Endogenous Choice of a Mediator: Inefficiency of Bargaining*

Jin Yeub Kim†

December 10, 2014

Abstract

I study 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 mechanism that respects private information. I show that the very process of selecting a mediator exhibits an inherent inefficiency in interim bargaining due to the incentive of each party to avoid seeming weak. I investigate this result for a benchmark class of examples using a two-pronged approach, cooperative and noncooperative. In the cooperative approach, I find that there exists a *unique* neutral bargaining solution. Under the noncooperative approach, I model the mediator-selection problem as a two-stage game, in which the parties may disagree on the choice of a mediator. I define *threat-secure* mediators, which can resist such disagreement, and show that the threat-secure mediator exists and is *unique*. 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.

JEL Classification: C71; C72; C78; D82

Keywords: Cooperative and noncooperative games; Bargaining; Mediation; Mechanism design

*I am deeply grateful to Roger Myerson, Lars Stole, and Ethan Bueno de Mesquita for valuable advice, helpful conversations, and continuous support. I have benefited greatly from several discussions with Alex Frankel, Heung Jin Kwon, and Richard van Weelden. I also appreciate Sandeep Baliga, Mostafa Beshkar, Yeon-Koo Che, Peter Cramton, Thomas Gresik, Jong-Hee Hahn, Leslie Johns, Byung-Cheol Kim, Jinwoo Kim, Kyungmin (Teddy) Kim, Sun-Tak Kim, David Miller, Jaeok Park, Cheng-Zhong Qin, Ilya Segal, Hugo Sonnenschein, Yuki Takagi, Gustavo Torrens, and seminar participants at University of Chicago, Indiana University-Bloomington, Monash University, Sogang University, University of Tokyo, and various conferences for helpful feedback and suggestions.

†Department of Economics, University of Nebraska-Lincoln, NE 68588. E-mail: jkim43@unl.edu, Webpage: https://sites.google.com/site/jinyeubkim

1 Introduction

Two bargaining parties with private information often employ mediators as one of the primary tools of dispute resolution.[1] However, in many settings, two parties often choose a mediator that offers a higher risk of disagreement than do many of the available alternatives. 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 a mediator. Through this paper, I attempt to provide a richer understanding of the failure of efficiency in two-person bargaining with incomplete information.

I consider a bargaining problem in which two players with private information about their own types – strong or weak – can jointly reach an agreement ("peace") or else a disagreement ("war") occurs. In an attempt to extend the set of equilibria to include better outcomes, the players can seek to communicate with each other through a mediator.[2] In this paper, the definition of a mediator is a person who recommends a bargaining outcome depending on the players' reported types while respecting their private information. This setup allows me to consider a mediator to be equivalent to a communication-facilitation device, or a mechanism.[3]

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 for all of 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 mediators, the problem of "information leakage" arises: advocating for a particular mediator conveys information about the player's type. For this reason, the issue of which mediator gets 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 among 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 information leakage issue, either directly or indirectly, I can think about what are the desiderata for the solution concept. To fully explore this problem, I take two different

[1] In the context of international conflicts, mediation efforts are recurrent as a means for reducing the likelihood of disagreement and mitigating the potentially violent consequences. According to Wilkenfeld et al. (2003), 30 percent of the 419 cases coded as international crises in the International Crisis Behavior Project data set between 1918 to 1996 were mediated. They report that 64 percent of all international conflicts between 1990 and 1996 involved mediation efforts. Dispute resolution mechanisms (i.e., mediators) can indeed be useful in bargaining situations by mitigating the risk of bargaining failure.

[2] Another way to extend the set of equilibria is to allow renegotiation opportunities in a long-term relationship, which is not the focus of this paper. See Brennan and Watson (2013) for an environment with costly renegotiation.

[3] A formal justification is given in Online Appendix D. See Kim (2014, 3-12) for more discussions on bargaining, mediation, and mechanism design.

approaches: cooperative and noncooperative theories of bargaining with incomplete information. With both approaches, I address which mediator is most likely to arise in the mediator selection process, how the issue of information leakage impacts the endogenous selection of a mediator, and as a result how the endogenous selection process can lead to inefficiency in bargaining.

In Section 2, I identify a class of benchmark bargaining models within general Bayesian bargaining problems, where privately informed players have conflicting interests when they are of different types; and formally define the space of feasible mediators. Under the standard mechanism design approach, by the revelation principle, I take the set of mediators that are available to the players to be synonymous with the set of incentive compatible and individually rational mechanisms the players can agree on.[4]

Under the cooperative approach, I ask the following question: What is a reasonable set of mediator(s) we should expect to see arise as an outcome of an undefined selection process within the set of feasible mediators? To answer this question, I implicitly take into account the information leakage problem by imposing a set of relevant properties or axioms that must be true for a reasonable bargaining solution that captures the trade-off between the different types of players. In Sections 3 and 4, I refine the solution set with two 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 is an infinite number of interim incentive efficient mediators among which there is a unique ex ante incentive efficient one. In order to give a stronger prediction of which mediator would be chosen from this considerable range of interim incentive efficient mediators, I further 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 is there a *unique* neutral bargaining solution (Theorem 1), but the solution is not the ex ante incentive efficient choice. The suboptimal choice represents, of all the interim incentive efficient mediators, the *most ex ante inefficient*[5] mediator associated with the highest probability of disagreement (Theorem 2).

Section 4 also explores the intuitions behind these results and provides various implications. In my benchmark class of bargaining models, an agreement is symmetric and incentive efficient for the players if they have the same type. But if one player is strong while the other player is weak, then an agreement would be better for the weak player but worse for the strong player, compared to the disagreement outcome that either player can force. The weak player's gains are greater than the strong player's losses, and so ex ante efficiency calls for an agreement in this case. But without side-payments, the strong

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

[5]Precisely stated, the most ex ante inefficient mediator, contrary to the ex ante incentive efficient mediator, is a solution to the program of minimizing the players' ex ante expected payoffs over the set of all interim incentive efficient mediators. In other words, the most ex ante inefficient mediator is ex ante Pareto dominated by any other interim incentive efficient mediator.

player would prefer to force a disagreement when matched with a weak player. Thus in the mediator selection process, each player should be afraid of seeming weak, and so the players would choose the mediator that is favorable to the strong player. In particular, what is best for the strong type is the ex ante worst mediator. In this sense, the process of selecting a mediator itself generates an inherent inefficiency in interim bargaining.

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 itself be entirely convincing without also considering a more explicit noncooperative game. To this end, in Section 5, I determine the solution set differently than the previous approach by modeling the mediator selection process as a two-stage game: a first stage in which each player votes for a certain mediator, followed by a second stage in which the players revert to playing the default game noncooperatively if they disagree on the choice of a mediator. I ask the following question: Does there exist an equilibrium in this game in which the players always unanimously match their choices, allowing the possibility of learning from the players' non-matching voting decisions? Under the noncooperative approach, I develop and illustrate a new concept, *threat-security*, that takes account of inferences from non-matching votes in a consistent way. What I find in this context is that there exists only one threat-secure and interim incentive efficient mediator (Theorem 3), which is the one that is ex ante Pareto dominated by any other mediator.

In Section 6, I synthesize the results from two approaches: Both approaches not only offer ex ante incentive inefficiency, but also prescribe the same unique solution. That is, the threat-secure mediator is exactly the same as the neutral bargaining solution (Theorem 4). In general, the neutral bargaining solutions and the threat-secure set are not nested. However, in my setting, two distinct ways of looking at the problem lead to the same suboptimal choice of a mediator. I examine the logical foundations to this equivalence result. In the class of benchmark examples, players pick (in the cooperative sense) the ex ante worst mediator effectively pooling in a way that reveals no information. Of note is that the nature of that pooling is reminiscent of signaling in the noncooperative game. The equivalence result is not yet posed as a formal theorem in a general framework; yet, I attempt to lay the noncooperative foundations to the cooperative ideas of Myerson (1984*b*)'s neutral bargaining solution. Section 7 concludes with briefly mentioning a few of future research directions. All proofs can be found in Appendices A, B, and Online Appendix C. A number of additional online appendices provide supplementary materials including an illustrative example.[6]

1.1 Related Literature and Contributions

I conclude the introduction by highlighting the contribution of this paper in relation to various strands of literature. A significant amount of the theoretical work on mediation in the international conflict literature focuses on when and how mediation improves on unmediated communication (e.g., Kydd, 2003; Hafer, 2008; Fey and Ramsay, 2009; Fey and Ramsay, 2010). Including Fey and Ramsay (2009), the recent research on the topic of

[6]Link to Online Appendices: `https://sites.google.com/site/jinyeubkim/files//Online_Appendices.pdf?attredirects=0`

mediation adopts mechanism design tools to identify the conditions under which mediation can be effective in preventing conflicts (e.g., Fey and Ramsay, 2011; Hörner, Morelli and Squintani, 2011; Meirowitz et al., 2012). Despite the extensive literature on the subject, most has focused on the analysis of effective mediators and thus has ascribed the sources of bargaining failure (i.e., disagreement) to particular attributes of mediators. But of note is that these mediators are selected endogenously by the bargaining parties. Focusing on the question of how mediators are selected, my paper contributes a new idea to the debate over the sources of bargaining failure: the process of bargaining over mediators itself can increase the risk of disagreement.

Starting with Nash (1950), a large literature studies the two-person bargaining problems, including classical bargaining literature by 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)). However, there is no commonly accepted interim bargaining solution concept in the literature. The neutral bargaining solution (Myerson, 1984*b*) has not been generally accepted as a standard for analysis of bargaining with incomplete information mainly due to the complicacy of the concept. Although its application involves some subtle analysis, the general conditions for a neutral bargaining solution are indeed well determined. The contribution of my paper– as a revisit to the neutral bargaining solution–is twofold. First, this paper identifies the right class of bargaining models to yield good insights into the problem of how ex ante efficiency can be seriously misleading as a solution concept for interim bargaining; and to further underline the analytical power of the neutral bargaining solution as an interim bargaining solution concept. Second, my paper proves the uniqueness of the neutral bargaining solution for a class of models. Myerson (1984*b*) establishes the existence of neutral bargaining solutions for any class of two-person Bayesian bargaining problems; yet there is no general uniqueness theorem.

Finally, I note that several authors have addressed the issue of information leakage in mechanism selection games or the robustness of the optimal mechanisms (e.g., Holmström and Myerson, 1983; Lagunoff, 1995; Cramton and Palfrey, 1995; Laffont and Martimort, 2000; Celik and Peters, 2011). The methodological approaches there are somewhat similar to the noncooperative approach in the current paper in that those authors also model the bargaining process of selecting a mechanism as extensive form noncooperative games, where equilibrium play in the "outside" option depends on the players' revised beliefs; however, they impose different assumptions on the exact conduits of how the game is played.[7] Set aside any technical similarity or dissimilarity with the literature, the novelty of my paper is that, by studying the underlying incentives that arise in bargaining over mediators, I combine the cooperative ideas from Myerson (1984*b*)'s neutral bargaining solution with the strategic intuitions behind a noncooperative game. It is my opinion that this paper should be taken as only a first step in an attempt to build a bridge

[7]For example, their concepts rely on different equilibria solution concepts, different restrictions on off path beliefs, or different assumptions on voting procedures and "outside" options. The key difference from my paper lies on the basic structure of the game: the above mentioned papers formalize the idea of which mechanism is feasible or implementable in a *pairwise* comparison.

between cooperative and noncooperative game theories in the context of bargaining with incomplete information.

2 The Model

I consider a model of mediator selection game in which two players with private information in a Bayesian bargaining problem negotiate the choice of a mediator, through whom the players share information and reach a bargaining outcome. The definitions of what follows are done in full generality following the basic setup in Myerson (1984*b*), although I focus on a simplified problem as described subsequently. I then identify a specific class of the underlying bargaining problems and define the set of feasible mediators, which taken together constitute a mediator selection game.

2.1 General Bayesian Bargaining Problems

A general two-person Bayesian 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 finite set of feasible bargaining *outcomes* available to the two players from which they can choose. $d_0 \in D$ denotes the disagreement outcome that the two players must get if they fail to coordinate. For each $i \in \{1, 2\}$, T_i is the finite set of possible types t_i for player i.[8] The players are uncertain about each other's type, and the players' types are unverifiable. Let $T = T_1 \times T_2$ denote the set of all possible type combinations $t = (t_1, t_2)$. 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 $u_i(d, t_1, t_2)$ denotes the payoff to player i if $d \in D$ is the outcome and (t_1, t_2) is the true vector of the players' types. The payoffs are in von Neumann-Morgenstern utility scale. 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. 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 $-i$'s type being t_{-i}. For simplicity, I assume that the types are independently distributed, with the probability distribution of i's type denoted $\bar{p}_i$ in $\Delta(T_i)$, which is common knowledge. That is, $\bar{p}_i(t_i)$ denotes the prior marginal probability that player i's type will be t_i and $p_i(t_{-i}|t_i) = \bar{p}_{-i}(t_{-i})$, $\forall i \in N$, $\forall t_{-i} \in T_{-i}$, $\forall t_i \in T_i$.

Hereafter, I restrict attention to two symmetric players, each with two discrete types, and two outcomes; and describe the setup in the language of international conflicts. Each player $i = \{1, 2\}$ has private information about his type $t_i \in T_i = \{s, w\}$, where s denotes the strong type and w denotes the weak type. For the sake of simplicity and tractability, I assume symmetry in probabilities – the prior marginal probabilities of the strong type are the same for both players; namely, $\bar{p}_1(s) = \bar{p}_2(s) \equiv p$. Relaxing this

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

assumption complicates some details of the analysis without adding additional insights. There are two possible outcomes – "war" and "peace" – that are jointly feasible for the players together.[9] A central facet of international conflict is that no enforcement body exists to permit binding contracts and that a country has inalienable control over its action. Thus a country can at any time unilaterally choose to initiate a war even if the mediator's recommendation is peace; and there is no way a country can commit not to do so.[10] Because war can be forced by either player, with two outcomes – war or peace – war is the only threat-point in bargaining. Therefore, I can designate "war" to be the disagreement outcome d_0 that will occur if the players fail to agree on a mediator.[11] Thus let $D = \{d_0, d_1\}$, where d_0 is "war" and d_1 is "peace."

This parsimonious setup still captures many situations that the two players face in bargaining and yields a rich set of theoretical implications. It is straightforward to extend this stylized setup in a number of ways to suit different applications. I mention a few of these directions. First, calling d_0 and d_1 war and peace, respectively, is only a matter of labeling; the analysis in this paper could be equally applicable in any kind of bargaining situations with a specification of one designated outcome d_0. Also, restricting attention to two outcomes substantially simplifies the exposition while conveying all the key insights. Suitable versions of the results continue to hold when more than two outcomes are allowed as long as the disagreement outcome d_0 is fixed. Finally, the framework I have developed here is particularly amenable to allow for the possibility of multiple threat-points. The logic of my analysis suggests that the main insights of the results in this paper also apply to this extension but with some added nuances.

2.2 A Class of Benchmark Bargaining Models

I focus attention on a particular class of examples. Let Γ^* denote a two-person Bayesian bargaining problem Γ such that $(u_i)_{i\in\{1,2\}}$ satisfies the following assumptions:

Assumption **(A1)**. $\sum_i u_i(d_1, t) > 0$, $\forall t$.

Assumption **(A2)**. $u_i(d_1, t) < 0$ when $t_i = s$ and $t_{-i} = w$, $\forall i$.

Assumption **(A3)**. The payoffs are symmetric in the sense that $T_1 = T_2$ and

$$u_1(d_1, (\alpha, \beta)) = u_2(d_1, (\beta, \alpha)),\ \forall \alpha \in T_1,\ \forall \beta \in T_1.$$

[9] For example, two states are involved in a dispute over a divisible item, an area of territory, or an allocation of resources that could possibly lead to war.

[10] In many other situations besides international conflicts, mediation also cannot be implemented without the prior consent of the players.

[11] Then, as long as a mediator offers each type of each player an expected utility that is not less than the expected utility from the disagreement outcome, each type of each player should be willing to participate in mediation and voluntarily make a binding commitment to obey a mediator's recommendation (and subsequently follow the recommendation). Thus the issue of imperfect commitment is not present in my setting. See Bester and Strausz (2001) for the environments with limited commitment in a principal-agent contracting problem.

In words, **(A1)** peace (agreement) is socially better than war (disagreement), but **(A2)** a strong player prefers to force war when matched with a weak player. Assumption **(A3)** simply states that the payoffs for a particular type of a player depend only on the other player's type, not on the identities of the players. Because of **(A3)**, I focus on $i = 1$ in the analysis that follows.

Then Γ^* can be described as follows: An agreement is better than the disagreement outcome for both players if they have the same type. But if one player is strong while the other player is weak, then an agreement would be better for the weak player but worse for the strong player, compared to the disagreement outcome that either player can force. The weak player's gains are greater than the strong player's losses in the latter case. In this paper, I study mediator selection in this class of two-person Bayesian bargaining problems Γ^* within the general class of Bayesian bargaining problems Γ.

2.3 The Set of Feasible Mediators

In any Bayesian bargaining problem Γ^*, the players may agree on some mediator who specifies how the choice $d \in D$ should depend on the players' types. That is, they may bargain over the selection of a mediator who, if chosen, mediates by the following form: first, each player is asked to separately and confidentially report his type to the mediator; then, after getting these reports, the mediator recommends an outcome to the players.[12] Then, allowing randomized strategies, a *mediator* (or *mediation mechanism*) can be defined as a function $\mu : D \times T \to \mathbb{R}$ such that

$$\sum_{c \in D} \mu(c|t) = 1 \text{ and } \mu(d|t) \geq 0, \ \forall d \in D, \ \forall t \in T. \tag{2.1}$$

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

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

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

is the conditional expected utility for player i, given that he is of type t_i and both players report their types honestly, if the mediator μ is selected.

Because the types are unverifiable, players will not reveal their types honestly unless they are given incentives to do so. Also, because players can force the disagreement outcome, they will not participate in mediation unless they are given incentives to do so. Thus each mediator must incorporate the players' incentive constraints that they may not

[12] My modeling choice of mediator as *mechanism*, which recommends a bargaining *outcome* to the players depending on the players' reports, is justified formally in Online Appendix D. By defining a mediator as a mechanism that chooses an outcome, I can characterize the (ex ante) probability of each outcome that is obtainable with different mediators while leaving potentially endless variation in any bargaining procedure unspecified.

want to tell the truth or they may not want to participate.[13] Let $U_i^*(\mu, s_i|t_i)$ denote the expected utility for type t_i of player i if he lies about his type and reports s_i while player $-i$ is honest. That is,

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

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

$$U_i(\mu|t_i) \geq U_i^*(\mu, s_i|t_i),\ \forall i,\ \forall t_i \in T_i,\ \forall s_i \in T_i. \tag{2.2}$$

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

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

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:

$$U_i(\mu|t_i) \geq 0,\ \forall i,\ \forall t_i \in T_i. \tag{2.3}$$

Therefore, taking relevant incentive constraints into account, I define the *incentive feasible mediator* for a Bayesian bargaining problem Γ^* to be a mediator μ that is both incentive compatible and individually rational in the sense of conditions (2.2) and (2.3). Then, by the revelation principle, there is no loss of generality in focusing on incentive feasible mediators, and thus I can naturally assume that the rational intelligent players themselves should be able to bargain over the set of all incentive feasible mediators, denoted as $F(\Gamma^*)$.

Then, a mediator selection game induced from the original Bayesian bargaining problem is characterized by a pair $(\Gamma^*, F(\Gamma^*))$, where Γ^* is the Bayesian bargaining problem in the benchmark class and $F(\Gamma^*)$ is the set of feasible mediators for Γ^*.

3 The Interim Incentive Efficient Mediators

Given the set of incentive feasible mediators, concepts of efficiency under the mechanism design approach can be applied to identify a set of "optimal" mediators – mediators that are efficient in either an ex ante or interim sense – among which the players can choose from. An incentive feasible mediator μ is *interim incentive efficient* if and only if there exists no other incentive feasible mediator 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. Formally, a mediator μ' is interim Pareto superior to μ if and only if $U_i(\mu'|t_i) \geq U_i(\mu|t_i),\ \forall i,\ \forall t_i \in T_i$, and this inequality

[13] The revelation principle (Myerson, 1979) asserts 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, where relevant, individually rational)" (Myerson, 1991, 487).

is strict for at least one type of one player. An incentive feasible mediator μ is *ex ante incentive efficient* if and only if there exists no other incentive feasible mediator that is ex ante Pareto superior to μ. Let $U_i(\mu)$ denote the ex ante expected utility for player i, if the mediator μ is selected. That is, for any $i \in \{1, 2\}$,

$$U_i(\mu) = \sum_{t \in T} \sum_{d \in D} \bar{p}_i(t_i)\bar{p}_{-i}(t_{-i})\mu(d|t)u_i(d, t).$$

Then, a mediator μ' is ex ante Pareto superior to μ if and only if $U_i(\mu') \geq U_i(\mu)$, $\forall i$, and this inequality is strict for at least one player. Note that for $F(\Gamma^*)$, the set of ex ante incentive efficient mediators in $F(\Gamma^*)$ is a subset of the set of interim incentive efficient mediators in $F(\Gamma^*)$.[14]

One might naively consider ex ante incentive efficiency to be a requirement for a bargaining solution. But, if a player expresses his preference for the ex ante incentive efficient solution, then it would be giving information to the other side that might be detrimental. Therefore, there is no reason a priori to expect the ex ante incentive efficient solution to be negotiated by privately informed parties. However, interim incentive efficiency is clearly a minimal requirement in a setting in which each player already knows only his own type at the initial decision-making stage.[15] That is, when each player's type is private information, once the players know their types, then they should choose a mediator from the set of interim incentive efficient mediators.

Hereafter, I restrict my attention to the *symmetric* mediators that treat the players symmetrically, but the result continues to hold when non-symmetric mediators are allowed. Let $S(\Gamma^*)$ denote the set of all symmetric *interim incentive efficient* (IIE) mediators for Γ^*. In this section, I fully characterize $S(\Gamma^*)$ and thereby refining a mediator selection game $(\Gamma^*, F(\Gamma^*))$ such that the players bargain over $S(\Gamma^*)$. Note that any mediator can be entirely defined by the probability weights he puts on the outcomes of war and peace for each type realization. The following structure will then prove useful in characterizing $S(\Gamma^*)$.

Remark 1. Let μ_y be a mediator defined as a function $\mu : D \times T \to \mathbb{R}$ such that (2.1) holds; $\mu(d_0|t) = 0$ if $t = (w, w)$, $\mu(d_0|t) = y$ if $t \in \{(s, w), (w, s)\}$, and $\mu(d_0|t) = z$ if $t = (s, s)$, where $y \geq 0$ and z is uniquely determined given y.

That is, μ_y implements d_0 (war) with probability zero when $t = (w, w)$, with some non-negative probability y when $t \in \{(s, w), (w, s)\}$, and with probability z when $t = (s, s)$, where z is a function of y. The following lemma gives three thresholds that separate four cases along the range of the probability of the strong types, p. For notational purposes, I write (t_1t_2) interchangeably with (t_1, t_2) for the vector of types t.

[14] See Holmström and Myerson (1983, 1806).

[15] The notion of interim incentive efficiency was first introduced in Holmström and Myerson (1983). Ledyard and Palfrey (2007) use this concept to fully characterize interim efficient mechanisms for the class of linear independent environments.

Lemma 1 (Thresholds). *For any given Γ^*, there exist unique $p^{'} \in (0,1)$, $p^* \in (0,1)$, and $p^{**} \in (0,1)$ such that each satisfies the following, respectively:*

$$
\begin{aligned}
&p^{'}u_1(d_1, ss) + (1-p^{'})u_1(d_1, sw) = 0;\\
&p^*(1-y)u_1(d_1, ws) + (1-p^*)u_1(d_1, ww)\\
&\qquad = p^*u_1(d_1, ws) + (1-p^*)(1-y)u_1(d_1, ww),\ \forall y \in (0,1];\\
&p^{**}(1-z(y,p^{**}))u_1(d_1, ss) + (1-p^{**})(1-y)u_1(d_1, sw)\\
&\qquad = p^{**}u_1(d_1, ss) + (1-p^{**})u_1(d_1, sw),\\
&\qquad \textit{where } z(y,p^{**}) = y\left[1 - \frac{(1-p^{**})u_1(d_1, ww)}{p^{**}u_1(d_1, ws)}\right],\ \forall y \in (0,1].
\end{aligned}
$$

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

Lemma 1 implies that there exists a unique $p^{'}$ such that the participation constraints bind for the strong type given μ_0 with $z = 0$; there exists a unique p^* such that the informational incentive constraints bind for the weak type given μ_y for any $y \in (0,1]$ and $z = 0$; and there exists a unique p^{**} such that the strong type is indifferent between μ_y, for any $y \in (0,1]$ and $z(y,p^{**})$, and μ_0 with $z = 0$. The closed form solutions for these thresholds are given in Appendix A.1.

Note that I must distinguish between the two instances: $p^{'} \leq p^*$ and $p^{'} > p^*$. Along with Lemma 2, Proposition 1 gives a complete characterization of symmetric IIE mediators, for Γ^* when $p^{'} \leq p^*$, by μ_y with further restrictions on y and z depending on the probability of the strong types.

Lemma 2. *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 $\bar{z}(p) \equiv z(1,p)$ is increasing in p.*

Proposition 1 (When $\mathbf{p^{'} \leq p^*}$). *For any Γ^*, $S(\Gamma^*) = \{\mu_y\}$ if and only if*

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

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

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

The key point of Proposition 1 is that, for given p, symmetric IIE mediators can be identified by μ_y with restrictions on y – the probability of choosing war if $t \in \{(s,w),(w,s)\}$); The probability of choosing war if $t = (s,s)$, z, is pinned down entirely by y for each case. For example, when $p \in [p^{'}, p^*)$, $S(\Gamma^*) = \{\mu_y | y \in [0,1] \text{ and } z = 0\}$. That is, symmetric mediators who recommend war with probability zero when both players are of the same

type and recommend war with any probability between zero and one when players are of different types are IIE. These mediators cannot be interim Pareto dominated by any other incentive feasible mediator, and therefore all and only these mediators are interim incentive efficient when $p \in [p^{'}, p^{*})$.

There are three issues to note for ***Case 1***, ***Case 3***, and ***Case 4***, respectively.[16] First, when $p < p^{'}$ (***Case 1***), any symmetric IIE mediator must choose war with some positive probability of at least $\underline{y}(p)$ to induce the strong type to participate. Second, when $p \in [p^{*}, p^{**})$ (***Case 3***), a symmetric IIE mediator must also put some non-negative 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 the type combinations is interim Pareto dominated by μ_0.

The following Proposition 2 together with Lemma 2 fully characterizes the set of symmetric IIE mediators for Γ^{*} when $p^{'} > p^{*}$. The only difference from Proposition 1 is that when $p \in [p^{*}, p^{'})$, a symmetric IIE mediator must respect both the facts that the weak type has an incentive to report dishonestly if the mediator does not put some positive probability on war when $t = (s, s)$, and that the strong type has an incentive not to participate if the mediator does not choose war with sufficiently high probability when $t \in \{(s, w), (w, s)\}$.

Proposition 2 (When $\mathbf{p^{'} > p^{*}}$). *For any Γ^{*}, $S(\Gamma^{*}) = \{\mu_y\}$ if and only if*

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

Case 2′. *if $p \in [p^{*}, p^{'})$: $y \in \left[\underline{\underline{y}}(p), 1\right]$ and $z := z(y, p) \in \left[\underline{\underline{z}}(p), \bar{z}(p)\right]$, where $\underline{\underline{y}}(p) > 0$ and the corresponding $\underline{\underline{z}}(p) > 0$ are defined by* (A.4) *and* (A.5) *in Appendix A.2;*

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

Case 4. *if $p \geq p^{**}$: $y = 0$ and $z = 0$, as in Proposition 1. Case 4.*

Propositions 1 and 2 both imply that there are multiple symmetric IIE mediators if $p < p^{**}$. If $p \geq p^{**}$, then only one mediator μ_0 exists in $S(\Gamma^{*})$.[17] Therefore, the intriguing cases occur only when $p < p^{**}$ in which there is an infinite number of symmetric IIE mediators from which the players can select.

Remark 2. In what follows, I only focus on when $p < p^{**}$.

[16] More detailed explanations of these issues in relation to the thresholds and incentive feasibility can be found in Appendix A.2.

[17] In fact the unique symmetric IIE mediator simulates the unique Bayesian equilibrium of the underlying Bayesian game without communication when $p \geq p^{**}$. In this case, the expected payoff allocation in such equilibrium is exactly the payoff allocation that can be achieved through μ_0, and no higher expected payoff allocation can be achieved. Thus when $p \geq p^{**}$, the players would not need to seek a mediator initially. See Online Appendix D.2.

In fact, for each p, I can order the symmetric IIE mediators according to the ex ante welfare criterion.

Lemma 3. *Within $S(\Gamma^*)$ for each p, $y < y^{'}$ if and only if $U_i(\mu_y) > U_i(\mu_{y^{'}})$ for all i.*

Proposition 3. *The unique ex ante incentive efficient mediator in $S(\Gamma^*)$ is:*

(When $\mathbf{p}^{'} \leq \mathbf{p}^*$). *(a) $\mu_{\underline{y}(p)}$ for **Case 1**; (b) μ_0 for **Cases 2 & 3**.*

(When $\mathbf{p}^* < \mathbf{p}^{'}$). *(a) $\mu_{\underline{y}(p)}$ for **Case 1**; (b) $\mu_{\underline{\underline{y}}(p)}$ for **Case 2'**; (c) μ_0 for **Case 3**.*

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 symmetric IIE mediators. The following Lemma 4 gives the interim welfare ordering in $S(\Gamma^*)$.

Lemma 4. *Within $S(\Gamma^*)$ 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 4 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 a continuum of symmetric IIE mediators in $S(\Gamma^*)$ that are not ex ante incentive efficient. Moreover, μ_1 is ex ante Pareto inferior to any other mediators in $S(\Gamma^*)$; gives the highest interim payoff for the strong type and the lowest interim payoff for the weak type; and is associated with the highest probability of the disagreement outcome.*

Corollary 1 immediately follows from the previous results. The interim but not ex ante incentive efficient mediators sometimes allow players to go to war with a higher probability than the ex ante incentive efficient mediator. The "worst" mediator μ_1, who puts a higher probability on war than any other IIE mediator, is the one that can be described as being *the farthest away from* the ex ante incentive efficient mediator. The probability of war associated with such mediator is the highest among all symmetric IIE mediators.

4 Bargaining over Mediators: 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 predict which mediator two privately informed parties might select. In the previous section, I characterized the set of "optimal" mediators – mediators that are efficient in an interim sense – among which the players would reasonably choose from. However the notion of interim incentive efficiency still identifies too large a set of attainable mediators. This range then begs the question of

whether we can determine a smaller set. In order to answer this question, I use a cooperative solution concept that reasonably refines the set of interim incentive efficient mediators, called a neutral bargaining solution (Myerson, 1984*b*).[18] In particular, I characterize the neutral bargaining solutions for mediator selection by the players, and show inefficiency in interim bargaining.

4.1 Cooperative Approach and the Inscrutability Principle

Before continuing, I invoke the advantage of using the cooperative approach particularly for games of mediator selection. If I build mediator selection process explicitly into an ad hoc extensive form game, then I should be able to describe the equilibrium correspondence of this game. Unfortunately, the results of this analysis might depend very strongly on the precise form of the game and might be driven by the details in the game. Also, in many settings, the procedures of negotiations over the choice of a mediator are often amorphous and the exact conduits of making offers and counter-offers are not always precisely stipulated.[19] The cooperative framework abstracts away from these procedural details, and instead delineates the properties of bargaining outcomes that are robust to variations in the procedure. Then the cooperative approach, without having to model the details of the mediator selection process, can ably make some kind of prediction of where the negotiations and bargaining might reasonably lead to. Thus the cooperative approach may be better suited than the noncooperative approach to analyzing the problem of mediator selection by the parties who have private information.

At some level, the potentially endless variation in the bargaining process for selecting a mediator should in principle be describable. If we do a careful job of modeling a noncooperative mediator selection game, then we can take a Bayesian Nash equilibrium of the game. This equilibrium should ultimately be an equilibrium mapping from some distributions of the players' types to some distributions over the mediators. Then, for any outcome achievable via any equilibrium of any bargaining procedure under some noncooperative game, there should be an equivalent "grand" mediator that accomplishes the same outcome and is incentive feasible.[20] Thus there is no loss of generality in assuming that all types of all players would agree to choose the same grand mediator without sharing any information during the bargaining process. This idea of rolling any information leakage that could occur in any process of mediator selection into the grand mediator is captured by the *inscrutability principle* (Myerson, 1983).[21]

[18] Attempts to formalize notions of "reasonableness" were explored in the seminal work of Myerson (1983, 1984*a*,*b*).

[19] For example, in international relations, the exact protocols of mediator selection stage are not defined; it is unclear how negotiations are conducted, who makes an offer to whom, how the order of making offers or counter-offers is determined, and so on.

[20] The revelation principle implies that the set of all equilibria under some noncooperative game – once well-defined – is attainable as the set of all grand mediators. Moreover, for any incentive feasible grand mediator for some bargaining problem, there is a noncooperative game that generates the grand mediator as an equilibrium.

[21] I thank Roger Myerson for his comments on connecting the inscrutability principle to a mediator selection game. See Myerson (1984, 463) and Myerson (1991, 504-505) for more exposition of the in-

Thus in light of Myerson (1983), I can take the underlying Bayesian bargaining problem and the existence of incentive feasible mediators as primitives; and, without referring to a specific game form, impose some desiderata for a reasonable mediator choice that the privately informed players would most likely inscrutably agree on. If the incentive feasible mediator that is best for each player depends on what his type is, then no matter what type each player might be he cannot choose the incentive feasible mediator that is best for him unless the other player believes that both types would have selected the same mediator. Therefore, a player must make some sort of compromise between what he really wants and what he might have wanted if his type had been different. Due to the conflicting incentives of different possible types of the same player, there might be some interim incentive efficient mediators that are not expected to be chosen by the players in the bargaining process. The concept of Myerson (1984*b*)'s neutral bargaining solution exactly captures the idea of this compromise between the different incentives of different types, as well as between the players.

4.2 The Neutral Bargaining Solution

Myerson (1984*b*) develops a generalization of the Nash bargaining solution for two-person bargaining problems with incomplete information from a simple set of axioms that takes into account the inscrutable intertype compromise. "A *neutral bargaining solution* is defined to be any mechanism that is a solution for *every* bargaining solution concept that satisfies [the two axioms: an extension axiom and a random-dictatorship axiom]" (Myerson, 1991, 518).[22] I omit detailed expositions of these axioms, which can be found either in Myerson (1984*b*) or Myerson (1991, 516-517). More importantly, letting the "bargaining solution concept" be the set of all symmetric IIE mediators would satisfy both axioms. Thus I can think of refining the set of neutral bargaining solutions within the set of all symmetric IIE mediators.

Let $NS(\Gamma^*)$ denote the set of all neutral bargaining solutions of Γ^*. To characterize $NS(\Gamma^*)$, every $\mu_y \in S(\Gamma^*)$ has to be checked whether it is a solution for every bargaining solution concept that satisfies the two axioms. This search is simplified by reproducing the characterization theorems in Myerson (1984*b*) that are generated from the two axioms, suitably modified to fit into my class of bargaining models. This result, stated as Theorem C.1 in Online Appendix C, offers the most tractable set of conditions for computing neutral bargaining solutions. In fact, only one mediator in $S(\Gamma^*)$ satisfies such conditions. This result is formally stated in Theorem 1, and Proposition 4 fully characterizes $NS(\Gamma^*)$ that will be the bargaining solution in $(\Gamma^*, F(\Gamma^*))$.[23]

Theorem 1. *For any two-person Bayesian bargaining problem* Γ^**, the neutral bargaining solution is unique.*

---

scrutability principle.

[22]The neutral bargaining solutions form the smallest set satisfying two axiom. Myerson (1984*b*) proves that the set of neutral bargaining solutions is nonempty for any finite two-player Bayesian bargaining problem.

[23]I relegate the proofs for these results to Online Appendix C due to mere technicality involved.

Proposition 4. *$NS(\Gamma^*) = \{\mu_1\}$, where $z = 0$ if $p < p^*$ and $z = \bar{z}(p)$ if $p \in [p^*, p^{**})$.*

Theorem 1 asserts that, for the class of environments in my framework, the concept of the neutral bargaining solution gives a unique prediction to which mediator is chosen inscrutably in $(\Gamma^*, F(\Gamma^*))$. Proposition 4 implies that when the probability of the strong type is sufficiently small, i.e., $p < p^{**}$, the unique ex ante incentive efficient mediator who puts the lowest probability on war is not in the solution set refined by requiring the axioms of the neutral bargaining solution. Instead, the interim neutral bargaining solution selects the "worst" mediator among all IIE mediators that treat the players symmetrically. Therefore, my cooperative approach ably makes a unique prediction of which mediator should reasonably arise as an outcome of the mediator selection process: the one who is associated with the highest probability of disagreement. Online Appendix E gives detailed explanations regarding the underlying properties of the neutral bargaining solutions characterized in Proposition 4 in relation to the incentive constraints.

4.3 Failure of Ex Ante Efficiency

Before stating the key result of this paper, I examine the intuition under how the issue of information leakage (indirectly) impacts the endogenous selection of such mediator in the cooperative approach. The inscrutability principle implies that the players with private information bargain over mediators in an inscrutable way, without sharing any information during the mediator selection process. If a player bargains in a scrutable way, then the other player might learn something about him that could be detrimental. A reasonable interim bargaining solution concept should then be a non-revealing solution by itself.

However, the players might somehow be constrained by the fact that they could reveal their types. In my cooperative approach, every mediator that were being excluded from the set of neutral bargaining solutions is excluded by the logic that took into account the possibility of revealing information during the bargaining process. That is, there might be some interim incentive efficient mediators that are not expected to be selected by the players in such a process precisely because some player might choose to reveal information about his type instead of letting some mediators be selected.

In the class of models that I consider, expressing a preference for the ex ante incentive efficient mediator, who is known to be the best at implementing an agreement, might convey information that such player is in a weaker position. The strong player would immediately be convinced to force a disagreement outcome when matched with such weak player. Therefore, each player – whether he is strong or weak – would not want the other player to infer via his mediator choice that he is weak, even if the probability of the strong type is fairly small. In a sense, the strong player is very eager to reveal its type, whereas the weak player wants to conceal its type in bargaining; and so, to maintain inscrutability, the weak player would have to mimic the strong type.[24] Thus each player, being afraid

[24]Even when both players are weak, they act as though they know their types and are strong; and lean towards what they would do if they were strong, which is choosing the mediator that is better for the strong type, to avoid disclosing that they are weak.

of seeming weak in the mediator selection process, would choose the mediator that favors the strong type, never revealing their types during bargaining.

This intuition implies that the inscrutable intertype compromise between the two types gets resolved in favor of the strong type. Even if the probability of the strong type is rather small, the strong type would implicitly be more influential on the players' behavior in the bargaining process. What is best for the strong type is the farthest away from ex ante efficiency among all symmetric IIE mediators. The following theorem illustrates the key result of this paper.

Theorem 2. *For any two-person Bayesian bargaining problem Γ^*, the unique neutral bargaining solution is ex ante Pareto inferior to any other mediator in $S(\Gamma^*)$.*

The proof is immediate from Proposition 4 and Corollary 1. My cooperative approach using the neutral bargaining solution yields the conclusion that the players endogenously choose the most ex ante inefficient mediator. Such mediator is associated with the highest probability of disagreement, and so over-implements the disagreement payments.[25] An important implication of Theorem 2 is that the selection of a mediator can be twisted toward ex ante inefficiency by the incentive of each player to avoid seeming weak to their adversary. That is, in a model where ex ante efficient outcomes are technically feasible, the players do not choose the mediator that maximizes the ex ante efficient gains, and so they systematically do worse than ex ante efficiency. The class of benchmark models in this paper is well-identified for showing clearly how ex ante efficiency can go wrong in interim bargaining. Moreover, the uniqueness of the chosen mediator in such class further justifies the neutral bargaining solution as a powerful interim bargaining solution concept.

Another interesting implication of Theorem 2 concerns the sources of bargaining failure (i.e., disagreement). Much of the literature on mediation ascribes the sources of bargaining failure to particular "exogenous" attributes of mediation.[26] In contrast, the key insight of my paper comes from noticing the fundamental nature of bargaining itself – that mediators are actually chosen endogenously by the disputing parties themselves. Then my cooperative theory of how privately informed parties bargain over mediators ably offers another yet novel explanation for the sources of inefficiency in bargaining: the bargaining process itself. Therefore, I argue that the very process of selecting a mediator may exhibit an inherent inefficiency in interim bargaining.

Before I conclude with this conclusion, I should examine its logical foundations more closely by carefully analyzing the mediator selection process as a noncooperative game and

[25]This result evokes Myerson and Satterthwaite (1983) in which the status quo disagreement payments are likewise over-implemented.

[26]For example, Kydd (2003) argues that for mediation to be effective, the mediator must be biased and endowed with some independent knowledge on the private information of the disputants. Fey and Ramsay (2010) show that 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. Other sources that could affect the risk of disagreement include particular information structures (e.g., the types of uncertainty states face, or whether asymmetric information is significant or not), the intensity of conflict between the disputants, whether a mediator has a strong enforcement power, reputation issues, etc.

describing what the rational players should actually do. Despite the analytical power of the cooperative approach in predicting the outcome of bargaining among the players, one might wonder the exact conduit for which informational concerns lead to the mediator that favors the strong type. In the next section, I further attempt to build a noncooperative theory of mediator selection and formalize a solution concept by analyzing a specific two-stage game of the mediator selection process under the noncooperative perspective.

5 Bargaining over Mediators: Noncooperative Approach

The essential idea to be developed in this section is that a mediator μ_{y^*} should be considered *threat-secure* iff the players would always unanimously choose (or vote for) μ_{y^*} among many available mediators. In order to formulate this idea, I want to establish whether the players would ever disagree on their votes, where beliefs following non-matching votes are required to satisfy consistency conditions.

Let us assume that at the start of the interaction players are faced with a continuum of candidate mediators in $S(\Gamma^*)$. As noted in Remark 2, I restrict attention to when $p < p^{**}$. For any given $p < p^{**}$, every mediator in $S(\Gamma^*)$ can be completely identified by $y \in [y_{min}(\Gamma^*), 1]$ alone, where $y_{min}(\Gamma^*) \geq 0$ depends on the primitives of Γ^*. Hence, let $Y = [y_{min}(\Gamma^*), 1]$ represent the continuum of candidate mediators. If the players do not agree on a mediator, then they would play a default game – the Bayesian game without communication that underlies the class of Bayesian bargaining problems considered in this paper. I relegate the details of this game to Online Appendix D; in this section, let the default game be denoted by G with finite action space. The set of possible actions consists of "war" (disagreement) and "peace" (agreement) for both players.

Let us consider the following two-stage *ratification game*: In the first stage, the players each simultaneously casts one vote to a mediator in the set of candidate mediators Y. A strategy for i in the first stage specifies the vote if his type were t_i, i.e., $v_i : T_i \to Y$. Let $v_i(t_i)$ denote the voting strategy for player i of type t_i. An outcome to the first stage indicates the *vote outcome* $\kappa = (\kappa_i)_{i\in\{1,2\}}$ where $\kappa_i \in Y$ denotes player i's voting decision. For example, $\kappa = (\kappa_1, \kappa_2) = (0.5, 0.7)$ if player 1 votes for $\mu_{0.5}$ and player 2 votes for $\mu_{0.7}$. After the first stage, players observe each other's voting decision.[27] In the second stage, a mediator μ_{y^*} is implemented if the vote outcome is a vector composed of y^*; otherwise, the default game is played noncooperatively with the observation of the realized vector κ. In such a ratification game, I want to know whether there exists an equilibrium of voting strategies such that the votes always match; that is, the players unanimously vote for the same mediator in all information states.

A player's optimal voting strategy in this ratification game should depend on how he would expect the default game to be played and what he would believe about the other player's type, if their votes did not match. Thus I shall need the following notation.

[27]Players are not told the details of the vote in Holmström and Myerson (1983) because they assume a secret ballot, whereas Cramton and Palfrey (1995) and Celik and Peters (2011) assume a non-anonymous voting procedure. The latter assumption is more appealing in my setting where one type of a player might have an incentive to publicly announce displeasure with a particular mediator because of the information the announcement conveys when the players are negotiating.

Let K denote the set of vote outcomes under which the votes do not match, i.e., $K = \{(\kappa_1, \kappa_2)|\kappa_1 \neq \kappa_2\}$. Upon observing the non-matched vote outcome $\kappa \in K$, the players revert to playing the default game noncooperatively under updated beliefs on each other.[28] For any player i and any type t_i, let $\psi_i^\kappa(t_i)$ denote the probability that i would choose disagreement if t_i were his type and G is played given the knowledge $\kappa \in K$. For all $t_i \in T_i$, $\bar{q}_{i,\kappa_i}(t_i)$ represents the belief probability that player $-i$ would assign to the event that t_i is the other player's type if $\kappa_i \neq \kappa_{-i}$ is observed. Let $\bar{q}^\kappa = (\bar{q}_{i,\kappa_i})_{i\in\{1,2\}}$ be a vector of updated beliefs conditional on the event $\kappa = (\kappa_i)_{i\in\{1,2\}} \in K$.

From these definitions, the quantities $(v, \psi^\kappa, \bar{q}^\kappa)$ for any given $\kappa \in K$ must be non-negative and must satisfy:

$$v_i(t_i) \in [y_{min}(\Gamma^*), 1],\ \psi_i^\kappa(t_i) \leq 1,\ \sum_{t_i} \bar{q}_{i,\kappa_i}(t_i) = 1,\ \forall i,\ \forall t_i. \tag{5.1}$$

To show that μ_{y^*} is threat-secure, I want to show that there is a Nash equilibrium of this ratification game in which the votes always match to be y^* and the voted mediator μ_{y^*} is played honestly. The votes match to be y^* in all information states if and only if:

$$v_1(t_1) = v_2(t_2) = y^*,\ \forall t \in T. \tag{5.2}$$

For each vote outcome $\kappa \in K$ leading to the noncooperative play of G, let $\Sigma(\bar{q}^\kappa)$ denote a Nash equilibrium in the continuation game of G under the belief system $\bar{q}^\kappa$. Then, for any given $\kappa \in K$, the strategies $(\psi_i^\kappa)_{i\in\{1,2\}}$ which the players would use in G form an equilibrium $\Sigma(\bar{q}^\kappa)$ of G with respect to the posterior beliefs $(\bar{q}_{i,\kappa_i})_{i\in\{1,2\}}$ if and only if:

$$\begin{aligned} U_i(G|t_i, \Sigma(\bar{q}^\kappa)) &\equiv \sum_{t_{-i}\in T_{-i}} \bar{q}_{-i,\kappa_{-i}}(t_{-i})(1-\psi_i^\kappa(t_i))(1-\psi_{-i}^\kappa(t_{-i}))u_i(d_1,t) \\ \geq U_i(G, \hat{\psi}_i|t_i, \Sigma(\bar{q}^\kappa)) &\equiv \sum_{t_{-i}\in T_{-i}} \bar{q}_{-i,\kappa_{-i}}(t_{-i})(1-\hat{\psi}_i)(1-\psi_{-i}^\kappa(t_{-i}))u_i(d_1,t), \\ &\qquad \forall i,\ \forall t_i \in T_i,\ \forall \hat{\psi}_i \in [0,1]. \end{aligned} \tag{5.3}$$

Condition (5.3) asserts that player i with type t_i should not expect any other strategy $\hat{\psi}_i$ to be better for him in G than the strategy selected by his ψ_i^κ when G is played following $\kappa \in K$ and when player $-i$ is expected to use her equilibrium strategy ψ_{-i}^κ.

If (5.2) and (5.3) hold, then the voting strategies $(v_i)_{i\in\{1,2\}}$ and the continuation strategies $(\psi_i^\kappa)_{i\in\{1,2\}}$ in G together with honest behavior in μ_{y^*} form a Nash equilibrium of the ratification game if and only if:

$$\begin{aligned} \sum_{t_{-i}\in T_{-i}} \sum_{d\in D} \bar{p}_{-i}(t_{-i})\mu_{y^*}(d|t)u_i(d,t) \geq \sum_{t_{-i}\in T_{-i}} \sum_{d\in D} \bar{p}_{-i}(t_{-i})\mu_{y^*}(d|t_{-i},\hat{t}_i)u_i(d,t), \\ \forall i,\ \forall t_i \in T_i,\ \forall \hat{t}_i \in T_i; \end{aligned} \tag{5.4}$$

[28] In this game, an off-the-equilibrium-path entails a non-matched vote outcome, not a unilateral deviation by one player. Therefore both players must form updated beliefs if K occurs.

and

$$\sum_{t_{-i}\in T_{-i}}\sum_{d\in D}\bar{p}_{-i}(t_{-i})\mu_{y^*}(d|t)u_i(d,t)$$
$$\geq \sum_{t_{-i}\in T_{-i}}\bar{p}_{-i}(t_{-i})(1-\psi_i^{(\kappa_i,y^*)}(t_i))(1-\psi_{-i}^{(\kappa_i,y^*)}(t_{-i}))u_i(d_1,t), \tag{5.5}$$
$$\forall \kappa_i \in Y\backslash\{y^*\},\ \forall i,\ \forall t_i \in T_i.$$

Since μ_{y^*} is incentive compatible under the prior beliefs, (5.4) is satisfied on the equilibrium path – no player can expect to gain by lying in μ_{y^*} when it is unanimously voted for and implemented. Condition (5.5) asserts that player i cannot gain by casting a non-matching vote $\kappa_i \neq \kappa_{-i} = y^*$ when t_i is his true type, given player $-i$ voting strategy v_{-i} and the continuation strategies ψ^κ. If conditions (5.1) through (5.5) hold, then always unanimous vote for μ_{y^*} followed by truthful revelation to μ_{y^*} is a sequential equilibrium in the two-stage ratification game.

5.1 Credible Non-Matching Beliefs

As long as some player j different from i is expected to always vote for $y^{'}$ (so that $v_j \equiv y^{'}$), then there is always a sequential equilibrium in which player i as well always vote for $y^{'}$. Then I would get the extreme result that every symmetric IIE mediator would be threat-secure. However I need to take into account the possibility of learning from the vote outcome. Because I focus on each player's decision to match his vote to the other player, a non-matched vote outcome may reveal information to the players. This observation suggests that I should test the robustness of equilibria, $v_i \equiv y^{'}$ for all i, to credible non-matched vote outcomes.

Note that voting decisions convey information only when these decisions are different, i.e., $\kappa_1 \neq \kappa_2$. K denotes the set of these events that the players' votes do not match. Therefore, I must impose some restrictions on beliefs in the ratification game when $\kappa \in K$ occurs. That is, if a non-matched vote outcome $\kappa \in K$ occurs, then each player i's non-matching voting decision $\kappa_i \neq \kappa_{-i}$ should be rationalized by identifying a posterior belief $\bar{q}_{i,\kappa_i}$ that is consistent with i's incentive to cast a vote κ_i that is different from κ_{-i} . The posterior beliefs $(\bar{q}_{i,\kappa_i})_{i\in\{1,2\},\kappa\in K}$ are induced from the priors $\bar{p}$ and the voting strategies v. Based on the refinement of perfect sequential equilibria proposed by Grossman and Perry (1986), let us first define a credible non-matching belief system:

Definition 1. Conditional on the event $\kappa \in K$, a belief system $\bar{q}^\kappa = (\bar{q}_{1,\kappa_1}, \bar{q}_{2,\kappa_2})$ on T is a *credible non-matching belief system* that supports κ if there exist a noncooperative equilibrium $\Sigma(\bar{q}^\kappa)$ of G and voting strategies $v = (v_i)_{i\in\{1,2\}}$ such that together satisfy:

1. $v_i(t_i) \neq \kappa_{-i}$ with positive probability for some $t_i \in T_i$ of any i;
2. $v_i(t_i) = \kappa_i$ with probability one for all $t_i \in T_i$ of any i such that

$$\sum_{t_{-i}\in T_{-i}}\bar{q}_{-i,\kappa_{-i}}(t_{-i})\sum_{d\in D}\mu_{\kappa_{-i}}(d|t)u_i(d,t) < U_i(G|t_i,\Sigma(\bar{q}^\kappa));$$

3. $v_i(t_i) = \kappa_{-i}$ with probability one for all $t_i \in T_i$ of any i such that

$$\sum_{t_{-i} \in T_{-i}} \bar{q}_{-i,\kappa_{-i}}(t_{-i}) \sum_{d \in D} \mu_{\kappa_{-i}}(d|t) u_i(d,t) > U_i(G|t_i, \Sigma(\bar{q}^{\kappa}));$$

4. The posterior beliefs $(\bar{q}_{i,\kappa_i})_{i \in \{1,2\}}$ satisfy Bayes theorem, given the priors $\bar{p}$ and the voting strategies v:

$$\bar{q}_{i,\kappa_i}(t_i) = \frac{\bar{p}_i(t_i) \cdot \mathbf{1}_i(t_i)}{\sum_{\hat{t}_i \in T_i} \bar{p}_i(\hat{t}_i) \cdot \mathbf{1}_i(\hat{t}_i)}$$

where $\mathbf{1}_i : T_i \to \{0,1\}$ is an indicator function defined as, for all i,

$$\mathbf{1}_i(t_i) := \begin{cases} 1 & \text{for } t_i \text{ such that } v_i(t_i) \neq \kappa_{-i} \text{ with positive probability,} \\ 0 & \text{for } t_i \text{ such that } v_i(t_i) = \kappa_{-i} \text{ with probability one.} \end{cases}$$

Condition (1) states that for each player i there is a set of types that cast non-matching votes with positive probability. If so, condition (4) then implies that Bayes theorem should be used to compute $\bar{q}^{\kappa}$. Conditions (2) and (3) require that if, conditional on the event $\kappa \in K$, player i with type t_i would get higher expected utility by casting a non-matching vote (and playing G) than casting a matching vote, then player i with type t_i must cast a non-matching vote with probability one; and if player i with type t_i loses from casting a non-matching vote, then player i with type t_i must cast a matching vote with probability one. Thus if player i has a posterior $\bar{q}_{-i,\kappa_{-i}}$ after observing $\kappa \in K$, then his best response ψ_i^{κ} will lead all types of $-i$ that voted differently to be better off than by voting the same as i and those types who are indifferent "participate in voting differently" with a probability that leads player i's posterior to be $\bar{q}_{-i,\kappa_{-i}}$. In other words, if $\bar{q}^{\kappa}$ is a credible non-matching belief system, then each player i can rationalize $-i$'s non-matching voting decision by believing $-i$'s type is distributed according to $\bar{q}_{-i,\kappa_{-i}}$.

These restrictions essentially ask whether each player, upon voting differently, can induce the other player to reason that the non-matching vote must have come from the subset of types who might possibly gain from the non-matched vote outcome. A pair of such set of types for each player breaks an equilibrium with out-of-equilibrium "messages" if all types in such set for each player improve their payoffs by voting differently as long as the other player's beliefs put all weight on this particular subset. For an equilibrium to survive the credibility refinement, either **(i)** any possible non-matched vote outcome must not be supported by a credible non-matching belief system, or **(ii)** "credibly" revealed non-matching types must be indifferent between matching and non-matching. If there is some types that strictly gain from the non-matched vote outcome, then an equilibrium is not robust in the sense that there exists voting strategies that are supported by a credible belief system. In the next subsection, I characterize a unique sequential equilibrium that survives such refinement.

5.2 Threat-Secure Mediators

If μ_{y^*} is threat-secure then there is a rational voting equilibrium in which the players always unanimously vote for μ_{y^*}, where beliefs following non-matched votes are required to satisfy consistency conditions. Let K_{y^*} denote the set of non-matched vote outcomes such that one player's voting decision is y^* while the other player's voting decision is different from y^*, i.e., $K_{y^*} = \{\kappa \in K | \kappa_i = y^* \text{ for some } i\}$. The definition of threat-security is formally stated as follows.[29]

Definition 2. A mediator μ_{y^*} is *threat-secure* if and only if there exists v, $\bar{q}^\kappa$, and $\Sigma(\bar{q}^\kappa)$ for all $\kappa \in K_{y^*}$ such that the conditions (5.1) through (5.5) are all satisfied; and either

(i) for any $\kappa \in K_{y^*}$, there does not exist a credible non-matching belief system; or

(ii) for every credible non-matching belief system $\bar{q}^\kappa$ associated with $\kappa \in K_{y^*}$, the corresponding equilibrium $\Sigma(\bar{q}^\kappa)$ satisfies:

$$\sum_{t_{-i} \in T_{-i}} \bar{q}_{-i,y^*}(t_{-i}) \sum_{d \in D} \mu_{y^*}(d|t) u_i(d,t) = U_i(G|t_i, \Sigma(\bar{q}^\kappa)),$$

for all t_i such that $v_i(t_i) \neq \kappa_{-i}$ with positive probability, for i such that $\kappa_i \neq y^*$.

Threat-security captures the idea that no type of any player is willing to credibly cast a non-matching vote against the play of μ_{y^*}. In other words, if μ_{y^*} is threat-secure, then there does not exist a credible non-matching belief system for any possible non-matched vote outcome such that some type in that non-matching belief strictly prefers the default game to μ_{y^*}. To establish that a mediator is threat-secure, every pair of non-matching voting decisions needs to be checked whether those decisions can be rationalized. This search is simplified by establishing the following lemma.

Lemma 5. *If $y^* < 1$, then for some $\kappa \in K_{y^*}$ there exists a credible non-matching belief system $\bar{q}^\kappa$ and a corresponding equilibrium $\Gamma(\bar{q}^\kappa)$ (that do not satisfy part (ii) of Definition 2).*

Lemma 5 states that when $y^* < 1$, there is some $\kappa \in K_{y^*}$ such that after observing such κ, the players' beliefs are restricted to a credible non-matching belief system that can rationalize κ. This result implies that there are no equilibria in which the players match their votes to be $y^* < 1$ in all information states. The idea behind the proof is straightforward. Suppose that after the first stage the realized vector of voting decisions were $(y^*, 1)$, where $y^* < 1$. Noticing Lemma 4, intuition suggests that player 1 will be suspected of being the weak type. Such an inference from the non-degenerate voting decisions will lead player 2 of the strong type – believing player 1 is weak – to force disagreement much more

[29] The definition of threat-security resembles that of strong ratifiability in Cramton and Palfrey (1995); however, the structure of the game and the underlying updating rule are fundamentally different from theirs.

in the default game than it would under its original priors; thereby making it profitable for player 2 of the strong type to cast a non-matching vote while destroying any benefit player 1 of the weak type could get from casting a non-matching vote. In fact the weak types of all players would be indifferent between matching and not matching their votes to the other player. Note that there is no restriction on v_i for types that are indifferent, reflecting the idea that such types may randomly choose whether to match or not; thus matching or non-matching by such types can be rationalized. Regardless, the strong type of player 2 strictly benefits from voting differently than y^* while the weak type of player 2 participates in voting differently than y^* with a probability that together lead player 2's non-matching voting decision be rationalized by player 1's posterior beliefs that remain the same as interim beliefs. Also the strong type of player 1 strictly benefits from voting for "1," i.e. matching its vote to player 2, while the weak type of player 1 participates in voting differently than "1" with a probability that together lead player 1's non-matching voting decision be rationalized by player 2's beliefs putting all weights on the weak type. Each player anticipates what the other player would think (and do) upon observing the non-matched vote outcome, in turn justifying the non-matched vote outcome $(y^*, 1)$. Thus always unanimous vote on $y^* < 1$ would be vulnerable to the credibility refinement because the "credibly" revealed strong type of the player whose vote did not match would strictly prefer non-matching to matching.

Now, I can establish the existence of the symmetric IIE and threat-secure mediator in the class of benchmark bargaining models discussed in this paper. The subsequent proposition characterizes the set of threat-secure mediators on the interim Pareto frontier, denoted as $TS(\Gamma^*)$.

Theorem 3. *For any two-person Bayesian bargaining problem Γ^*, there exists a unique threat-secure and symmetric interim incentive efficient mediator.*

Proposition 5. *$TS(\Gamma^*) = \{\mu_1\}$, where $z = 0$ if $p < p^*$ and $z = \bar{z}(p)$ if $p \in [p^*, p^{**})$.*

Proposition 5 states that a threat-secure mediator is uniquely defined in the set $TS(\Gamma^*)$. If $p < p^*$, then the threat-secure mediator chooses disagreement with probability one when two players are of different types and never chooses disagreement when they are the same type. If $p \in [p^*, p^{**})$, then the threat-secure mediator chooses disagreement with probability one when two players are of different types and with a positive probability when both players are the strong type.

The results suggest that no other mediator than the "worst" one will be always unanimously chosen in the ratification game. The intuition continues from the idea behind Lemma 5. Consider any $\kappa \in K_1$. Suppose that, without loss of generality, $\kappa = (\kappa_1, 1)$ for any $\kappa_1 < 1$. We need to check whether there is a credible non-matching belief system such that if, after observing κ, the players follow the play of G prescribed by evaluating their noncooperative equilibrium strategies at their new beliefs, then only the types of players that are believed to have voted differently with positive probability could possibly gain by voting differently. In fact there is a unique pair of such types that could possibly

gain by voting differently when the other player reacts as if only such types could be voting differently with positive probability: $\{w\}$ for player 1 and $\{s, w\}$ for player 2. Such inference will lead the strong types to force disagreement much more in G, which makes casting any vote $\kappa_1 < 1$ unprofitable and thus makes all types of any player (weakly) prefer casting a vote on "1" to the default game. The weak types of any player earn zero profits in either case (so that the "credibly" revealed weak type of player 1 is indifferent between matching and non-matching), and so the players' beliefs are restricted to the unique credible non-matching belief system that satisfies part (ii) of Definition 2.[30] Thus the voting strategies such that $v_i(t_i) = 1$ for all t_i, for all i, would be sustainable as an equilibrium in which the credibility restrictions are satisfied. These criteria reject equilibria consisted of unanimous vote on $y^* < 1$ in which the players are "threatened" by a credible non-matching belief system.

The conduits that lead to this unique equilibrium can be described as follows. Upon observing non-degenerate voting decisions, it is credible for a player to infer that the other player is weak as a result of the other player's lower vote, which in turn makes casting a lower vote unprofitable. In contrast, a player can send a credible signal that she is of the strong type by casting a higher vote, and benefit from subsequently playing the default game. That is, a player, if his type is weak, would (weakly) lose from casting a lower vote than the other player's; on the other hand, a player, if her type is strong, would benefit from casting a higher vote than the other player's. Therefore, the weak type tends to match the vote with the strong type, whereas the strong type tends to not match the vote with the weak type. In other words, the weak types constantly pursue the behaviors of the strong types and the strong types repeatedly avoid captures by voting higher. Such behaviors are reminiscent of "cat-and-mouse games," except that this "contest" ends when the strong types "escape" to casting a vote on its most preferred mediator, in which case the "cat" (weak type) has actually not won a victory over the "mouse" (strong type) despite its successful "capture" (matching vote). As a result, the players would always unanimously choose the most ex ante inefficient mediator in the two-stage ratification game described. The suboptimal choice of such mediator is associated with the highest probability of disagreement (or bargaining failure) among all symmetric IIE mediators.[31]

[30] Of note is that, in my setting, there is at most one credible non-matching belief system for each event $\kappa \in K$. That is, when $\kappa \in K$ occurs, there is a unique pair of set of types for each player who would be credibly identified by casting a non-matching vote with positive probability. Then there is only one way to rationalize each non-matched vote outcome.

[31] It is worth comparing the paper's relationship to other papers on noncooperative bargaining models of "implementable" mechanisms in terms of the predictions about the choice of a mediator. Holmström and Myerson (1983) consider *durability* that formalizes the idea of a mechanism being invulnerable to proposals of alternative mechanisms in a pairwise comparison. A similar idea has been discussed under the name *resilient allocation rule* (Lagunoff, 1995) in a buyer-seller bargaining problem. The concept of *ratifiability* in Cramton and Palfrey (1995) is a mirror image of durability in the sense that it specifies what alternative mechanisms can be unanimously approved against a status quo mechanism. Laffont and Martimort (2000) show that the optimal collusion-proof mechanism is strongly ratifiable in the sense of Cramton and Palfrey (1995). While Holmström and Myerson (1983), Lagunoff (1995), Cramton and Palfrey (1995), and Laffont and Martimort (2000) focus on refinements of off the equilibrium path beliefs in their solution concepts to find which mechanism is feasible, Celik and Peters (2011) study a larger class of equilibria and characterize a condition under which all the implementable allocation rules are truthfully

6 Noncooperative Foundations of the Cooperative Approach

In this section, I discuss the connection between the solution set refined by the cooperative approach and the equilibrium choice in the noncooperative approach. Propositions 4 and 5 are significant because they imply the following core result of this paper.

Theorem 4. *For any two-person Bayesian bargaining problem* Γ^*, $TS(\Gamma^*) = NS(\Gamma^*) \subset S(\Gamma^*)$.

Theorem 4 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 predictions of cooperative game theory and noncooperative game theory are not nested, but in the benchmark class of examples considered in this paper, they coincide. This result begs the question of why the threat-security generates the same outcome as the neutral bargaining solution even though they stem from such different approaches.

Because I have modeled the cooperative bargaining theory of mediator selection only as a particular noncooperative game in extensive form, the above equivalence theorem cannot now be posed as a general theorem that the solution concepts from two approaches are equivalent. Nonetheless, one way to lay the noncooperative foundations for the solution that arises under the cooperative approach is to contrast the reasoning for the information leakage problem behind the cooperative solution with the strategic incentives behind the noncooperative solution.

In the ratification game, each player can send a signal that he is of a certain type indirectly through his strategy, e.g., voting for a particular mediator. In particular, the strong type benefits from revealing her type by voting differently while there is a disadvantage to the weak type if he gives away his identity. That is, the strong type can send a credible signal whereas the weak type cannot. Then only the equilibrium in which every player strategically pools on the strong action – voting for the "worst" mediator – survives the credibility restrictions on beliefs.

Under my cooperative approach, when two privately informed players bargain over the choice of a mediator, the inscrutability and the intertype compromise together lead to the unique neutral bargaining solution that implicitly puts more weight on the strong type. This notable distortion inherent in the neutral bargaining solution concept has a *signaling* component: Every player "pretends" to be strong and "picks" in a cooperative sense the mediator that is favorable to the strong type – inscrutably agreeing on the "worst" mediator; but this is effectively asking every player to pool in a way such that no information is revealed. The nature of this pooling is exactly reminiscent of a pooling equilibrium of a signaling outcome in the noncooperative game. It is as if we built the

implementable. Note that directly imposing the concept of durability or security (non-ratifiability) does not rule out any symmetric IIE mediators for the benchmark class of examples that I study in this paper; not only is the neutral bargaining solution durable and secure, but in fact all of the symmetric IIE mediators are. Thus those concepts would not give any stronger prediction over the notion of interim incentive efficiency in my framework.

signaling distortion in the noncooperative game directly into the cooperative mathematics. In this sense, my noncooperative approach can be interpreted as being the justification for the neutral bargaining solution within the benchmark class.

Table 1: Connecting the intuitions of "information leakage"

Cooperative approach	Noncooperative approach
Inscrutability principle	Pooling equilibrium
Intertype compromise	Signaling distortion
(Lagrange multipliers)	(credible beliefs)
⇓	⇓
Neutral bargaining solution	Threat-secure mediator

Both approaches take into account, either directly or indirectly, the information leakage problem, and prescribe the same unique solution. Thus the noncooperative foundations of the cooperative approach rest on the following conjecture that I call the *equivalence hypothesis*:

> The cooperative outcomes of mediator selection games with incomplete information in which the players bargain inscrutably and make intertype compromise should be the same as the equilibrium outcomes of signaling games with incomplete information in which the players credibly signal their types indirectly via their choices of mediators.

The informal justification comes from the fact that the strategic intuition to why the most ex ante inefficient mediator is the only threat-secure mediator underscores the intuition behind the information leakage problem in the cooperative bargaining theory with incomplete information. The equivalence hypothesis allows us to carry the intuitions behind one approach over the other. This hypothesis is an assertion about the relation between cooperative and noncooperative solution concepts rather than a formal theorem about any broader equivalence, which I have not yet been able to prove but hope to rigorously formalize in future research.

7 Concluding Comments

In this paper, I build a theory of how players with private information might agree on a mediator. The general tenor of the results suggest that the two players in a mediator selection game endogenously choose the ex ante worst mediator, and the information leakage problem is part of the answer as to how the mediator is chosen. My key insight is that, when players advocate for a particular mediator, they reveal information about their types. In the class of benchmark bargaining models, in which the agreement outcome is ex ante efficient but the strong player always prefers the disagreement outcome when matched with a weak player, players are afraid of seeming weak to their adversary in the mediator selection process. These incentives imply that the ex ante incentive efficient

mediator, which is most attractive to the weak player, will not be chosen. Further, both concepts of the interim neutral bargaining solution and threat-security select the ex ante worst mediator among all interim incentive efficient mediators that treat the players symmetrically. The suboptimal choice of such mediator is associated with the highest probability of disagreement. The result suggests a novel idea to the long-held debate over the sources of bargaining failure: that the endogenous choice of a mediator inherently leads to a higher risk of disagreement (i.e., bargaining failure) and to inefficiency in bargaining if the players already know their own types.

There are a number of substantial issues that require some discussion. I conclude this paper by mentioning a few that I hope will be studied in future work. Although I have described my model in the context of international conflicts, the model is equally applicable to other forms of bargaining games, such as collective bargaining between firms and unions, or mergers negotiations between two firms. Also, the framework I have developed here is particularly amenable to allow the use of side-payments by the players. However, because my interest is in understanding how informational incentives influence the players' endogenous selection of a mediator prior to their decisions over the bargaining outcomes, I have studied only bargaining games without transfers.

On another note, Myerson (1984*b*) establishes existence of neutral bargaining solutions for any class of two-person Bayesian bargaining problems, but there is no general uniqueness theorem. The present paper is a direct continuation on this work in the sense that I establish existence and uniqueness for a particular class. The challenge is to identify a larger class of examples that shows both uniqueness and ex ante inefficiency in interim bargaining.

Another important implication concerns a more general connection between the cooperative and noncooperative solution concepts. Under the equivalence hypothesis, I can only carry the strategic intuitions behind the noncooperative approach over the cooperative ideas of intertype compromise that arises in the neutral bargaining solutions. One can fully justify the neutral bargaining solutions, without having to model the details of the noncooperative mediator selection games, by finding the exact counterparts underlying the two approaches. In any case, it is my opinion that the equivalence hypothesis should be taken as only a first step in an attempt to build a bridge between cooperative and noncooperative game theories in the context of bargaining with incomplete information.

A Appendix A: Proofs for Section 3

By restricting analysis to symmetric mediators who treat the same type of all players symmetrically along with the symmetry assumption on payoffs, I can focus on $i = 1$ without loss of generality.

A.1 Proof of Lemma 1 (Thresholds).

It is easy to see that

$$\begin{aligned} pu_1(d_1, ss) + (1-p)u_1(d_1, sw) &< 0 \text{ if } p = 0; \\ &> 0 \text{ if } p = 1, \end{aligned}$$

because $u_1(d_1, ss) > 0$ and $u_1(d_1, sw) < 0$. Note that $pu_1(d_1, ss) + (1-p)u_1(d_1, sw)$ is a strictly increasing (continuous) function of p because $u_1(d_1, ss) - u_1(d_1, sw) > 0$. Therefore, by the single crossing property, there exist a unique $p^{'} \in (0,1)$ such that $p^{'}u_1(d_1, ss) + (1-p^{'})u_1(d_1, sw) = 0$, which gives

$$p^{'} = \frac{-u_1(d_1, sw)}{u_1(d_1, ss) - u_1(d_1, sw)} = \frac{-u_2(d_1, ws)}{u_2(d_1, ss) - u_2(d_1, ws)} \text{ (by } \textbf{(A3)}\text{)}.$$

Also, we have that

$$p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww) - [pu_1(d_1, ws) + (1-p)(1-y)u_1(d_1, ww)] \quad \text{(A.1)}$$

is a strictly decreasing (continuous) function of p for any $y \in (0,1]$, and

$$\begin{aligned} \text{(A.1)} &> 0 \text{ if } p = 0; \\ &< 0 \text{ if } p = 1. \end{aligned}$$

Again by the single crossing property, there exist a unique $p^* \in (0,1)$ such that (A.1) $= 0$ for any $y \in (0,1]$, that is,

$$p^*(1-y)u_1(d_1, ws) + (1-p^*)u_1(d_1, ww) = p^*u_1(d_1, ws) + (1-p^*)(1-y)u_1(d_1, ww), \quad \text{(A.2)}$$

which gives

$$p^* = \frac{u_1(d_1, ww)}{u_1(d_1, ww) + u_1(d_1, ws)} = \frac{u_2(d_1, ww)}{u_2(d_1, ww) + u_2(d_1, sw)} \text{ (by } \textbf{(A3)}\text{)}.$$

Now, let z be a function of $y \in (0,1]$ and $p \in (0,1)$ such that

$$z := z(y,p) = y\left[1 - \frac{(1-p)u_1(d_1, ww)}{pu_1(d_1, ws)}\right],$$

which is increasing in y (given p) and in p (given y). Then,

$$p(1 - z(y,p))u_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw) - [pu_1(d_1, ss) + (1-p)u_1(d_1, sw)] \quad \text{(A.3)}$$

is a strictly decreasing (continuous) function of p for any $p \geq p^*$ given any $y \in (0,1]$, and

$$\begin{aligned} \text{(A.3)} &> 0 \text{ if } p = p^*, \\ &< 0 \text{ if } p = 1. \end{aligned}$$

Thus by the single crossing property, there is a unique $p^{**} \in (p^*, 1)$ such that:

$$\begin{aligned} &p^{**}(1 - z(y, p^{**}))u_1(d_1, ss) + (1-p^{**})(1-y)u_1(d_1, sw) \\ &\quad = p^{**}u_1(d_1, ss) + (1-p^{**})u_1(d_1, sw), \ \forall y \in (0,1], \end{aligned}$$

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 also write p^{**} in terms of player 2's payoffs. The existence of p^{**} already implies that $p^{**} > p^*$. Lastly, $u_1(d_1, ss)u_1(d_1, ww) > 0$ implies $p^{**} > p^{'}$. For future reference,

$$\begin{aligned} &u_1(d_1, ss)u_1(d_1, ww) \geq -u_1(d_1, sw)u_1(d_1, ws) \quad \text{if and only if } p^* \geq p^{'}; \\ &u_1(d_1, ss)u_1(d_1, ww) < -u_1(d_1, sw)u_1(d_1, ws) \quad \text{if and only if } p^{'} > p^*. \ \square \end{aligned}$$

A.2 The Interpretations of the Three Thresholds and Incentive Feasibility

Before proceeding to show Propositions 1 and 2, the following discussions (with proof of Lemma 2) will be useful. In order to prevent confusion in the discussions and proofs hereafter, I use double subscripts on μ specifying both y and z, i.e., $\mu_{y,z}$.

First note that, from Lemma 1, the first threshold $p^{'}$ is such that $p^{'}u_1(d_1, ss) + (1 - p^{'})u_1(d_1, sw) = 0$. This condition is equivalent to

$$U_i(\mu_{0,0}|s)|_{p=p^{'}} = 0, \ \forall i;$$

that is, the participation constraints bind for the strong type given that the players are associated with mediator $\mu_{0,0}$. Then for any $p < p^{'}$, $U_i(\mu_{0,0}|s) < 0$, $\forall i$; i.e., $\mu_{0,0}$ is not individually rational to the strong types. Noticing $u_1(d_1, ss) > 0$ and $u_1(d_1, sw) < 0$, I can find, for each $p < p^{'}$, some $\underline{y}(p) > 0$ such that $pu_1(d_1, ss) + (1-p)(1-\underline{y}(p))u_1(d_1, sw) = 0$, which is equivalent to $U_i(\mu_{\underline{y}(p),0}|s) = 0$, $\forall i$. This implies that when $p < p^{'}$ a mediator who chooses war with any probability lower than $\underline{y}(p)$ for $t \in \{(s, w), (w, s)\}$ would not be incentive feasible; in particular, not individually rational for the strong type to participate with. The intuition is straightforward for p very near zero: If a player is actually strong and expects the other player to be weak almost surely, then the disagreement payoff (or the expected utility of "always war") for the strong player would be strictly higher than the expected utility that she could get through a mediator who chooses war with "not-so-high" probability, and thus she would rather not participate. To give the strong type an incentive to participate, any incentive feasible mediator must choose war with some positive probability of at least $\underline{y}(p)$ if $t \in \{(s, w), (w, s)\}$ when $p < p^{'}$. Thus $\underline{y}(p)$ for $p < p^{'}$ when $p^{'} \leq p^*$ (and similarly, $\underline{y}(p)$ for $p < p^*$ and $\underline{\underline{y}}(p)$ for $p \in [p^*, p^{'})$ when $p^* < p^{'}$) can be interpreted as the lower bound on the probability with which a mediator must put on war if $t \in \{(s, w), (w, s)\}$ in order to be incentive feasible when $p < p^{'}$; and $p^{'}$ can be interpreted as the cutoff for which a strong type player stops wanting disagreement against an unknown opponent under $\mu_{\underline{y}(p),0}$ when $p^{'} \leq p^*$ (or $\mu_{\underline{y}(p),0}$ when $p^* < p^{'}$).

The second threshold p^* is such that (A.2) for any $y \in (\bar{0}, 1]$, i.e., the weak type's

informational incentive constraint binds given $\mu_{y,0}$ with any $y \in (0,1]$:

$$U_i(\mu_{y,0}|w)|_{p=p^*} = U_i(\mu_{y,0}, s|w)|_{p=p^*}, \ \forall i.$$

This implies that when $p > p^*$, $\mu_{y,0}$ with $y \in (0,1]$ 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 he is the strong type. That is, I can find, for each $p \in (p^*, p^{**})$ and for each $y \in (0,1]$, some $z > 0$ such that $p(1-y)u_1(d_1, ws)+(1-p)u_1(d_1, ww) = p(1-z)u_1(d_1, ws) + (1-p)(1-y)u_1(d_1, ww)$, which is equivalent to $U_i(\mu_{y,z}|w) = U_i(\mu_{y,z}, s|w)$, $\forall i$, for any given $y \in (0,1]$. Solving for z, we have

$$z := z(y,p) = y\left[1 - \frac{(1-p)u_1(d_1, ww)}{pu_1(d_1, ws)}\right], \ \forall y \in (0,1].$$

This $z(y,p)$ is the probability of choosing war if $t = (s,s)$ when $p \in [p^*, p^{**})$.[32] For $y = 1$, let

$$\bar{z}(p) \equiv z(1,p) = 1 - \frac{(1-p)u_1(d_1, ww)}{pu_1(d_1, ws),}$$

which is increasing in p. This proves Lemma 2.

Note that when $p^* < p^{'}$, then for the case of $p \in [p^*, p^{'})$ in Proposition 2 there is a similar lower bound requirement on the probability of choosing war when $t \in \{(s,w),(w,s)\}$ for a mediator to be individually rational for the strong type, as in the case of $p < p^{'}$ in Proposition 1 (when $p^{'} \leq p^*$). Moreover, as discussed in the previous paragraph, such mediator must also put some positive probability on war if $t = (s,s)$ in order to be incentive compatible for the weak type. Because the probability of choosing war when $t = (s,s)$ is uniquely determined by y, there is also a corresponding lower bound on z. 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 both the binding strong type's participation constraints and the binding weak type's informational incentive constraints:

$$\begin{aligned} &U_i(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|s) = 0, \ \forall i, \\ \text{and} \ \ &U_i(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|w) = U_i(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w), \ \forall i. \end{aligned}$$

Focusing on $i = 1$, the above binding constraints respectively give:

$$p(1-\underline{\underline{z}}(p))u_1(d_1, ss) + (1-p)(1-\underline{\underline{y}}(p))u_1(d_1, sw) = 0,$$

and

$$\begin{aligned} &p(1-\underline{\underline{y}}(p))u_1(d_1, ws) + (1-p)u_1(d_1, ww) \\ &\quad = p(1-\underline{\underline{z}}(p))u_1(d_1, ws) + (1-p)(1-\underline{\underline{y}}(p))u_1(d_1, ww), \end{aligned}$$

[32] When $p = p^*$, $z(y,p) = 0$ for any $y \in [0,1]$.

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

$$\underline{\underline{y}}(p) = \frac{pu_1(d_1, ss)u_1(d_1, ws) + (1-p)u_1(d_1, sw)u_1(d_1, ws)}{pu_1(d_1, ss)u_1(d_1, ws) + (1-p)(u_1(d_1, sw)u_1(d_1, ws) - u_1(d_1, ss)u_1(d_1, ww))} \tag{A.4}$$

$$\underline{\underline{z}}(p) = 1 - \frac{(1-p)^2 u_1(d_1, sw)u_1(d_1, ww)}{p(pu_1(d_1, ss)u_1(d_1, ws) + (1-p)(u_1(d_1, sw)u_1(d_1, ws) - u_1(d_1, ss)u_1(d_1, ww)))}. \tag{A.5}$$

Regardless of whether $p^{'} \leq p^*$ or $p^{'} > p^*$, if $p \in [p^*, p^{**})$, non-negative y and z together entail $\mu_{y,z}$ to be individually rational for the strong types and incentive compatible for the weak types.

Finally, the third threshold p^{**} is the cutoff at which the strong types become indifferent between $\mu_{0,0}$ and $\mu_{y,z}$ with any $y \in (0, 1]$ (and the corresponding $z \in (0, \bar{z}(p)]$):

$$U_i(\mu_{y,z}|s)|_{p=p^{**}} = U_i(\mu_{0,0}|s)|_{p=p^{**}}, \ \forall i.$$

Then, when $p \geq p^{**}$, any mediator who puts a positive probability on war regardless of the type combination is interim Pareto dominated by $\mu_{0,0}$.

Lemma A.1 (Incentive Feasibility). *All of the mediators characterized in Propositions 1 and 2 are incentive feasible.*

Proof:
(i) Incentive Feasibility for Proposition 1: ***Case 1.*** $p < p^{'}$**:** $U_1(\mu_{y,0}|s) \geq 0$ if and only if $y \geq \underline{y}(p)$ because

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

Also, $U_1(\mu_{y,0}|s) \geq 0 > U_1(\mu_{y,0}, w|s)$ because

$$\begin{aligned} U_1(\mu_{y,0}, w|s) &= p\mu_{y,0}(d_1, ws)u_1(d_1, ss) + (1-p)\mu_{y,0}(d_1, ww)u_1(d_1, sw) \\ &\leq p\frac{pu_1(d_1, ss)}{-(1-p)u_1(d_1, sw)}u_1(d_1, ss) + (1-p)u_1(d_1, sw) < 0, \end{aligned}$$

where the first weak inequality follows from $\mu_{y,0}(d_1, ws) \equiv 1-y \leq 1-\underline{y}(p)$ and $\mu_{y,0}(d_1, ww) \leq 1$; and the second inequality follows from $p < p^{'}$. Also,

$$U_1(\mu_{y,0}|w) = p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww) > 0,$$

$$\begin{aligned} U_1(\mu_{y,0}, s|w) &= pu_1(d_1, ws) + (1-p)(1-y)u_1(d_1, ww) \\ &< p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww), \end{aligned}$$

where the last inequality follows from $p < p^{'} \leq p^{*}$. Therefore, $\mu_{y,0}$ for any $y \in [\underline{y}(p), 1]$ is incentive feasible when $p < p^{'}$ in the sense of conditions (2.2) and (2.3).

Case 2. $p \in [p^{'}, p^{*})$***:*** Recall that p^{*} is such that $U_1(\mu_{y,0}|w)|_{p=p^{*}} = U_1(\mu_{y,0}, s|w)|_{p=p^{*}}$ and $U_1(\mu_{y,0}|w) - U_1(\mu_{y,0}, s|w)$ is a strictly decreasing function of p, for any $y \in (0, 1]$. So when $p < p^{*}$, $U_1(\mu_{y,0}|w) \geq U_1(\mu_{y,0}, s|w)$, where the equality holds when $y = 0$. Also, $U_1(\mu_{y,0}|w) > 0$ for any $y \in [0, 1]$ when $p < p^{*}$. And $U_1(\mu_{y,0}|s) \geq 0$ if and only if

$$\begin{aligned} & pu_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw) \geq 0 \\ \longleftrightarrow\ & y \geq 1 + \frac{pu_1(d_1, ss)}{(1-p)u_1(d_1, sw)}, \end{aligned}$$

which holds because $1 + \frac{pu_1(d_1,ss)}{(1-p)u_1(d_1,sw)} \leq 0$ for $p \geq p^{'} = \frac{-u_1(d_1,sw)}{u_1(d_1,ss)-u_1(d_1,sw)}$ (where $U_1(\mu_{y,0}|s) = 0$ when $p = p^{'}$); and $U_1(\mu_{y,0}|s) \geq U_1(\mu_{y,0}, w|s)$ because the inequality simply follows from $u_1(d_1, sw) \leq 0$ and $y \in [0, 1]$:

$$\begin{aligned} U_1(\mu_{y,0}|s) &= pu_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw) \\ &\geq p(1-y)u_1(d_1, ss) + (1-p)u_1(d_1, sw),\ \forall y \in [0, 1]. \end{aligned}$$

Therefore, $\mu_{y,0}$ for any $y \in [0, 1]$ is incentive feasible when $p \in [p^{'}, p^{*})$.

Case 3. $p \in [p^{*}, p^{**})$***:*** Note from the previous proof of Lemma 1 that y and z are such that $U_1(\mu_{y,z}|w) = U_1(\mu_{y,z}, s|w)$. Trivially, $U_1(\mu_{y,z}|w) = p(1-y)u_1(d_1, ws) + (1-p)u_1(d_1, ww) > 0$ for any $y \in [0, 1]$. Also,

$$\begin{aligned} & U_1(\mu_{y,z}|s) = p(1-z)u_1(d_1, ss) + (1-p)(1-y)u_1(d_1, sw) > 0 \\ \longleftrightarrow\ & y > 1 + \frac{p(1-z)u_1(d_1, ss)}{(1-p)u_1(d_1, sw)}, \end{aligned}$$

where the last inequality is true by the construction of y and z and $p < p^{**}$. And

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

because $y \geq z$ and $u_1(d_1, sw) < 0$. Therefore, $\mu_{y,z}$ for any $y \in (0, 1]$ and for a corresponding $z \in (0, \bar{z}(p)]$ is incentive feasible when $p \in (p^{*}, p^{**})$.

Note that when $p \geq p^{'}$, no type should have any incentive to report dishonestly to $\mu_{0,0}$ who puts probability one on d_1 regardless of t. So, we have $U_1(\mu_{0,0}|s) = U_1(\mu_{0,0}, w|s)$ and $U_1(\mu_{0,0}|w) = U_1(\mu_{0,0}, s|w)$. Also, the following holds:

$$\begin{aligned} & U_1(\mu_{0,0}|s) = pu_1(d_1, ss) + (1-p)u_1(d_1, sw) \geq 0 \\ \longleftrightarrow\ & p \geq \frac{-u_1(d_1, sw)}{u_1(d_1, ss) - u_1(d_1, sw)} = p^{'}, \\ & \text{and } U_1(\mu_{0,0}|w) = pu_1(d_1, ws) + (1-p)u_1(d_1, ww) > 0. \end{aligned}$$

(i) Incentive Feasibility for Proposition 2: ***Case 2′.*** $p \in [p^*, p^{'})$**:** Because $\underline{\underline{z}}(p) > 0$ and $\underline{\underline{y}}(p) > 0$ are implicitly defined by $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|s) = 0$ and $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|w) = U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w)$, we only need to check that $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w) \geq 0$ and $U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, w|s) \leq 0$:

$$\begin{aligned}
U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w) &= p(1-\underline{\underline{z}}(p))u_1(d_1, ws) + (1-p)(1-\underline{\underline{y}}(p))u_1(d_1, ww) > 0;\\
U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, w|s) &= p(1-\underline{\underline{y}}(p))u_1(d_1, ss) + (1-p)u_1(d_1, sw)\\
&< p(1-\underline{\underline{z}}(p))u_1(d_1, ss) + (1-p)(1-\underline{\underline{y}}(p))u_1(d_1, sw)\\
&= U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|s) = 0,
\end{aligned}$$

where the first inequality trivially holds because $u_1(d_1, ws) > 0$ and $u_1(d_1, ww) > 0$; and the second inequality holds because $\underline{\underline{y}}(p) > \underline{\underline{z}}(p)$, $u_1(d_1, ss) > 0$, and $u_1(d_1, sw) < 0$. Therefore, $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}$ is incentive feasible. For all of the other mediators in Proposition 2, the proofs are analogous to that for Proposition 1. □

A.3 Computing Symmetric Interim Incentive Efficient Mediators

A mediator μ is interim incentive efficient iff μ is incentive feasible (i.e., satisfies (2.2) and (2.3)) and $\nexists\ \mu^{'}$ such that $\mu^{'}$ is incentive feasible and satisfies $U_i(\mu^{'}|t_i) \geq U_i(\mu|t_i)$ for all i, t_i and this inequality is strict for at least one type of one player. By the supporting hyperplane theorem, a mediator μ is interim incentive efficient if and only if there exist some positive utility weights $\lambda_i(t_i)$ for each type t_i of each player i such that μ is an optimal solution to the optimization problem: $\max_{\mu: D\times T\to\mathbb{R}} \Sigma_{i\in N}\Sigma_{t_i\in T_i}\lambda_i(t_i)U_i(\mu|t_i)$ subject to (2.1), (2.2), and (2.3). Because the objective and constraints are all linear in μ, this optimization problem is a linear programming problem. Therefore, a Lagrangean function can be formed. Let $\alpha_i(s_i|t_i)$ denote the Lagrange multiplier for the incentive compatibility constraints (2.2) and $\beta_i(t_i)$ denote the Lagrange multiplier for the individual rationality constraints (2.3).

Remark A.1. By restricting analysis to symmetric mediators who treat the same type of all players symmetrically along with the symmetry assumption on 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 $\lambda(s)+\lambda(w) = 1$. Similarly, $\alpha_1(w|s) = \alpha_2(w|s) \equiv \alpha(w|s)$, $\alpha_1(s|w) = \alpha_2(s|w) = \alpha(s|w)$, $\beta_1(s) = \beta_2(s) \equiv \beta(s)$, 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, the duality theorem of linear programming implies the following theorem, which is a modified version of the well-known result,[33] stated below without proof. The theorem gives the most tractable conditions for computing the symmetric IIE mediators, which are used to prove Propositions 1 and 2 along with Lemma 2.

Theorem A.1 (Symmetric Interim Incentive Efficient Mediators). *An incentive feasible mediator μ is symmetric interim incentive efficient if and only if there exist vectors*

$$\lambda = (\lambda(t_i))_{t_i \in T_i},\ \alpha = (\alpha(s_i|t_i))_{s_i \in T_i, t_i \in T_i},\ \text{and}\ \beta = (\beta(t_i))_{t_i \in T_i}$$

such that

$$\begin{aligned} \lambda(t_i) > 0,\ \alpha(s_i|t_i) \geq 0,\ \beta(t_i) \geq 0,\ \forall t_i \in T_i,\ \forall s_i \in T_i, \\ \alpha(s_i|t_i)(U_i(\mu|t_i) - U_i^*(\mu, s_i|t_i)) = 0,\ \forall t_i \in T_i,\ \forall s_i \in T_i, \\ \beta(t_i)U_i(\mu|t_i) = 0,\ \forall t_i \in T_i, \\ \sum_{d \in D} \mu(d|t) \sum_{i \in N} v_i(d,t,\lambda,\alpha,\beta) = \max_{d \in D} \sum_{i \in N} v_i(d,t,\lambda,\alpha,\beta),\ \forall t \in T. \end{aligned} \tag{A.6}$$

A.4 Proofs of Propositions 1 and 2

The proofs of Propositions 1 and 2 involve linear programming problems characterized by Theorem A.1, which can be verified by any one of many widely available computer programs. Because of tedious technicality involved in computing interim incentive efficient mediators, I omit many parts of the proofs and only give the proofs of sufficiency; the full proofs are available upon request. The key idea for the proofs of sufficiency is as follows: First, note that all of the mediators characterized in Propositions 1 and 2 are incentive feasible by Lemma A.1. Then, for any $\mu_{y,z}$ where y and z satisfy the restrictions in the propositions for each case, if I can find vectors λ, α, and β such that all of the conditions in (A.6) are satisfied for $\mu_{y,z}$, then such $\mu_{y,z}$ is symmetric IIE.

For Proposition 1:
Case 1. $p < p'$**:** 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 $y(p) \in (0,1)$ and is a decreasing function of $p \in (0,p')$. For this mediator, $U_1(\mu_{\underline{y}(p),0}|s) = 0 > U_1(\mu_{\underline{y}(p),0}, w|s)$,

[33] See Holmström and Myerson (1983) and Theorem 10.1 (Myerson, 1991, 498).

$U_1(\mu_{\underline{y}(p),0}|w) > 0$, and $U_1(\mu_{\underline{y}(p),0}|w) > U_1(\mu_{\underline{y}(p),0}, s|w)$. (See the proof of Lemma A.1.) 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 $t \in \{(s,w),(w,s)\}$, it must be:

$$\begin{aligned}\sum_{d\in D} \mu_{\underline{y}(p),0}(d|t) \sum_{i\in N} v_i(d,t,\lambda,\alpha,\beta) &= \max_{d\in D} \sum_{i\in N} v_i(d,t,\lambda,\alpha,\beta) \\ &= 0, \text{ for } t \in \{(s,w),(w,s)\},\end{aligned} \tag{A.7}$$

where the last equality follows because $v_i(d_0,t,\lambda,\alpha,\beta) = 0$ for any t, λ, α, and β; and $\mu_{\underline{y}(p),0}$ puts a positive probability on both outcomes. The left-hand side of (A.7) for $t = (s,w)$ and $t = (w,s)$ are, respectively:

$$\begin{aligned}\mu_{\underline{y}(p),0}(d_1|sw) \sum_{i\in N} v_i(d_1,sw,\lambda,\alpha,\beta) =&(1-\underline{y}(p))[(\lambda(s)+\beta(s))u_1(d_1,sw)/p \\ &+ \lambda(w)u_2(d_1,sw)/(1-p)],\end{aligned} \tag{A.8}$$

$$\begin{aligned}\mu_{\underline{y}(p),0}(d_1|ws) \sum_{i\in N} v_i(d_1,ws,\lambda,\alpha,\beta) =&(1-\underline{y}(p))[\lambda(w)u_1(d_1,ws)/(1-p) \\ &+ (\lambda(s)+\beta(s))u_2(d_1,ws)/p].\end{aligned} \tag{A.9}$$

Equations (A.8) and (A.9) equal zero only when

$$\begin{aligned}\lambda(s) &< \frac{pu_1(d_1,ws)}{(1-p)(-u_1(d_1,sw)) + pu_1(d_1,ws)} \equiv \lambda_p^*, \\ \text{and } \beta(s) &= \frac{-p(1-\lambda(s))u_1(d_1,ws) - (1-p)\lambda(s)u_1(d_1,sw)}{(1-p)u_1(d_1,sw)} \equiv \beta_p^* > 0.\end{aligned}$$

This implies that for $\mu_{\underline{y}(p),0}$, any $\lambda(s) < \lambda_p^*$, $\alpha = (0,0)$ and $\beta = (\beta_p^*, 0)$, where β_p^* depends on $\lambda(s)$, satisfy (A.6). Second, for $\mu_{y,0}$ with any $y > \underline{y}(p)$, all of the constraints are not binding and thus it must be $\alpha = (0,0)$ and $\beta = (0,0)$. Then for $\mu_{y,0}$ with $y \in \big(\underline{y}(p), 1\big)$, only $\lambda(s)$ such that $\lambda(s) = \lambda_p^*$ makes (A.8) and (A.9) equal zero, and thus $\lambda(s) = \lambda_p^*$, $\alpha = (0,0)$, and $\beta = (0,0)$ together satisfy (A.6). Lastly for $\mu_{1,0}$, any $\lambda(s) > \lambda_p^*$, $\alpha = (0,0)$, and $\beta = (0,0)$ satisfy (A.6). Thus by Theorem A.1, when $p < p^{'}$, any $\mu_{y,z}$ such that $y \in \big[\underline{y}(p), 1\big]$ and $z = 0$ is interim incentive efficient.

Case 2. $p \in [p^{'}, p^*)$***:*** For $\mu_{0,0}$, the informational incentive constraints bind and the participation constraints do not bind for both types. Therefore, it must be $\alpha = (\alpha_1, \alpha_2)$ and $\beta = (0,0)$ for some $\alpha_1 > 0$ and $\alpha_2 > 0$. In particular, any $\lambda(s) < \lambda_p^*$, $\alpha = (\alpha_1, \alpha_2)$, and $\beta = (0,0)$ satisfy the conditions in (A.6), where λ_p^* is defined above; and $\alpha_1 > 0$ and $\alpha_2 > 0$, which depend on $\lambda(s)$, together satisfy

$$\begin{aligned}&(\lambda(s)+\alpha_1)u_1(d_1,ss) - \alpha_2 u_1(d_1,ws) > 0, \\ \text{and } &(1-p)((\lambda(s)+\alpha_1)u_1(d_1,sw) - \alpha_2 u_1(d_1,ww)) \\ &+ p((1-\lambda(s)+\alpha_2)u_1(d_1,ws) - \alpha_1 u_1(d_1,ss)) > 0.\end{aligned} \tag{A.10}$$

For $\mu_{y,0}$ such that $y \in (0,1)$, all of the constraints are not binding; so, $\lambda(s) = \lambda_p^*$, $\alpha = (0,0)$, and $\beta = (0,0)$ satisfy (A.6). Lastly for $\mu_{1,0}$, any $\lambda(s) > \lambda_p^*$, $\alpha = (0,0)$, and $\beta = (0,0)$ satisfy (A.6). Thus by Theorem A.1, when $p \in [p^{'}, p^*)$, any $\mu_{y,z}$ such that $y \in [0,1]$ and $z = 0$ is interim incentive efficient.

Case 3. $p \in [p^*, p^{**})$***:*** By the same logic as above, $\mu_{0,0}$ is interim incentive efficient. For $\mu_{y,z}$ where $y \in (0,1)$, z is computed such that $U_1(\mu_{y,z}|w) = U_1(\mu_{y,z}, s|w)(> 0)$. That is, there is a unique corresponding $z \in (0, \bar{z}(p))$ for each $y \in (0,1)$ such that

$$z := z(y,p) = y\left[1 - \frac{(1-p)u_1(d_1, ww)}{pu_1(d_1, ws)}\right].$$

If otherwise, the weak type has an incentive to lie, and so $\mu_{y,z}$ would not be incentive feasible. Also noting that $U_1(\mu_{y,z}|s) > 0$ and $U_1(\mu_{y,z}|s) > U_1(\mu_{y,z}, w|s)$, it must be $\alpha(w|s) = 0$, $\alpha(s|w) > 0$, $\beta(s) = \beta(w) = 0$. In order for $\mu_{y,z}$ to randomize between d_0 and d_1 when $t \in \{(s,s),(s,w),(w,s)\}$, it must be:

$$\begin{aligned}\sum_{d\in D}\mu_{y,z}(d|t)\sum_{i\in N}v_i(d,t,\lambda,\alpha,\beta) &= \max_{d\in D}\sum_{i\in N}v_i(d,t,\lambda,\alpha,\beta)\\ &= 0, \text{ for } t \in \{(s,s),(s,w),(w,s)\}.\end{aligned} \tag{A.11}$$

For $t = (s,s)$, the left-hand side of (A.11) is

$$\mu_{y,z}(d_1|ss)\sum_{i\in N}v_i(d_1, ss, \lambda, \alpha, \beta) = (1-z)\left[2\cdot\left(\lambda(s)u_1(d_1,ss) - \alpha(s|w)u_1(d_1,ws)\right)/p\right];$$

and for $t = (s,w)$, the left-hand side of (A.11) is

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

The left-hand side for $t = (w,s)$ basically yields the same thing because of symmetry. Setting both to zero, I get

$$\begin{aligned}\lambda(s) &= \frac{-pu_1(d_1,ws)}{p(u_1(d_1,ss) - u_1(d_1,ws)) + (1-p)\left(u_1(d_1,sw) - u_1(d_1,ss)u_1(d_1,ww)/u_1(d_1,ws)\right)}\\ &\equiv \lambda_p^{**},\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}\alpha(s|w) &= \lambda_p^{**}\frac{u_1(d_1,ss)}{u_1(d_1,ws)}\\ &\equiv \alpha_p^{**} > 0.\end{aligned}$$

Therefore, for $\mu_{y,z}$ such that $y \in (0,1)$ and $z \in (0, \bar{z}(p))$, $\lambda(s) = \lambda_p^{**}$, $\alpha = (0, \alpha_p^{**})$, and $\beta = (0,0)$ satisfy (A.6). Lastly for $\mu_{1,\bar{z}(p)}$, where $\bar{z}(p) = z(1,p)$ is increasing in p, any

$\lambda(s) > \lambda_p^{**}$, $\alpha = (0, \alpha^{'})$, where $\alpha^{'} \equiv \lambda(s)\frac{u_1(d_1,ss)}{u_1(d_1,ws)}$, and $\beta = (0,0)$ satisfy (A.6). Thus by Theorem A.1, when $p \in [p^*, p^{**})$, any $\mu_{y,z}$ such that $y \in [0,1]$ with a corresponding $z \in [0, \bar{z}(p)]$ is interim incentive efficient.

Case 4. $p \geq p^{**}$***:*** For $\mu_{0,0}$, the informational incentive constraints bind and the participation constraints do not bind for both types. In this case, any $\lambda(s) \in (0,1)$, $\alpha = (\alpha_1^{'}, \alpha_2^{'})$, and $\beta = (0,0)$, for $\alpha_1^{'} > 0$ and $\alpha_2^{'} > 0$ such that $(\lambda(s)+\alpha_1^{'})u_1(d_1, ss) - \alpha_2^{'}u_1(d_1, ws) > 0$ and $(1-p)((\lambda(s)+\alpha_1^{'})u_1(d_1, sw) - \alpha_2^{'}u_1(d_1, ww)) + p((1-\lambda(s)+\alpha_2^{'})u_1(d_1, ws) - \alpha_1^{'}u_1(d_1, ss)) > 0$, satisfy (A.6). Note that for all $\lambda(s) \in (0,1)$, the optimization problem yields a unique solution $\mu_{0,0}$. This implies that when $p \geq p^{**}$, $\mu_{0,0}$ is interim Pareto superior to any other mediator; and no other mediator interim Pareto dominates $\mu_{0,0}$. Thus, when $p \geq p^{**}$, $\mu_{0,0}$ is the only interim incentive efficient mediator.

For Proposition 2:

The proof is analogous to the proof of Proposition 1. It suffices to prove that there is a lower bound in the range of feasible y (and the corresponding z) for $\mu_{y,z}$ to be interim incentive efficient in ***Case 2′.*** $p \in [p^*, p^{'})$. Note that $\underline{\underline{y}}(p)$ and $\underline{\underline{z}}(p)$ are already defined by (A.4) and (A.5) on page 31. For $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}$, it is easy to check the following:

$$U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|s) = 0 > U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, w|s);$$

$$U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}|w) = U_1(\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}, s|w) > 0,$$

because $\underline{\underline{y}}(p)$ and $\underline{\underline{z}}(p)$ are the probabilities on war respectively for $t \in \{(s,w),(w,s)\}$ and for $t = (s,s)$ that simultaneously make the participation constraints bind for the strong type and the informational incentive constraints bind for the weak type. So, it must be that:

$$\beta(s) > 0,\ \alpha(w|s) = 0,\ \beta(w) = 0,\ \text{and } \alpha(s|w) > 0.$$

In order for $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}$ to randomize between d_0 and d_1 when $t \in \{(s,s),(s,w),(w,s)\}$, it must be:

$$\begin{aligned}\sum_{d\in D}\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}(d|t)\sum_{i\in N}v_i(d,t,\lambda,\alpha,\beta) &= \max_{d\in D}\sum_{i\in N}v_i(d,t,\lambda,\alpha,\beta)\\ &= 0,\ \text{for } t \in \{(s,s),(s,w),(w,s)\}\end{aligned} \tag{A.13}$$

For $t = (s,s)$, the left-hand side of (A.13) is

$$\begin{aligned}\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}(d_1|ss)\sum_{i\in N}v_i(d_1,ss,\lambda,\alpha,\beta) =& (1-\underline{\underline{z}}(p))[2\cdot\{(\lambda(s)+\beta(s))u_1(d_1,ss)\\ &- \alpha(s|w)u_1(d_1,ws)\}/p];\end{aligned}$$

and for $t = (s,w)$, the left-hand side (A.13) is

$$\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}(d_1|sw)\sum_{i\in N} v_i(d_1,sw,\lambda,\alpha,\beta) = (1-\underline{\underline{y}}(p))[\{(\lambda(s)+\beta(s))u_1(d_1,sw) - \alpha(s|w)u_1(d_1,ww)\}/p + (\lambda(w)+\alpha(s|w))u_2(d_1,sw)/(1-p)].$$

The equation for $t=(w,s)$ basically yields the same thing because of symmetry. Setting both of the above to zero, I get

$$\alpha(s|w) = \frac{(1-\lambda(s))pu_1(d_1,ws)}{(1-p)(u_1(d_1,ww)-u_1(d_1,sw)u_1(d_1,ws)/u_1(d_1,ss))-pu_1(d_1,ws)} \equiv a_p^{***} > 0$$

and

$$\beta(s) = \frac{a_p^{***}u_1(d_1,ws)-\lambda(s)u_1(d_1,ss)}{u_1(d_1,ss)} \equiv \beta_p^{***} > 0.$$

Therefore, for $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}$, $\lambda(s) < \lambda_p^{**}$ with λ_p^{**} as defined in (A.12), $\alpha = (0, a_p^{***})$, and $\beta = (\beta_p^{***}, 0)$, where α_p^{***} and β_p^{***} depend on $\lambda(s)$, satisfy (A.6). Thus by Theorem A.1, when $p \in [p^*, p^{'})$, $\mu_{\underline{\underline{y}}(p),\underline{\underline{z}}(p)}$ is interim incentive efficient. □

A.5 All Other Proofs for Section 3

Proof of Lemma 3:
(*Only if*) Suppose that $y^{'} > y$ with some corresponding $z^{'} \geq z$ such that the restrictions for each case hold. Then, the ex ante evaluations of a mediator $\mu_{y',z'}$ by any player is given by

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

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

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

Then,

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

since $z^{'} \geq z$, $y^{'} > y$, $u_1(d_1,ss) > 0$, and $u_1(d_1,sw)+u_1(d_1,ws) > 0$ by **(A1)** – **(A3)**. Thus, $U_i(\mu_{y,z}) > U_i(\mu_{y',z'})$ for all i; 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 $U_i(\mu_{y,z}) > U_i(\mu_{y',z'})$ for

all i. Suppose to the contrary that $y \geq y^{'}$. Then, the corresponding z and $z^{'}$ are such that $z \geq z^{'}$, which, together with $y \geq y^{'}$, imply $pp(z^{'} - z)u_1(d_1, ss) + p(1-p)(y^{'} - y)\left[u_1(d_1, sw) + u_1(d_1, ws)\right] \leq 0$. This is a contradiction. Thus, it must be $y < y^{'}$. □

Proof of Proposition 3:
An incentive feasible μ in $S(\Gamma^*)$ 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, $\nexists \delta \in S(\Gamma^*)\backslash\{\mu\}$ such that $U_i(\delta) \geq U_i(\mu),\ \forall i$, with strict inequality for at least one player. Because of symmetry, strict inequality must hold for both players. Lemma 3 states that symmetric IIE mediators in $S(\Gamma^*)$ can be *strictly* ordered by the ex ante evaluations in terms of the probability weight on war. The lemma implies that the mediator associated with the (unique) lowest value of y for each case would give both players the maximum ex ante expected utilities. Thus Proposition 3 immediately follows, and all of the other mediators in $S(\Gamma^*)$ are ex ante Pareto dominated by the unique ex ante incentive efficient mediator. □

Proof of Lemma 4:
Suppose that $y < y^{'}$ with a corresponding $z \leq z^{'}$ such that the restrictions for each case hold. Then, 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 given by

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

For ***Cases 1*** & ***2***, since $z = z^{'} = 0$, 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 $z^{'} = y^{'}\left[1 - \frac{(1-p)u_1(d_1,ww)}{pu_1(d_1,ws)}\right]$, we have

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

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

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

for any p; and the interim evaluation of $\mu_{y',z'}$ by any player of the weak type is given by

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

for any p. Then, for any y' and y such that $y' - y > 0$, we have

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

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

Proof of Corollary 1:
The proof follows directly from the previous results. By Proposition 3, there is exists a mediator who is uniquely ex ante incentive efficient. Therefore, excluding the case where $S(\Gamma^*)$ is a singleton, Propositions 1 and 2 imply that there is a continuum of symmetric IIE mediators that are ex ante Pareto inferior to the unique ex ante incentive efficient mediator. Moreover, by Lemma 3, the mediator that minimizes the ex ante expected payoffs of the players over the set of symmetric IIE mediators is μ_1. By Lemma 4, this mediator gives the highest interim expected utility to the strong type and the lowest interim expected utility to the weak type. □

B Appendix B: Proofs for Section 5

Proof of Lemma 5:
It suffices to find a $\kappa \in K_{y^*}$ such that a credible non-matching belief system $\bar{q}^\kappa$ that supports such κ exists; and $\bar{q}^\kappa$ together with a corresponding noncooperative equilibrium $\Sigma(\bar{q}^\kappa)$ do not satisfy part (ii) of Definition 2. Consider $\kappa \in K_{y^*}$, where $y^* < 1$, such that one player's voting decision is y^* and the other player's voting decision is "1," i.e., either $(y^*, 1)$ or $(1, y^*)$. Let's focus on $(\kappa_1, \kappa_2) = (y^*, 1)$ without loss of generality.
Note that by Lemma 4, under the priors, the interim expected utility from μ_1 for the weak type is strictly lower than the interim utility from μ_{y^*}; and the interim expected utility from μ_1 for the strong type is strictly greater than the interim utility from μ_{y^*}. That is, at the interim stage, a strong type of any player prefers the ex ante Pareto inferior μ_1 to μ_{y^*} when she knows only her type, while a weak type of any player prefers the ex ante Pareto superior μ_{y^*} to μ_1 when he knows only his type. Therefore if a non-matched vote outcome resulted in passive inferences (i.e., no updating), then unanimous vote on μ_{y^*} followed by truthful revelation in μ_{y^*} is a sequential equilibrium in the two-stage ratification game.
When $\kappa = (y^*, 1)$ occurs, the following war strategies

$$\psi^\kappa = ((\psi_1^\kappa(s) = 1, \psi_1^\kappa(w) = 0), (\psi_2^\kappa(s) = 1, \psi_2^\kappa(w) = 0)) \tag{B.1}$$

constitute an equilibrium $\Sigma(\bar{q}^\kappa)$ of G with respect to the posterior beliefs $\bar{q}^\kappa$ such that[34]

$$\bar{q}^\kappa = (\bar{q}_{1,y^*}(w) = 1, \bar{q}_{2,1}(s) = \bar{p}_2(s)). \tag{B.2}$$

[34]This is the unique equilibrium to G under the $\bar{q}^\kappa$-revised beliefs.

The rest of the proof consists of showing that $\bar{q}^{\kappa}$ is a credible non-matching belief system, i.e., consistent with the players' incentives to cast non-matching votes. To do so, we need to identify the set of types for each player that cast a non-matching vote with positive probability and check whether those types credibly cast a non-matching vote. Suppose a pair of such set of types are: $\{w\}$ for player 1 and $\{s, w\}$ for player 2, i.e., $v_1(w) \neq \kappa_2 = 1$ with positive probability (while $v_1(s) = 1$), and $v_2(s) \neq \kappa_1 = y^*$ with positive probability and $v_2(w) \neq \kappa_1 = y^*$ with positive probability. If so, (B.2) satisfy condition (4) in Definition 1. Now condition (2) in Definition 1 must hold for all t_i of any i such that $v_i(t_i) = \kappa_i \neq \kappa_{-i}$ given (B.2). We can easily see that the following holds for the strong type of player 2:

$$\begin{aligned}
&\bar{q}_{1,y^*}(s)\sum_{d\in D}\mu_{y^*}(d|ss)u_2(d,ss) + \bar{q}_{1,y^*}(w)\sum_{d\in D}\mu_{y^*}(d|ws)u_2(d,ws)\\
&= (1-y^*)u_2(d_1,ws)\\
<&U_2(G|s,\Sigma(\bar{q}^{\kappa})\\
&= \bar{q}_{1,y^*}(s)(1-\psi_2^{(y^*,1)}(s))(1-\psi_1^{(y^*,1)}(s))u_2(d_1,ss)\\
&+ \bar{q}_{1,y^*}(w)(1-\psi_2^{(y^*,1)}(s))(1-\psi_1^{(y^*,1)}(w))u_2(d_1,ws) = 0,
\end{aligned} \tag{B.3}$$

where the inequality follows because $(1-y^*) > 0$ and $u_2(d_1, ws) < 0$. Therefore, it must be $v_2(s) = \kappa_2 = 1 \neq \kappa_1 = y^*$ with probability one by condition (2). Similarly, for the strong type of player 1:

$$\begin{aligned}
&\bar{q}_{2,1}(s)\sum_{d\in D}\mu_1(d|ss)u_1(d,ss) + \bar{q}_{2,1}(w)\sum_{d\in D}\mu_1(d|sw)u_1(d,sw)\\
&= \bar{p}_2(s)\mu_1(d_1|ss)u_1(d_1,ss) + \bar{p}_2(w)\mu_1(d_1|sw)u_1(d_1,sw)\\
&= \bar{p}_2(s)\mu_1(d_1|ss)u_1(d_1,ss)\\
>&U_1(G|s,\Sigma(\bar{q}^{\kappa})\\
&= \sum_{t_2\in T_2}\underbrace{\bar{q}_{2,1}(t_2)}_{=\bar{p}_2(t_2)}\underbrace{(1-\psi_1^{(y^*,1)}(s))}_{=0}(1-\psi_2^{(y^*,1)}(t_2))u_1(d_1,(s,t_2)) = 0
\end{aligned} \tag{B.4}$$

where the first equality follows because $\bar{q}_{2,1}(s) = \bar{p}_2(s)$, the second equality follows because $\mu_1(d_1|sw) = 0$, the inequality follows because $\mu_1(d_1|ss) > 0$ for any p and $u_1(d_1, ss) > 0$; thereby justifying $v_1(s) = \kappa_2$ with probability one by condition (3).
Now for the weak type of player 2, given (B.1) and (B.2), we have:

$$\begin{aligned}
&\bar{q}_{1,y^*}(s)\sum_{d\in D}\mu_{y^*}(d|sw)u_2(d,sw) + \bar{q}_{1,y^*}(w)\sum_{d\in D}\mu_{y^*}(d|ww)u_2(d,ww)\\
&= u_2(d_1,ww)\\
=&U_2(G|w,\Sigma(\bar{q}^{\kappa})) = \bar{q}_{1,y^*}(s)(1-\psi_2^{(y^*,1)}(w))(1-\psi_1^{(y^*,1)}(s))u_2(d_1,sw)\\
&+ \bar{q}_{1,y^*}(w)(1-\psi_2^{(y^*,1)}(w))(1-\psi_1^{(y^*,1)}(w))u_2(d_1,ww) = u_2(d_1,ww);
\end{aligned} \tag{B.5}$$

and for the weak type of player 1, given (B.1) and (B.2), we have:

$$\begin{aligned}
&\bar{q}_{2,1}(s)\sum_{d\in D}\mu_1(d|ws)u_1(d,ws)+\bar{q}_{2,1}(w)\sum_{d\in D}\mu_1(d|ww)u_1(d,ww)\\
&=\bar{p}_2(w)u_1(d_1,ww)\\
=&U_1(G|w,\Sigma(\bar{q}^\kappa))=\bar{q}_{2,1}(s)(1-\psi_1^{(y^*,1)}(w))(1-\psi_2^{(y^*,1)}(s))u_1(d_1,ws)\\
&+\bar{q}_{2,1}(w)(1-\psi_1^{(y^*,1)}(w))(1-\psi_2^{(y^*,1)}(w))u_1(d_1,ww)\\
&=\bar{p}_2(w)u_1(d_1,ww).
\end{aligned}\tag{B.6}$$

The weak type of any player earns zero expected utility in either case. Thus matching or non-matching by such a type can be rationalized in the sense that the conditions (2) and (3) do not impose any restriction on the voting decisions for types that are indifferent. In other words, a weak type of player 1 participates in $v_1(w)\neq\kappa_2=1$ with positive probability (while $v_1(s)=\kappa_2=1$) that leads player 2's updated belief to be $\bar{q}_{1,y^*}(w)=1$; a weak type of player 2 participates in $v_2(w)\neq\kappa_1=y^*$ with positive probability (together with $v_2(s)\neq\kappa_1$) that leads player 1's updated beliefs remain the same as their interim values, i.e., $\bar{q}_{2,1}(t_2)=\bar{p}_2(t_2)$ for all $t_2\in T_2$.
Thus a belief system $\bar{q}^\kappa=(\bar{q}_{1,y^*}(w)=1,\bar{q}_{2,1}(s)=\bar{p}_2(s))$ is a credible non-matching belief system in the sense of Definition 1; however for this credible belief system, part (ii) of Definition 2 does not hold for some type, which casts a different vote than y^* with positive probability, of player 2 whose voting decision is non-matching, i.e., $\kappa_2\neq y^*$. In particular, whereas a weak type of player 2 is indifferent from matching and non-matching, satisfying part (ii); a strong type of player 2 strictly prefers the default game to μ_{y^*} and credibly casts a non-matching vote with probability one. Therefore, if $y^*<1$, there exists a credible non-matching belief system $\bar{q}^\kappa$ for $\kappa=(y^*,1)$ (and $\kappa=(1,y^*)$) and a corresponding equilibrium $\Sigma(\bar{q}^\kappa)$ under $\bar{q}^\kappa$ that together do not satisfy part (ii) of Definition 2. □

Proof of Theorem 3:
Lemma 5 implies that a sequential equilibrium of always unanimous vote on μ_{y^*} for any $y^*<1$ fails the credibility refinement. That is, such equilibrium is not robust in the sense that, even if the off-the-path event $\kappa\in K_{y^*}$ has zero probability, a credible non-matched vote outcome is possible: A strong type can credibly signal what her type is by casting a non-matching vote on her most preferred mediator μ_1 when it is to the strong type's advantage for the other player to recognize the signal, and the other player's belief successfully recognizes that the strong type's non-matching voting decision is a credible signal about her type. Therefore when $p<p^{**}$ any mediator $\mu_{y'}$ such that $y'<1$ is not threat-secure.
We are left to check whether there exists an equilibrium in which the players unanimously vote for μ_1 in all information states, where beliefs following any non-matched vote outcome are required to satisfy consistency conditions. Suppose that $\kappa\in K_1$ occurs. Without loss of generality, let $\kappa_1\neq 1$ and $\kappa_2=1$. Then the equilibrium strategies ψ^κ in G and the posterior belief system $\bar{q}^\kappa$ such that (B.1) and (B.2) for $\kappa=(\kappa_1,1)$ for any $\kappa_1\neq 1$, and

the voting strategies $v(\cdot)$ together satisfy the following: **(a)** $v_1(w) \neq \kappa_2 (= 1)$ with positive probability, $v_2(s) \neq \kappa_1$ with positive probability, and $v_2(w) \neq \kappa_1$ with positive probability; **(b)** condition (3) in Definition 1 is satisfied for the strong type of player 1, i.e., $v_1(s) = \kappa_2$ with probability one; **(c)** condition (2) in Definition 1 is satisfied for the strong type of player 2, i.e., $v_2(s) = \kappa_2$ with probability one; **(d)** the weak types of both players are indifferent between casting a matching vote and a non-matching vote; and **(e)** $\bar{q}^{\kappa}$ satisfies condition (4) in Definition 1, consistent with both types of player 2's incentives to cast a non-matching vote with a probability that lead player 1's posteriors remain the same as interim beliefs, and also consistent with the strong type of player 1's incentive to cast a matching vote with probability one and the weak type of player 1's incentive to cast a non-matching vote with positive probability that lead player 2's posteriors be $\bar{q}_{1,\kappa_1}(w) = 1$. The proofs for conditions **(a)** through **(e)** analogously follow from the proof of Lemma 5. Then for any $\kappa \in K_1$, there is a unique credible non-matching belief system that can be formed over a pair of "non-matching" types for each player; e.g., in the above case, $\{w\}$ for player 1 and $\{s, w\}$ for player 2 for any $\kappa \in K_1$ such that $\kappa_1 \neq 1$ and $\kappa_2 = 1$. Also for such credible non-matching belief system, part (ii) of Definition 2 holds for player i such that $\kappa_i \neq 1$; e.g., for player 1 whose voting decision is $\kappa_1 \neq 1$, all types such that $v_1(t_1) \neq \kappa_2 = 1$ with positive probability is indifferent between matching and non-matching, where the weak type is the only type such that $v_1(t_1) \neq \kappa_2$ with positive probability by **(a)**, **(b)**, and **(d)**. The posterior beliefs following any non-matched vote outcome $\kappa \in K_1$ satisfy consistency conditions along with $\bar{q}^{\kappa}$, ψ^{κ}, and equilibrium voting strategies $v_1(t_1) = v_2(t_2) = 1$ for all $t \in T$ together satisfying conditions (5.1) through (5.5). Thus by Definition 2, μ_1 is threat-secure, which exists because $\mu_1 \in S(\Gamma^*)$ for any $p < p^{**}$. □

Proof of Proposition 5:
Directly follows from the proof of Theorem 3. □

References

Bester, Helmut and Roland Strausz. 2001. "Contracting with Imperfect Commitment and the Revelation Principle: The Single Agent Case." *Econometrica* 69(4):1077–1098.

Brennan, James R. and Joel Watson. 2013. "The Renegotiation-Proofness Principle and Costly Renegotiation." *Games* 4(3):347–366.

Celik, Gorkem and Michael Peters. 2011. "Equilibrium Rejection of a Mechanism." *Games and Economic Behavior* 73(2):375–387.

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

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.

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.

Kim, Jin Yeub. 2014. "Models to Understand Incentives in Conflict." PhD diss., University of Chicago.

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

Laffont, Jean-Jacques and David Martimort. 2000. "Mechanism Design with Collusion and Correlation." *Econometrica* 68(2):309–342.

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

Ledyard, John O. and Thomas R. Palfrey. 2007. "A General Characterization of Interim Efficient Mechanisms for Indepedent Linear Environments." *Journal of Economic Theory* 133(1):441–466.

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

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.

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.

Wilkenfeld, Jonathan, Kathleen Young, Victor Asal and David Quinn. 2003. "Mediating International Crises: Cross-National and Experimental Perspectives." *Journal of Conflict Resolution* 47(3):279–301.