Endogenous Choice of a Mediator: Inefficiency of Bargaining*

[Job Market Paper]

Jin Yeub Kim†

January 20, 2014

Abstract

In this paper, I build a theory of how privately informed parties choose mediators, which is a topic that is not well developed in the literature. I consider a bargaining problem in which two parties with private information about their types negotiate the choice of a mediator. A mediator, in my context, is equivalent to a communication mechanism that respects private information. I use a two-pronged approach, cooperative and noncooperative, to show that the selection of a mediator is endogenous and is driven by the "information leakage" problem. With the noncooperative approach, I consider a *threat-secure* mediator who survives a vote against any alternative mediators. For a benchmark class of examples, I show that the threat-secure mediator exists and is unique. In the cooperative approach, I find that there exists a unique neutral bargaining solution – an incomplete information version of the Nash bargaining solution. I establish that the selected mediators in both approaches are the same. Moreover, the selected mediator is – among all the interim incentive efficient mediators – the *worst* for the parties ex ante and the least peaceful mediator. Therefore, I argue that the very process of selecting a mediator may exhibit an inherent inefficiency in bargaining if the parties already know their types. This paper represents a first attempt to build a bridge between cooperative and noncooperative game theories in the context of bargaining with incomplete information.

Keywords: Cooperative game, Noncooperative game, Bargaining, Mechanism design.
JEL Classification: C71, C72, C78, D82.

*I am greatly indebted to Roger Myerson, Lars Stole, and Ethan Bueno de Mesquita for valuable advice and continuous support. I also appreciate Sandeep Baliga, Yeon-Koo Che, Peter Cramton, Alex Frankel, Thomas Gresik, Jong-Hee Hahn, Jinwoo Kim, Kyungmin (Teddy) Kim, Heung Jin Kwon, David Miller, Cheng-Zhong Qin, Ilya Segal, Hugo Sonnenschein, Richard van Weelden, and seminar participants at the University of Chicago Economic Theory Working Group Meetings, the 24th International Conference on Game Theory at Stony Brook, Summer 2013 Yonsei Economic Research Institute Seminar, 2013 Asian Meeting of Econometric Society at Singapore, and Fall 2013 Midwest Economic Theory Meetings, for helpful suggestions and discussions. First version: November, 2012.

†Department of Economics, University of Chicago, Chicago, IL 60637. E-mail: jinyeub@uchicago.edu, Webpage: http://home.uchicago.edu/~jinyeub.

1 Introduction

Two parties with private information often employ mediators as one of the primary tools of dispute resolution. However, in many economic, political, and social situations, two parties often fail to reach a consensus in their selection of a mediator that resolves conflict peacefully. Despite the significant amount of theoretical work on mediation, the theory of how privately informed parties choose mediators is not well developed. My interest is in understanding the endogenous selection of mediation. Through this paper, I attempt to build a theory of how parties with private information might agree on a mediator and provide a richer understanding of the failure of efficiency in bargaining.

I consider a simple bargaining problem in which two players with private information about their own types – strong or weak – can each choose "war" or "peace." There are also mediators that the players can negotiate over which to choose. In this paper, the definition of a mediator is a person who is not informed about the players' types but who is trying to negotiate settlement between the players while respecting their private information. This setup allows us to consider a mediator to be equivalent to a communication-settlement device, or a mechanism.[1] Therefore, by the revelation principle, I take the space of mediators that are available to the players to be synonymous with the space of incentive compatible and individually rational[2] mechanisms the players can agree on.

Which mediator should the players choose? One might be tempted to think that the players would bargain for an ex ante Pareto dominant solution. That is, the players could possibly select the mediator that is incentive efficient given the other's type and that maximizes the ex ante payoffs of all the players. Indeed, in my setting, there is a unique ex ante incentive efficient mediator that both players might find it focal to choose. However, this naive idea that the ex ante incentive efficient mediator would be chosen is problematic. In particular, if players already know their types at the time they bargain over choosing a mediator, the problem of "information leakage" arises: the fact that a player expresses his preference for a particular mediator conveys information about his type. For this reason, the issue of which mediator would get selected is far from trivial.

My main insight is that the selection of the mediator is endogenous, and the selection reflects this informational concern. Not surprisingly, it is not at all obvious that the players amongst

[1] A mediator is "a person or machine that can help the players communicate and share information" (Myerson, 1991, 250).

[2] No player could gain by being the only one to lie to the mediator about his type or to not participate in the mediation.

themselves would be able to get to the ex ante incentive efficient mediator; and if they do not, it is also not clear which one should be chosen. Taking into account the endogenous information leakage issue, either directly or indirectly, we can think about what are the desiderata for the solution concept. This paper is about understanding how this information leakage issue impacts the endogenous selection of a mediator and which mediator is most likely to arise in the negotiation over mediation. To fully explore this problem, I take two different approaches: cooperative and noncooperative games.

Cooperative Approach – A Non-Strategic Analysis A cooperative approach is about putting some structure on what we think a bargaining solution must satisfy. The cooperative approach asks the following question: What is a reasonable set of mediators we should expect to see arise as an outcome of an undefined communication process within the set of incentive feasible mediators? In this approach, I explicitly take into account the information leakage problem by imposing a set of relevant properties or axioms that must be true for a reasonable solution that captures the tradeoff between the different types of players. I refine the solution set in the negotiating process with various notions of "reasonableness," following the seminal work of Holmström and Myerson (1983) and Myerson (1984*b*).

As a first cut, I use the notion of interim incentive efficiency, which is a minimal requirement in a setting with incomplete information. In my model, there are an infinite number of interim incentive efficient mediators but a single unique ex ante incentive efficient one. In order to select from within this large set of interim incentive efficient mediators, I pose a set of arguably reasonable axioms. These axioms, in particular, imply a solution concept that follows Myerson (1984*b*), called a neutral bargaining solution. The neutral bargaining solution can be thought of as an incomplete information generalization of the Nash bargaining solution. In my setting, not only does there exist a unique neutral bargaining solution, but the solution is not the ex ante incentive efficient choice.

Noncooperative Approach – A Two-Stage Ratification Game The cooperative approach, although illuminating why we should expect to see an ex ante inefficient mediator arise endogenously in the mediator selection process, may not by itself be entirely convincing without also considering a noncooperative game. In the noncooperative approach, I determine the solution set differently than the previous approach by considering an extensive form ratification game that is similar to the approach of Cramton and Palfrey (1995). In this game, I take some candidate mediator as a status quo mechanism. The noncooperative approach asks the following: Given some

status quo mechanism, would the players both agree to switch to another mediator in some sequential equilibrium, noting that information might be revealed by the players' votes, which could affect the continuation payoffs of the players?

Therefore, I must allow for the fact that if one of the players votes in favor and one votes against, the players are going to learn from disagreement, consequently affecting their payoffs. I consider a pairwise test in a ratification game where some uninformed third party can announce an alternative mechanism and the players can switch. From the noncooperative approach, a reasonable restriction on the selection of a mediator is that such a mediator should survive a vote against any alternative options. Whether such a mediator exists is unclear, but if it does exist, we would have a compelling reason to select it. The notion of survivability against all alternatives in a pairwise voting game will be formalized as a concept of what I call *threat-security*. What I find in this context is that there exists only one threat-secure mediator that will always survive the vote against any alternatives, and it is the one that is ex ante Pareto dominated by any alternative mediator.

Equivalence of the Two Approaches In both approaches, the selected mediator is not the ex ante incentive efficient choice. Both approaches not only offer ex ante incentive inefficiency, but also lead to the same inefficient mediator. That is, the threat-secure mediator is exactly the same as the neutral bargaining solution. In general, the neutral bargaining solutions and the threat-secure set are not nested. However, in the class of bargaining games I consider, two distinct ways of looking at the problem lead to the same suboptimal choice of a mediator, which represents, of all the interim incentive efficient mediators, the *most ex ante inefficient*[3] and the least peaceful mediator, whereas the ex ante incentive efficient one is the most peaceful mediator. In this sense, I argue that the process of selecting a mediator itself has an inherent inefficiency in bargaining.

The strategic intuition as to why the ex ante incentive efficient mediator does not survive in the noncooperative game underscores the intuition behind the information leakage problem in the cooperative game. In the cooperative sense, players pick a mediator, effectively pooling in a way that reveals no information. But the nature of that pooling is exactly like signaling in the noncooperative game. The novelty of this paper is that it provides a first attempt to connect a cooperative bargaining theory[4] to a noncooperative ratification game.

[3]Precisely stated, contrary to the ex ante incentive efficient mediator, the most ex ante inefficient mediator is a solution to the program of minimizing the players' ex ante expected payoffs over the set of all interim incentive efficient mediators.

[4]For classical bargaining literature, see Nash (1950), Nash (1953) (the canonical solution concept for two-person bargaining problems), Harsanyi and Selten (1972) (an extension of the Nash bargaining solution for two-person games with incomplete information), and Myerson (1979) (a modified version of Harsanyi and Selten (1972)).

1.1 Literature on Mediation

A significant amount of the theoretical work on mediators and mediation focuses on when and how mediation improves on unmediated communication in the context of international bargaining situations (Kydd, 2003; Hafer, 2008; Fey and Ramsay, 2009; Fey and Ramsay, 2010).[5] Including Fey and Ramsay (2009), recent research on mediation adopts mechanism design tools to study the effectiveness of mediation on preventing conflicts. Fey and Ramsay (2011) show the relation between the kinds of uncertainty states face and the possibility of a peaceful resolution of conflict. In a simple model of conflict, Hörner, Morelli and Squintani (2011) prove that mediation improves the ex ante probability of peace with respect to unmediated peace talks when the intensity of conflict is high or when asymmetric information is significant. Also, in their working paper, Meirowitz et al. (2012) claim that mediation is just as effective as a hypothetical optimal institution with strong enforcement power and the ability to internalize the long-term consequences of its behavior. Much of the literature has focused on the analysis of the effectiveness of mediation, and not on the question of how mediators are selected. My paper is about the endogenous choice of a mediator, combining the cooperative ideas from Myerson (1984*b*)'s neutral bargaining solution with the noncooperative approach based on Cramton and Palfrey (1995).

The remainder of the paper goes as follows: In Section 2, I introduce the benchmark model. Section 3 and Section 4 present the interim incentive efficient set and the neutral bargaining solution under a cooperative approach. In Section 5, I develop and illustrate a concept of threat-security under a noncooperative approach. Section 6 discusses the equivalence between the two approaches and the intuition behind the distortions, and Section 7 concludes. All proofs can be found either in Appendix A or Online Appendix B. Online Appendix C provides an example to illustrate the core ideas of the symmetric interim incentive efficient set, the neutral bargaining solution, and the threat-secure mediators.

[5] Kydd (2003) identifies sufficient conditions for effective mediation. He argues that for a mediator's communication to be credible, the mediator must be biased and endowed with some independent knowledge on the private information of the disputants. Fey and Ramsay (2010) endogenize the process by which information is gained, and show that information mediation has no significant effect on the likelihood of ending a dispute if mediators do not have access to exogenous sources of information, beyond what the disputants relay to them. These results hold in cases where the potential source of conflict is private information of private values. Hafer (2008), in her working paper analyzing a model of contested political authority, shows that neither cheap talk communication nor mediation promote peaceful resolutions or lead to improvements in social welfare. On the other hand, Fey and Ramsay (2009) identify conditions under which mediation achieves peace with probability one.

2 The Model

I consider a simple Bayesian bargaining problem as a mechanism design problem of mediators, in which two players with private information can bargain over selecting a mediator. In my context, I consider a mediator to be represented as a mechanism. The definitions of what follows are done in full generality, although I later come to a specific problem in the context of international conflict in Section 2.2, in which I will be interested in the case of two players, each with two types, and two outcomes.

2.1 The Base Setup

The two-person bargaining problem is characterized by the following structures:

$$\Gamma = (D, d_0, T_1, T_2, u_1, u_2, p_1, p_2),$$

whose components are interpreted as follows. D is the set of feasible bargaining *outcomes* available to the two players from which they can choose. $d_0 \in D$ denotes the disagreement outcome, which the two players get if they fail to coordinate. For each $i \in \{1, 2\}$, T_i is the set of possible types t_i for player i.[6] The players are uncertain about each other's type, and the players' types are unverifiable. For each $i \in \{1, 2\}$, u_i is player i's utility payoff function from $D \times T_1 \times T_2$ into $\mathbb{R}$, such that $u_i(d, t)$ denotes the payoff to player i if $d \in D$ is the outcome and $t \in T$ is the true vector of the players' types. The payoffs are in von Neumann-Morgenstern utility scale. Each p_i is a conditional probability distribution function that represents player i's beliefs about the other player's type as a function of his own type. That is, $p_i(t_{-i}|t_i)$, $i \in \{1, 2\}$, denotes the conditional probability that player i of type t_i would believe about the other player's type being t_{-i}.

Let $T = T_1 \times T_2$ denote the set of all possible type combinations $t = (t_1, t_2)$. For mathematical simplicity, I assume D and T are finite sets. Without loss of generality, I assume the utility payoff scales are normalized so that $u_i(d_0, t) = 0$ for all i and all t. That is, each player could guarantee himself a payoff of zero by refusing to involve a mediator.

For simplicity, I assume that the players' beliefs are consistent with some prior probability distribution $\bar{p}$ in $\Delta(T)$, which is common knowledge, under which the players' types are independent random variables. That is, I assume that for every i, there exists a probability distribution $\bar{p}_i$ in

[6] Each $t_i \in T_i$ represents player i's characteristics, such as preferences, strengths, capabilities, or endowments.

$\Delta(T_i)$ such that $\bar{p}_i(t_i)$ is the prior marginal probability that player i's type will be t_i and

$$p_i(t_{-i}|t_i) = \bar{p}_{-i}(t_{-i}),\ \forall i \in N,\ \forall t_{-i} \in T_{-i},\ \forall t_i \in T_i.$$

Because they have incomplete information, the players in a bargaining problem do not have to agree on a specific outcome in D. Instead, two players may agree on some communication mechanism, which is a decision rule specifying how the outcome $d \in D$ depends on the players' types. As the players choose a mechanism and not an outcome, they can conceal their true types to get a better bargaining deal. In my context, they would bargain over the selection of a mediator who mediates according to a mechanism μ, which is a communication mechanism of the following form: Each player is asked to confidentially report his type to the mediator; then, after getting these reports, the mediator recommends an outcome to the players.[7]

Allowing randomized strategies, a *mediator* (or *mediation mechanism*) can be defined as a function $\mu : D \times T \to \mathbb{R}$ such that

$$\sum_{c \in D} \mu(c|t) = 1 \text{ and } \mu(d|t) \geq 0,\ \forall d \in D,\ \forall t \in T.$$

That is, $\mu(d|t)$ is the probability that d is the bargaining outcome chosen by the mediator with the mechanism μ, if t_1 and t_2 are the players' types.

Given any mechanism μ, for any $i \in \{1, 2\}$ and any $t_i \in T_i$,

$$U_i(\mu|t_i) = \sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_i(t_{-i}|t_i)\mu(d|t)u_i(d, t)$$

is the conditional expected utility for player i given that he is of type t_i, if the mechanism μ is implemented. The expected utility for type t_i of player i if he lies about his type and reports s_i while the other player is honest is

$$U_i^*(\mu, s_i|t_i) = \sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_i(t_{-i}|t_i)\mu(d|t_{-i}, s_i)u_i(d, t).$$

[7] In the literature on third party intervention, the essence of intermediaries is to convey information, and "mediated" communication involves a nonstrategic communication device that receives and transmits messages. Then, a mediator recommends a bargaining outcome to the players, and his recommendations depend on the players' reports. I consider a setting in which the players can follow the chosen mediator's recommendation or go to the disagreement outcome. Although this might seem limiting, it is the simplest way to find out which mediator the players are likely to choose. With this approach, I can analyze the outcomes of bargaining games and characterize the probability of peace obtainable with different mediators, while leaving potentially endless variation in the extensive form of Bayesian bargaining games unspecified.

A mediation mechanism μ is *incentive compatible* if and only if it satisfies the following informational incentive constraints:

$$U_i(\mu|t_i) \geq U_i^*(\mu, s_i|t_i),\ \forall i \in N,\ \forall t_i \in T_i,\ \forall s_i \in T_i.$$

A mediation mechanism μ is *individually rational* if and only if it satisfies the following participation constraints:

$$U_i(\mu|t_i) \geq \sum_{t_{-i} \in T_{-i}} p_i(t_{-i}|t_i) u_i(d_0, t),\ \forall i \in N,\ \forall t_i \in T_i.$$

Since the disagreement payments are normalized such that $u_i(d_0, t) = 0$ for all i and all t, the participation constraints reduce to

$$U_i(\mu|t_i) \geq 0,\ \forall i \in N,\ \forall t_i \in T_i.$$

The revelation principle (Myerson, 1979) implies that a mechanism cannot be implemented by any equilibrium of a communication game induced by any communication system unless the mechanism is incentive compatible and individually rational. Thus, there is no loss of generality in focusing on such direct revelation mechanisms.[8]

Therefore, taking relevant incentive constraints into account, I define the *incentive feasible mediator* to be whoever mediates according to the incentive compatible, individually rational mediation mechanism for the above Bayesian bargaining problem. That is, by the revelation principle, I can naturally assume that players bargain over the set of incentive feasible mediators, denoted as F. As I take a mediator to be synonymous with a mechanism, I use the terms mediator and its corresponding mediation mechanism interchangeably throughout the paper.

2.2 The Benchmark Model

In this subsection, I consider the context of international conflicts in which two symmetric players, each with two discrete types, must make a decision $d \in D$. For example, two states are involved in a dispute over a divisible item, area of territory, or an allocation of resources that could lead to war.

There are two possible decisions called d_0 and d_1. Let $D = \{d_0, d_1\}$, where d_0 can be interpreted as going to war, and d_1 as peace (e.g., negotiated settlement).[9] Each $i = \{1, 2\}$ has private

[8] See Holmström and Myerson (1983, 1804) and Myerson (1991, 487).

[9] In this paper, I restrict attention to two outcomes; this substantially simplifies the exposition while conveying all

information $t_i \in T_i = \{s, w\}$, where s denotes the strong type and w denotes the weak type. I assume that the types are independent; that is, t_i is drawn from the distribution $\bar{p}_i$ independent of t_{-i}, and this is common knowledge. For the sake of simplicity and tractability, I assume symmetry in probability – that the prior marginal probability of the strong types are the same for both players; namely, $\bar{p}_1(s) = \bar{p}_2(s) \equiv p$.[10]

A central facet of international conflict is that no enforcement body exists to permit binding contracts. Thus, a country can, at any time, unilaterally choose to initiate a war, and there is no way a country can commit not to do so. I call this feature the *unilateral war assumption*: War can be forced by either player. With two outcomes to choose from – war or peace, the disagreement outcome is clearly war and the following common theoretical assumption is necessary.[11]

Assumption. War is the only threat in bargaining.

With this assumption, I can designate $d_0 \in D$ to be the disagreement outcome. In the context of international conflict, the payoffs are natural examples of interdependence.

Assumptions. I maintain the following assumptions on the payoffs:

(A1) $\sum_i u_i(d_1, t) > 0$ for all t

(A2) $u_1(d_1, sw) < 0$

(A3) $u_1(d_1, t) = u_2(d_1, t)$ when $t_1 = t_2$;

$u_1(d_1, sw) = u_2(d_1, ws)$ and $u_2(d_1, sw) = u_1(d_1, ws)$.

(A1) implies that war is socially wasteful and is a destructive option, and the peace outcome is socially better than war. Although the war outcome is always socially Pareto inferior to some peaceful settlement, according to **(A2)**, a stronger country can take most of the stake in the dispute when war occurs and get a payoff larger than it can expect to get from a negotiated settlement. That is, a strong type prefers going to war when it is sure of facing a weak type. **(A3)** simply

the main insights. Also, calling d_0 and d_1 war and peace, respectively, is only a matter of labeling. The results continue to hold when more than two outcomes are allowed given that the disagreement outcome d_0 is fixed. Therefore, the analysis of this work is robust to extending the number of outcomes and could be equally applicable in any kind of bargaining situations as long as the following payoff assumptions are satisfied.

[10] Although I can relax the assumptions, this parsimonious benchmark model still captures many situations that the two players face in bargaining and yields a rich set of theoretical implications.

[11] This may be unrealistic in international relations, because countries always have a stalemate option. Because it might be interesting to say something more about the more realistic possibility of multiple threats (either going to war and staying in an unproductive stalemate), I consider the extension of a three-option model in Online Appendix E. However, the results qualitatively do not change in the extension.

assumes that payoffs are symmetric, depending only on the type combinations. Because of **(A3)**, I can focus on $i = 1$ in the analysis that follows.

3 The Interim Incentive Efficient Set

In this section, I characterize the interim incentive efficient set based on Holmström and Myerson (1983); the notion of interim incentive efficiency is a minimal requirement in a setting with incomplete information. With private information about each player's type, of all the mediators that players can choose, it should be that, once the players know their types, the mediator is interim incentive efficient. Otherwise, a third party can offer an interim Pareto superior one instead and the players will all be better off. As the players are rational, they would pick a mediator from the set of interim incentive efficient mediators.

An incentive feasible mediator associated with a mechanism μ is *interim incentive efficient* if and only if there exists no other incentive feasible mediator with a mechanism that is interim Pareto superior to μ. If every player would surely prefer μ' over μ when he knows his own type, whatever his type might be, then μ' is interim Pareto superior to μ. Interim welfare is evaluated after each player learns his own type but before he learns the other player's type. An incentive feasible mediator associated with a mechanism μ is *ex ante incentive efficient* if and only if there exists no other incentive feasible mediator with a mechanism that is ex ante Pareto superior to μ. Ex ante welfare is evaluated before any types are specified. If a mediator is incentive efficient, then another mediator who does not know any player's actual type could not propose any other incentive feasible mediation mechanism that every player is sure to prefer according to an appropriate welfare criterion.[12]

One might naively consider ex ante incentive efficiency to be a requirement for a bargaining solution. But if a player expresses a preference for the ex ante incentive efficient outcome, it would be giving information to the other side that might be detrimental. Therefore, we have no reason a priori to expect the ex ante incentive efficient outcome to be negotiated with privately informed parties. Therefore, as a first cut, I use the notion of interim incentive efficiency, which is a minimal requirement in a game in which each player knows his own type but does not know the other's type at the initial decision making stage.

Let $S(\Gamma)$ denote the set of incentive feasible mediators for Γ, each of whom is ready to offer to

[12]Note that for any given set of incentive feasible mechanisms, the set of ex ante incentive efficient mechanisms is a subset of the set of interim incentive efficient mechanisms.

mediate through some incentive compatible and individually rational mediation mechanism that is on the *interim incentive efficient* (IIE) frontier. Here, I focus on the *symmetric* interim incentive efficient mediation mechanisms. That is, I restrict attention to mechanisms that treat $t = \{sw\}$ and $t = \{ws\}$ cases symmetrically, but the result continues to hold when non-symmetric mediation mechanisms are allowed.

In fact, each interim incentive efficient mediator can be entirely defined by the probability weight he puts on the outcome of war and peace for each type realization. Therefore, the following lemma is useful because I am going to use it routinely throughout the paper.

Lemma 1. *Every symmetric interim incentive efficient mechanism is a function* $\mu_y : D \times T \to \mathbb{R}$ *such that*

$$\begin{aligned} &\mu_y(d_0|t) = 0,\ \mu_y(d_1|t) = 1 && \text{for } t \in \{ww\} \\ &\mu_y(d_0|t) = y,\ \mu_y(d_1|t) = 1 - y && \text{for } t \in \{sw, ws\} \\ &\mu_y(d_0|t) = z,\ \mu_y(d_1|t) = 1 - z && \text{for } t \in \{ss\}, \end{aligned}$$

where $y \geq 0$ *and* z *is uniquely determined given* y.

That is, μ_y implements d_0 (war) with probability zero when $t = \{ww\}$, with the probability $y \geq 0$ when $t = \{sw, ws\}$, and with the probability $z \geq 0$ when $t \in \{ss\}$, where z is uniquely defined by any given y. Therefore, each symmetric interim incentive efficient mediator is indexed by y and the following lemma gives three thresholds that separate four cases along the range of p.

Lemma 2 (Thresholds). *Let* $p^{'}$, p^*, *and* p^{**} *be such that*

$$\begin{aligned} p^{'} &\equiv \frac{-u_1(d_1, sw)}{u_1(d_1, ss) - u_1(d_1, sw)}, \\ p^* &\equiv \frac{u_1(d_1, ww)}{u_1(d_1, ww) + u_1(d_1, ws)}, \ \textit{and} \\ p^{**} &\equiv \frac{u_1(d_1, ss)u_1(d_1, ww) - u_1(d_1, sw)u_1(d_1, ws)}{u_1(d_1, ss)u_1(d_1, ww) - u_1(d_1, sw)u_1(d_1, ws) + u_1(d_1, ss)u_1(d_1, ws)}. \end{aligned}$$

Then, $p^{**} > p^*$ *and* $p^{**} > p^{'}$.

The interpretations of these thresholds are given in Appendix A. However, I must distinguish between the two instances: $p^{'} \leq p^*$ and $p^{'} > p^*$. Along with Lemma 3, Proposition 1 characterizes $S(\Gamma)$, putting further restrictions on y and z depending on the probability of the strong types when $p^{'} \leq p^*$. Proposition 2 is for when $p^{'} > p^*$.

Lemma 3. *When* $p \in [p^*, p^{**})$, $z := z(y,p) = y\left[1 - \frac{(1-p)\cdot u_1(d_1,ww)}{p\cdot u_1(d_1,ws)}\right]$ *and* $\bar{z}(p) \equiv z(1,p)$, *which is increasing in* p.

Proposition 1. *When* $p^{'} \leq p^*$, *the set of symmetric interim incentive efficient mediators* $S(\Gamma) = \{\mu_y\}$ *can be completely described such that:*

Case 1. $p < p^{'}$**:** $y \in \left[\underline{y}(p), 1\right]$ *and* $z = 0$, *where* $\underline{y}(p) = 1 + \frac{p\cdot u_1(d_1,ss)}{(1-p)\cdot u_1(d_1,sw)}$ *which is decreasing in* p;

Case 2. $p \in [p^{'}, p^*)$**:** $y \in [0,1]$ *and* $z = 0$;

Case 3. $p \in [p^*, p^{**})$**:** $y \in [0,1]$ *and* $z := z(y,p) \in [0, \bar{z}(p)]$.

Case 4. $p \geq p^{**}$**:** $y = 0$ *and* $z = 0$.

The key point of Proposition 1 is that each symmetric interim incentive efficient mediator can be identified by a one-dimensional object μ_y with restrictions on y (the probability on war for $t = \{sw, ws\}$); z (the probability on war for $t = \{ss\}$) is pinned down entirely by y for each case. For example, when $p \in [p^{'}, p^*)$ (***Case 2***), μ_0 always puts probability one on peace; $\mu_{0.3}$ puts probability one on peace when both players are of the same type, but puts a positive probability of 0.3 on war when players are of different types; and μ_1 puts probability one on peace when both players are of the same type, but puts probability one on war when players are of different types. These mediators cannot be interim Pareto dominated by any other incentive feasible mediator, and so all of these mediators are interim incentive efficient.

There are three issues to note for ***Case 1***, ***Case 3***, and ***Case 4***, respectively. First, when $p < p^{'}$ (***Case 1***), the lowest probability that a symmetric interim incentive efficient mediator can possibly put on war for $t = \{sw, ws\}$ must be bounded below by $\underline{y}(p)$, because if some mediator puts an even lower probability than $\underline{y}(p)$ on war, then it would not be incentive feasible. Therefore, when the strong type is very rare, a mediator must put some minimum probability on war to induce a strong type to participate and not deviate to unilaterally forcing war. Otherwise, "always war" gives the strong type a strictly higher expected payoff than when participating in mediation, because he expects with high probability that the other player is weak. Second, when $p \in [p^*, p^{**})$ (***Case 3***), an interim incentive efficient mediator must also put some positive probability on war when both players are of the strong type to prevent the weak type from reporting dishonestly. Third, when $p \geq p^{**}$ (***Case 4***), any mediator who puts a positive probability on war regardless of type realization will be interim Pareto dominated by μ_0.

The following Proposition 2 along with Lemma 3 characterizes the set of symmetric interim incentive efficient mediators when $p^{'} > p^{*}$. The only difference from Proposition 1 is that when $p \in [p^{*}, p^{'})$ (***Case 2'***), then an interim incentive efficient mediator has to be associated with a mechanism that puts some positive probability on war for $t = \{sw, ws\}$ with a lower-bound requirement (as in Proposition 1. ***Case 1***), along with some corresponding positive probability on war for $t = \{ss\}$ (as in Proposition 1. ***Case 3***).

Proposition 2. *When $p^{*} < p^{'}$, the set of symmetric interim incentive efficient mediators $S(\Gamma) = \{\mu_y\}$ can be completely described such that:*

Case 1'. *$p < p^{*}$: $y \in \left[\underline{y}(p), 1\right]$ and $z = 0$, as in Proposition 1. Case 1;*

Case 2'. *$p \in [p^{*}, p^{'})$: $y \in \left[\underline{\underline{y}}(p), 1\right]$ and $z := z(y,p) \in \left[\underline{\underline{z}}(p), \bar{z}(p)\right]$, where $\underline{\underline{y}}(p) > 0$ and a corresponding $\underline{\underline{z}}(p) > 0$ are defined in Appendix A;*

Case 3. *$p \in [p^{'}, p^{**})$: $y \in [0,1]$ and $z := z(y,p) \in [0, \bar{z}(p)]$, as in Proposition 1. Case 3; and*

Case 4. *$p \geq p^{**}$: $y = 0$ and $z = 0$.*

Proposition 1 and Proposition 2 imply that there are multiple symmetric interim incentive efficient mediators when $p < p^{**}$ (***Case 1***, ***1'***, ***2***, ***2'***, and ***3***). When $p \geq p^{**}$ (***Case 4***), then only one mediator μ_0 exists in $S(\Gamma)$. Therefore, the intriguing cases are only when $p < p^{**}$, in which there are an infinite number of symmetric interim incentive efficient mediators from which players can select. In fact, within each case, I can order the mechanisms according to the ex ante welfare criterion.

Lemma 4. *For each p, $y < y^{'}$ if and only if μ_y is ex ante Pareto superior to $\mu_{y'}$, i.e., $U_i(\mu_y) > U_i(\mu_{y'})$ for all i.*

Proposition 3. *The unique ex ante incentive efficient mediator in $S(\Gamma)$ is*
*When $p^{'} \leq p^{*}$, (a) $\mu_{\underline{y}(p)}$ for **Case 1**; (b) μ_0 for **Case 2**, **3** & **4**.*
*When $p^{*} < p^{'}$, (a) $\mu_{\underline{y}(p)}$ for **Case 1'**; (b) $\mu_{\underline{\underline{y}}(p)}$ for **Case 2'**; (c) μ_0 for **Case 3** & **4**.*

Proposition 3 gives the best mediator in an ex ante sense. The unique ex ante incentive efficient mediator is the one who puts the lowest probability on war among all of the interim incentive efficient mediators. Note that μ_0 is such that $\mu_0(d_1|t) = 1$, $\forall t \in T$, which can be interpreted

as a peaceful mediator who always implements a peace outcome regardless of the players' types. Therefore, a peaceful mediator is the one who never creates an impasse that prevents an agreement.

The following Lemma 5 gives the interim welfare ordering in $S(\Gamma)$.

Lemma 5. *For each p, $y < y^{'}$ if and only if $U_i(\mu_y|s) < U_i(\mu_{y'}|s)$ and $U_i(\mu_y|w) > U_i(\mu_{y'}|w)$ for all i.*

Lemma 5 implies that if (and only if) μ_y gives higher ex ante expected utilities to the players than $\mu_{y'}$, then μ_y gives a strictly lower interim expected utility to the strong type and a strictly higher interim expected utility to the weak type than $\mu_y^{'}$.

Corollary 1. *Suppose that $p < p^{**}$. Then, there exist multiple interim incentive efficient mediators in $S(\Gamma)$ who are not ex ante incentive efficient. Moreover, the one that is ex ante Pareto inferior to any other mediators in $S(\Gamma)$ is μ_1, which gives the highest interim payoff for the strong type.*

Corollary 1 follows from the previous results. It suggests that there are multiple mediators in $S(\Gamma)$ who will, if chosen, mediate players according to some incentive feasible mechanism that is not necessarily ex ante incentive efficient. The interim but not ex ante incentive efficient mediator can be considered "*bad*" in the sense that it sometimes allows players to fail to reach a negotiated settlement and to go to war with higher probability than the ex ante incentive efficient mediator. The "*worst*" mediator, who puts a higher probability on war than any other interim incentive efficient mediator, is the one that can be described as being *the farthest away from the ex ante incentive efficient mediator* but gives the highest interim utility to the strong type and the lowest interim utility to the weak type. Alternatively, the worst one also can be thought of as the unique solution to the minimization problem of the ex ante expected payoffs.

4 Cooperative Approach

My goal is to develop a formal argument of endogenous choice of a mediator that determines the smallest possible set of solutions, so as to get the strongest possible predictions. At some level, there is potentially an endless variation to an extensive form communication game of the negotiation process over selecting mediators, which I am not explicitly modeling. Instead, I take payoffs and the existence of incentive feasible mediators as primitives. The revelation principle implies that if all incentive compatible direct mechanisms have some property, then every equilibrium of every

game form has this property. With a cooperative approach, I am explicitly taking into account the information leakage problem by imposing relevant properties or axioms that must be satisfied for a reasonable solution. In this section, I characterize the neutral bargaining solution using the solution concept in Myerson (1984*b*); and I show the inefficiency in bargaining.

4.1 The Neutral Bargaining Solution

The Inscrutability Principle (Myerson, 1983) Before continuing, I invoke the *inscrutability principle* (Myerson, 1983). The inscrutability principle states that "...when we consider a mechanism selection game, in which individuals can bargain over mechanisms, there should be no loss of generality restricting our attention to equilibria in which there is one incentive feasible mechanism that is selected with probability 1, independently of anyone's type" (Myerson, 1991, 504). The inscrutability principle can be justified by viewing the mechanism selection process itself as a communication game induced from the original Bayesian bargaining game and by applying the revelation principle.

At some level, there is potentially endless variation in the communication process for selecting a mediator. Without referring to a specific game form, once a noncooperative game of the negotiation process over mediators is well defined, we can take a Bayesian Nash equilibrium of it. This equilibrium should ultimately be an equilibrium mapping from some distributions of players' types to some distributions over the mediators playing the mechanisms. For example, suppose there is an equilibrium of some mediator selection game in which some mechanism δ would be selected if the profile of players' types were in some set $A \subseteq T$, and some other mechanism γ would be selected otherwise. Then, there should be an equivalent equilibrium of the mediator selection game in which the players always select a "grand" mechanism μ, which coincides with mechanism δ when the players report a profile of types in A and which coincides with γ when they report a profile of types that is not in A. Myerson (1983) shows that there is no loss of generality in assuming that all types of players agree to play this same grand mechanism μ without any communication during the negotiating process.[13] This idea of rolling all of the information leakage that could occur in any process of mediator selection into the grand mechanism is the inscrutability principle.

Therefore, whatever the equilibrium outcome of any bargaining procedure is, without loss of

[13]That is, for any outcome achievable via any equilibrium under some noncooperative game, I can think of the equilibrium outcome as an incentive feasible grand mechanism that basically everyone selects. Moreover, for any incentive feasible grand mechanism for some bargaining problem, there is some noncooperative game that generates it as an equilibrium.

generality, I can reduce that equilibrium into a grand mechanism that accomplishes the same thing and is incentive feasible. I call this grand mechanism an inscrutable mechanism, or an inscrutable mediator. Under the inscrutability principle, the set of all equilibria under a noncooperative game once defined is attainable as the set of all inscrutable mechanisms. That is, if all inscrutable mechanisms have some property, then every equilibrium of a noncooperative game has this property. In light of Myerson (1983), I assume that the players bargain inscrutably in the negotiating process.[14]

The Neutral Bargaining Solution Now, I can ask further what properties must be satisfied for the mediators that the players inscrutably agree on. The concept of interim incentive efficiency implicitly uses $U(\mu) = ((U_i(\mu|t_i)_{t_i \in T_i})_{i \in \{1,2\}}$ given any mechanism μ as the relevant utility allocation vector. Because each player already knows his true type at the time of bargaining, the average of each player's expected utility over his various types might not be appropriate. It might seem that just the components corresponding to the two actual types are significant in determining whether a mediator with such μ is chosen. However, all of the components of $U(\mu)$ are relevant in determining whether a mediator is chosen because we need to take into account the question of inscrutable intertype compromise that arises in the theory of bargaining with incomplete information.

Any inscrutable bargaining theory must implicitly make some trade-off or compromise between the two possible types of each player, if only because there are many interim incentive efficient mechanisms that differ in the expected payoffs that they give to the strong and the weak types. That is, even though each player already knows his true type at the time of bargaining, to bargain inscrutably, each player must use a bargaining strategy that maintains a balance between the conflicting goals he would have if he were of a different type. Therefore, a bargaining solution must specify criteria for an equitable compromise between different possible types of each player, as well as between two players.

With a further refinement, I create a subset within the set of interim incentive efficient mediators from a simple set of axioms that takes into account the inscrutable intertype compromise using Myerson (1984*b*). Myerson (1984*b*) develops a generalization of the Nash bargaining solution for two-person bargaining games with incomplete information from some properties that a fair bargaining solution should satisfy. The *neutral bargaining solutions* (Myerson, 1984*b*) form the smallest

[14] In other words, neither player ever deliberately reveals any information about his true type until the mediator is agreed on in the negotiating process. If two players agree, then they simultaneously and confidentially report their types to the mediator, and then the decision is adopted according to μ. "In effect, any information that the players could have revealed during the mechanism selection process can be revealed instead to the mediator after it has been selected, and the mediator can use this information in the same way that it would have been used in the mechanism selection process" (Myerson, 1991, 504). Also see Myerson (1984, 463).

set satisfying two axioms: an extension axiom and a random-dictatorship axiom.[15]

The first axiom is a stronger version of Nash's independence of irrelevant alternatives axiom. If the Bayesian bargaining problem Γ could be extended in a way that the set of decision options is increased without changing the disagreement outcome such that both players would be willing to settle for the expected utility payoffs that are arbitrarily close to what they can get from the feasible mechanism μ, then the players would be willing to settle for the mechanism μ even when these extra decision options are not available. The second axiom is related to Nash's symmetry axiom. In cases in which the mechanism each player would select if he were the principal who selects the mechanism dictatorially could be predicted, and if giving each player a probability .5 of acting as the principal (the random dictatorship) is interim incentive efficient, then the random dictatorship mechanism should be a fair bargaining solution. Although the hypotheses of this axiom are quite restrictive, the axiom means that only in a bargaining game where the random dictatorship is both well understood and interim incentive efficient, then it should be a fair claim to being a bargaining solution.

A neutral bargaining solution of a bargaining problem Γ is defined to be any mediator associated with a mechanism with the property that it is a solution for every bargaining solution concept which satisfies these two axioms. Myerson (1984*b*) proves that the set of neutral bargaining solutions is nonempty for any finite two-player Bayesian bargaining problem. I use the characterization theorem in Myerson (1984*b*) that is generated by the two axioms, slightly modified for my case, to get the most tractable conditions for computing neutral bargaining solutions.

Theorem (Myerson, 1984*b*). *A mechanism μ is a neutral bargaining solution for Γ if and only if, for each positive number ε, there exist vectors λ, α, β, and ϖ (which may depend on ε) such that*

$$\left(\left(\lambda_i(t_i) + \sum_{s_i \in T_i} \alpha_i(s_i|t_i) + \beta_i(t_i)\right)\varpi_i(t_i) - \sum_{s_i \in T_i} \alpha_i(t_i|s_i)\varpi_i(s_i)\right)/\bar{p}_i(t_i)$$

$$= \sum_{t_{-i} \in T_{-i}} \bar{p}_{-i}(t_{-i}) \max_{d \in \{d_0, d_1\}} \sum_{j \in \{1,2\}} \frac{v_i(d, t, \lambda, \alpha, \beta)}{2}, \quad \forall i,\ \forall t_i \in T_i;$$

$$\lambda_i(t_i) > 0,\ \alpha_i(s_i|t_i) \geq 0,\ \beta_i(t_i) \geq 0, \quad \forall i,\ \forall s_i \in T_i,\ \forall t_i \in T_i;$$

$$\text{and}\ \ U_i(\mu|t_i) \geq \varpi_i(t_i) - \varepsilon, \quad \forall i,\ \forall t_i \in T_i,$$

[15] See Myerson (1984*b*) for a detailed exposition of these axioms.

where $v_i(\cdot)$ is the virtual utility payoff to player i from outcome d, defined by

$$v_i(d,t,\lambda,\alpha,\beta) = ((\lambda_i(t_i) + \sum_{s_i \in T_i} \alpha_i(s_i|t_i) + \beta_i(t_i))u_i(d,t) - \sum_{s_i \in T_i} \alpha_i(t_i|s_i)u_i(d,(t_{-i},s_i)))/\bar{p}_i(t_i);$$

$\alpha_i(s_i|t_i)$ denotes the Lagrange multiplier for the informational incentive constraints; and $\beta_i(t_i)$ denotes the Lagrange multiplier for the participation constraints. Moreover, (λ,α,β) satisfy the interim incentive efficiency conditions for the mechanism μ.

I explain more about the virtual utility payoffs later in this subsection, but for now, note that the virtual utility differs qualitatively from the actual utility. The above theorem asserts that a neutral bargaining solution can be characterized as an incentive feasible mediator who is not only efficient in terms of maximizing the sum of the players' virtual utility payoffs, but is also equitable in terms of balancing out intertype compromise. Any allocation vector ϖ that satisfies the conditions in the theorem for some positive λ and some non-negative α and β can be called *virtually equitable* for Γ. The first equation in the theorem defines the intertype-equity conditions such that the weighted sums of players' possible conditionally expected utilities should be equal. The third condition says that a neutral bargaining solution should generate type-contingent expected utilities that are equal to or interim Pareto superior to a limit of virtually equitable utility allocations. The theorem implies that a player makes equity comparisons in virtual utility terms rather than in actual utility terms.[16]

The theorem gives a set of conditions that must hold for every mediator in the set of neutral bargaining solutions. Moreover, because the set of all interim incentive efficient mechanisms would satisfy the two axioms, I can think about picking a neutral bargaining solution over the set of all symmetric interim incentive efficient mediators $S(\Gamma)$. Let $NS(\Gamma)$ denote the set of neutral bargaining solutions identified for the class of two-person Bayesian bargaining game Γ that satisfies assumptions **(A1)** through **(A3)**. Then, for the bargaining problem considered in this paper, only one mediator exists among $S(\Gamma)$ that satisfies the conditions in the above theorem.

Theorem 1. *For any two-person Bayesian bargaining problem Γ that satisfies **(A1)** – **(A3)** with independent and symmetric priors, the neutral bargaining solution is unique.*

[16]The players can negotiate over the set of virtual utility maximizing allocations without revealing information about their types.

Proposition 4. *The neutral bargaining solution in* $S(\Gamma)$ *is*

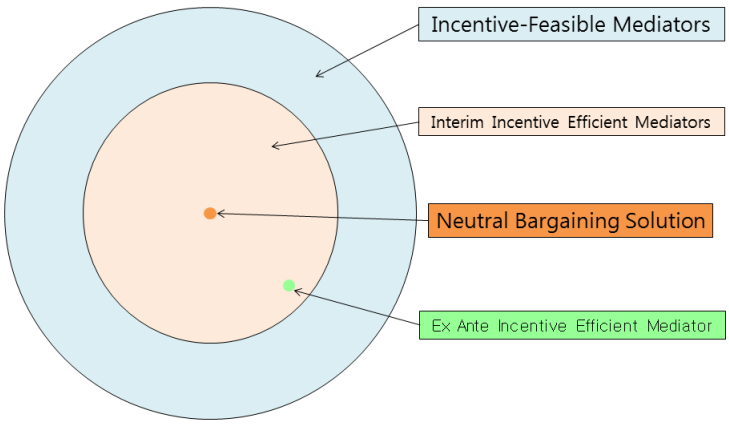
If $p < p^*$: $NS(\Gamma) = \{\mu_1\}$ *with* $z = 0$.

If $p \in [p^*, p^{**})$: $NS(\Gamma) = \{\mu_1\}$ *with* $z = \bar{z}(p)$.

If $p \geq p^{**}$: $NS(\Gamma) = \{\mu_0\}$.

Theorem 1 asserts that, for the class of environments in my framework, it identifies the smallest solution set and gives a unique prediction. Proposition 4 implies that when the probability of the strong type is sufficiently small, the unique ex ante incentive efficient mediator who puts the lowest probability on war should not be in the solution set refined by requiring the axioms of the neutral bargaining solution. Instead, the notion of a neutral bargaining solution picks a unique mediator in $S(\Gamma)$ who is associated with the highest probability of war and is the farthest away from ex ante incentive efficiency, which I call the "worst" mediator.

Figure 4.1: Structure of Solution Set Refinements in the Cooperative Approach



The reasoning behind this result is that the weak types want to mimic the strong types, because expressing a preference for the ex ante incentive efficient mediator will immediately convince the other player to go to war. That is, because a player would actually be very eager to reveal information about his type if he were strong, and only the weak type would have some incentive to conceal his type in bargaining, the inscrutable intertype compromise tends to get resolved in favor of the strong type. In this sense, the concept of the neutral bargaining solution provides the answer

to which mediator, or communication system, is chosen inscrutably. As a result, any negotiating process should lead to the neutral bargaining solution characterized in Proposition 4, which is the reasonable expectation in any negotiating process.

ARROGANCE OF STRENGTH Now, I discuss a property of the neutral bargaining solution in relation to the virtual utility payoffs. Borrowing from Myerson (1984*b*), consider a fictitious game of the (λ, α, β) virtual bargaining problem, in which the players' types are verifiable, each player's payoffs are in the virtual utility scales $v_i(\cdot, \cdot, \lambda, \alpha, \beta)$ with respect to λ, α, and β, instead of $u_i(\cdot, \cdot)$, and these virtual utility payoffs are transferable among the players. If (λ, α, β) satisfy the interim incentive efficiency conditions for μ and $\alpha_i(s_i|t_i) > 0$, then we say that type t_i *jeopardizes* type s_i. A positive Lagrange multiplier $\alpha_i(s_i|t_i)$ implies that t_i can only jeopardize s_i in μ if the informational incentive constraint is binding for μ.

In the bargaining problem Γ considered in this paper, I obtain the following corollaries.

Corollary 2. *When $p < p^*$, for the neutral bargaining solution $\mu_{1,0}$, the conditions in Myerson (1984*b*)'s Theorem are satisfied for all $\varepsilon \geq 0$ by any $\lambda_i(s) > \lambda^*$ for some $\lambda^* > p$, $\alpha_i(s|w) = \alpha_i(w|s) = 0$, and $\beta_i(s) = \beta_i(w) = 0$ for all i.*

When $p < p^*$, in the (λ, α, β) virtual bargaining problem with transferable virtual utility payoffs, no incentive constraints are binding. In other words, the constraints do not matter. The fact that there are no binding constraints is exactly what makes $\lambda_i(t_i)$ weights different from the probability weights. In particular, we have $\lambda_i(s) > p$ and $\lambda_i(w) < 1 - p$ for all i. For the neutral bargaining solution μ_1 when $p < p^*$, the type-dependent weights $\lambda_i(t_i)$ are necessarily distorted in a way that they scale up the actual utility of the strong type and scale down the actual utility of the weak type, as if the strong type were more important. Then, the neutral bargaining solution can be interpreted as if some principal maximizes the ex ante social welfare putting extra welfare weights on the strong types in the virtual utility characterization.[17]

The more interesting cases in the neutral bargaining solutions should be when the incentive constraints bind; for the benchmark class in my framework, these cases would correspond to when $p \in [p^*, p^{**})$.

[17] Otherwise, the mediator would not have been an equilibrium in any communication game defined. This can be seen as an analog to being renegotiation-proof. Putting more weight on the strong type's utility allows the equilibrium to be interim incentive efficient and inscrutable.

Corollary 3. *When $p \in [p^*, p^{**})$, for the neutral bargaining solution μ_1, the conditions in Myerson (1984b)'s Theorem are satisfied for all $\varepsilon \geq 0$ by any $\lambda_i(s) > \lambda^{**}$ for some $\lambda^{**} > p$, some $\alpha_i(s|w) > 0$, $\alpha_i(w|s) = 0$, and $\beta_i(s) = \beta_i(w) = 0$ for all i.*

When $p \in [p^*, p^{**})$, the constraint that a player of a weak type should not be tempted to pretend that her type is strong is a binding constraint for the neutral bargaining solution μ_1. Because the weak type wants to imitate the strong type, the weak type jeopardizes the strong type. Then, the strong type of a player begins to act according to virtual preferences which exaggerate the difference from the weak type that he needs to distinguish from. Therefore, the virtual utility of the strong type should differ from the actual utility of the strong type in that the virtual utility exaggerates the difference from the weak type that jeopardizes the strong type. That is, the virtual utility for the strong type is a positive multiple of his actual utility minus a positive multiple of the "false" utility of the weak type; and the virtual utility for the weak type is a positive multiple of her actual utility. Then, the neutral bargaining solution μ_1 when $p \in [p^*, p^{**})$ suggests that a player bargains as if he wanted to maximize some virtual utility scale that exaggerates the difference from the other type in response to the pressure he, in particular the strong type, feels from the other player's distrust of any statements he could make about his type.

In both cases, we can describe a player of a strong type to be in a "surprisingly strong" position in a sense that the probability of the strong type is rather small, and the probability of war in the neutral bargaining solution is the same as the probability of war that would occur if the player were the principal with all the negotiating ability and he were of the strong type. Myerson (1991, 523) calls this property of the neutral bargaining solution *arrogance of strength*. Such arrogance of strength is reasonable to expect in the neutral bargaining solution of Γ considering the inscrutable intertype compromise.

4.2 Failure of Ex Ante Efficiency

The neutral bargaining solution characterized in Proposition 4 is the "worst" mediator, who has the highest probability of failing to make a peaceful settlement. Why is this mediator the unique neutral bargaining solution, and how do these properties relate to the distortion in the neutral bargaining solution? Ultimately, I argue that the cooperative idea of the neutral bargaining solution leads to inefficiency in bargaining.

I assume that the players in the bargaining process with private information bargain over a mediator in an inscrutable way, without any communication during the mediator selection stage.

If a player did anything different, the other player might learn something about him. Thus, both types do not want to reveal their true types and nothing is learned in the negotiating process. Therefore, despite the type of bargaining theory, the focus, without loss of generality, can be on the inscrutable theory of bargaining in the choice of a mediator.

However, the inscrutability principle does not imply that the possibility of information leakage during a mediator selection process is irrelevant.[18] There might be some interim incentive efficient mediators that are not selected by the players in such a process, precisely because some player might choose to reveal or conceal information about his type instead of letting some mediators be selected.

So, in the class of bargaining games that I described, what is it that a player does not want the other to learn about him? Among all of the interim incentive efficient mediators, the players might prefer the ex ante incentive efficient solution before they learn their own types from the ex ante point of view. But, because players already know their own types, what each player does not want the other to learn via his choice of mediator is whether he is weak. Expressing preference for the ex ante incentive efficient mediator, who is known to be better at implementing a peaceful settlement, might convey information that he is in a weaker position. Whether a player is strong or weak, he would not want the other player to infer that he is weak, even if the probability of the strong type is small. Thus, even when both players are weak, they will tilt toward what they would do if they were strong, choosing the mediator who is better for the strong type to avoid disclosing that they are weak.

Therefore, the players in a bargaining process inscrutably agree on a neutral bargaining solution: the players act as though they know their types and are strong, and pick the mediator in a way that would be better for when they are strong, never revealing their types during bargaining. This notable distortion in the neutral bargaining solution concept has a "signaling" component. In a cooperative sense, players pick a mediator such that no information is revealed, but this is effectively asking the players to pool in a way such that no information is revealed. That nature of pooling is in the spirit of a pooling equilibrium of a signaling outcome in a noncooperative game.[19] That is, the "flavor" of signaling implicitly comes into the argument of Myerson (1984*b*)'s neutral bargaining solution, in which all of the distorted welfare weights together with the inscrutability

[18] See Myerson (1991, 505).

[19] Although formally I do not have a signaling game in which one player is choosing a strategy that signals his intention or strength, it is as if we built the signaling distortion in a noncooperative game directly into the cooperative mathematics.

principle lead to a result that the strong types get more weight in the final mechanism.

This result implies that the inscrutable intertype compromise between the strong type and the weak type gets resolved in favor of the strong type. Even if the probability of the strong type is fairly small, the strong type is going to end up being more influential on the players' behavior in the bargaining process, the property that we call arrogance of strength. What is best for the strong type is typically very far from ex ante efficiency for the fundamental reason that the strong type wants to separate itself from the weak type.

As a result, in a model where ex ante efficient outcomes are technically feasible, there exists an equilibrium of any mediator selection game in which players systematically do not choose the mediator who could maximize the ex ante efficient gains because they want to "signal" in a cooperative sense that they are the strong type. Hence, they systematically do worse than the ex ante efficiency. This result explains why interim bargaining theory is not similar to ex ante bargaining theory, as well as how ex ante bargaining can get exactly the wrong answer.

Therefore, I argue that the very process of selecting a mediator exhibits an inherent inefficiency of bargaining for the following reasons. The mediator who is good at achieving peace gets defeated, and the players end up choosing a mediator who over-implements the war payments, implicitly putting more weight on the payoffs for the strong type. The neutral bargaining solution results in too much war, and the players' expected gains from trade are less than the optimal (the second best). This result evokes Myerson and Satterthwaite (1983) in which the status quo disagreement payments are likewise over implemented. The result also implies a new way to think about distortions that occur in political conflicts in the context of international relations: taking the countries that are the most difficult to negotiate with and making them more likely to arise in equilibrium.

5 Noncooperative Approach

The cooperative approach, while illuminating which mediator we should reasonably expect to arise endogenously, by itself is not satisfactory without also considering a more explicit noncooperative game.

Consider a ratification game in which some uninformed third party can announce an alternative mechanism, and players can switch from some status quo mechanism if both players *ratify* (or, unanimously vote in favor of) the alternative. I assume that if one votes in favor and one votes against, then they are going to learn from disagreement. Considering such a pairwise ratification

game, I ask the following: Given some status quo mechanism, can we find another mechanism that can beat it in some sequential equilibrium in a ratification game, where beliefs following disagreement are required to satisfy credibility conditions?[20]

I now introduce the concept of *threat-security*, based on the extended concept of ratifiability in Cramton and Palfrey (1995), which I compare to Holmström and Myerson (1983) and my setting.

5.1 Durability, Ratifiability, and Assumptions

Holmström and Myerson (1983) consider *durability* that also formalizes the idea of a mechanism being invulnerable to proposals of alternative mechanisms in a pairwise comparison.[21] They show that the set of durable interim incentive efficient mechanisms is non-empty for any finite Bayesian bargaining problem and that with one-sided incomplete information, all interim incentive efficient mechanisms are durable. However, I argue that durability does not give a strong prediction and is limiting in my framework. For the games that I study in this paper, not only is the neutral bargaining solution durable, but in fact the entire interim incentive efficient set is durable.

The difference arises because of the assumptions about passive beliefs, voting, and the mechanisms. First, durability relies on the assumption that passive beliefs hold whenever an alternative is rejected that allows for only learning from agreement to the alternative.[22] So when the players go back to play the status quo mechanism, they play as they could have in the first place. However, in my games, the focus is on the information that gets leaked when there is disagreement in the negotiations over selecting a mediator. This distinction reflects somewhat different implicit assumptions. Durability tries to capture the idea of one type of player proposing an alternative at the interim stage (although it is not literally talking about the player's proposal stage), but I consider the proposed alternative as selected by an uninformed third party. Therefore, no information is revealed by a proposal itself. Thus, I assume that if an alternative is ratified, the alternative mediator is played with the original beliefs, and learning only comes from disagreement.

Second, players are not told the details of the vote in Holmström and Myerson (1983) because they assume a secret ballot. I assume a non-anonymous voting procedure. This assumption is

[20] Whereas a cooperative approach is to think about a reasonable solution set compared to the status quo of an unmediated Bayesian game, this approach looks at whether one mediator survives in comparison to another.

[21] A similar idea has been discussed under the name *resilient allocation rule* (Lagunoff, 1995) in a buyer-seller bargaining problem.

[22] A mechanism is durable if there exists a sequential equilibrium such that if an alternative is offered, at least one player rejects the alternative for every equilibrium path. If a player strictly prefers the alternative, he votes for it even if he thinks the other is not voting for it. And if a player is indifferent (the trembling point), then he is allowed to vote against the alternative.

realistic because I am concerned with international relations settings in which one type of country might have an incentive to publicly announce displeasure with the possible alternative when the countries are negotiating because of the information the announcement conveys.

Third, Holmström and Myerson (1983) assume a fixed direct revelation status quo mechanism and arbitrary mechanisms as proposals. Therefore, if players commit themselves to a status quo mechanism, then they cannot alter the outcome selected by the status quo mechanism afterward. However, my status quo is not necessarily a direct mechanism, and I only consider direct mechanisms as alternatives.

For these reasons, I do not pursue durability, but proceed with the concept of *ratifiability* in Cramton and Palfrey (1995). Ratifiability is a mirror image of durability but its setting is much more compelling to relate to my game because it also allows public voting and takes into account that the beliefs are changed in the status quo after the unsuccessful ratification of an alternative. Further, the players voluntarily decide to participate in the alternative mediation at the interim stage; and if participation is unanimous, then the players are bound by the alternative and cannot renegotiate during implementation.

However, they also consider the status quo mechanism to be fixed as in Holmström and Myerson (1983), although not necessarily a direct mechanism, as a reduced-form game of Bayesian Cournot competition. That is, whatever beliefs from disagreement are injected into the exogenously imposed status quo, the players are ultimately committed to whatever the fixed status quo is with possibly altered beliefs only about the vetoer.

My innovations over Holmström and Myerson (1983) and Cramton and Palfrey (1995) are that the status quo is designed that by assuming public voting with two players, *both* players learn about each other after disagreement. Therefore, beliefs following disagreement are required to satisfy *the credibility conditions*.

5.2 The Threat-Secure Mediator

An interim incentive efficient mediator δ should be considered *threat-secure* if and only if there does not exist an alternative mediator that can be unanimously ratified in favor of the alternative. In this section, I establish whether there is at least one type in the unanimous ratification game that shuts down the alternative offer so the renegotiation does not happen. In order to develop this idea, I first formulate the modified concept of a ratifiable alternative mediator.

Different from the standard mechanism design approach,[23] I focus on each player's decision to switch to an alternative mediator γ, where a failure to ratify might reveal information. What is learned from the unilateral veto with public voting alters both players' status quo payoffs after non-ratification. Therefore, in deciding whether to veto or not veto the mediator γ, a player considers how the other player's belief about him and his belief about the other might change as a result of the unilateral veto. I link together a player's decision to veto with rational expectations about the outcome in the continuation game, taking into account the appropriate individual rationality constraints.

An (uninformed) alternative mediator proposes to implement an incentive compatible direct-revelation mechanism $\gamma : D \times T \to \mathbb{R}$. Participation constraints require that every type of every player to prefer to participate in γ than to play the *status quo mechanism* G^δ with finite message space and outcome function. I assume that a status quo G^δ depends on strategic actions and is identified with a mediator that players are initially involved with who implements an incentive compatible direct-revelation mechanism $\delta : D \times T \to \mathbb{R}$.

For each $i \in \{1, 2\}$, let $A_i = \{0\} \times R_i$, where 0 denotes going to war, and each set R_i represents a nonempty finite set of possible reports r_i as an informational input into G^δ. I write $R = R_1 \times R_2$ and $r = (r_1, r_2)$. Then, G^δ is a function of the form $G^\delta : D \times A_1 \times A_2 \to \mathbb{R}$ such that

$$\begin{aligned} G^\delta(d_0|a_1, a_2) &= 1 \qquad \text{if } a_1 = 0 \text{ or } a_2 = 0 \\ G^\delta(d|a_1, a_2) &= \delta(d|r) \qquad \text{otherwise.} \end{aligned}$$

The strategies in the status quo mechanism G^δ consist of the following: Each player reports (possibly deceptively) or chooses to go to war, given its type and its belief about the other player's type. With prior beliefs, the equilibrium outcome in the status quo mechanism G^δ is to hire mediator δ and both players honestly reporting their types.[24]

So, the participation constraints are considered by analyzing a two-stage ratification game. I want to establish whether there exists an equilibrium of vetoing strategies such that γ is unanimously ratified over G^δ for some equilibrium path. An individual's optimal vetoing strategy in this

[23] "In many mechanism design settings, learning from disagreement is not an issue, since the status quo mediator is not affected by a change in the players' beliefs" (Cramton and Palfrey, 1995, 257).

[24] If one of the players rejects γ, then the inference drawn from the vote outcome alters both the vetoer and ratifier's equilibrium plays in such a way that they play G^δ (mediator δ appended to the war option) differently than with prior beliefs, since with updated beliefs, mediator δ may not be individually rational anymore, and the players may instead prefer to go to war. If either side chooses war, then the players get the war outcome. Therefore, individual rationality must embody the altered equilibrium play of G^δ following unsuccessful ratification.

ratification game should depend on what he would believe about the other's type and what the other would believe about his type, conditioned on the (unexpected) failure to ratify the alternative, and how he would expect the status quo to be played as a result of the veto.

Thus, I shall need the following notations. I maintain the independence and symmetry assumptions on the probability of types. In the first stage, the players each vote simultaneously for or against the alternative mediator γ. A strategy for i in the first stage specifies the probability that each player i vetoes γ if his type were t_i. Let $v_i(t_i)$ denote the vetoing strategy of player i of type t_i. An outcome to the first stage indicates the players $M \subseteq \{1,2\}$ that vetoed γ. In the second stage, γ is implemented if $M = \emptyset$; otherwise, the status quo mechanism G^δ is played with the knowledge that players M vetoed γ.

Because ratification requires a unanimous vote, I must specify the beliefs of all players if i unilaterally deviates from this ratification equilibrium by vetoing. For each i vetoing, for all $j \in \{1,2\}$ and for all $t_j \in T_j$, let $\bar{q}_{j,i}(t_j)$ denote player $-j$'s belief about player j's type, conditional on the knowledge that $M = \{i\}$. In particular, suppose that if i alone vetoes γ and the players believe that the other's type is given by the distribution $\bar{q}_{j,i}$, then the status quo mechanism G^δ is played with the updated beliefs about the other player's type to be $\bar{q}_{j,i}(t_j)$. For example, if player 1 vetoes ($i = 1$), then $\bar{q}_{1,1}(t_1)$ is the probability that player 2 (ratifier) believes about player 1's type being t_1, and $\bar{q}_{2,1}(t_2)$ is the probability that player 1 (vetoer) believes about player 2's type being t_2. Let $\bar{q}_{\cdot,i} = (\bar{q}_{2,i}(t_2), \bar{q}_{1,i}(t_1))$ be a vector of updated beliefs conditional on the event $M = \{i\}$, in which $M = \{i\}$ denotes the vote outcome that i vetoes the alternative while the other does not.

Now, for any r_j in R_j, let $\sigma_{j,i}(r_j|t_j)$ denote the probability that j would use the report r_j and let $\psi_{j,i}(t_j)$ denote the probability that j would go to war, if t_j were his type and if G^δ is played following a veto by i. For any j and any t_j, a strategy $(\sigma_{j,i}(r_j|t_j), \psi_{j,i}(t_j))$ in G^δ represents a plan for player j if t_j were his type and γ did not win as a result of i's veto.

From these definitions, the quantities $(v, \sigma, \psi, \bar{q})$ must be non-negative and must satisfy

$$\sum_{t_j} \bar{q}_{j,i}(t_j) = 1, \ \sum_{r_j \in R_j} \sigma_{j,i}(r_j|t_j) = 1, \ \psi_{j,i}(t_j) \leq 1, \ v_j(t_j) \leq 1, \ \forall i, \ \forall j, \ \forall t_j \in T_j. \tag{T1}$$

Given any types vector $t = (t_1, t_2)$ in T and given any possible reports vector $r = (r_1, r_2)$ in R, let $\sigma_i(r|t) = \prod_{j=1}^{2} \sigma_{j,i}(r_j|t_j)$. Thus, $\sigma_i(r|t)$ is the probability that the individuals would give reports (r_1, r_2), if their types were (t_1, t_2) and if G^δ is played by i's veto.

I must impose some restrictions on the strategies $\sigma_{j,i}$ and $\psi_{j,i}$, and the posteriors $\bar{q}_{j,i}$ when G^δ is

played as a result of i's veto. For any i, let $\Sigma(\bar{q}_{\cdot,i})$ denote a Nash equilibrium in the subgame when γ does not win by i's veto with respect to the revised beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{2,i}(t_2), \bar{q}_{1,i}(t_1))$. Then, the strategy profile $(\sigma_{\cdot,i}, \psi_{\cdot,i}) = ((\sigma_{1,i}(r_1|t_1), \sigma_{2,i}(r_2|t_2)), (\psi_{1,i}(t_1), \psi_{2,i}(t_2)))$ form an equilibrium $\Sigma(\bar{q}_{\cdot,i})$ of G^δ with respect to the posterior beliefs $(\bar{q}_{2,i}(t_2),\ \bar{q}_{1,i}(t_1))$ if and only if:

$$\begin{aligned} U_j(G^\delta|t_j, \Sigma(\bar{q}_{\cdot,i})) = & \sum_{t_{-i}} \bar{q}_{-j,i}(t_{-j})\{\psi_{j,i}(t_j)u_j(d_0,t) + (1-\psi_{j,i}(t_j)) \cdot (\psi_{-j,i}(t_{-j})u_j(d_0,t) \\ & + (1-\psi_{-j,i}(t_{-j})) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r)u_j(d,t))\} \\ & \geq U_j(G^\delta, \hat{r}_j, \psi'_{j,i}|t_j, \Sigma(\bar{q}_{\cdot,i})) \\ & = \sum_{t_{-i}} \bar{q}_{-j,i}(t_{-j})\{\psi'_{j,i}(t_j)u_j(d_0,t) + (1-\psi'_{j,i}(t_j)) \cdot (\psi_{-j,i}(t_{-j})u_j(d_0,t) \\ & \quad + (1-\psi_{-j,i}(t_{-j})) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r_{-j}, \hat{r}_j)u_j(d,t))\}, \\ & \qquad \forall j,\ \forall t_j \in T_j,\ \forall \hat{r}_j \in R_j,\ \forall \psi'_{j,i} \in [0,1]. \end{aligned} \tag{T2}$$

Condition (T2) asserts that player j with type t_j should not expect any other report $\hat{r}_j$ or any other war strategy $\psi'_{j,i}$ to be better for him in G^δ than the strategies selected by his $\sigma_{j,i}$ and $\psi_{j,i}$ when i vetoes and G^δ is played.

Let $\bar{q} = \{\bar{q}_{\cdot,1}, \bar{q}_{\cdot,2}\}$ be a collection of *vote beliefs vectors* following any veto by a single player when the other player votes for ratification. Let $\Sigma(\bar{q}) = \{\Sigma(\bar{q}_{\cdot,1}), \Sigma(\bar{q}_{\cdot,2})\}$ denote a corresponding collection of equilibria in the resulting play of G^δ with those vote beliefs.

Unanimous ratification of γ followed by truthful revelation in γ is a sequential equilibrium in the two-stage ratification game if and only if γ is incentive compatible and γ is *individually rational relative to* G^δ, that is, there exists $\bar{q}$ and $\Sigma(\bar{q})$ such that for each i and each t_i:

$$\begin{aligned} \sum_{t_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d\in D} \gamma(d|t)u_i(d,t) & \geq \sum_{t_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d\in D} \gamma(d|t_{-i}, \hat{t}_i)u_i(d,t); \\ \text{and } \sum_{t_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d\in D} \gamma(d|t)u_i(d,t) & \geq \sum_{t_{-i}} \bar{p}_{-i}(t_{-i})\{\psi'_{i,i}(t_i)u_i(d_0,t) \\ & \quad + (1-\psi'_{i,i}(t_i)) \cdot (\psi_{-i,i}(t_{-i})u_i(d_0,t) \\ & \quad + (1-\psi_{-i,i}(t_{-i})) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r_{-i}, \hat{r}_i)u_i(d,t))\}, \\ & \qquad \forall \hat{t}_i \in T_i,\ \forall \psi'_{i,i} \in [0,1],\ \forall \hat{r}_i \in R_i. \end{aligned} \tag{T3}$$

That is, (T3) asserts that player i cannot gain by vetoing the alternative γ (and then going to war with probability $\psi'_{i,i}$ and reporting $\hat{r}_i$ if G^δ is played) when t_i is his true type and the other player is expected to use her equilibrium $\psi_{-i,i}$ and $\sigma_{-i,i}$ strategies (in G^δ characterized by an equilibrium $\Sigma(\bar{q}_{\cdot,i})$ given $\bar{q}_{\cdot,i}$), together with honest reporting in γ. The left-hand side is t_i's interim payoff with the mediator γ, which corresponds to the equilibrium path for the equilibrium in which all types of all players vote for ratification of γ.

I now impose restrictions on beliefs in the ratification game that require an equilibrium to be supported by all credible updating rules for rationalizing observations off the equilibrium path, based on Cramton and Palfrey (1995). If i vetoes, even if that event has zero probability, $-i$ will try to rationalize the veto by identifying a veto belief that is consistent with i's incentive to veto. In addition, vetoer i would have some posterior subjective probability distribution over T_{-i} upon observing he was the only one to veto and learning that the other player did not veto. Define the veto set $V_i \subseteq T_i$ as those types that veto with positive probability: $V_i = \{t_i \in T_i | v_i(t_i) > 0\} \neq \varnothing$, and the non-veto set $W_{-i} = T_{-i} \backslash V_{-i}$ as those types that never veto.

Credibility Conditions. A probability distribution $\bar{q}_{i,i}(t_i)$ on T_i is a *credible veto belief* about vetoer i and a probability distribution $\bar{q}_{-i,i}(t_{-i})$ on T_{-i} is a *credible non-veto belief* about ratifier $-i$ if there exists a continuation equilibrium $\Sigma(\bar{q}_{\cdot,i})$ $(= (\sigma_{\cdot,i}, \psi_{\cdot,i}))$ in G^δ with respect to the beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{i,i}(t_i), \bar{q}_{-i,i}(t_{-i}))$ and veto probabilities $v_i(\cdot)$ such that $\bar{q}_{\cdot,i}$, $\Sigma(\bar{q}_{\cdot,i})$, and $v_i(\cdot)$ together satisfy the conditions $(T4)$ through $(T9)$:

$$v_i(t_i) > 0 \text{ for some } t_i \in T_i. \tag{T4}$$

If i (vetoer) of type t_i would benefit from a veto, then he must veto γ:

$$\begin{aligned} \text{if } \sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \sum_{d \in D} \gamma(d|t) u_i(d,t) < & \sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \{ \psi_{i,i}(t_i) u_i(d_0, t) \\ & + (1 - \psi_{i,i}(t_i)) \cdot (\psi_{-i,i}(t_{-i}) u_i(d_0, t) \\ & + (1 - \psi_{-i,i}(t_{-i})) \sum_{r \in R} \sigma_i(r|t) \sum_{d \in D} \delta(d|r) u_i(d,t)) \}, \\ & \text{then } v_i(t_i) = 1. \end{aligned} \tag{T5}$$

Those types t_i that lose from a veto must not veto γ:

$$\begin{aligned}\text{if } \sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \sum_{d\in D} \gamma(d|t) u_i(d,t) > \sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \{ & \psi_{i,i}(t_i) u_i(d_0,t) \\ & + (1-\psi_{i,i}(t_i)) \cdot (\psi_{-i,i}(t_{-i}) u_i(d_0,t) \\ & + (1-\psi_{-i,i}(t_{-i})) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r) u_i(d,t))\}, \\ \text{then } v_i(t_i) = 0.\end{aligned} \tag{T6}$$

Also, $-i$ (ratifier) of type t_{-i} who benefit from a vote on γ when i vetoes must not veto γ:

$$\begin{aligned}\text{if } & \sum_{t_i} \bar{q}_{i,i}(t_i) \{ \psi_{-i,i}(t_{-i}) u_{-i}(d_0,t) + (1-\psi_{-i,i}(t_{-i})) \cdot (\psi_{i,i}(t_i) u_{-i}(d_0,t) \\ & + (1-\psi_{i,i}(t_i)) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r) u_{-i}(d,t))\} \\ & > \sum_{t_i} \bar{q}_{i,i}(t_i) \sum_{d\in D} \delta(d|t) u_{-i}(d,t) \\ & \text{then } v_{-i}(t_{-i}) = 0.\end{aligned} \tag{T7}$$

Furthermore, $\bar{q}_{\cdot,i} = (\bar{q}_{i,i}(t_i), \bar{q}_{-i,i}(t_{-i}))$ satisfies Bayes' rule, given $\bar{p}$ and v:

$$\bar{q}_{i,i}(t_i) = \begin{cases} \frac{\bar{p}_i(t_i) v_i(t_i)}{\sum_{t_i \in V_i} \bar{p}_i(t_i) v_i(t_i)} & for\ t_i \in V_i \\ 0 & for\ t_i \notin V_i. \end{cases} \tag{T8}$$

$$\bar{q}_{-i,i}(t_{-i}) = \begin{cases} \frac{\bar{p}_{-i}(t_{-i})(1-v_{-i}(t_{-i}))}{\sum_{t_{-i}\in W_{-i}} \bar{p}_{-i}(t_{-i})(1-v_{-i}(t_{-i}))} & for\ t_{-i} \in W_{-i} \\ 0 & for\ t_{-i} \notin W_{-i}. \end{cases} \tag{T9}$$

The set of types V_i that veto with positive probability in $\bar{q}_{i,i}(t_i)$ is called a *credible veto set*. The set of types W_{-i} that do not veto with positive probability in $\bar{q}_{-i,i}(t_{-i})$ is called a *credible non-veto set*.

Conditions (T8) and (T9) require that ratifier $-i$ and vetoer i update their beliefs about the other by using Bayes theorem to compute their posteriors $\bar{q}_{i,i}(t_i)$ and $\bar{q}_{-i,i}(t_{-i})$, respectively. Let $\bar{q}_{\cdot,i} = (\bar{q}_{i,i}(t_i), \bar{q}_{-i,i}(t_{-i}))$ be called a vector of *credible vote beliefs* if (T4) through (T9) are all satisfied. The credibility conditions have a flavor similar to the motivation behind Grossman and Perry (1986) and Farrell (1985). Roughly speaking, a set of types V_i breaks an equilibrium with

an off-the-equilibrium-path "message" if all types in V_i improve their payoff by vetoing as long as the ratifier believes that only all of the types in V_i would always deviate and veto. Here, $-i$ is specifically assumed to update her prior beliefs over V_i in accordance with Bayes' Rule. However, no restriction is on v_i for types that are indifferent between vetoing and not vetoing, which reflects the idea that such types may randomly choose whether to deviate.

The credibility conditions ask whether, upon deviating, the vetoer can induce the other player to reason that the "message" must have been sent by a type within V_i. That is, whether a set of types can credibly distinguish themselves by explicitly sending an off-the-path-equilibrium message of "we are the types in V_i." This is justified when V_i is precisely the set of vetoer's types that would benefit from deviating whenever, in response to deviating, the other player plays any best response to i's veto with her beliefs restricted to V_i. Thus, the types that do not belong to V_i are worse off (compared to their equilibrium payoff) under such best response. An equilibrium fails the credibility conditions if the types in V_i are precisely the ones who gain by vetoing, when the ratifier's beliefs are a Bayesian update of her prior beliefs on V_i.

The credibility condition (T7) assumes away the case of i not vetoing, in which case $-i$'s veto would switch from γ to G^δ. But as long as $-i$'s beliefs are restricted to V_i such that the other is expected to veto, $-i$ must expect that her vote cannot make any difference because G^δ will be implemented in any case. (T7) requires that if $-i$ with type t_{-i} would get higher expected utility from the equilibrium $\Sigma(q_{\cdot,i})$ of G^δ than the expected utility from her also vetoing, then $-i$ must divulge her desire for γ by voting for γ. Note that the left-hand side of (T7) is the interim payoff for player $-i$ of type t_{-i} when $-i$ ratifies but i vetoes, and the right-hand side of (T7) is the interim payoff to player $-i$ of type t_{-i} when $-i$ also vetoes and G^δ is played under the original beliefs (hiring δ and reporting honestly). The condition assumes that the likely vetoer i will have the prior belief about $-i$ if $-i$ does veto, which makes sense as all $t_{-i} \in W_{-i}$ are expected to not veto (as $-i$ of t_{-i} would benefit from revealing her non-veto). This assumption also justifies that in (T5) and (T6), any possible change of behavior of $-i$ in G^δ when $-i$ also vetoes but learns that i vetoed is assumed away.

Remark 1. I do not entertain the possibility of multiple ways to rationalize a veto or a non-veto. In my setting, there is at most one credible vote belief for any i and so, at most one credible veto set that satisfies (T4) through (T9), and thus there is no issue of multiple credible vote beliefs.

Definition 1. $(v, \sigma, \psi, \bar{q})$ is an *equilibrium ratification* of γ when the status quo is G^δ, if and only if γ satisfies the conditions (T1) – (T3) and for all i either

(i) there does not exist a credible veto belief, or

(ii) for every credible vote belief $\bar{q}_{\cdot,i}$ such that the conditions (T4) – (T9) are all satisfied,

$$\begin{aligned}\sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \sum_{d\in D} \gamma(d|t) u_i(d,t) = &\sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \{\psi_{i,i}(t_i) u_i(d_0,t) \\ &+ (1-\psi_{i,i}(t_i)) \cdot (\psi_{-i,i}(t_{-i}) u_i(d_0,t) \\ &+ (1-\psi_{-i,i}(t_{-i})) \sum_{r\in R} \sigma_i(r|t) \sum_{d\in D} \delta(d|r) u_i(d,t))\}, \ \forall t_i \in V_i.\end{aligned}$$

Definition 2. An alternative mediator γ is *ratifiable against* the mediator δ if and only if there exists an equilibrium ratification of γ when the status quo is G^δ.

An alternative mediator γ is ratifiable against the mediator δ if and only if there exists a sequential equilibrium of the two-stage game in which γ is unanimously voted for at some profile of types, where beliefs following non-ratification are required to satisfy the credibility conditions. If after a veto, the players' beliefs are restricted to credible vote beliefs, then the alternative mediator is ratifiable.

An alternative mediator γ is *not* ratifiable against δ if and only if there is no equilibrium to the two-stage game in which γ is unanimously approved over δ along some equilibrium path. This definition suggests that rejection of γ is not so severe. Given the definition of a ratifiable alternative mediator, I can now define a secure status quo mechanism and a threat-secure mediator.

Definition 3. A status quo mechanism G^δ is *secure* if and only if every interim incentive efficient mediator that is ratifiable against δ is an equilibrium of G^δ under the prior beliefs.

A secure mechanism cannot be overturned by a ratifiable alternative. Moreover, if δ is ratifiable against itself, then such a status quo mechanism G^δ is not vulnerable to allowing players to send preplay cheap talk messages that might communicate information about their types.

Definition 4. An interim incentive efficient mediator δ is *threat-secure* if and only if there does not exist an alternative mediator that is ratifiable against δ.

I say that a threat-secure mediator is immune to the ratification of any alternatives and endures a unilateral war threat between the players, given reasonable updating in the event of a veto.

To establish that a mechanism is secure, every other symmetric incentive feasible direct mechanism in $S(\Gamma)$ has to be checked whether there is a credible vote belief in each case. This search is simplified by establishing the following Propositions 5 and 6, which are also used to prove an existence result.

Proposition 5. *If γ is ex ante Pareto superior to δ, then γ is not ratifiable against δ.*

Proposition 6. *If γ is ex ante Pareto inferior to δ, then unanimous ratification of γ, when the status quo is G^δ, is a sequential equilibrium of the ratification game where beliefs following disagreement are required to satisfy the credibility conditions.*

Proposition 5 implies that an alternative mediator is not ratifiable against any ex ante Pareto inferior mediator, and Proposition 6 implies that an alternative mediator is ratifiable against any ex ante Pareto superior mediator. Now, I can establish the existence of the interim incentive efficient and threat-secure mediator in the class of environments discussed in this paper. The following Proposition 7 characterizes the set of threat-secure mediators on the interim Pareto frontier, denoted as $TS(\Gamma)$.[25]

Theorem 2. *For any two-person Bayesian bargaining problem Γ that satisfies **(A1)** – **(A3)** with independent and symmetric priors, there exists a unique interim incentive efficient and secure status quo mechanism.*

Proposition 7. *The threat-secure mediator in $S(\Gamma)$ is*

If $p < p^*$: $TS(\Gamma) = \{\mu_1\}$ *with* $z = 0$.

If $p \in [p^*, p^{**})$: $TS(\Gamma) = \{\mu_1\}$ *with* $z = \bar{z}(p)$.

If $p \geq p^{**}$: $TS(\Gamma) = \{\mu_0\}$.

[25] We can also think about a ratification game between no mediation versus mediation in which the status quo G^δ is an unmediated Bayesian game. I consider this possibility in Online Appendix D. I show that all of the interim incentive efficient mediators are ratifiable against the underlying disagreement game, and so "no mediation" is not secure.

Proposition 7 states that a threat-secure mediator is uniquely defined to be in the set $TS(\Gamma)$. If $p < p^*$, then the threat-secure mediator chooses war with probability one when two players are of different types. If $p \in [p^*, p^{**})$, then the threat-secure mediator chooses war with probability one when two players are of different types and with a positive probability when both players are the strong type. If $p \geq p^{**}$, then the peaceful mediator who always chooses war with probability zero is trivially the threat-secure mediator because there are no other alternatives.

In this ratification game, Proposition 7 together with Propositions 5 and 6 imply that, if the players start with the ex ante incentive efficient mediator, then there is a sequential equilibrium in which the players immediately ratify an alternative mediator to replace the ex ante incentive efficient one. On the other hand, if the players start with the threat-secure mediator, then there is no sequential equilibrium in which the players unanimously agree to an alternative ex ante Pareto superior one. That is, once the players insist on the threat-secure mediator, they are stuck with the mediator and can never renegotiate for a more ex ante efficient one.

This result implies that the unique threat-secure mediator is the one that is not ex ante incentive efficient, and indeed is the most ex ante inefficient. The suboptimal choice of such a mediator is in fact the most war-inducing mediator associated with the highest probability of failing to reach a peaceful settlement among all the other interim incentive efficient mediators.

6 Discussions: Equivalence of the Two Approaches

In this section, I look at the relation between the ex ante incentive efficient mediator, the neutral bargaining solution, and the threat-secure mediator, and discuss the implications of the solution set refined by the cooperative approach and the equilibrium choice in the noncooperative approach.

The following is the main result of this paper: the neutral bargaining solution coincides with the threat-secure mediator in the stylized game in my framework.

Theorem 3. *For any two-person Bayesian bargaining problem* Γ *that satisfies assumptions* ***(A1)*** *–* ***(A3)*** *with independent and symmetric priors,* $TS(\Gamma) = NS(\Gamma) \subset S(\Gamma)$*. Moreover, the symmetric interim incentive efficient and threat-secure mediator is not ex ante incentive efficient.*

Theorem 3 implies that there is a unique symmetric mediator that is interim incentive efficient, the neutral bargaining solution, and threat-secure; and is distinct from the ex ante incentive efficient solution. In general, the neutral bargaining solution and the threat-secure set are not nested, but

in the class of bargaining games considered in this paper, they are the same. Then, we can ask the deeper game-theoretic question of why the threat-secure set generates the same outcome as the neutral bargaining solution even though they stem from such different approaches – one from a mostly noncooperative approach and the latter from a completely cooperative approach.

Proposition 5 implies that, in the pairwise ratification game between mediators, there is no equilibrium where an alternative ex ante Pareto superior mediator (who puts lower probability on war) is chosen in the first stage if the probability of the strong type is sufficiently small. By voting, a player could send a signal that he is of a certain type. In any case, the strong type benefits from revealing his type when the possibility of other mediators is considered; thus, the strong type's signal is credible and satisfies the credibility requirements. However, the weak type's veto does not give a credible signal because if he gives away his identity there is a disadvantage to vetoing. In equilibrium, every player strategically pools on the strong action. Thus, I can argue that an ex ante Pareto better mediator gets defeated and players end up choosing an ex ante inefficient mediator who puts more probability on war than the ex ante efficient mediator. If this pairwise ratification game is iteratively played between all possible interim incentive efficient mediators with any order, an ex ante Pareto superior mediator is not sustainable against ratification of any other ex ante Pareto inferior mediator. Thus, the threat-secure mediator defined in Proposition 7 is the only one that survives ratification of any other alternative mediator.

The strategic intuition to why the ex ante incentive efficient mediator cannot survive in the noncooperative game underscores the reasoning for the information leakage problems behind the cooperative game. Proposition 4 implies that any negotiating process should lead to the neutral bargaining solution. Both types do not want to reveal their types in the negotiating process, and the inscrutable intertype compromise between the two types leads to implicitly putting more weight on the strong type. Thus, both types pretend to be strong in order to conceal their types, and so, the "worst" mediator turns out to be the reasonable solution we should expect to see endogenously arising in a mediator selection game.

In a cooperative sense, players should pick a mediator, effectively *pooling* in a way that no information is revealed. This nature of pooling is exactly reminiscent of *signaling* in a noncooperative game. In the ratification game, every player pretends to be strong to signal their types, and so the "worse" mediator is chosen in equilibrium. The combination of the inscrutability principle and the requirements in the neutral bargaining solution generate this kind of built-in signaling in a cooperative game. Likewise, I can argue that the observation of the players mimicking other players

in the noncooperative game comes directly from a cooperative approach.

Therefore, both approaches take into account, either directly or indirectly, the information leakage problem. By either approach, I can conclude that the selection of a mediator by privately informed players is endogenous and depends on the information that the players reveal by expressing their preferences for mediation. Consequently, with both approaches, I can define the smallest solution set for the given bargaining problem: When two privately informed players bargain over the choice of a mediator, the reasonable solution of a mediator in a cooperative sense is the neutral bargaining solution; and the equilibrium solution in a noncooperative sense is the threat-secure mediator. Remarkably, I obtain the same unique solution that is farthest away from ex ante incentive efficiency. Moreover, in the class of bargaining games considered in this paper, the intuition behind a noncooperative game can be founded on the very intuition behind a cooperative bargaining theory.

7 Concluding Comments

This paper is an attempt to build a theory of how players with private information might agree on a mediator. The general tenor of the results suggest that two parties in potential conflict endogenously choose the "worst" mediator, and the information leakage problem is part of the answer as to how the mediator is chosen. The "worst" mediator is the one who is not ex ante incentive efficient and is indeed the farthest away. The suboptimal choice of such a mediator is associated with the highest probability of failing to reach a peaceful settlement. Moreover, two different approaches lead to the same outcome.

The novelty of my paper is that it points towards a more general connection between cooperative and noncooperative game theories. Essentially, I merge a cooperative idea with a noncooperative idea. That is, I first describe what the reasonable solution should look like in a cooperative sense, with the relevant properties satisfied – the unique neutral bargaining solution. Then, I ask whether there exists a threat-secure mediator that survives a vote against any other mediator in an extensive form ratification game. The neutral bargaining solution is equivalent the threat-secure mediator, and in fact, is the only threat-secure mediator. In a sense, this paper represents a first attempt to build a bridge between a cooperative bargaining theory and a noncooperative ratification game.

In this paper, I assume that a player cannot pay off another player to prevent war. Without this assumption, the analysis becomes even more complicated, but it would be desirable to relax the

assumption. On another note, if I take the model and raise the payoffs such that choosing the ex ante efficient mediator becomes more valuable for the weak types; then the choice of mediator does not change, and the outcome actually gives a higher interim payoff for the weak types. But in fact, the inefficient mediator is chosen by a wider range of the probability of the strong type. This result implies that what causes players to more likely move toward an ex ante efficient mediator renders the "signaling" problem more perverse. Thus, even though players' types are close to wanting to have an ex ante efficient outcome in which the world should be a better place; in fact, things get worse.

Through this paper, I hope to make a contribution to wider applications of third party intervention to explain how and why a third party is chosen inefficiently in various economic, social, and political interactions. One of many applications of my model is to international conflict.[26] Mediation, arbitration, and peace talks are thought of as optimal international conflict resolution institutions that attempt to improve the efficiency of the negotiation. However, when adversaries attempt peace negotiations over issues such as territorial boarders or security, they do not always choose efficient third parties. Although I have focused my application of the model to international relations, the model could be equally applicable to other forms of bargaining games, such as collective bargaining negotiations between firms and unions in an industrial relations context. In addition, in an economic setting, the research could look into situations in which two or more firms jointly bargain over selecting an investment bank to advise them on mergers.

References

Cramton, Peter C. and Thomas R. Palfrey. 1995. "Ratifiable Mechanisms: Learning from Disagreement." *Games and Economic Behavior* 10(2):255–283.

Esö, Péter and James Schummer. 2009. "Credible Deviations from Signaling Equilibria." *International Journal of Game Theory* 48(3):411–430.

Farrell, Joseph. 1985. "Communication, Coordination and Nash Equilibrium." *Economics Letters* 27(3):209–214.

Fey, Mark and Kristopher W. Ramsay. 2009. "Mechanism Design Goes to War: Peaceful Outcomes with Interdependent and Correlated Types." *Review of Economic Design* 13(3):233–250.

[26] Mediation efforts are a recurrent and potentially important feature of international conflict, and international mediation is an intriguing facet of international conflict.

Fey, Mark and Kristopher W. Ramsay. 2010. "When is Shuttle Diplomacy Worth the Commute? Information Sharing through Mediation." *World Politics* 62(4):529–560.

Fey, Mark and Kristopher W. Ramsay. 2011. "Uncertainty and Incentives in Crisis Bargaining: Game-Free Analysis of International Conflict." *American Journal of Political Science* 55(1):149–169.

Grossman, S.J. and M. Perry. 1986. "Perfect Sequential Equilibrium." *Journal of Economic Theory* 39(1):97–119.

Hafer, Catherine. 2008. "Conflict over Political Authority." Unpublished.

Harsanyi, John C. and Reinhard Selten. 1972. "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information." *Management Science* 18(5):80–106.

Holmström, Bengt and Roger B. Myerson. 1983. "Efficient and Durable Decision Rules with Incomplete Information." *Econometrica* 51(6):1799–1819.

Hörner, Johannes, Massimo Morelli and Francesco Squintani. 2011. "Mediation and Peace." Unpublished.

Kydd, Andrew. 2003. "Which Side Are You On? Bias, Credibility, and Mediation." *American Journal of Political Science* 47(4):597–611.

Lagunoff, Roger D. 1995. "Resilient Allocation Rules for Bilateral Trade." *Journal of Economic Theory* 66(2):463–487.

Maskin, Eric and Jean Tirole. 1990. "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values." *Econometrica* 58(2):379–409.

Meirowitz, Adam, Massimo Morelli, Kristopher W. Ramsay and Francesco Squintani. 2012. "Mediation and Strategic Militarization." Unpublished.

Milgrom, Paul and Nancy Stokey. 1982. "Information, Trade and Common Knowledge." *Journal of Economic Theory* 26(1):17–27.

Myerson, Roger B. 1979. "Incentive Compatibility and the Bargaining Problem." *Econometrica* 47(1):61–74.

Myerson, Roger B. 1983. "Mechanism Design by an Informed Principal." *Econometrica* 51(6):1767–1797.

Myerson, Roger B. 1984*a*. "Cooperative Games with Incomplete Information." *International Journal of Game Theory* 13(2):69–96.

Myerson, Roger B. 1984*b*. "Two-Person Bargaining Problems with Incomplete Information." *Econometrica* 52(2):461–488.

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict*. Cambridge, M.A.: Harvard University Press.

Myerson, Roger B. and Mark A. Satterthwaite. 1983. "Efficient Mechanisms for Bilateral Trading." *Journal of Economic Theory* 29(2):265–281.

Nalebuff, Barry. 1987. "Credible Pretrial Negotiation." *The RAND Journal of Economics* 18(2):198–210.

Nash, John F. 1950. "The Bargaining Problem." *Econometrica* 18(2):155–162.

Nash, John F. 1953. "Two-Person Cooperative Games." *Econometrica* 21(1):128–140.

Rosenthal, Robert W. 1978. "Arbitration of Two-Party Disputes under Uncertainty." *Review of Economic Studies* 45(3):595–604.

Appendix A: Proofs

Supplementary Explanations and Proofs of Section 3

An interim incentive efficient mechanism in the incentive feasible set F is any mechanism that is an optimal solution to an optimization problem of the form

$$\max_{\mu^j \in F} \sum_{t \in T} p(t) \sum_{d \in D} \mu(d|t) \sum_{i \in N} \lambda_i(t_i) u_i(d,t)$$

where $p(t) = \prod_{i=1}^{2} \bar{p}_i(t_i)$ and the utility weight $\lambda_i(t_i)$ is a positive number for each player i and each type t_i. The type sets and outcomes sets are all finite. Under the finiteness assumption, the set of incentive feasible mechanisms is a convex polyhedron. By the supporting hyperplane theorem, an incentive feasible mechanism μ is incentive efficient if and only if there exist some positive numbers $\lambda_i(t_i)$ for each type t_i of each player i such that μ is an optimal solution to the optimization problem

$$\max_{\mu: D \times T \to \mathbb{R}} \sum_{i \in N} \sum_{t_i \in T_i} \lambda_i(t_i) U_i(\mu|t_i)$$

$$\text{s.t. } U_i(\mu|t_i) \geq U_i^*(\mu, s_i|t_i), \ \forall i \in N, \ \forall t_i \in T_i, \ \forall s_i \in T_i,$$

$$U_i(\mu|t_i) \geq 0, \ \forall i \in N, \ \forall t_i \in T_i,$$

$$\sum_{d \in D} \mu(d|t) = 1 \text{ and } \mu(d|t) \geq 0 \ \forall d \in D, \ \forall t \in T,$$

where $U_i^*(\mu, s_i|t_i) = \sum_{t_{-i} \in T_{-i}} \sum_{d \in D} \bar{p}_{-i}(t_{-i}) \mu(d|t_{-i}, s_i) u_i(d,t)$. Because the objective and constraints are all linear in μ, this optimization problem is a linear programming problem. Because players are symmetric in type-dependent payoffs, I can set $\lambda_1(s) = \lambda_2(s) \equiv \lambda(s)$ and $\lambda_1(w) = \lambda_2(w) \equiv \lambda(w)$; and normalize such that $\lambda(s) + \lambda(w) = 1$ without loss of generality. For the above optimization problem, a Lagrangean function can be formed. Let $\alpha_i(s_i|t_i)$ denote the Lagrange multiplier for the incentive compatibility constraints and $\beta_i(t_i)$ denote the Lagrange multiplier for the individual rationality constraints. Also, I can set $\alpha_1(w|s) = \alpha_2(w|s) \equiv \alpha(w|s)$, $\alpha_1(s|w) = \alpha_2(s|w) = \alpha(s|w)$, $\beta_1(s) = \beta_2(s) \equiv \beta(s)$, and $\beta_1(w) = \beta_2(w) \equiv \beta(w)$. Let $\alpha = (\alpha(w|s), \alpha(s|w))$ and $\beta = (\beta(s), \beta(w))$. Then the Lagrangean function can be written as

$$\sum_{i \in N} \sum_{t_i \in T_i} \lambda(t_i) U_i(\mu|t_i) + \sum_{i \in N} \sum_{t_i \in T_i} \alpha(s_i|t_i) \left(U_i(\mu|t_i) - U_i^*(\mu, s_i|t_i) \right) + \sum_{i \in N} \sum_{t_i \in T_i} \beta(t_i) U_i(\mu|t_i).$$

Let $v_i(d,t,\lambda,\alpha,\beta) = [(\lambda(t_i) + \sum_{s_i \in T_i} \alpha(s_i|t_i) + \beta(t_i))u_i(d,t) - \sum_{s_i \in T_i} \alpha(t_i|s_i)u_i(d,(t_{-i},s_i))]/\bar{p}_i(t_i)$. This $v_i(d,t,\lambda,\alpha,\beta)$ is called the virtual utility payoff to player i from outcome d, when the type profile is t, with respect to the utility weights λ and the Lagrange multipliers α and β. With this setup, I arrive at the most tractable conditions for computing the interim incentive efficient mediators, which are used to prove Lemma 3, Proposition 1, and Proposition 2. Without loss of generality, I can focus on $i = 1$.

Theorem (Theorem 10.1, Myerson (1991)). *An incentive feasible mediator who mediates according to an incentive compatible and individually rational mediation mechanism μ is incentive efficient if and only if there exist vectors $\lambda = (\lambda(t_i))_{t_i \in T_i}$, $\alpha = (\alpha(s_i|t_i))_{s_i \in T_i, t_i \in T_i}$, and $\beta = (\beta(t_i))_{t_i \in T_i}$ such that*

$$\begin{aligned} \lambda(t_i) > 0,\ \alpha(s_i|t_i) \geq 0,\ \beta(t_i) \geq 0,\ \forall t_i \in T_i,\ \forall s_i \in T_i \\ \alpha(s_i|t_i)(U_i(\mu|t_i) - U_i^*(\mu, s_i|t_i)) = 0,\ \forall t_i \in T_i,\ \forall s_i \in T_i, \\ \beta(t_i)U_i(\mu|t_i) = 0,\ \forall t_i \in T_i, \\ \sum_{d \in D} \mu(d|t) \sum_{i \in N} v_i(d,t,\lambda,\alpha,\beta) = \max_{d \in D} \sum_{i \in N} v_i(d,t,\lambda,\alpha,\beta),\ \forall t \in T. \end{aligned} \tag{7.1}$$

Remark. Henceforth, in order to prevent any confusion, I will use double subscripts on μ specifying both y and z, i.e., $\mu_{y,z}$.

Lemma (Thresholds). *There exists non-negative $p^{'}$, p^*, and p^{**} such that*

$$\begin{aligned} p^{'} &\equiv \frac{-u_1(d_1,sw)}{u_1(d_1,ss) - u_1(d_1,sw)} = \frac{-u_2(d_1,ws)}{u_2(d_1,ss) - u_2(d_1,ws)}, \\ p^* &\equiv \frac{u_1(d_1,ww)}{u_1(d_1,ww) + u_1(d_1,ws)} = \frac{u_2(d_1,ww)}{u_2(d_1,ww) + u_2(d_1,sw)}, \\ p^{**} &\equiv \frac{u_1(d_1,ss)u_1(d_1,ww) - u_1(d_1,sw)u_1(d_1,ws)}{u_1(d_1,ss)u_1(d_1,ww) - u_1(d_1,sw)u_1(d_1,ws) + u_1(d_1,ss)u_1(d_1,ws)} \\ &= \frac{u_2(d_1,ss)u_2(d_1,ww) - u_2(d_1,ws)u_2(d_1,sw)}{u_2(d_1,ss)u_2(d_1,ww) - u_2(d_1,ws)u_2(d_1,sw) + u_2(d_1,ss)u_2(d_1,ws)}, \end{aligned}$$

$p^{**} > p^*$, and $p^{**} > p^{'}$.

Proof of Lemma 2. $p^{'}$ is computed such that the participation constraints bind for the strong type

given a mediator $\mu_{0,0}$, where $\mu_{0,0}(d_1|t) = 1$ for all t, that is: $U_1(\mu_{0,0}|s)|_{p=p'} = 0$

$$\longleftrightarrow p^{'}\mu_{0,0}(d_1|ss)u_1(d_1, ss) + (1-p^{'})\mu_{0,0}(d_1|sw)u_1(d_1, sw) = 0,$$

which gives $p^{'} = \frac{-u_1(d_1,sw)}{u_1(d_1,ss)-u_1(d_1,sw)}$. By **(A3)**, $\frac{-u_1(d_1,sw)}{u_1(d_1,ss)-u_1(d_1,sw)} = \frac{-u_2(d_1,ws)}{u_2(d_1,ss)-u_2(d_1,ws)}$. p^* is such that the informational incentive constraints bind for the weak type given mediator $\mu_{1,0}$, where $\mu_{1,0}(d_1|ss) = \mu_{1,0}(d_1|ww) = 1$ and $\mu_{1,0}(d_1|sw) = \mu_{1,0}(d_1|ws) = 0$, i.e., $U_1(\mu_{1,0}|w)|_{p=p^*} = U_1(\mu_{1,0}, s|w)|_{p=p^*}$

$$\begin{aligned}\longleftrightarrow & p^*\mu_{1,0}(d_1|ws)u_1(d_1, ws) + (1-p^*)\mu_{1,0}(d_1|ww)u_1(d_1, ww)\\ & = p^*\mu_{1,0}(d_1|ss)u_1(d_1, ws) + (1-p^*)\mu_{1,0}(d_1|sw)u_1(d_1, ww),\end{aligned}$$

which gives $p^* = \frac{u_1(d_1,ww)}{u_1(d_1,ww)+u_1(d_1,ws)}$. By **(A3)**, $\frac{u_1(d_1,ww)}{u_1(d_1,ww)+u_1(d_1,ws)} = \frac{u_2(d_1,ww)}{u_2(d_1,ww)+u_2(d_1,sw)}$. Let y and z be computed such that the informational incentive constraints bind for the weak type given mediator $\mu_{y,z}$ where $\mu_{y,z}(d_1|ss) = 1-z$, $\mu_{y,z}(d_1|sw) = \mu_{y,z}(d_1|ws) = 1-y$, and $\mu_{y,z}(d_1|ww) = 1$, that is: $U_1(\mu_{y,z}|w) = U_1(\mu_{y,z}, s|w)$

$$\begin{aligned}\longleftrightarrow & p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww)\\ & = p(1-z)u_1(d_1, ws) + (1-p)(1-y)u_1(d_1, ww)\\ \longleftrightarrow & z := z(y,p) = y\left[1 - \frac{(1-p)\cdot u_1(d_1, ww)}{p\cdot u_1(d_1, ws)}\right],\end{aligned}$$

where $y \in [0,1]$. When $y = 1$, we let $\bar{z}(p) \equiv z(1,p) = 1 - \frac{(1-p)\cdot u_1(d_1,ww)}{p\cdot u_1(d_1,ws)}$. Then, p^{**} is such that the strong type is indifferent between mediator $\mu_{0,0}$ and mediator $\mu_{1,\bar{z}(p)}$, where $\mu_{1,\bar{z}(p)}(d_1|ss) = 1-\bar{z}(p)$, $\mu_{1,\bar{z}(p)}(d_1|sw) = \mu_{1,\bar{z}(p)}(d_1|ws) = 0$, and $\mu_{1,\bar{z}(p)}(d_1|ww) = 1$:

$$\begin{aligned}& U_1(\mu_{1,\bar{z}(p)}|s) = U_1(\mu_{0,0}|s)|_{p=p^{**}}\\ \longleftrightarrow & p^{**}\mu_{1,\bar{z}(p)}(d_1|ss)u_1(d_1, ss) + (1-p^{**})\mu_{1,\bar{z}(p)}(d_1|sw)u_1(d_1, sw)\end{aligned}$$

which gives $p^{**} = \frac{u_1(d_1,ss)u_1(d_1,ww)-u_1(d_1,sw)u_1(d_1,ws)}{u_1(d_1,ss)u_1(d_1,ww)-u_1(d_1,sw)u_1(d_1,ws)+u_1(d_1,ss)u_1(d_1,ws)}$. By **(A3)**, we can write p^{**} in terms of player 2's utilities. The threshold p^{**} is interpreted as the upper bound of p in order for $\mu_{y,z}$ to be interim incentive efficient. Note that $U_1(\mu_{1,\bar{z}(p)}|s) > U_1(\mu_{0,0}|s)$ at $p = p^*$, both $U_1(\mu_{1,\bar{z}(p)}|s)$ and $U_1(\mu_{0,0}|s)$ are decreasing in p, and $U_1(\mu_{1,\bar{z}(p)}|s) < U_1(\mu_{0,0}|s)$ at $p = 1$. Thus, by the single crossing property, there is a unique p^{**} such that $U_1(\mu_{1,\bar{z}(p^{**})}|s) = U_1(\mu_{0,0}|s)|_{p=p^{**}}$. Therefore, $p^{**} > p^* \longleftrightarrow -u_1(d_1, ws)u_1(d_1, sw) > 0$ since $u_1(d_1, ws) > 0$ and $u_1(d_1, sw) < 0$; and

$p^{**} > p^{'} \longleftrightarrow u_1(d_1, ss)u_1(d_1, ww) > 0$. However, $u_1(d_1, ss)u_1(d_1, ww) \geq -u_1(d_1, sw)u_1(d_1, ws)$ if and only if $p^* \geq p^{'}$; and $u_1(d_1, ss)u_1(d_1, ww) < -u_1(d_1, sw)u_1(d_1, ws)$ if and only if $p^{'} > p^*$. □

The Interpretations of the Three Thresholds. The first threshold $p^{'}$ is such that the participation constraint binds for the strong type given that the players are associated with mediator $\mu_{0,0}$, i.e., $U_1(\mu_{0,0}|s) = 0$. That is, when $p < p^{'}$, mediator $\mu_{0,0}$ is not individually rational to the strong types. So, the lowest probability some mediator can put on the war outcome when $p < p^{'}$ should be computed such that the participation constraint binds for the strong type given that mediator. The lower bound for the probability on war, denoted by $\underline{y}(p)$, is such that $U_1(\mu_{\underline{y}(p),0}|s) = 0$. This probability implies that when $p < p^{'}$, some mediators who put comparably lower probability than $\underline{y}(p)$ on the war outcome are not incentive feasible; in particular, not individually rational for the strong type to participate with. That is, when the strong type is relatively rare, if a player happens to be strong, then he would rather not participate and deviate to unilaterally forcing war because "always war" gives the strong player a strictly higher expected payoff than participating in mediation when he expects with high probability the other player is weak. To give the strong type an incentive to participate, the mechanism has to put at least some minimum level of probability on war when $t \in \{sw, ws\}$ in order for it to be incentive feasible.

The second threshold p^* is such that the weak type's informational incentive constraint binds for the given mediator $\mu_{1,0}$, that is, $U_1(\mu_{1,0}|w) = U_1(\mu_{1,0}, s|w)$. This implies that when $p > p^*$, mediator $\mu_{1,0}$ would no longer be incentive compatible for the weak type. Therefore, when $p \in (p^*, p^{**})$, a mediator must also put some positive probability on war when both players are of the strong type to prevent the weak type from reporting dishonestly that she is the strong type, such that $\mu_{y,z}(d_0|ss) = z > 0$, where z must satisfy $U_1(\mu_{y,z}|w) = U_1(\mu_{y,z}, s|w)$ for any given $y \in (0, 1]$.

The third threshold p^{**} is such that $U_1(\mu_{1,\bar{z}(p)}|s) = U_1(\mu_{0,0}|s)$. p^{**} can be interpreted as the upper bound of p for $\mu_{y,z}$ to be interim incentive efficient. When $p \geq p^{**}$, any mediator who puts a positive probability on war regardless of the type combination is interim Pareto dominated by the mediator $\mu_{0,0}$.

Incentive Feasibility, $\underline{\underline{y}}(\bar{p}(s))$, and $\underline{\underline{z}}(\bar{p}(s))$ in Proposition 2. For Proposition 2, when $p \in (p^*, p^{'})$, for a mediator to be individually rational and incentive compatible for both types, it must put some positive probability on war for $t = \{sw, ws\}$ with a lower bound requirement and some corresponding positive probability on war for $t = \{ss\}$. In particular, $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}(d_0|sw) \equiv \underline{\underline{y}}(p)$ and

$\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}(d_0|ss) \equiv \underline{\underline{z}}(p)$ are simultaneously determined by the binding strong type's participation constraint and the binding weak type's informational incentive constraint. That is, $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|s) = 0$ and $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|w) = U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w)$ give, respectively,

$$\begin{aligned} &p(1-\underline{\underline{z}}(p))u_1(d_1,ss) + (1-p)(1-\underline{\underline{y}}(p))u_1(d_1,sw) = 0, \\ &\qquad \text{and} \quad p(1-\underline{\underline{y}}(p))u_1(d_1,ws) + (1-p)u_1(d_1,ww) \\ &= p(1-\underline{\underline{z}}(p))u_1(d_1,ws) + (1-p)(1-\underline{\underline{y}}(p))u_{1'}(d_1,ww), \end{aligned}$$

that in turn give $\underline{\underline{y}}(p) > \underline{\underline{z}}(p) > 0$ for all $p \in (p^*, p')$ such that

$$\begin{aligned} \underline{\underline{y}}(p) &= \frac{pu_1(d_1,ss)u_1(d_1,ws) + (1-p)u_1(d_1,sw)u_1(d_1,ws)}{pu_1(d_1,ss)u_1(d_1,ws) + (1-p)(u_1(d_1,sw)u_1(d_1,ws) - u_1(d_1,ss)u_1(d_1,ww))} \\ \underline{\underline{z}}(p) &= 1 - \frac{(1-p)^2 u_1(d_1,sw)u_1(d_1,ww)}{p(pu_1(d_1,ss)u_1(d_1,ws) + (1-p)(u_1(d_1,sw)u_1(d_1,ws) - u_1(d_1,ss)u_1(d_1,ww)))}. \end{aligned}$$

Remark. The proofs of Lemma 1 and Lemma 3 are inclusive in the proof of Proposition 1.

Lemma 6 (Incentive Feasibility). *The specified mediators in Proposition 1 and Proposition 2 are incentive feasible.*

Proof of Lemma 6. See Online Appendix B.

Proof of Proposition 1 (Characterization of interim incentive efficient mediators):
First, note that all of the mediators characterized in Proposition 1 are incentive feasible by Lemma 6. I only provide the proof for ***Case 1*** and relegate the proofs for other cases to Online Appendix B because they are analogous.

Case 1. $p < p'$***: then*** $S(\Gamma) = \left\{\mu_{y,z} | y \in \left[\underline{y}(p), 1\right], z = 0\right\}$**.** For any $\mu_{y,z} \in S(\Gamma)$, if I can find vectors λ, α, and β such that all of the conditions in (7.1) are satisfied, then $\mu_{y,z}$ for any $y \in \left[\underline{y}(p), 1\right]$ and $z = 0$ are interim incentive efficient. First, for $\mu_{\underline{y}(p),0}$, where $\underline{y}(p) = 1 + \frac{pu_1(d_1,ss)}{(1-p)u_1(d_1,sw)}$, $\underline{y}(p)$ is computed such that $U_1(\mu_{\underline{y}(p),0}|s) = 0$ $\left(> U_1(\mu_{\underline{y}(p),0}, w|s)\right)$ given $\mu_{\underline{y}(p),0}$. Note that $\bar{y}(p) \in (0,1)$ and is a decreasing function of $p \in (0, p')$. For this mechanism, I have $U_1(\mu_{\underline{y}(p),0}|s) = 0 > U_1(\mu_{\underline{y}(p),0}, w|s)$, $U_1(\mu_{\underline{y}(p),0}|w) > 0$, and $U_1(\mu_{\underline{y}(p),0}|w) > U_1(\mu_{\underline{y}(p),0}, s|w)$. (See the proof of Lemma 6 in Online Appendix B.) So, it must be $\alpha(w|s) = \alpha(s|w) = 0$, $\beta(s) > 0$, $\beta(w) = 0$ because the multipliers must be zero for the constraints that do not bind and positive for those that bind. In order for $\mu_{\underline{y}(p),0}$ to

randomize between d_0 and d_1 when $t \in \{sw, ws\}$, it must be that, for $t \in \{sw, ws\}$,

$$\sum_{d \in D} \mu_{\underline{y}(p),0}(d|t) \sum_{i \in N} v_i(d, t, \lambda, \alpha, \beta) = \max_{d \in D} \sum_{i \in N} v_i(d, t, \lambda, \alpha, \beta) = 0, \tag{7.2}$$

where the last equality follows because $v_i(d_0, t, \lambda, \alpha, \beta) = 0$ for any t, λ, α, and β, noting that the incentive efficient mechanism $\mu_{\underline{y}(p),0}$ puts a positive probability on both outcomes, that is,

$$\max_{d \in D} \sum_{i \in N} v_i(d, t, \lambda, \alpha, \beta) = \sum_{i \in N} v_i(d_1, t, \lambda, \alpha, \beta) = \sum_{i \in N} v_i(d_0, t, \lambda, \alpha, \beta) = 0.$$

The left-hand side of (7.2) for each $t \in \{sw, ws\}$ is

$$\begin{aligned} \mu_{\underline{y}(p),0}(d_1|sw) \sum_{i \in N} v_i(d_1, sw, \lambda, \alpha, \beta) =& (1 - \underline{y}(p)[(\lambda(s) + \beta(s))u_1(d_1, sw)/p \\ &+ \lambda(w)u_2(d_1, sw)/\bar{p}(w)], \end{aligned} \tag{7.3}$$

$$\begin{aligned} \mu_{\underline{y}(p),0}(d_1|ws) \sum_{i \in N} v_i(d_1, ws, \lambda, \alpha, \beta) =& (1 - \underline{y}(p)[\lambda(w)u_1(d_1, ws)/(1 - p) \\ &+ (\lambda(s) + \beta(s))u_2(d_1, ws)/p]. \end{aligned} \tag{7.4}$$

(7.3) and (7.4) are zero only when

$$\begin{aligned} \lambda(s) &< \frac{pu_1(d_1, ws)}{(1-p)(-u_1(d_1, sw)) + pu_1(d_1, ws)} \equiv \lambda^*(p), \\ \text{and } \beta(s) &= \frac{-p(1 - \lambda(s))u_1(d_1, ws) - (1-p)\lambda(s)u_1(d_1, sw)}{(1-p)u_1(d_1, sw)} \equiv \beta^*(s), \end{aligned}$$

Therefore, for $\mu_{\underline{y}(p),0}$, any $\lambda(s) < \lambda^*(p)$, $\alpha = (0, 0)$ and $\beta = (\beta^*(s), 0)$ satisfy (7.1). For $\mu_{y,0}$ such that $y \in \left(\underline{y}(p), 1\right)$, all of the constraints are not binding. Therefore, it must be $\alpha = (0, 0)$ and $\beta = (0, 0)$. Then, only $\lambda(s)$ such that $\lambda(s) = \lambda^*(p)$ makes (7.3) and (7.4) equal zero. Therefore, for $\mu_{y,0}$ where $y \in \left(\underline{y}(p), 1\right)$, $\lambda(s) = \lambda^*(p)$, $\alpha = (0, 0)$, and $\beta = (0, 0)$ satisfy (7.1). Lastly, for $\mu_{1,0}$, any $\lambda(s) > \lambda^*(p)$, $\alpha = (0, 0)$, and $\beta = (0, 0)$ satisfy (7.1). Because $\mu_{y,z}$ for any $y \in \left[\underline{y}(p), 1\right]$ and $z = 0$ is interim incentive efficient, there is no other incentive feasible mechanism that is interim Pareto superior to such $\mu_{y,z}$. However, it is necessary to check if there is no interim incentive efficient mediator other than those in $\{\mu_{y,z} | y \in \left[\underline{y}(p), 1\right], z = 0\}$. But for any incentive feasible $\delta \notin \left\{\mu_{y,z} | y \in \left[\underline{y}(p), 1\right], z = 0\right\}$, there does not exist vectors λ, α, and β that satisfy (7.1). That is, any weights $\lambda(s) < \lambda^*(p)$ generates only $\mu_{\underline{y}(p),0}$, any weights $\lambda(s) \geq \lambda^*(p)$ generate only $\mu_{y,0}$

with $y \in (\underline{y}(p), 1]$. Therefore, $S(\Gamma) = \left\{\mu_{y,z} | y \in \left[\underline{y}(p), 1\right], z = 0\right\}$ completely describes the set of incentive efficient mediators in ***Case 1.*** $p < p^{'}$. □

Proof of Proposition 2 (Characterization of interim incentive efficient mediators):
The proof basically follows from the proof of Proposition 1. See Online Appendix B. The proofs of Proposition 1 and Proposition 2 also confirm Lemma 1 and Lemma 3. □

Proof of Lemma 4 (Ex ante welfare ordering):
(*Only if*) Suppose that $y^{'} > y$ with some corresponding $z^{'} > z$ such that the restrictions for each case hold. Then, the ex ante evaluations of a mechanism $\mu_{y',z'}$ by any player is given by

$$\begin{aligned} U_1(\mu_{y',z'}) =& p\left[p(1-z^{'})u_1(d_1, ss) + (1-p)(1-y^{'})u_1(d_1, sw)\right] \\ &+ (1-p)\left[p(1-y^{'})u_1(d_1, ws) + (1-p)u_1(d_1, ww)\right], \end{aligned}$$

and the ex ante evaluations of a mechanism $\mu_{y,z}$ by any player is given by

$$\begin{aligned} U_1(\mu_{y,z}) =& p\left[p(1-z)u_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw)\right] \\ &+ (1-p)\left[p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww)\right]. \end{aligned}$$

Then, $U_1(\mu_{y,z}) - U_1(\mu_{y',z'})$

$$= pp(z^{'} - z)u_1(d_1, ss) + p(1-p)(y^{'} - y)\left[u_1(d_1, sw) + u_1(d_1, ws)\right] > 0,$$

because $z^{'} \geq z$, $y^{'} > y$, $u_1(d_1, ss) > 0$, and $u_1(d_1, sw) + u_1(d_1, ws) > 0$ by assumptions **(A1)** – **(A3)**. Thus, $U_i(\mu_{y,z}) > U_i(\mu_{y',z'})$ for all $i \in \{1, 2\}$, that is, $\mu_{y,z}$ is ex ante Pareto superior to $\mu_{y',z'}$.
(*If*) If $\mu_{y,z}$ is ex ante Pareto superior to $\mu_{y',z'}$, then it must be

$$U_1(\mu_{y,z}) - U_1(\mu_{y',z'}) = pp(z^{'} - z)u_1(d_1, ss) + p(1-p)(y^{'} - y)\left[u_1(d_1, sw) + u_1(d_1, ws)\right] > 0.$$

Suppose to the contrary that $y \geq y^{'}$. Then, the corresponding z and $z^{'}$ are such that $z \geq z^{'}$, which implies that the above must be non-positive. This is a contradiction. Thus, it must be $y < y^{'}$. □

Proof of Proposition 3 (Uniqueness of ex ante incentive efficient mediator):
I want to prove that the mediator specified in Proposition 3 is uniquely ex ante incentive efficient for each range of p. An incentive feasible μ is ex ante incentive efficient if and only if there is no other mechanism that is in $S(\Gamma)$ and is ex ante Pareto superior to μ, that is, $\delta \in S(\Gamma) \backslash \{\mu\}$ such

that

$$\sum_{t\in T}\sum_{d\in D} p(t)\delta(d|t)u_i(d,t) \geq \sum_{t\in T}\sum_{d\in D} p(t)\mu(d|t)u_i(d,t),\ \forall i \in N, \tag{7.5}$$

with strict inequality for at least one player. I can look for an ex ante incentive efficient mediator within the set of interim incentive efficient mediators $S(\Gamma)$, because for the given set of incentive feasible mediators F, the set of ex ante incentive efficient mechanisms in F is a subset of the set of interim incentive efficient mechanisms in F. Notice that the right-hand side of (7.5) equals $\sum_{t_i\in T_i}\bar{p}_i(t_i)U_i(\mu|t_i)$.

[For Proposition 1: When $p^{'} \leq p^{*}$] For ***Case 1*** $p < p^{'}$, I claim that $\mu_{\underline{y}(p),0}$ is the unique ex ante incentive efficient mediator in $S(\Gamma) = \left\{\mu_{y,z}|y \in \left[\underline{y}(p),1\right], z=0\right\}$. Suppose, to the contrary, that $\mu_{y,0}$ for any $y \in (\underline{y}(p),1]$ is ex ante Pareto superior to $\mu_{\underline{y}(p),0}$. Because of **(A3)**, without loss of generality, I focus on player 1. Then, for any $y \in (\underline{y}(p),1]$, it must be

$$\sum_{t_1\in T_1}\bar{p}_1(t_1)U_1(\mu_{y,0}|t_1) > \sum_{t_1\in T_1}\bar{p}_1(t_1)U_1(\mu_{\underline{y}(p),0}|t_1),$$

$$\longleftrightarrow p(1-p)(\underline{y}(p)-y)u_1(d_1,sw) + (1-p)p(\underline{y}(p)-y)u_1(d_1,ws) > 0$$
$$\longleftrightarrow (\underline{y}(p)-y)(u_1(d_1,sw)+u_1(d_1,ws)) > 0.$$

Because $y > \underline{y}(p)$, it must be $u_1(d_1,sw)+u_1(d_1,ws) < 0$. However, assumptions **(A1)** and **(A3)** imply $u_1(d_1,sw)+u_1(d_1,ws) > 0$. Contradiction. Therefore, there does not exist any mechanism in $S(\Gamma)\backslash\{\mu_{\underline{y}(p),0}\}$ that is ex ante Pareto superior to $\mu_{\underline{y}(p),0}$, where $S(\Gamma) = \left\{\mu_{y,z}|y \in \left[\underline{y}(p),1\right], z=0\right\}$ for ***Case 1*** $p < p^{'}$. This also proves uniqueness, since all the other mechanisms are ex ante Pareto dominated by $\mu_{\underline{y}(p),0}$, as $\mu_{\underline{y}(p),0}$ gives both players strictly better ex ante expected payoffs. For ***Case 2*** $p \in [p^{'},p^{*}]$, I claim that $\mu_{0,0}$ is the unique ex ante incentive mediator in $S(\Gamma) = \{\mu_{y,z}|y \in [0,1], z=0\}$. Suppose, to the contrary, that $\mu_{y,0}$ for any $x \in (0,1]$ is ex ante Pareto superior to $\mu_{0,0}$. Then, $\sum_{t_1\in T_1}\bar{p}_1(t_1)U_1(\mu_{y,0}|t_1) > \sum_{t_1\in T_1}\bar{p}_1(t_1)U_1(\mu_{0,0}|t_1)$

$$\longleftrightarrow p(1-p)(-y)u_1(d_1,sw) + (1-p)p(-y)u_1(d_1,ws) > 0.$$

Again, this is a contradiction because $y > 0$ and $u_1(d_1,sw)+u_1(d_1,ws) > 0$. Therefore, there does not exist any mechanism in $S(\Gamma)\backslash\{\mu_{0,0}\}$ that is ex ante Pareto superior to $\mu_{0,0}$, where $S(\Gamma) = \{\mu_{y,z}|y \in [0,1], z=0\}$ for ***Case 2*** $p \in [p^{'},p^{*}]$; and in fact, $\mu_{y,0}$ for all $y \in (0,1]$ are ex ante Pareto dominated by $\mu_{0,0}$. For ***Case 3*** $p \in (p^{*},p^{**})$, I also claim that $\mu_{0,0}$ is the unique ex ante incentive mediator in $S(\Gamma) = \{\mu_{y,z}|y \in [0,1], z := z(y,p) \in [0,\bar{z}(p)]\}$. Suppose, to the contrary,

that $\mu_{y,z}$ for any $y \in (0,1]$ and for $z = z(y,p) \in (0, \bar{z}(p)]$ is ex ante Pareto superior to $\mu_{0,0}$. Then, $\sum_{t_1 \in T_1} \bar{p}_1(t_1) U_1(\mu_{y,z}|t_1) > \sum_{t_1 \in T_1} \bar{p}_1(t_1) U_1(\mu_{0,0}|t_1)$

$$\longleftrightarrow p(-z(y,p))u_1(d_1, ss) + (1-p)(-y)(u_1(d_1, sw) + u_1(d_1, ws)) > 0. \tag{7.6}$$

But because $z(y,p) > 0$, $u_1(d_1, ss) > 0$, and $u_1(d_1, sw) + u_1(d_1, ws) > 0$, the left-hand side of (7.6) must be negative, which is a contraction. Therefore, there does not exist any mechanism in $S(\Gamma)\backslash\{\mu_{0,0}\}$ that is ex ante Pareto superior to $\mu_{0,0}$, where $S(\Gamma) = \{\mu_{y,z} | y \in [0,1], z := z(y,p) \in [0, \bar{z}(p)]\}$ for ***Case 3*** $p \in (p^*, p^{**})$, and $\mu_{y,z}$ for all $y \in (0,1]$ and for $z = z(y,p) \in (0, \bar{z}(p)]$ are ex ante Pareto dominated by $\mu_{0,0}$. For ***Case 4*** $p \geq p^{**}$, $\mu_{0,0}$ is the unique interim incentive efficient mediator and thus is the unique ex ante incentive efficient mediator.

[For Proposition 2: When $p^{'} > p^*$] All the interim incentive efficient mediators characterized in Proposition 2, ***Case 1***$^{'}$, ***Case 3***, and ***Case 4*** are identically defined as in Proposition 1, ***Case 1***, ***Case 3***, and ***Case 4***, respectively. Thus, I only need to prove that $\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}$ is the unique ex ante incentive efficient mediator in $S(\Gamma) = \left\{\mu_{y,z} | y \in \left[\underline{\underline{y}}(p), 1\right], z := z(y,p) \in \left[\underline{\underline{z}}(p), \bar{z}(p)\right]\right\}$ for ***Case 2***$^{'}$ $p \in (p^*, p^{'})$. Suppose, to the contrary, that $\mu_{y,z}$ for some $y \in (\underline{\underline{y}}(p), 1]$ and for $z = z(y,p) \in (\underline{\underline{z}}(p), \bar{z}(p)]$, where $\underline{\underline{z}}(p) = z(\underline{\underline{y}}(p), p)$ and $\bar{z}(p) = z(1,p)$, is ex ante Pareto superior to $\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}$. Then, it must be

$$\sum_{t_1 \in T_1} \bar{p}_1(t_1) U_1(\mu_{y,z}|t_1) > \sum_{t_1 \in T_1} \bar{p}_1(t_1) U_1(\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}|t_1)$$

$$\longleftrightarrow p(\underline{\underline{z}}(p) - z(y,p))u_1(d_1, ss) + (1-p)(\underline{\underline{y}}(p) - x)(u_1(d_1, sw) + u_1(d_1, ws)) > 0.$$

Because $z(y,p) > \underline{\underline{z}}(p)$ and $y > \underline{\underline{y}}(p)$ for $y \in (\underline{\underline{y}}(p), 1]$, along with $u_1(d_1, ss) > 0$ and $u_1(d_1, sw) + u_1(d_1, ws) > 0$, the above inequality yields a contradiction. Therefore, there does not exist any mechanism in $S(\Gamma)\backslash\{\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}\}$ that is ex ante Pareto superior to $\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}$, where $S(\Gamma) = \left\{\mu_{y,z} | y \in \left[\underline{\underline{y}}(p), 1\right], z := z(y,p) \in \left[\underline{\underline{z}}(p), \bar{z}(p)\right]\right\}$ for ***Case 2***$^{'}$ $p \in (p^*, p^{'})$. This also proves uniqueness, because all of the other mechanisms are ex ante Pareto dominated by $\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}$, as $\mu_{\underline{\underline{y}}(p), \underline{\underline{z}}(p)}$ gives both players strictly better ex ante expected payoffs. Thus, the mediators characterized in Proposition 3 are ex ante incentive efficient mediators and, in fact, unique for each specified range of p . □

Proof of Lemma 5 (Interim welfare ordering):

Suppose that $\mu_{y,z}$ is ex ante Pareto superior to $\mu_{y^{'},z^{'}}$. Then, by Lemma 4, $y < y^{'}$ with a corre-

sponding $z < z^{'}$ such that the restrictions for each case hold. The interim evaluation of $\mu_{y,z}$ by any player of the strong type is given by

$$U_1(\mu_{y,z}|s) = p(1-z)u_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw),$$

and the interim evaluation of $\mu_{y^{'},z^{'}}$ by any player of the strong type is give by

$$U_1(\mu_{y^{'},z^{'}}|s) = p(1-z^{'})u_1(d_1, ss) + (1-p)(1-y^{'})u_1(d_1, sw).$$

For ***Case 1***, ***1***$^{'}$ & ***2***, because $z = z^{'} = 0$, we have

$$U_1(\mu_{y,z}|s) - U_1(\mu_{y^{'},z^{'}}|s) = (1-p)(y^{'} - y)u_1(d_1, sw) < 0$$

by $y^{'} - y > 0$, and $u_1(d_1, sw) < 0$.
For ***Case 3***, taking into account $z = y\left[1 - \frac{(1-p)u_1(d_1,ww)}{pu_1(d_1,ws)}\right]$ and $z^{'} = y^{'}\left[1 - \frac{(1-p)u_1(d_1,ww)}{pu_1(d_1,ws)}\right]$,

$$\begin{aligned} U_1(\mu_{y,z}|s) - U_1(\mu_{y^{'},z^{'}}|s) =& (y^{'} - y)[p(u_1(d_1, ss)u_1(d_1, ww) \\ & - u_1(d_1, sw)u_1(d_1, ws) + u_1(d_1, ss)u_1(d_1, ws)) \\ & - (u_1(d_1, ss)u_1(d_1, ww) - u_1(d_1, sw)u_1(d_1, ws))] < 0, \end{aligned}$$

since $p < p^{**}$ and $y^{'} - y > 0$. For ***Case 2***$^{'}$, the above holds if I restrict attention to $y \in [\underline{\underline{y}}(p), 1]$ and a corresponding $z := z(y,p) \in [\underline{\underline{z}}(p), \bar{z}(p)]$.
For the weak type, the interim evaluation of $\mu_{y,z}$ by any player of the weak type is given by

$$U_1(\mu_{y,z}|w) = p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww),$$

and the interim evaluation of $\mu_{y^{'},z^{'}}$ by any player of the weak type is give by

$$U_1(\mu_{y^{'},z^{'}}|w) = p(1-y^{'})u_1(d_1, ws) + (1-p)u_1(d_1, ww).$$

Then, for any $y^{'}$ and y such that $y^{'} - y > 0$

$$U_1(\mu_{y,z}|w) - U_1(\mu_{y^{'},z^{'}}|w) = p(y^{'} - y)u_1(d_1, ws) > 0,$$

because $u_1(d_1, ws) > 0$. The converse trivially follows. □

Proof of Corollary 1 (Multiple interim but not ex ante incentive efficient mediators):
The proof follows directly from Proposition 1, Proposition 2, Lemma 4, Proposition 3, and Lemma 5. By Proposition 3, there is exists a mediator who is uniquely ex ante incentive efficient. So, excluding the case where $S(\Gamma)$ is a singleton, that is, when $p \geq p^{**}$, all of the other mediators characterized by Proposition 1 and Proposition 2 are ex ante Pareto inferior to the unique ex ante incentive efficient mediator characterized by Proposition 3. Moreover, by Lemma 4 and by tracing out the interim Pareto frontier, the mediator that minimizes the ex ante expected payoffs of the players over the set of interim incentive efficient mediators is $\mu_{1,0}$ when $p \leq p^*$, and $\mu_{1,\bar{z}(p)}$ when $p \in (p^*, p^{**})$. These mediators give the highest interim expected utilities for the strong type among all mediators in $S(\Gamma)$ according to Lemma 5. □

Proofs of Section 4

Proof of Theorem 1 (Uniqueness of neutral bargaining solution):
The proof follows directly from the proof of Proposition 4. □

Proof of Proposition 4 (Characterization of neutral bargaining solutions):
The proofs use linear programming techniques to compute the set of all possible welfare weights. The set of symmetric bargaining weights has the property that the weight on the strong type is at least as great as the distorted probability. I can characterize the neutral bargaining solution within the set of interim incentive efficient mechanisms by showing that any weight in this interval generates a neutral bargaining solution, and any neutral bargaining solution has weights in this interval. In varying the welfare weights, I trace out the Pareto frontier for distorted weights. Taking symmetry into account, I can suppress the subscripts i for the vectors λ, α, β, and ϖ that must satisfy the conditions, for all $\varepsilon > 0$:

$$\begin{aligned}
&\left(\left(\lambda(t_i) + \sum_{s_i \in T_i} \alpha(s_i|t_i) + \beta(t_i)\right)\varpi(t_i) - \sum_{s_i \in T_i} \alpha(t_i|s_i)\varpi(s_i)\right)/\bar{p}_i(t_i) \\
&= \sum_{t_{-i} \in T_{-i}} \bar{p}_{-i}(t_{-i}) \max_{d \in \{d_0, d_1\}} \sum_{j \in \{1,2\}} \frac{v_j(d, t, \lambda, \alpha, \beta)}{2}, \quad \forall i,\ \forall t_i \in T_i; \\
&\quad \lambda(t_i) > 0,\ \alpha(s_i|t_i) \geq 0,\ \beta(t_i) \geq 0, \quad \forall i,\ \forall s_i \in T_i,\ \forall t_i \in T_i; \\
&\text{and}\ \ U_i(\mu|t_i) \geq \varpi(t_i) - \varepsilon, \quad \forall i,\ \forall t_i \in T_i,
\end{aligned} \tag{7.7}$$

I only provide the proof for ***Case 1*** of Proposition 1 and relegate the proofs for other cases of Proposition 1 and for Proposition 2 to Online Appendix B because the proofs are analogous.

Case 1. $p < p^{'}$***: then*** $NS(\Gamma) = \{\mu_{1,0}\}$**.** Proposition 1 already shows that any $\lambda(s) < \lambda^*(p)$ generates only $\mu_{\underline{y}(p),0}$ as the interim incentive efficient mediator together with $\alpha = (0,0)$ and $\beta = (\beta^*(s), 0))$; $\lambda(s) = \lambda^*(p)$ generates only $\mu_{y,0}$ for $y \in (\underline{y}(p), 1)$ together with $\alpha = (0,0)$ and $\beta = (0,0)$; and any $\lambda(s) > \lambda^*(p)$ generates only $\mu_{1,1}$ together with $\alpha = (0,0)$ and $\beta = (0,0)$. Thus, to characterize the neutral bargaining solution within $S(\Gamma)$, it is necessary to check whether there exist vectors $\varpi = (\varpi(s), \varpi(w))$ that satisfy (7.7) for each interim incentive efficient mediator. For $\mu_{\underline{y}(p),0}$, the conditions in (7.7) for the strong type are

$$\begin{aligned}
(\lambda(s) + \beta^*(s))\,\varpi(s)/p &= \sum_{t_{-i} \in T_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d \in D} \mu_{\underline{y}(p),0}(d|(s, t_{-i})) \sum_{j \in \{1,2\}} \frac{v_j(d, (s, t_{-i}), \lambda, \alpha, \beta)}{2} \\
&= p(\lambda(s) + \beta^*(s))u_1(d_1, ss)/p \\
&+ (1-p)(1-\underline{y}(p))\frac{1}{2}\{(\lambda(s) + \beta^*(s))u_1(d_1, sw)/p \\
&+ \lambda(w)u_2(d_1, sw)/(1-p)\} \\
&= p(\lambda(s) + \beta^*(s))u_1(d_1, ss)/p,
\end{aligned}$$

because $\beta^*(s) = \frac{-p(1-\lambda(s))u_1(d_1,ws)-(1-p)\lambda(s)u_1(d_1,sw)}{(1-p)u_1(d_1,sw)}$, and so it follows that $\varpi(s) = pu_1(d_1, ss)$. Then, it must be

$$\begin{aligned}
U_1(\mu_{\underline{y}(p),0}|s) =& pu_1(d_1, ss) + \bar{p}(w)(1-\underline{y}(p))u_1(d_1, sw) \\
&\geq pu_1(d_1, ss),
\end{aligned}$$

which is a contradiction because $u_1(d_1, sw) < 0$. For the weak type, it must be

$$\begin{aligned}
\lambda(w)\varpi(w)/(1-p) &= \sum_{t_{-i} \in T_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d \in D} \mu_{\underline{y}(p),0}(d|(s, t_{-i})) \sum_{j \in \{1,2\}} \frac{v_j(d, (s, t_{-i}), \lambda, \alpha, \beta)}{2} \\
&= p(1-\underline{y}(p))\frac{1}{2}\{\lambda(w)u_1(d_1, ws)/(1-p) \\
&+ (\lambda(s) + \beta^*(s))u_2(d_1, ws)/p\} + (1-p)\lambda(w)u_1(d_1, ww)/(1-p) \\
&= (1-p)\lambda(w)u_1(d_1, ww)/(1-p).
\end{aligned}$$

This condition gives $\varpi(w) = (1-p)u_1(d_1, ww)$, and $\varpi(w)$ satisfies: $U_1(\mu_{\underline{y}(p),0}|w) = p(1-\underline{y}(p))u_1(d_1, ws) + (1-p)u_1(d_1, ww) \geq \varpi(w)$. Therefore, $\varpi(w) = (1-p)u_1(d_1, ww)$ satisfies (7.7) for each positive

number ε, but there does not exist $\varpi(s)$ that satisfies (7.7) for

$$\varepsilon \in \left(0, (1-p)(1-\underline{y}(p))\left[-u_1(d_1, sw)\right]\right)$$

for any $\lambda(s) < \lambda^*(p)$. Thus, $\mu_{\underline{y}(p),0}$ is not a neutral bargaining solution. For $\mu_{y,0}$ such that $y \in (\underline{y}(p), 1)$, taking into account $\lambda(s) = \lambda^*(p)$, $\alpha = (0,0)$, and $\beta = (0,0)$, the conditions in (7.7) for the strong type are

$$\begin{aligned}
\lambda^*(p)\varpi(s)/p &= \sum_{t_{-i}\in T_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d\in D} \mu_{y,0}(d|(s,t_{-i})) \sum_{j\in\{1,2\}} \frac{v_j(d,(s,t_{-i}),\lambda,\alpha,\beta)}{2} \\
&= p\lambda^*(p)u_1(d_1,ss)/p \\
&+ (1-p)(1-y)\frac{1}{2}\{\lambda^*(p)u_1(d_1,sw)/p \\
&+ (1-\lambda^*(p)u_2(d_1,sw)/(1-p)\} \\
&= p\lambda^*(p)u_1(d_1,ss)/p,
\end{aligned}$$

where the third equation follows from plugging in $\lambda^*(p) \equiv \frac{pu_1(d_1,ws)}{(1-p)(-u_1(d_1,sw))+pu_1(d_1,ws)}$ or by noticing that $\mu_{y,0}$ randomizes between d_0 and d_1 when $t = \{sw, ws\}$. It follows that $\varpi(s) = pu_1(d_1, ss)$. Then, it must be $U_1(\mu_{y,0}|s) = pu_1(d_1,ss) + (1-p)(1-y)u_1(d_1,sw) \geq pu_1(d_1,ss)$, which is a contradiction since $u_1(d_1,sw) < 0$. For the weak type's $\varpi(w)$ such that $\varpi(w) = (1-p)u_1(d_1,ww)$, the following is true: $U_1(\mu_{y,0}|w) = p(1-y)u_1(d_1,ws) + (1-p)u_1(d_1,ww) \geq (1-p)u_1(d_1,ww)$. However, there does not exist $\varpi(s)$ that satisfies (7.7) for all $\varepsilon < (1-p)(1-y)\left[-u_1(d_1,sw)\right]$, where $y \in (\underline{y}(p), 1)$. Thus, $\mu_{y,0}$ for any $y \in (\underline{y}(p), 1)$ is not a neutral bargaining solution. There is only one candidate left in $S(\Gamma)$, which is $\mu_{1,o}$; and because the set of neutral bargaining solutions is nonempty (within the set of interim incentive efficient mechanisms, Theorem 2, Myerson (1984*b*)), it must be that $\mu_{1,o}$ is the unique neutral bargaining solution for ***Case 1*** $p < p'$. To confirm, it is necessary to check whether there exist vectors $\varpi = (\varpi(s), \varpi(w))$ together with $\alpha = (0,0)$ and $\beta = (0,0)$, and any $\lambda(s) > \lambda^*$, that satisfy (7.7).

$$\begin{aligned}
\lambda(s)\varpi(s)/p &= \sum_{t_{-i}\in T_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d\in D} \mu_{1,0}(d|(s,t_{-i})) \sum_{j\in\{1,2\}} \frac{v_j(d,(s,t_{-i}),\lambda,\alpha,\beta)}{2} \\
&= p\lambda(s)u_1(d_1,ss)/p.
\end{aligned}$$

This gives $\varpi(s) = pu_1(d_1,ss)$. Then, it must be $U_1(\mu_{1,0}|s) = pu_1(d_1,ss) \geq \varpi(s) = pu_1(d_1,ss)$ which holds with equality. For the weak type's $\varpi(w)$ such that $\varpi(w) = (1-p)u_1(d_1,ww)$, the following is also trivially true: $U_1(\mu_{1,0}|w) = (1-p)u_1(d_1,ww) \geq \varpi(w) = (1-p)u_1(d_1,ww)$. Because for any

$\lambda(s) > \lambda^*(p)$, $\alpha = (0,0)$, $\beta = (0,0)$, and $\varpi = (U_1(\mu_{1,0}|s), U_1(\mu_{1,0}|w))$ satisfy the conditions (7.7) for the case of $\varepsilon = 0$, the same λ, α, β, and ϖ satisfy the Theorem for every positive ε. Therefore, the neutral bargaining solution is characterized by $NS(\Gamma) = \{\mu_{1,o}\}$ and is unique for ***Case 1*** $p < p'$ in Proposition 1. □

Proofs of Corollaries 2 and 3 (Welfare weights and multipliers for the neutral bargaining solutions): For $\mu_{1,0}$ in ***Case 1*** of Proposition 1, as shown in the proof of Proposition 3, the conditions in Myerson (1984b)'s Theorem are satisfied for all $\varepsilon \geq 0$ by any $\lambda(s) > \lambda^*(p)$ where $\lambda^*(p) = \frac{pu_1(d_1,ws)}{(1-p)(-u_1(d_1,sw))+pu_1(d_1,ws)} > p$ because of **(A1)** and **(A3)**, and by $\alpha(s|w) = \alpha(w|s) = 0$ and $\beta(s) = \beta(w) = 0$ because all the constraints do not bind. For all the other cases, see Online Appendix B. □

PROOFS OF SECTION 5 AND 6

Proof of Proposition 5 (Non-Ratification of ex ante Pareto superior mediator):
Assume that γ is ex ante Pareto superior to δ. Note that $\gamma \in S(\Gamma)$ and $\delta \in S(\Gamma)$ are both interim incentive efficient, where $S(\Gamma)$ is characterized by Proposition 1 (and Proposition 2). Then, by Lemma 5, under the prior beliefs, the interim utility from γ for the weak type is strictly greater than the interim utility from δ; and the interim utility from γ for the strong type is strictly lower than the interim utility from δ. That is, a strong type of any player prefers the ex ante Pareto inferior δ to γ when he knows only his type, while a weak type of any player prefers the ex ante Pareto superior γ to δ when she knows only her type. The proof consists of showing that the posterior $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(w) = 1, \bar{q}_{i,i}(s) = 1)$ is a unique credible vote belief for all i with a corresponding credible veto set $V_i = \{s\}$ and credible non-veto set $W_{-i} = \{w\}$. Without loss of generality, I can focus on $i = 1$ and because of the symmetry assumption, the same holds for $i = 2$. The reporting strategies $\sigma = ((\sigma_{i,i}(s|s) \in [0,1], \sigma_{i,i}(w|w) = f(\sigma_{-i,i}(w|w))), (\sigma_{-i,i}(t_{-i}|t_{-i}) \in [0,1], \forall t_{-i}))$ and the war strategies $\psi = ((\psi_{i,i}(s) = 1, \psi_{i,i}(w) = 0), (\psi_{-i,i}(t_{-i}) \in (0,1), \forall t_{-i}))$, which the players would use in G^δ, form a unique equilibrium in the subgame when γ does not win with respect to the posterior beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(w) = 1, \bar{q}_{i,i}(s) = 1)$. Note that the vetoer i of weak type's reporting strategy $\sigma_{i,i}(w|w)$ depends on the ratifier $-i$ of weak type's completely mixed equilibrium reporting strategy $\sigma_{-i,i}(w|w)$ such that $\sigma_{i,i}(w|w) = 1$ if $\sigma_{-i,i}(w|w) > \frac{\delta(d_1|ss)-\delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$, $\sigma_{i,i}(w|w) \in (0,1)$ if $\sigma_{-i,i}(w|w) = \frac{\delta(d_1|ss)-\delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$, and $\sigma_{i,i}(w|w) = 0$ if $\sigma_{-i,i}(w|w) < \frac{\delta(d_1|ss)-\delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$. Note that $\delta(d_1|ws) = \delta(d_1|sw)$ and $\delta(d_1|sw) < \delta(d_1|ss)$. For $\bar{q}_{\cdot,i}$ to be credible, under the equilibrium

reporting strategies σ and war strategies ψ given the posterior beliefs $\bar{q}_{\cdot,i}$, it must be for the vetoer i of strong type:

$$\sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \sum_{d\in D} \gamma(d|(s,t_{-i}))u_i(d,(s,t_{-i})) = \bar{q}_{-i,i}(s) \sum_{d\in D} \gamma(d|ss)u_i(d,ss) + \bar{q}_{-i,i}(w) \sum_{d\in D} \gamma(d|sw)u_i(d,sw) = \gamma(d_1|sw)u_i(d_1,sw)$$

$$\begin{aligned} <&\bar{q}_{-i,i}(s)\{\psi_{i,i}(s)u_i(d_0,ss) + (1-\psi_{i,i}(s))(\psi_{-i,i}(s)u_i(d_0,ss) \\ &+ (1-\psi_{-i,i}(s)) \sum_{r\in R} \sigma_i(r|ss) \sum_{d\in D} \delta(d|r)u_i(d,ss))\} \\ &+ \bar{q}_{-i,i}(w)\{\psi_{i,i}(s)u_i(d_0,sw) + (1-\psi_{i,i}(s))(\psi_{-i,i}(w)u_i(d_0,sw) \\ &+ (1-\psi_{-i,i}(w)) \sum_{r\in R} \sigma_i(r|sw) \sum_{d\in D} \delta(d|r)u_i(d,sw))\} = u_i(d_0,sw) = 0, \end{aligned}$$

where the inequality follows from $u_i(d_1,t) < 0$ when $t = (t_i, t_{-i}) = (s,w)$. Because a strong type strictly benefits from vetoing, $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(w) = 1, \bar{q}_{i,i}(s) = 1)$ is a credible vote belief. This result holds for any $i = \{1,2\}$. Moreover, there is no other vote belief that satisfies (ii) of Definition 1. To see this, suppose to the contrary that $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$ is a credible vote belief with a corresponding credible veto set $V_i = \{w\}$ and a credible non-veto set $W_{-i} = \{s\}$. Then, given these beliefs, the vetoer of weak type strictly lose from a veto, and thus he should not have vetoed the ex ante Pareto superior mechanism. That is, with the equilibrium strategies σ and ψ in G^δ with respect to $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$, where it is $\psi_{-i}(s) = 1$ in particular,

$$\begin{aligned} \sum_{t_{-i}} \bar{q}_{-i,i}(t_{-i}) \sum_{d\in D} \gamma(d|(w,t_{-i}))u_i(d,(w,t_{-i})) &= \bar{q}_{-i,i}(s) \sum_{d\in D} \gamma(d|ws)u_i(d,ws) \\ &+ \bar{q}_{-i,i}(w) \sum_{d\in D} \gamma(d|ww)u_i(d,ww) \\ &= \gamma(d_1|ws)u_i(d_1,ws) \end{aligned}$$

$$\begin{aligned} >&\bar{q}_{-i,i}(s)\{\psi_i(w)u_i(d_0,ws) + (1-\psi_i(w))(\psi_{-i}(s)u_i(d_0,ws) \\ &+ (1-\psi_{-i}(w)) \sum_{r\in R} \sigma_i(r|ws) \sum_{d\in D} \delta(d|r)u_i(d,ws))\} \\ &+ \bar{q}_{-i,i}(w)\{\psi_i(w)u_i(d_0,ww) + (1-\psi_i(w))(\psi_{-i}(w)u_i(d_0,ww) \\ &+ (1-\psi_{-i}(w)) \sum_{r\in R} \sigma_i(r|ww) \sum_{d\in D} \delta(d|r)u_i(d,ww))\} = u_i(d_0,ws) = 0. \end{aligned}$$

Therefore, $V_i = \{w\}$ is not credible. Also, $V_i = \{s, w\}$ is not credible, because then the status quo is played with the prior beliefs for both the vetoer and the ratifier, and thus, the weak type vetoer would strictly lose from the veto, which implies that the weak type's veto is not credible. Therefore, for all i, there exists a unique credible vote belief $\bar{q}_{\cdot,i}$ such that $\bar{q}_{-i,i}(w) = 1$ and $\bar{q}_{i,i}(s) = 1$ that does not satisfy (ii) of Definition 1 for any γ. So, γ is not ratifiable against δ because there is no equilibrium to the two-stage game in which an ex ante Pareto superior γ is unanimously approved over G^δ along every equilibrium path. □

Proof of Proposition 6 (Ratification of ex ante Pareto inferior mediator):
Assume that γ is ex ante Pareto inferior to δ. Again, by Lemma 5, under the priors, the interim utility from γ for the weak type is strictly lower than the interim utility from δ; and the interim utility from γ for the strong type is strictly greater than the interim utility from δ. That is, a strong type of any player prefers the ex ante Pareto inferior γ to δ when he knows only his type, while a weak type of any player prefers the ex ante Pareto superior δ to γ when she knows only her type. The proof consists of showing that for all i, either there does not exist a credible veto belief; or for every credible vote belief, it is rationalizable under the credibility conditions in (ii) of Definition 1. Without loss of generality, assume $i = 1$ and the same holds for $i = 2$. Let's suppose that the credible veto set is $V_i = \{w\}$ and the credible non-veto set $W_{-i} = \{s\}$ under the posterior beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$. The reporting strategies $\sigma = ((\sigma_{i,i}(t_i|t_i) \in [0,1], \forall t_i), (\sigma_{-i,i}(s|s) \in [0,1], \sigma_{-i,i}(w|w) = g(\sigma_{i,i}(w|w))))$ and the war strategies $\psi = ((\psi_{i,i}(t_i) \in (0,1), \forall t_i), (\psi_{-i,i}(s) = 1, \psi_{-i,i}(w) = 0))$, which the players would use in G^δ form a unique equilibrium in the subgame when γ does not win with respect to the posterior beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$ for any i vetoing. Here, the ratifier $-i$ of weak type's reporting strategy $\sigma_{-i,i}(w|w)$ is a function of the vetoer $-i$ of weak type's completely mixed equilibrium reporting strategy $\sigma_{i,i}(w|w)$ such that $\sigma_{-i,i}(w|w) = 1$ if $\sigma_{i,i}(w|w) > \frac{\delta(d_1|ss) - \delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$, $\sigma_{-i,i}(w|w) \in (0,1)$ if $\sigma_{i,i}(w|w) = \frac{\delta(d_1|ss) - \delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$, and $\sigma_{-i,i}(w|w) = 0$ if $\sigma_{i,i}(w|w) < \frac{\delta(d_1|ss) - \delta(d_1|sw)}{1+\delta(d_1|ss)-2\delta(d_1|sw)}$. Denoting it as $\Sigma(\bar{q}_{\cdot,i})$, these strategies form the only equilibrium of G^δ with respect to $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$, which satisfy (T2). Because γ is incentive compatible, for unanimous ratification of γ followed by truthful revelation in γ to be a sequential equilibrium, it must also be that γ is individually rational relative to G^δ. $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$ and the corresponding equilibrium $\Sigma(\bar{q}_{\cdot,i})$ satisfy (T3) for $i = 1$ of the strong type, for all $\psi'_{i,i} \in [0,1]$, and for all $\hat{r}_i \in T_i$:

$$\sum_{t_{-i}} \bar{p}_{-i}(t_{-i}) \sum_{d \in D} \gamma(d|(s, t_{-i}))u_i(d, (s, t_{-i})) = \bar{p}_2(s)\gamma(d_1|ss)u_1(d_1, ss) + \bar{p}_2(w)\gamma(d_1|sw)u_1(d_1, sw)$$

$$\geq \sum_{t_{-i}} \bar{p}_{-i}(t_{-i})\{\psi'_{i,i}(s)u_i(d_0,(s,t_{-i})) + (1-\psi'_{i,i}(s))\cdot(\psi_{-i,i}(t_{-i})u_i(d_0,(s,t_{-i})$$
$$+ (1-\psi_{-i,i}(t_{-i}))\sum_{r\in R}\sigma_i(r|(s,t_{-i}))\sum_{d\in D}\delta(d|r_{-i},\hat{r}_i)u_i(d,(s,t_{-i})))\}$$
$$= \bar{p}_2(w)(1-\psi'_{1,1}(s))\sum_{r\in R}\sigma_1(r|sw)\delta(d_1|r_2,\hat{r}_1)u_1(d_1,sw),$$

because $\bar{p}_2(s)\gamma(d_1|ss)u_1(d_1,ss) + \bar{p}_2(w)\gamma(d_1|sw)u_1(d_1,sw) \geq 0$ for any given incentive feasible γ and $u_1(d_1,sw) < 0$. Also, for $i = 1$ of the weak type, (T3) is satisfied:

$$\sum_{t_{-i}}\bar{p}_{-i}(t_{-i})\sum_{d\in D}\gamma(d|(w,t_{-i}))u_i(d,(w,t_{-i})) = \bar{p}_2(s)\gamma(d_1|ws)u_1(d_1,ws) + \bar{p}_2(w)\gamma(d_1|ww)u_1(d_1,ww)$$

$$\geq \sum_{t_{-i}} \bar{p}_{-i}(t_{-i})\{\psi'_{i,i}(w)u_i(d_0,(w,t_{-i})) + (1-\psi'_{i,i}(w))\cdot(\psi_{-i,i}(t_{-i})u_i(d_0,(w,t_{-i})$$
$$+ (1-\psi_{-i,i}(t_{-i}))\sum_{r\in R}\sigma_i(r|(w,t_{-i}))\sum_{d\in D}\delta(d|r_{-i},\hat{r}_i)u_i(d,(w,t_{-i})))\}$$
$$= \bar{p}_2(w)(1-\psi'_{1,1}(w))\sum_{r\in R}\sigma_1(r|ww)\delta(d_1|r_2,\hat{r}_1)u_1(d_1,ww),$$

because $u_1(d_1,ws) > 0$, $u_1(d_1,ww) > 0$, and $\gamma(d_1|ww) = 1$ (regardless of which γ).

For i of weak type's veto to be credible, under the equilibrium reporting strategies σ and war strategies ψ in $\Sigma(\bar{q}_{\cdot,i})$ given the posterior beliefs $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$, I must have:

$$\sum_{t_{-i}}\bar{q}_{-i,i}(t_{-i})\sum_{d\in D}\gamma(d|(w,t_{-i}))u_i(d,(w,t_{-i})) = \bar{q}_{-i,i}(s)\sum_{d\in D}\gamma(d|ws)u_i(d,ws)$$
$$+ \bar{q}_{-i,i}(w)\sum_{d\in D}\gamma(d|ww)u_i(d,ww)$$
$$= \gamma(d_1|ws)u_i(d_1,ws)$$

$$< \bar{q}_{-i,i}(s)\{\psi_{i,i}(w)u_i(d_0,ws) + (1-\psi_{i,i}(w))(\psi_{-i,i}(s)u_i(d_0,ws)$$
$$+ (1-\psi_{-i,i}(s))\sum_{r\in R}\sigma_i(r|ws)\sum_{d\in D}\delta(d|r)u_i(d,ws))\}$$
$$+ \bar{q}_{-i,i}(w)\{\psi_{i,i}(w)u_i(d_0,ww) + (1-\psi_{i,i}(w))(\psi_{-i,i}(w)u_i(d_0,ww)$$
$$+ (1-\psi_{-i,i}(w))\sum_{r\in R}\sigma_i(r|ww)\sum_{d\in D}\delta(d|r)u_i(d,ww))\}$$
$$= (1-\psi_{i,i}(w))\psi_{-i,i}(s)u_i(d_0,ws) = 0,$$

which is a contradiction because $u_i(d_1,ws) > 0$ for any γ that puts probability $\gamma(d_0|ws) < 1$ on

d_0. Note that it might be the case that $\gamma(d_1|ws)u_i(d_1, ws) = 0$, in which case γ is associated with a mechanism γ such that $\gamma(d_1|ws) = 0$ or $\gamma(d_0|ws) = 1$.

For any alternative interim Pareto inferior γ such that $\gamma(d_0|sw) = \gamma(d_0|ws) < 1$, the vetoer of weak type who is believed to have vetoed is strictly worse off if he vetoes, and so he must not have vetoed γ, that is, $v_i(w) = 0$. Therefore, $V_1 = \{w\}$ is not credible. For an alternative interim Pareto inferior γ such that $\gamma(d_0|sw) = \gamma(d_0|ws) = 1$, the vetoer of weak type is indifferent between vetoing or not. Moreover, the ratifier of strong type does have an incentive to not veto satisfying (T7), and thus, the player's beliefs are restricted to credible vote belief $\bar{q}_{\cdot,i} = (\bar{q}_{-i,i}(s) = 1, \bar{q}_{i,i}(w) = 1)$ satisfying (ii) of Definition 1. Also, it is easy to check that $V_i = \{s\}$ and $V_i = \{s, w\}$ are not credible in either cases. Depending on $\gamma(d_0|ws)$, for all i, it is either the case that (i) there does not exist a credible veto belief; or (ii) for every credible vote belief $\bar{q}_{\cdot,i}$ (in which case there is only a unique credible belief with $\bar{q}_{-i,i}(s) = 1$ and $\bar{q}_{i,i}(w) = 1$), it is restricted under the condition in (ii) of Definition 1. Therefore, because there exists an equilibrium ratification of γ when the status quo is G^δ, unanimous ratification of ex ante Pareto inferior γ (relative to δ) followed by truthful revelation to γ, together with the equilibrium reporting strategies σ and the equilibrium war strategies ψ in $\Sigma(\bar{q}_{\cdot,i})$ specified above, forms a sequential equilibrium of the two stage ratification game, where beliefs following disagreement satisfy the credibility conditions.

In fact, there is only one sequential equilibrium that survives the credibility refinement, in a sense that the equilibrium is unique up to equivalence in the equilibrium outcome. That is, when γ is associated with $\gamma(d_0|sw) = \gamma(d_0|ws) < 1$, there is no other sequential equilibrium where it does get voted down (so $v_i(t_i) = 0$ for all i and for all t_i in the unique sequential equilibrium); and when γ is associated with $\gamma(d_0|sw) = \gamma(d_0|ws) = 1$, then there is no other sequential equilibrium than the one with $v_i(s) = 0$ and $v_i(w) \in (0, 1)$ for all i, with the weak type randomly choosing between veto and not veto, where the beliefs satisfy the credibility conditions. My formulation of the refinement criterion defining beliefs "off the equilibrium path" eliminates other equilibria such that it selects a unique sequential equilibrium with credible beliefs. □

Proof of Theorem 2 (Existence and uniqueness of secure mechanism):

(*Existence*) By Lemma 5 and Corollary 1, there exists an incentive feasible mechanism in $S(\Gamma)$ that is on the interim incentive efficient frontier and that is ex ante Pareto inferior to every other mechanisms in $S(\Gamma)$. (When $p \geq p^{**}$, there is a unique ex ante and interim incentive efficient mechanism.) Denote a mediator associated with this mechanism to be δ^*. Then, Proposition 5

proves that any mediator $\gamma \in S(\Gamma)\backslash\delta^*$ is not ratifiable against δ^* when the status quo is G^{δ^*} since γ is ex ante Pareto superior to δ^*. That is, there does not exist an alternative mechanism that could be unanimously ratified against δ^*. Moreover, it is easy to check that the conditions in Definition 1 are satisfied so that δ^* is the only mediator ratifiable against itself, that is, G^{δ^*} would be impervious to permitting some sort of preplay communication before G^{δ^*} is carried out. Because the only interim incentive efficient mediator that is ratifiable against G^{δ^*} – which is δ^* – is an equilibrium outcome of G^{δ^*} under the prior beliefs, by Definition 3, mechanism G^{δ^*} is secure. Because δ^* exists in the set $S(\Gamma)$, then G^{δ^*} exists.

(*Uniqueness*) Moreover, G^{δ^*} is the only secure status quo mechanism. That is, all other status quo mechanisms G^{γ} characterized by $\gamma \in S(\Gamma)\backslash\delta^*$ are not secure. To see this, consider a status quo mechanism G^{γ} where $\gamma \in S(\Gamma)\backslash\delta^*$. Then, by Proposition 6 and noticing that there exists some ex ante Pareto inferior mechanism $\gamma^{'}$ relative to γ, $\gamma^{'}$ is ratifiable against γ for every $\gamma^{'}$ that is ex ante Pareto inferior to γ. However, those interim incentive efficient mediators $\gamma^{'}$ that are ratifiable against γ are not the equilibrium outcome of G^{γ} under the prior beliefs. Therefore, for any $\gamma \in S(\Gamma)\backslash\delta^*$, a mechanism G^{γ} is not secure by Definition 3. Thus, there exists a unique interim incentive efficient and secure mechanism G^{δ^*}. □

Proof of Proposition 7 (Characterization of threat-secure mediators):

Notice from the previous proof that $\delta^* \in S(\Gamma)$ is ex ante Pareto inferior to every other mechanisms in $S(\Gamma)$. Therefore, any $\gamma \in S(\Gamma)\backslash\delta^*$ is ex ante Pareto superior to δ^*; and by Proposition 5, any $\gamma \in S(\Gamma)\backslash\delta^*$ is not ratifiable against δ^*. Then, by Definition 4, δ^* is threat-secure, and it is the only threat-secure mediator among all of the interim incentive efficient mediators. To characterize δ^*, note that δ^* is constructed to be the one that is interim incentive efficient and that is ex ante Pareto inferior to every other mediator in $S(\Gamma)$ when there is another mediator. Therefore, by Corollary 1, $\delta^* = \mu_{1,o}$ when $p \leq p^*$ and $\delta^* = \mu_{1,\tilde{z}(p)}$ when $p \in (p^*, p^{**})$. When $p \geq p^{**}$, there is a unique interim incentive efficient mediator $\mu_{0,0}$, which could trivially be δ^*. These constitute the set of threat-secure mediators $TS(\Gamma)$. □

Proof of Theorem 3 (Equivalence Theorem):

Directly follows from Propositions 1, 2, 3, 4, and 7. □