consistent behaviour, and would be the equilibrium even when less general deviations are permitted. Notable contributions to this literature include Rubinstein (1982); Chatterjee et al. (1993) on bargaining. Recently, a working paper of Ismail (2021) looks at strategic cooperation in the view of this agenda, where an extensive form is constructed to allow cooperation to be a choice. In $\mathcal{C}$-Negotiated Binding Agreement, the underlying incentive cooperation is taken to be implicit rather than explicit.

A number of papers consider a form of communication for equilibrium selection. Notable examples are of Bernheim et al. (1987)'s coalition proof equilibrium, where coalitions are permitted to deviate, but can only do so in a non-binding way and therefore deviations must be self-enforcing a la Nash. Farrell and Maskin (1989) and developed simultaneously by Bernheim and Ray (1989) develop a similar concept of renegotiation proof equilibrium, where the grand coalition may deviate to a preferred SPE at a point in a repeated game. The closest work within this literature to this paper is that of Rabin (1994). Rabin (1994) explicitly models a negotiation over the choice of equilibrium, rather than the implicit process by the above papers.

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. It may also embed a concept similar to that of the overtaking criteria of Rubinstein (1979). The sufficient conditions within the paper are also similar to the sufficient conditions used within infinitely repeated games where player specific punishment is used, for instance, in Fudenberg and Maskin (1986) and Abreu et al. (1994). Further relation to these results is discussed in the appendix.

9 Conclusion

In this paper, I propose a tractable model of negotiation. This is represented by proposals within a period, where if agents all propose the same action in two consecutive periods a binding agreement is made to play said action profile. The payoff of perpetual disagreement is taken to be between the $\liminf$ and $\limsup$ of the payoffs induced by the proposals. I study a form of subgame perfect equilibrium where agents only propose the actions they expect to take upon agreeing at some history, therefore there is no babbling. I refer to this as a *Negotiated Binding Agreement*. I provide necessary and sufficient conditions for the outcomes and strategies of a Negotiated Binding Agreement and go on to explore two key applications. Firstly, I show that in a public goods game cooperation can be supported by a Negotiated Binding Agreement. 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 any Negotiated Binding Agreement. In both examples, I fully characterise the Negotiated Binding Agreements.

I go on to show that the results provided are robust to perturbations in the negotiation process. Specifically, if the proposals are sequential rather than simultaneous, then the sufficient conditions hold, as well as necessary conditions tightly related to those of the baseline case. Further, if agents make proposals of all agents' actions, rather than just their own, and the payoff for infinite histories is only pinned down by the proposal of their own action, then the necessary and sufficient conditions hold. If the result of perpetual disagreement is that each agent believes that the worst agreement for them will be played by all other players, but they may deviate, then necessary and sufficient conditions hold. Finally, if the payoff of an infinite stream of proposals is taken to be worse than any agreement, the sufficient conditions are still valid.

Finally, I show how the necessary and sufficient conditions for the case where agents may only act unilaterally naturally generalise to the case where agents may act jointly, and show the link of these conditions to a perturbed version of the β-core of Aumann (1961).

A number of questions remain open. Firstly, there is an opportunity for applied theory to use such a concept where appropriate. 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). Due to the easy-to-use conditions, the negotiation process, as well as the corresponding results of this paper, may also provide some interesting insights in some applied theoretical settings.

On a broader level, a further examination of reasonable deviations, and the expansion of the set of equilibria that this would allow calls for further attention. Additionally, extending the model to allow for asymmetric information, and therefore agreements occur without full knowledge of the outcome, provides an interesting, albeit challenging, question. I leave these for future work.

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.

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.

Bolton, P. and Dewatripont, M. (2004). *Contract theory*. MIT press.

Börgers, T. (1993). Pure strategy dominance. *Econometrica: Journal of the Econometric Society*, pages 423–430.

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.

Caruana, G. and Einav, L. (2008). A Theory of Endogenous Commitment. *The Review of Economic Studies*, 75(1):99–116.

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.

Coase, R. H. (1960). The problem of social cost. *The Journal of Law and Economics*, 3:1–44.

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., Xue, L., et al. (2007). Coalitions, agreements and efficiency. *Journal of Economic Theory*, 136(1):105–125.

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.

Friedman, J. W. (1971). A non-cooperative equilibrium for supergames. *The Review of Economic Studies*, 38(1):1–12.

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554.

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

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.

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.

Kalai, E. (1985). Solutions to the bargaining problem. *Social goals and social organization: Essays in memory of Elisha Pazner*, pages 77–106.

Kamada, Y. and Kandori, M. (2020). Revision games. *Econometrica*, 88(4):1599–1630.

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.

Mailath, G. J. and Samuelson, L. (2006). *Repeated games and reputations: long-run relationships*. Oxford university press.

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.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

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. (2019). Maximality in the farsighted stable set. *Econometrica*, 87(5):1763–1779.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. *Journal of economic theory*, 21(1):1–9.

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

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109.

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

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

Scarf, H. E. (1971). On the existence of a coopertive solution for a general class of n-person games. *Journal of Economic Theory*, 3(2):169–181.

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

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

Shubik, M. (2012). What is a solution to a matrix game.

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

Vartiainen, H. (2014). Endogenous agenda formation processes with the one-deviation property. *Theoretical Economics*, 9(1):187–216.

Yamada, A. (2003). Efficient equilibrium side contracts. *Economics Bulletin*, 3(6):1–7.

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 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 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 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 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 so important for driving the results.

A.1. Robustness to Order of Proposals

As in section 2, let G be a 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 equilibria for this model here:

Definition (Subgame Perfect Equilibria). *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 equilibria.

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 generically, 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 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 D.

Further to this, the following proposition shows that the sufficient conditions 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 game with bounded payoffs.

Define the negotiation game with order as follows.

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 Equilibria). *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 equilibria.

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 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. With this, I present one last robustness result for Negotiated Binding Agreement.

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 of any agreement. 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 remain to be true. However, the sufficient conditions 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 agreement 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 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 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 Extended Literature Review

This paper falls is related to a number of areas within game theory. I outline those areas here, and discuss the most related papers within each strand of literature.

B.1. Models of Pre-Play Agreement

In terms of modelling choices for such negotiations, Mariotti (1997) is close to this paper. Specifically, the model of Mariotti (1997) is close to the analysis presented for strong Negotiated Binding Agreement in section 5. Mariotti (1997) studies an environment close to that of Greenberg (1990)'s coalitional contingent threats. Within Mariotti (1997), the model allows for strategies to be dependent on the history of proposals, in a similar sense to this paper. Both individuals and coalitions of players have a strategy of the actions they will propose, and are separate objects, while within this paper there is only a single joint strategy in equilibrium, while, in principle, the work of Mariotti (1997) may allow for many. Similarly to this model, a history terminates when a proposal is made twice sequentially for all agents and coalitions. When no proposal is made twice sequentially in an infinite history, Mariotti (1997) normalises the payoff to $-\infty$, ensuring that the payoff of any agreement is certainly better than no agreement. Instead, I allow for the payoff of no agreement to be pinned down within the game itself. The equilibrium concept of Mariotti (1997) also looks for short agreements in the flavour of no delay. Specifically, Mariotti (1997) prevents any agent or coalition from making a different proposal from the one they are currently making unless it is strictly in their benefit, given the strategy of others. This in, in essence, ensures that in equilibrium only one coalition can make a new proposal within a period. The set of equilibria satisfying these conditions are referred to as full coalitional equilibrium. I do not restrict to this case, and instead assume that equilibrium strategies reach a terminal history within finitely many periods and satisfy no babbling, but it may be that this involves more than one coalition deviating from a previous proposal. Mariotti (1997) provides some necessary conditions and sufficient conditions for two player games.

Recently, the paper of Kamada and Kandori (2020) studies two player revision games. Within this class of games there is a sequence of random revision opportunities where both players can simultaneously revise their proposal before some finite stopping time. Once the stopping time is reached, the action that has been most recently proposed is implemented without further deviation. They extend this idea to the case where agents may have opportunities to revise at different times. It is shown that players can sustain cooperation and look at necessary and sufficient conditions for cooperation to be sustained within a prisoners' dilemma. Within my model the opportunities to revise are deterministic and can potentially continue infinitely. Further, my model allows for results to be provided for n player games. Additionally, I provide conditions for a strong solution concept, where agents may jointly change their strategy. Similar to Kamada and Kandori (2020), my paper essentially uses trigger strategies for sufficiency.

Similarly, Caruana and Einav (2008) look at a game where players can unilaterally change their proposed action before some finite stopping time but have to pay a switching cost. These proposals are observed. As the deadline approaches the switching cost becomes higher and higher so there is almost full commitment at this point. Once the stopping time is reached, the action profile that was most recently proposed is implemented without further deviations. The model of this paper instead looks at a model where there are no switching costs. Further to this, time is potentially infinite. Additionally, the model within my paper additionally explores the possibility of joint deviations.

Kalai (1981) examines a negotiation procedure where players have rounds of negotiation. The procedure takes the following form. Let there be n possible periods. In period 1 all agents announce the action they intend to take. This is made public, and agents then may deviate. If any agent chooses to deviate from their proposal at the end of the period, they are then required to take this action, and may not deviate from this in further rounds. At the beginning of the next round, those who have not previously deviated at the end of a period are permitted to announce a new strategy. This process is iterated until no agent deviates at the end of the round, in which case the action profile that has previously been announced is implemented in the game. The solution concept used is perfect equilibrium introduced by Selten (1988). These are equilibria that are both subgame perfect and play weakly dominated strategies with probability 0 in all periods. Kalai (1981) shows that in a prisoner's dilemma, whenever there are more than 2 periods of possible announcements, only cooperation can be sustained. Within my model, those who deviate from their equilibrium strategy to be restricted in such a way, i.e. it may be that a deviation is non-constant and does not rule out the use of weakly dominated strategies.

B.2. Models of Bargaining and the Nash Agenda

Broadly speaking, this paper, especially within section 5 can be seen as providing a non-cooperative model to justify a cooperative solution concept. This falls in line with the literature of bargaining models and the Nash agenda (Nash, 1953). Most notably, Chatterjee et al. (1993) provides a similar model for a bargaining game with many players, where the key question is what agreements can be achieved without delay. In this model there are proposals and acceptance of what will be carried out. The payoff of infinite disagreement is normalised to 0, below the payoff of any agreement. This extends the model of Rubinstein (1982), however when more than 3 players are permitted many more outcomes may be equilibrium. My model does not strictly fall into the models of bargaining as the payoff of no agreement is not normalised to be worse than any agreement. However, as shown in appendix A, the sufficient conditions remain to hold if this were the case. Indeed, it is clear to see that the constructions used within this section would also work when coalitions may deviate.

Further to this, the philosophy of these questions is largely different, and looks for axiomatic reasons for the solution concepts for the bargaining problems, whereas within this paper it is merely the result of the equilibrium of the non-cooperative game that justifies the conditions. A further discussion of this philosophy can be found in Kalai (1985).

B.3. Cooperative Models

A number of papers have asked how to define notions of cooperative solutions for normal form games, I outline a few here and how they relate to this paper, and specifically how they relate to the conditions when we include coalitional concerns, in section 5.

Aumann (1961) defines two concepts of the core for normal form games, the α-core and the β-core. These concepts are earlier discussed in Aumann (1959). For an action profile to be the α-core, it is required that there is no action for a coalition C such that for all possible actions of all other players, would be profitable for all members of the coalition C. Further to this, Aumann (1961) defines the β-core, which requires that a single punishment can be used to prevent a coalition's deviation. There is no restriction on credibility of the actions of those outside of C. This is related to the concept of the β-core with respect to $ICIR$, where the actions used by other agents must be justified as they are part of a fully specified equilibrium, which is consistent at all histories, as discussed within section 5. This partially answers the question posed by Scarf (1971), who provides conditions for existence of the α-core, that is, can we define a notion of which actions are reasonable to support the core in normal form games. The sufficient conditions used for strong Negotiated Binding Agreement have a similar flavour to this, with the caveat that, again, they must be jointly justified as explained.

Aumann (1959) also defines the strong Nash equilibrium, which requires no coalition can deviate and improve the utility of all agents while leaving the action of other agents the same. In this paper, I show that, when coalitions are permitted to deviate, the concept of $\mathcal{C}$-Negotiated Binding Agreement is weaker than this. Similarly, Bernheim et al. (1987) define coalition proof Nash equilibrium. This allows agents to deviate as a group, so long as that deviation itself is self-enforcing, and no further private deviations within the group can be profitable for those deviating agents. As my concept allows for binding agreements, I do not restrict attention to self-enforcing deviations. However, an interesting question does arise: what can be achieved in $\mathcal{C}$-Negotiated Binding Agreement if he deviations themselves face the same criticism as equilibrium itself. I leave this for future work.

Chander and Tulkens (1997) explore a similar problem. They define the γ-core, where coalitions may deviate from some proposed equilibrium, and understand that all other agents will act in their own best interest. With this, the payoffs in the γ-core have to provide a higher payoff for all agents that they could receive under a pseudo-Nash equilibrium, where a coalition can attempt to improve all members utilities by jointly deviating, and all other individuals cannot improve their utility via a unilateral deviation. This is similar in some sense to the restriction of actions of others being justifiable within this paper, however I do not require that no coordination can take place once a deviation occurs.

Chander and Wooders (2020) also define a notion of coalitional subgame perfect equilibrium, however this is for games with transferable utility, where a coalition's deviation is pinned down with respect to the best subgame perfect equilibrium assuming all other players act without cooperation. In comparison to the solution of $\mathcal{C}$-subgame perfect equilibria, where agents' behaviour is pinned down by a single strategy, which is dependent on the proposals they observe, rather than knowledge of coalitions who have deviated, which is implicit in Chander and Wooders (2020). The paper of Chander and Wooders (2020) show that their solution concept is related to a perturbed version of Aumann (1959) α-core.

Ray and Vohra (1997) provide a concept of coalitional equilibrium, which is taken with respect to some predefined coalition configuration $\mathcal{C}$, which is a partition of players. This concept is such that a coalition cannot deviate, holding the action of all other agents constant, and improve the utility of all agents within the coalition. The definition of $\mathcal{C}$-subgame perfect equilibrium can be seen as the extensive form generalisation of coalitional equilibrium, especially when $\mathcal{C}$ is taken to be a partition. Ray and Vohra (1997) also examine equilibrium binding agreements, which explores which coalitions would form and the resulting agreements. They allow coalitions to break, resulting in a new partition that is a refinement of the previous configuration. Diamantoudi et al. (2007) explores a similar question, but allows for coalitions to reform, which is not permitted under Ray and Vohra (1997). This avenue of coalition formation is not explored within the framework of this paper, and rather opts to explore agreements when the coalitional configuration is given. Therefore this provides an open question.

Ray and Vohra (2019) explore when a concept of equilibrium of a negotiation game can be embedded within a farsighted stable set of Chwe (1994). This is similar in flavour to this paper in the sense I study a well-defined negotiation game and explore concepts of cooperation that are justifiable within this setting.

Application 1 of Negotiated Binding Agreement reflects the work of Nakanishi (2009). Within this work, Nakanishi (2009) shows that in a generalised prisoner's dilemma, in a version of Chwe (1994)'s farsighted stable set, which only allows for individual deviations, cooperation can be sustained. The key difference between the concept within this paper and that of Nakanishi (2009) is, that Negotiated Binding Agreement are pinned down by strategies in an underlying negotiation procedure, that are always optimal, rather than deviations leading to a sequence of deviations that improve utility after a deviation, but may not be fully optimal.

In this paper I define a notion of cooperative iterated dominance that is used for finding the possible outcomes of agreements. One related concept is that of coalitional rationalizability of Ambrus (2006) and the related concepts of Ambrus (2009). Coalitional rationalizability is a form of rationalazibility that is driven by individual behaviour and non-binding private communication. That is, coalitional rationazability let's agents who are seen as a coalition jointly rule out some actions to be played. If this restriction is improving for all agents in comparison to those action profiles excluded from the agreement, holding beliefs of those outside of the coalition constant, where an individuals beliefs are concentrated on the set of action profiles consistent with this agreement, then agents will jointly rule these action profiles out. Iterated Elimination of Coalitionally Irrational actions does not require that beliefs surrounding the actions those outside of the coalition is held constant. Further, the structure of Iterated Elimination of Coalitionally Irrational actions may imply that the resulting set does not have a product structure. Additionally, this process does not fall into the concept of sensible coalitional best response operators of Ambrus (2009), as the focus within that work is also on ensuring individuals within coalitions are satisfied, rather than a coalition as a whole. The simplest way to see this is restriction (ii) of sensible coalitional best response operators with Ambrus (2009), which requires at least some agent is best responding to the belief. This is not necessary within my work, as commitment may be made.

Herings et al. (2004) define a notion of coalition rationalizability for social situations. They do so by defining an extensive form game induced by the social situation, and provide a corresponding form of rationalizability. In my paper I do not consider social situations, but rather a less general model of games. I also do not use an extensive form game to define the notion of iterated elimination, as it is enough to focus purely on the game itself within this case.

Grandjean et al. (2017) define the strong curb set. This is a form of coalitional rationalizability for static games. This concept looks for a set $B \subset A$, such that a) B has a product structure and b) B_C contains all actions that any coalition C may take, given that the beliefs of the actions of agents outside the coalition are concentrated on B_{-C} and are kept constant when comparing between two possible coalitional actions. The concept iterated elimination of coalitionally irrational actions does not require that the beliefs of others actions are kept constant.

B.4. Contract Theory

Broadly speaking, as this paper is concerned with the agreements that can be made and the outcomes that they support, can be seen as related to the literature on contract theory. For the most part, this literature focuses on the case where transfers are permitted and a single agent can make a proposal for a contract, be that an agent with the game or a mechanism designer. A review of these canonical models can be found in Bolton and Dewatripont (2004). In this model, the contract, which in this paper is a naive agreement in the sense it only states the action that will be taken, and no other contingencies, is the result of agents acting in their own self-interest within a well-defined game.[14] A few papers have a similar flavour, where the contracts are the endogenous result of a game where all players have meaningful input.

Jackson and Wilkie (2005) consider an endogenous model of contracts that take the form of side payments. They study a two period process where, for a norm form game, they allow for agents to first commit to outcome dependent side payments that will be paid and received on top of the payment received from the outcome of the game. They examine the subgame perfect equilibrium of this game. They show that an efficient agreement may fail to exist in a two player game. Further, they show that many inefficient agreements may exist within 3 or more player games, while an efficient agreement certainly exists. They do not allow contracts to depend on contracts. This model is extended by Yamada (2003), where agents can commit to reject some transfers. This allows for efficient agreements to be made within 2 player games. Within my model, there are no transfers. Further, the structure of how an agreement is reached is different. In negotiated, the implicit contract, in the language of this paper the agreement, specifies only the action that will be taken, and does not allow for further deviations to take place after this. Further, I allow for a subset of agents to make an agreement amongst themselves, outlining how they will act within the negotiation it self. This is not permitted in Jackson and Wilkie (2005) nor Yamada (2003).

Ellingsen and Paltseva (2016) consider a model where agents may choose whether to take part in a negotiation, then, if they are part of the negotiation they propose contracts, if all members who face some commitment within the contract agree, this contract is implemented, and then the game that results from the contract is played. This, in essence, allows for agents to endogenously choose which contracts to be a part of and whether to take part in the negotiation at all. In this environment, with three or more players, efficiency may necessarily not be achieved. This can be seen as casting doubt on the so called Coase Theorem as posed in Coase (1960). In my model, a similar process exists within section 5, as coalitions require that all agents wish to be a part of the agreement to ensure an agreement takes place. However, I do not make the same

[14]Of course, a game that has an action space that allows for contingencies in the game itself can be negotiated upon.

restriction within the base model, and it is presumed from the outset that all agents are part of the negotiation exogenously.

The models of contracting over contracting, with notable examples in Kalai et al. (2010) and Peters and Szentes (2012), all for agents to write contracts, where the arguments that these contracts take are the contracts of others. A similar notion exists within Tennenholtz (2004) program equilibrium. This allows for the dispensing of a social planner. Further, a large set of potential outcomes can be the result of equilibrium for games with complete information, within these models, including efficient agreements. These models do not allow for communication over which contracts, or contracts over contracts, should be posed, and therefore do not allow for joint deviations.

B.5. Repeated Games

Both in terms of modelling and results, this paper has similarities to infinitely repeated games. Firstly, in terms of modelling, this paper takes a similar approach to the limit of the mean criteria for evaluating payoffs where no agreement is made. That is, payoffs are taken to the be the limit, taking T to infinity, of the mean of payoffs that would result from stopping at period T. If such a payoff is well defined, it is necessarily between the lim inf and lim sup. This is notably used in Aumann and Shapley (1994); Rubinstein (1994). In this setting, as in equilibrium all payoffs are defined by terminal histories that reach agreements, and therefore defined by a single action profile, rather than potentially many action profiles within these models and in terminal histories that reach no agreement in this model, this is equivalent to using the overtaking criteria of Rubinstein (1979), whenever the limit of the mean is well defined.

I now outline the related results within this literature. Further discussion of these results may be found in chapter 8 of Osborne and Rubinstein (1994).

Firstly, Folk theorems exist for infinitely repeated games using the limit of the mean criteria. Specifically, again subgame perfect equilibrium of an infinitely repeated game with either the overtaking criterion or limit of the mean criteria must necessarily give a payoff higher than the min max payoff of the stage game. Further, these are also sufficient when the limit of the mean criteria is taken: payoff that provides each agent with more than their min max payoff, when evaluated in the stage game, can be the result of a subgame perfect equilibrium, as shown in Rubinstein (1994) and Aumann and Shapley (1994). Rubinstein (1979) provides a similar result for infinitely repeated games with the overtaking criterion, with the caveat that the payoffs must be strictly higher than the min max value. Similar results exist for infinitely repeated games with hyperbolic discounting, notably, see Friedman (1971). In comparison to this paper, I show that all Negotiated Binding Agreement payoffs must be higher than the perturbed min max, when other agents are restricted. I show that this is weakly higher than the min max payoff. The structure of the game within this paper, which means that many finite histories are terminal, as well as the restriction on no babbling, jointly drives the disparity between the two models. Jointly, these restrict the strategies that players can use in two ways. Firstly, a strategy that typically could be used as a punishment within an infinitely repeated game, where the same action profile punishment for deviating occurs for a number of periods, cannot be used in the same way due to the structure of the game. This is as these strategies may cause the game to terminate, and these strategies do not offer future rewards for adhering to this punishment. This point is made more relevant by the no babbling condition, which ensures agents may only propose actions they would be willing to agree to.

Similarly, Rubinstein (1980) studies infinitely repeated games with both the overtaking or limit mean criteria where the solution concept is strong subgame perfect equilibrium. Within Rubinstein (1980) it is shown that, for the limit of the mean criteria, for any payoff available in the stage game such that no coalition can deviate and strictly improve the utility of all its members, regardless of what those outside of the coalition can do, then said payoff can be sustained in strong perfect equilibrium. Rubinstein (1980) show that the set is substantially changed when the overtaking criteria is used. Similar to the disparate between the results of this paper and Aumann and Shapley (1994); Rubinstein (1994) with Negotiated Binding Agreement, the same disparities arise in $\mathcal{C}$-Negotiated Binding Agreement, here the actions in the stage game that can be used to prevent deviations are limited. This is for identical reasoning to the previous paragraph.

Similar concepts of ensuring that coordination cannot entail an improvement for a group, while maintaining internal consistency in the solution concept, that is, ensuring that equilibrium is not only efficient and a subgame perfect equilibrium, but requiring that it is not sustained by threats that are not efficient, exist for repeated games. Weak renegotiation proof equilibrium given by Farrell and Maskin (1989) and developed simultaneously by Bernheim and Ray (1989) under the name of *consistency* is conceptually close to the case when individuals make deviations, and so may the grand coalition. That is, suppose that we are planning on using a subgame perfect equilibria that plays the path h by allowing agents to be punished with the subgame perfect equilibria that plays prescribes the history h' to be played from the point of deviation onwards. This is consistent if, once arriving to the strategies dictating that players should play along the path h', it would not be the case that all agents return to the strategy dictated by h. Note that this is a less stringent requirement than that of $\mathcal{C}$-subgame perfect equilibrium when $N \in \mathcal{C}$, as, in this case, to ensure the strategy was permissible we would need to ensure that agents would not want to deviate to *any* other strategy, not only those that would be the result of some equilibrium. Further discussion of this concept and the results surrounding it can be found in chapter 4 of Mailath and Samuelson (2006).

In terms of the structure of the sufficient conditions of Negotiated Binding Agreement and perfectly Negotiated Binding Agreement, the closest is that of the folk theorems shown in Fudenberg and Maskin (1986) and Abreu et al. (1994), which, essentially, rely on player specific punishment. These types of strategies are based on the premise that deviations from equilibrium of player i should be based on a punishment for a player that is invariant to the deviation they have made. Specifically, suppose that in the underlying game there is some collection of actions $\{a^*, a^1, ..., a^n\}$ such that a^i provides i with a payoff strictly higher than their min max. Further, for each i, the utility of the action a^* provides them with a strictly higher utility than a^i, and a^j provides a strictly higher utility than a^i for all $j \neq i$. Then, for a sufficiently high discount factor, there is some subgame perfect equilibrium for which a^* is played in every period. I additionally require that for each i a^i_i is a best response to a^i_{-i}, but otherwise only require weak relations, rather than strict relations. This is to ensure no deviation from punishment can take place. This is not needed within infinitely repeated games as the promise of returning to a^* is enough to prevent this, while within my model terminal histories may also be finite, and therefore there is no opportunity for future rewards. This is made more prevalent via no babbling.

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

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

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

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

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. By no babbling, we conclude that $s_i(h) \notin D_i(A_{-i})$ for any $h \in H$.

Now suppose by contradiction that, for all $j \in N$ $s_j(h') \in \tilde{A}_j^k$ $\forall k < m$ and $h' \in H$ but for some $i \in N$ $s_j(h') = a_i \notin \tilde{A}_j^{m+1}$ for some $h' \in H$. By no babbling it must be that a) $s_{-i}(h') \in \tilde{A}_i^m$ for all h' and b) by no babbling there is some $h \in H$ for which $a_i(s|h) = a_i$. Therefore it must be that $U_i(s|h) = u_i(a(s|h)) \leq \sup_{a'_{-i} \in \tilde{A}^m_{-i}} u_i(a_i, a'_{-i})$. 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. By no babbling, we conclude that $s_i(h) \notin D_i(\tilde{A}^m_{-i})$ for any $h \in H$ and therefore $s_i(h) \in \tilde{A}_i^{k+1}$, 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 $a_i \in B_i^k$ 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 $a_i \in B_i^k$ then $a_i \in \tilde{A}_i^k$. concluding the proof. □

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

1. $\underline{a}^i_i \in \operatorname{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}^i_{-i})$
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 $\underline{a}_i^i \in \operatorname{argmax}_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i) = \arg\sup_{a_i \in A_i} u_i(a_i, \underline{a}_{-i}^i)$ 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 $\underline{a}_i^i, \underline{a}_i^j, a_i^* \in \tilde{A}_i^{m+1}$. $\square$

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 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 $U_i(s'_i, s^*_{-i}|h) \geq \inf_{a_{-i} \in IIR_{-i}} u_i(\tilde{a}_i(a_{-i}), a_{-i}) = \inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i}) - \frac{\epsilon}{2} = \frac{\inf_{a_{-i} \in IIR_{-i}} \sup_{a_i \in A_i} u_i(a_i, a_{-i}) + U_i(s^*|h)}{2} > U_i(s^*|h)$. A contradiction, as therefore s^* is not a subgame perfect equilibrium and therefore not a Negotiated Binding Agreement. $\square$

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

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

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

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

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

Now to show that s^* is a subgame perfect equilibrium. 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. $\square$

Theorem. 5 *For any game G such that A_i is 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 $U_i(s'_C, s^*_{-C}|h) \geq \inf_{a_{-C} \in A_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in A_{-C}} u_i(a_C, a'_{-C}) \geq U_i(s^*|h)$ for all $i \in C$. Concluding that s^* is not a $\mathcal{C}$-subgame perfect equilibrium.

Now suppose by contradiction that $s(h') \in \tilde{A}^k(\mathcal{C})\ \forall k < m$ and $h' \in H$ but $s(h') = a \notin \tilde{A}^{m+1}(\mathcal{C})$ for some $h' \in H$. By definition, it must be that $a \in \bigcup_{C \in \mathcal{C}} [D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C}) \times A_{-C}]$. Therefore it must be that $a_C \in D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})$ for some $C \in \mathcal{C}$. By $\mathcal{C}$-no babbling we have that $\exists h \in H$ such that $a_C = a^*_C(s^*|h)$. By definition of coalition rationality given $\tilde{A}^{m-1}(\mathcal{C})_{-C}$, as $a_C \in D_C(\tilde{A}^{m-1}(\mathcal{C})_{-C})$ there must be some that there is some $a'_C : \tilde{A}^{m-1}(\mathcal{C})_{-C}$ such that $\inf_{a_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a'_C(a_{-C}), a_{-C}) > \sup_{a'_{-C} \in \tilde{A}^{m-1}(\mathcal{C})_{-C}} u_i(a_C, a'_{-C})$. Consider a deviation of C at history h such that $s_C(h') = a'_C(s_{-C}(h'))$

for all $h' \in H$. It follows that

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

. Therefore $U_i(s'_C, s^*_{-C}|h) > U_i(s^*|h)$ for all $i \in C$. Concluding that s^* is not a $\mathcal{C}$-subgame perfect equilibrium. 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, 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, and therefore cannot be a $\mathcal{C}$-Negotiated Binding Agreement. □

Theorem. *8 Take any 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. 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. □

D 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, 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.

By the inductive hypothesis, suppose that $s_i^*(h) \in \tilde{A}_i^m$ for all $h \in \tilde{H}_i$ and $i \in N$. Now suppose by contradiction that $s_i^*(h) = a_i \in D_i(\tilde{A}_{-i}^m)$. It must be that $a_i(s^*|h') = a_i$. Given this, $U_i(s^*|h') \leq \sup_{a_{-i} \in A_{-i}^m} u_i(a_i, a_{-i})$, as $s_{-i}^*(h'') \in \tilde{A}_{-i}^m$ for all $h'' \in \tilde{H}_i$. Further, s^* is a subgame perfect equilibrium, and therefore there is no profitable deviation for s_i^* at any history, including h'. Notice that as $a_i \in D_i(\tilde{A}_{-i}^m)$ then $\exists a_i' \in A_i$ such that $\inf_{a'_{-i} \in \tilde{A}_{-i}^m} u_i(a_i', a_{-i}') > \sup_{a_{-i} \in \tilde{A}_{-i}^m} u_i(a_i, a_{-i})$. 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 $s_{-i}^*(h'') \in \tilde{A}_{-i}^m$ that from this history on the only action profiles proposed are a_i', a_{-i}' such that $a_{-i}' \in \tilde{A}_{-i}^m$. Given this, we can conclude that $U_i(s_i', s_{-i}|h') \geq \inf_{a'_{-i} \in \tilde{A}_{-i}^m} u_i(a_i', a_{-i}') > \sup_{a_{-i} \in \tilde{A}_{-i}^m} u_i(a_i, a_{-i}) \geq U_i(s^*|h')$. A contradiction to s^* be a subgame perfect equilibrium. $\square$

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. 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. A contradiction. Concluding that s' is a Negotiated Binding Agreement with order $\mathcal{O}$, leading to the outcome a^*. $\square$

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, 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.

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, $U_i(s^*|h') \leq \sup_{a_{-i} \in A_{-i}^m} u_i(a_i, a_{-i})$, 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, 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.

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 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.. □