Optimal Mechanisms in Almost Ex-Ante Bargaining Problems*

Jin Yeub Kim†

March 14, 2017

Abstract

In the problem of mechanism design for bargaining with incomplete information, agents may agree on a mechanism behind the veil of ignorance that is to be implemented after they have observed their private information. Such bargaining process relies on the assumption that it is absolutely common knowledge that agents definitely do not know their types when bargaining over mechanisms. But what if it is too late to be sure that agents are uninformed ex ante? I consider a two-person bargaining problem in which bargaining over mechanisms takes place at the stage when there is some small probability that agents might have learned their types, called the *almost ex-ante* stage. I characterize optimal mechanisms that are robust to an ε-perturbation of the information structure at the bargaining stage.

JEL Classification Codes: C71; C78; D82.

Keywords: Bargaining; Incomplete information; Mechanism design; Mechanism selection stage.

*I am deeply grateful to Roger Myerson for his valuable advice, helpful conversations, and continuous support. I appreciate the feedback from the seminar participants at the University of Pittsburgh.

†Department of Economics, University of Nebraska-Lincoln, 1240 R Street, Lincoln, NE 68588-0489, USA. E-mail: shiningjin@gmail.com, Webpage: https://sites.google.com/site/jinyeubkim

1 Introduction

When agents bargain over contracts or mechanisms before observing their private information, standard arguments suggest that they can agree on a mechanism that is ex ante incentive efficient. If agents have private information at the stage of implementation of the mechanism, then the agents must be able to bind themselves to the mechanism that is selected before either has any private information or some arbitrator (who has also committed himself to the mechanism) must be able to enforce mechanism agreements. Otherwise, agents might try to renegotiate their mechanism after they have observed private information. Then the process of agreeing on a mechanism by privately informed agents can naturally be modeled as the application of a bargaining solution for a theory of bargaining with incomplete information to the set of possible mechanisms.[1]

The problem of ex ante bargaining raises yet another important conceptual issue. If agents retreat behind the veil of ignorance to bargain over mechanisms, there should be no doubt at all that agents do not know their types. But at the moment that agents bargain, what if they are no longer truly ignorant? The solutions for ex ante bargaining may be very sensitive to the assumption of the bargaining stage with absolute certainty about the information structure that nobody has any private information. The goal of this paper is to examine the robustness of ex ante bargaining solutions to adding a small perturbation to this assumption. In particular, I consider an *almost ex-ante* environment in which an agent might have been informed of his or her type with a small probability at the time of bargaining over mechanisms. In such bargaining problems, I provide characterizations of optimal mechanisms that are robust to an ε-perturbation of the ex ante bargaining stage.

One implication of the results in this paper relates to the ex ante welfare criterion that

[1] In two seminal papers, Myerson (1983, 1984) propose solutions for a theory of bargaining with incomplete information. See Kim (2017) for the application to the problem of third party selection with incomplete information. For a large body of related literature on mechanism selection, see Balkenborg and Makris (2015), Celik and Peters (2011), de Clippel and Minelli (2004), Cramton and Palfrey (1995), Holmström and Myerson (1983), Laffont and Martimort (2000), Lagunoff (1995), and Maskin and Tirole (1990, 1992), among many others.

is used in several studies to evaluate the performance of mechanisms. Invoking the ex ante measure is valid only under the assumptions that agents bargain over mechanisms before knowing their types and that agents must be able to commit themselves to implement the chosen mechanism even after learning their types. These assumptions are strong and hard to justify in terms of the relevance to the real bargaining situation where bargainers already have private information at the time they make a decision about the choice of possible agreements.

Further, a proper study of mechanism design for bargaining with incomplete information must account for different informational issues that may arise in the bargaining stage and in the implementation stage. The characterization of optimal mechanisms that are robust to a small perturbation of the ex ante bargaining stage provides a more solid grounding for applications of such solutions in many bargaining problems. The applications encompass court trials; commercial, labor, and employment disputes; disputes in the government and consumer sectors as well as international disputes.

2 The Model

2.1 Basic Setup of Bargaining Problems

To describe a bargaining problem with incomplete information, I use the concept of *Bayesian bargaining problem* proposed by Harsanyi (1967-8) and further analyzed for the fixed-threats case by Harsanyi and Selten (1972) and Myerson (1979, 1984). Formally, a two-person Bayesian bargaining problem G is an object of the form

$$G = (D, d_0, T_1, T_2, u_1, u_2, p)\,.$$

The D is the set of feasible outcomes or decisions that the players can jointly choose among, and $d_0 \in D$ is the conflict outcome which the players must get by default if they fail to

cooperate. For each player i, T_i is the set of possible types for player i and u_i is player i's utility payoff function from $D \times T_1 \times T_2$ into $\mathbb{R}$. The payoffs are in von Neumann-Morgenstern utility scale. Without loss of generality, I assume that utilities are normalized so that $u_i(d_0, t_1, t_2) = 0$ for all i, t_1 and t_2. Let $T = T_1 \times T_2$ denote the set of all possible type combinations $t = (t_1, t_2)$. The players' types are independent random variables under the common prior probability distribution $p \in \Delta(T)$ that determines the players' beliefs about other players' types. All types are assumed to have positive probability, $p_i(t_i) > 0$ for all i and all $t_i \in T_i$.

In any bargaining problem, the players can agree on some decision rule or mechanism that specifies how the choice should depend on the players' types. Formally, a (direct-revelation) mechanism is defined as a function $\mu : D \times T \to \mathbb{R}$ such that $\sum_{c \in D} \mu(c|t) = 1$ and $\mu(d|t) \geq 0$ for all $d \in D$ and for all $t \in T$. That is, $\mu(d|t)$ is the probability of choosing outcome d in the mechanism μ, if t_1 and t_2 are the players' types.

2.2 Information Structures at Bargaining and Implementation Stages

Unless there is only one sensible choice of mechanism μ to which players can commit themselves, a bargaining problem subsumes two stages: the stage of bargaining over mechanisms (or mechanism selection) and the stage of implementation of the mechanism. In a Bayesian bargaining problem, implementation takes place at the interim stage when each player has received his private information about his or her type t_i but does not know the other's information. When players' types are not verifiable, then attention should be restricted to mechanisms that incorporate different incentives of players' types.

At the moment that players bargain to agree on a mechanism, there is some small probability $\varepsilon > 0$ that each player has already learned his own type but not the other player's type, independently of the other player. I call this an *almost ex-ante* environment.[2] Then at the almost ex-ante stage of bargaining over mechanisms, there are effectively $|T_i| + 1$ number

[2] I thank Roger Myerson for suggesting this term.

of "types" of player i: $\varepsilon p_i(t_i)$ probability that player i already knows his type and the type is t_i for all $t_i \in T_i$, and $(1-\varepsilon)$ probability that player i is still waiting to learn his type.[3] I will call the latter type of player an "uninformed type." The choice of a mechanism will depend on the state of the players' information.

2.3 Discussion of Benchmark Cases

Two benchmark cases can be distinguished. The case of $\varepsilon = 1$ describes the situation in which bargaining takes place at the interim stage. In such case, players already have their private information even before they bargain over mechanisms. The choice of a mechanism can then be determined by an incomplete information bargaining solution applied to the set of possible mechanisms, such as Harsanyi and Selten (1972) solution or Myerson (1984) solution. The case of $\varepsilon = 0$ characterizes bargaining at the (absolutely) ex ante stage. In this second case, where bargaining over mechanisms takes place before players have received any private information, the choice of a mechanism will be an ex ante incentive-efficient mechanism. The two benchmark cases are analyzed in a simple example in Section 4.

The second case raises some conceptual issues. When a mechanism is chosen under the veil of ignorance, players must be able to bind themselves to the chosen mechanism that is to be implemented after they have private information. Otherwise, players might wish to renegotiate and replace the chosen mechanism with an alternative mechanism even when the mechanism is incentive feasible. More importantly, the process of selecting a mechanism at the ex ante stage relies on the assumption that it is absolutely common knowledge that players do not know their types. Hence, the choice of a mechanism that results in this case may be very sensitive to the absolute certainty about the bargaining stage being ex ante. This issues leads to consideration of another possible specification of the information structure: adding a small perturbation to the assumption that bargaining takes place at the ex ante stage.

[3]For exposition, I use male pronouns for a player if there is no confusion.

"A game with incomplete information is a game in which each player may have private information [$\cdots$] which the others do not know, at the time when the game is played" (Myerson, 1984, 461). But the information structure may differ in different stages of the game. A two-person Bayesian bargaining problem with the almost ex-ante stage of mechanism-selection captures the idea that there is incomplete information not only in the sense that players have different private information at the time when the mechanism is implemented, but also in the sense that some players may have private information that other players do not have at the time when the decisions about which mechanism to implement are made.

3 General Results

3.1 Optimal Mechanisms in Almost Ex-Ante Bargaining Problems

I characterize optimal mechanisms for two-person bargaining problems in which a mechanism is selected at the almost ex-ante stage and is implemented at the interim stage.

When players have private information about their types that are not verifiable at the stage of implementation of the mechanism, the set of possible mechanisms should be restricted to mechanisms that are incentive feasible—that is, incentive compatible and individually rational. A mechanism μ is incentive feasible if and only if it satisfies the following informational and participational incentive constraints for all $i \in \{1, 2\}$ and all $t_i \in T_i$:

$$\sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_{-i}(t_{-i}) \mu(d|t) u_i(d, t) \geq \sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_{-i}(t_{-i}) \mu(d|t_{-i}, s_i) u_i(d, t) \ \ \forall s_i \in T_i, \tag{1}$$

$$\sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_{-i}(t_{-i}) \mu(d|t) u_i(d, t) \geq 0. \tag{2}$$

The left-hand-side is the conditional expected utility for player i of type t_i if both players report their types honestly, given any mechanism μ at the implementation stage. The constraint (1) asserts that no type of any player would expect to gain by lying about his type in

implementing μ while the other player remains honest; the constraint (2) guarantees that no type of any player would expect to do worse in implementing μ than in the conflict outcome. If the players cannot commit themselves ex ante to honestly report their types and to not force the conflict outcome after they learn their types, then they should select mechanisms that respect the incentive feasibility constraints regardless of the information structure at the mechanism-selection stage.

Now at the stage of mechanism selection, the decision on which mechanism to select among the set of possible mechanisms that are incentive feasible will be based on aggregation of players' preferences over those mechanisms. Mechanisms are evaluated by their anticipated effects on players. How a mechanism should be evaluated depends crucially on what information, if any, a player possess at the time.

For a player who does not possess any private information (which happens with probability $(1-\varepsilon)$), mechanisms are evaluated according his ex ante preference. Players' types are not specified, but uninformed player i believes that he is type t_i with probability $p_i(t_i)$. With probability $\varepsilon p_{-i}(t_{-i})$, player i encounters player $-i$ who is informed and who is of type t_{-i}; and with probability $(1-\varepsilon)$, player i encounters player $-i$ who is uninformed and who is expected to be type t_{-i} with probability $p_{-i}(t_{-i})$. So the ex ante evaluation of a mechanism μ by uninformed player i is given by:

$$\sum_{t_i \in T_I} p_i(t_i) \Big[\sum_{t_{-i} \in T_{-i}} \varepsilon p_{-i}(t_{-i}) \mu(d|t) u_i(d,t) + (1-\varepsilon) \Big[\sum_{t_{-i} \in T_{-i}} p_{-i}(t_{-i}) \mu(d|t) u_i(d,t) \Big] \Big],$$

which reduces to

$$U_i(\mu|u) = \sum_{t_i \in T_i} p_i(t_i) \sum_{t_{-i} \in T_{-i}} \sum_{d \in D} p_{-i}(t_{-i}) \mu(d|t) u_i(d,t). \tag{3}$$

Technically this player does not know his type, but I condition the expected utility on being the "uninformed type" denoted by u for notational convenience to distinguish from a player whose type is privately known.

For a player who has received private information about his type (which happens with probability ε), mechanisms are evaluated according to his interim preference. Informed player i has the same probabilistic beliefs over the "types" of player $-i$ as an uninformed player, but informed player i exactly knows that he is type t_i. So the corresponding interim evaluation of a mechanism μ by player i given that he is of type t_i is:

$$\sum_{t_{-i}\in T_{-i}} \varepsilon p_{-i}(t_{-i})\mu(d|t)u_i(d,t) + (1-\varepsilon)\Big[\sum_{t_{-i}\in T_{-i}} p_{-i}(t_{-i})\mu(d|t)u_i(d,t)\Big],$$

which reduces to

$$U_i(\mu|t_i) = \sum_{t_{-i}\in T_{-i}} \sum_{d\in D} p_{-i}(t_{-i})\mu(d|t)u_i(d,t). \tag{4}$$

All of the conditional expected utility levels for each player given each of his possible types as well as the expected utility level for each player of being uninformed may be significant in determining whether mechanism μ is chosen. Whether informed or uninformed, players should bargain for mechanisms that respect their preferences over possible incentive feasible mechanisms. At the same time, because the incentive feasible mechanism that is best for a player generally depends on his information state, the players should not encode their private information, if any, in their mechanism proposals so as to prevent information leakage. Hence, the optimal choice of a mechanism at the almost ex-ante stage must be characterized based on consideration of all of the $U_i(\mu|t_i)$ and $U_i(\mu|u)$ numbers according to criteria of both efficiency and equity in some reasonable sense.

Efficiency. A mechanism μ is *almost ex-ante incentive-efficient* if μ is incentive feasible and there does not exist another incentive feasible mechanism μ' such that

$$U_i(\mu'|u) \geq U_i(\mu|u) \text{ for uninformed player } i,$$
$$U_i(\mu'|t_i) \geq U_i(\mu|t_i) \text{ for informed player } i \text{ of type } t_i \ \ \forall t_i \in T_i,$$

with at least one strict inequality.

The idea behind characterizing the set of almost ex-ante incentive-efficient mechanisms is similar to that in Myerson (1991, 497-498). Suppose that D and T are finite sets. Then the set of incentive feasible mechanisms is defined by a finite number of linear constraints. Thus by the supporting hyperplane theorem, an incentive feasible mechanism μ is almost ex-ante incentive-efficient iff there exist some positive number $\lambda_i(u)$ for an uninformed type of player i and some positive numbers $\lambda_i(t_i)$ for each type t_i of each player i such that μ is an optimal solution to the optimization problem

$$\max_{\mu:T\to\Delta(D)} \sum_i \Big[\lambda_i(u)U_i(\mu|u) + \sum_{t_i\in T_i} \lambda_i(t_i)U_i(\mu|t_i)\Big] \tag{5}$$

subject to (1) and (2) for all i and for all $t_i \in T_i$.

This optimization problem is a linear programming problem, so a Lagrangean function can be formed. Let $\alpha_i(s_i|t_i)$ denote the dual variable for the constraint (1). Because I can vary λ as a free parameter over $\mathbb{R}^{T_i\cup\{u\}}$, the dual variable for the constraint (2) can be suppressed. Then the Lagrangian function can be written as

$$\begin{aligned}&\sum_i \Big[\lambda_i(u)U_i(\mu|u) + \sum_{t_i\in T_i} \lambda_i(t_i)U_i(\mu|t_i)\Big] \\ &+ \sum_i \sum_{t_i\in T_i} \sum_{s_i\in T_i} \alpha_i(s_i|t_i)\Big[\sum_{t_{-i}\in T_{-i}} \sum_{d\in D} p_{-i}(t_{-i})\mu(d|t)u_i(d,t) - \sum_{t_{-i}\in T_{-i}} \sum_{d\in D} p_{-i}(t_{-i})\mu(d|t_{-i},s_i)u_i(d,t)\Big].\end{aligned}$$

This function can be simplified to

$$\sum_{t\in T} \sum_{d\in D} \sum_i \mu(d|t)V_i(d,t,\lambda,\alpha) \tag{6}$$

by letting

$$\begin{aligned}V_i(d,t,\lambda,\alpha) = p_{-i}(t_{-i})\Big[\Big(\lambda_i(u)p_i(t_i) + \lambda_i(t_i) + \sum_{s_i\in T_i} \alpha_i(s_i|t_i)\Big)u_i(d,t) \\ - \sum_{s_i\in T_i} \alpha_i(t_i|s_i)u_i(d,(t_{-i},s_i))\Big].\end{aligned} \tag{7}$$

I call this quantity $V_i(d,t,\lambda,\alpha)$ player i's *almost ex-ante virtual evaluation* of decision d in expectation of state t with respect to λ and α. This definition slightly differs from the original definition of virtual evaluation introduced in Myerson (1984). Player i at the time when the game begins does not know his actual type $t_i \in T_i$ if he is uninformed but believes ex ante to be type t_i with probability $p_i(t_i)$. Further player i, whether uninformed or informed, believes that the other player's type will be type t_{-i} with probability $\varepsilon p_{-i}(t_{-i}) + (1-\varepsilon)p_{-i}(t_{-i}) = p_{-i}(t_{-i})$. Hence, the almost ex-ante virtual evaluation incorporates the idea that player i has an expectation of the type profile being t when forming his evaluation of decision d.

By the duality theorem of linear programming, the dual problem for λ can be written as

$$\min_{\alpha} \sum_{t\in T} \max_{d\in D} \sum_{i} V_i(d,t,\lambda,\alpha),$$

where $\alpha = (\alpha_i(s_i|t_i))_{i\in\{1,2\},s_i\in T_i,t_i\in T_i}$. These arguments lead to the following characterization theorem for almost ex-ante incentive-efficient mechanisms.

Theorem 1. *For any two-person bargaining problem G with almost ex-ante mechanism-selection stage, an incentive feasible mechanism μ is almost ex-ante incentive-efficient if and only if there exist vectors $\lambda = (\lambda_i(u), (\lambda_i(t_i))_{t_i\in T_i})_{i\in\{1,2\}}$ and $\alpha = (\alpha_i(s_i|t_i))_{i\in\{1,2\},s_i\in T_i,t_i\in T_i}$ such that*

$$\lambda_i(u) > 0, \lambda_i(t_i) > 0, \quad \forall i \in \{1,2\},\ \forall t_i \in T_i,$$

$$\alpha_i(s_i|t_i) \geq 0, \quad \forall i \in \{1,2\},\ \forall s_i \in T_i,\ \forall t_i \in T_i,$$

$$\sum_{d\in D} \mu(d|t) \sum_{i} V_i(d,t,\lambda,\alpha) = \max_{d\in D} \sum_{i} V_i(d,t,\lambda,\alpha), \quad \forall t \in T,$$

and complementary slackness conditions of dual optima, where $V_i(d,t,\lambda,\alpha)$ is defined in (7).

Using the notion of incentive-efficiency, Theorem 1 characterizes "potentially optimal" selection of a mechanism for a two-person bargaining problem G in which bargaining over mechanisms takes place at the almost ex-ante stage. One can readily see that this theorem is very much like Theorem 10.1 in Myerson (1991). The idea behind the mathematical formulation of the conditions for characterizing incentive-efficient mechanisms is essentially the

same. However, an important point of departure of this paper is a variant of the assumption on the information structure of the bargaining stage; so that players' evaluations of mechanisms (in terms of virtual utilities) will differ from those in standard analysis of incentive efficiency in bargaining games with incomplete information.

Equity. The set of incentive-efficient mechanisms is not generally a singleton. A criterion for equity should also be used to delimit a possibly large set of almost ex-ante incentive efficient mechanisms that the players could reasonably consider. The equity criterion can be defined by the concept of transferable virtual-utility payoffs introduced in Myerson (1984). This concept captures the idea that players make some sort of equitable compromise between alternative "types" of the same player in order to not reveal the unknown state of their private information. Given my definition of almost ex-ante virtual evaluations, I can formulate the appropriately modified version of the characterization theorem given in Myerson (1984).

Theorem 2. *For any two-person bargaining problem G with almost ex-ante mechanism-selection stage, an incentive feasible mechanism μ is an almost ex-ante neutral mechanism if and only if μ is an almost ex-ante incentive-efficient mechanism and there exist sequences $\{\lambda^k\}_{k=1}^{\infty}$, $\{\alpha^k\}_{k=1}^{\infty}$, and $\{\omega^k\}_{k=1}^{\infty}$ such that:*

$$\lambda_i^k(u) > 0, \lambda_i^k(t_i) > 0, \quad \forall i \in \{1,2\},\ \forall t_i \in T_i,\ \forall k$$

$$\alpha_i^k(s_i|t_i) \geq 0, \quad \forall i \in \{1,2\},\ \forall s_i \in T_i,\ \forall t_i \in T_i,\ \forall k,$$

$$\Big(\lambda_i^k(u)p_i(t_i) + \lambda_i^k(t_i) + \sum_{s_i \in T_i} \alpha_i^k(s_i|t_i)\Big)\omega_i^k(t_i) - \sum_{s_i \in T_i} \alpha_i^k(t_i|s_i)\omega_i^k(s_i)$$

$$= \sum_{t_{-i} \in T_{-i}} \max_{d \in D} \sum_i V_i(d, t, \lambda^k, \alpha^k)/2, \quad \forall i \in \{1,2\},\ \forall t_i \in T_i,\ \forall k,$$

$$\limsup_{k \to \infty} \omega_i^k(t_i) \leq U_i(\mu|t_i), \quad \forall i \in \{1,2\},\ \forall t_i \in T_i,$$

where $V_i(d, t, \lambda^k, \alpha^k)$ is defined as in (7) with respect to λ^k and α^k. (The last two conditions immediately imply the corresponding conditions for uninformed player i.)

Although there is no general uniqueness theorem for the concept of the neutral bargaining solution, neutral solutions form the smallest set satisfying the three axioms: probability-invariance, extension, and random-dictatorship. I omit detailed expositions of these axioms, which can be found in Myerson (1984). What is essential is that Theorem 2 provides a complete characterization of optimal mechanisms for a two-person bargaining problem G where bargaining over mechanisms is at the almost ex-ante stage; the optimal mechanisms not only incorporate incentive-efficiency but also capture the idea of inscrutable intertype compromise.

3.2 Comparison with Ex Ante and Interim Incentive-Efficient Mechanisms

Following Holmström and Myerson's (1983) notations, I let Δ_A^* and Δ_I^* denote the sets of mechanisms that are respectively ex ante and interim incentive-efficient.

An incentive feasible mechanism μ is ex ante incentive-efficient iff there exist some positive numbers λ_i for each player i such that μ is an optimal solution to the optimization problem

$$\max_{\mu: T \to \Delta(D)} \sum_i \lambda_i U_i(\mu) \tag{8}$$

subject to (1) and (2) for all i and for all $t_i \in T_i$, where $U_i(\mu) \equiv \sum_{t \in T} \sum_{d \in D} p(t)\mu(d|t)u_i(d,t)$ is the expected utility for player i from the mechanism μ. Solving this problem gives us Δ_A^*.

An incentive feasible mechanism μ is interim incentive-efficient iff there exist some positive numbers $\lambda_i(t_i)$ for each type t_i of each player i such that μ is an optimal solution to the optimization problem:

$$\max_{\mu: T \to \Delta(D)} \sum_i \sum_{t_i \in T_i} \lambda_i(t_i) U_i(\mu|t_i) \tag{9}$$

subject to (1) and (2) for all i and for all $t_i \in T_i$, where $U_i(\mu|t_i)$ is the conditional expected utility for player i given that he is of type t_i from the mechanism μ, as defined in (4). Solving this problem gives us Δ_I^*. The characterization theorem for Δ_I^* is analogous to Theorem 1

except that the utility weights $(\lambda_i(u))_{i\in\{1,2\}}$ are eliminated both in the conditions and in the definition of virtual evaluations.

Let Δ^*_N denote the set of interim neutral mechanisms defined and characterized by Myerson (1984). By definition, $\Delta^*_N \subseteq \Delta^*_I$. Further, as Holmström and Myerson (1983) have described in their seminal paper, it is easy to see that $\Delta^*_A \subseteq \Delta^*_I$. The following result compares these sets to the set of almost ex-ante incentive-efficient mechanisms and to the set of almost ex-ante neutral mechanisms.

Theorem 3. *Let Δ^*_{AA} and Δ^*_{AAN} denote the sets of almost ex-ante incentive-efficient and almost ex-ante neutral mechanisms respectively. Then $\Delta^*_{AA} = \Delta^*_I$ and $\Delta^*_{AAN} = \Delta^*_N \subseteq \Delta^*_I$.*

Proof. Suppose that conditions for interim incentive-efficiency are satisfied for the incentive feasible mechanism μ together with some vectors $(\hat{\lambda}(t_i))_{i\in\{1,2\},t_i\in T_i}$ and $(\alpha_i(s_i|t_i))_{i\in\{1,2\},s_i\in T_i,t_i\in T_i}$. For any given $p_i(t_i)$ for every type t_i of any player i, there exist strictly positive numbers $\lambda_i(u)$ and $\lambda_i(t_i)$ such that $\lambda_i(u)p_i(t_i) + \lambda_i(t_i) = \hat{\lambda}_i(t_i)$. Then with these utility weights $(\lambda_i(u), (\lambda_i(t_i))_{t_i\in T_i})_{i\in\{1,2\}}$ together with $(\alpha_i(s_i|t_i))_{i\in\{1,2\},s_i\in T_i,t_i\in T_i}$, all the conditions in Theorem 1 are also satisfied for the same incentive feasible mechanism μ. The converse also holds by the same logic. Thus the set of interim incentive-efficient mechanisms and the set of almost ex-ante incentive-efficient mechanisms coincide. The equivalence for the sets of neutral mechanisms can be shown by letting $\lambda^k_i(u)p_i(t_i) + \lambda^k_i(t_i) = \hat{\lambda}^k_i(t_i)$ for all k, for all $t_i \in T_i$, and for all i, where the sequences $\{\lambda^k\}_k$ and $\{\hat{\lambda}^k\}_k$ satisfy respectively the conditions for almost ex-ante neutral mechanisms and those for interim neutral mechanisms. □

Theorem 3 immediately implies that $\Delta^*_A \subseteq \Delta^*_{AA}$. That is, ex ante incentive efficiency implies almost ex-ante incentive efficiency. However, the sets Δ^*_A and Δ^*_{AAN} are generally not equivalent. When players bargain over mechanisms at the almost ex-ante stage, there is a chance that a player might know his type as well as the other player might. So each player must use a strategy that maintains a fair compromise among the preferences of all of his possible types, including of being uninformed, in order to not reveal his true information state

during the bargaining process. Therefore, the players who must make such interpersonal-equity comparisons will choose among the set of almost ex-ante neutral mechanisms that may be ex ante incentive inefficient. The next section presents a simple example that illustrates this point.

4 An Example

There are two feasible outcomes: $D = \{d_0, d_1\}$ where d_1 represents an agreement outcome. Each player can be of type s or w, so that $T_i = \{s, w\}$ for all i, privately and independently drawn from the same distribution with probability p and $(1-p)$ respectively. Each type represents a player's preferences toward the agreement and conflict outcomes. Adopting a symmetric model, I simplify the notation by letting $\upsilon_{ss} \equiv u_1(d_1, s, s) = u_2(d_1, s, s)$, $\upsilon_{sw} \equiv u_1(d_1, s, w) = u_2(d_1, w, s)$, $\upsilon_{ws} \equiv u_1(d_1, w, s) = u_w(d_1, s, w)$, and $\upsilon_{ww} \equiv u_1(d_1, w, w) = u_2(d_1, w, w)$. These parameters satisfy $\upsilon_{ss} > 0$, $\upsilon_{ww} > 0$, $\upsilon_{ws} > 0$, $\upsilon_{sw} < 0$, and $\upsilon_{sw} + \upsilon_{ws} > 0$. Let $\Gamma \subset G$ denote the simple two-person bargaining problem.

The bargaining problem Γ describes conflict situations under uncertainty of the following kind: a type s prefers agreement to conflict with only with a type s opponent, whereas a type w always wants agreement regardless of the opponent's type; but the conflict outcome is socially inefficient in the sense that it shrinks the sum of the two players' payoffs relative to the agreement outcome for any type combination. In this class of games, I refer to type s as a strong type and to type w as a weak type, without the connotations of strength and weakness in material or financial capabilities.

At the stage of mechanism-selection, the players choose some direct-revelation mechanism $\mu : D \times T \to \mathbb{R}$. To simplify notation, I restrict attention to mechanisms that are symmetric across players. Hence, let $q_S = \mu(d_0|s, s)$, $q_M = \mu(d_0|s, w) = \mu(d_0|w, s)$, and $q_W = \mu(d_0|w, w)$. Any one of the possible mechanisms potentially available to the players can then be formally identified as a triplet $q = (q_S, q_M, q_W)$. Then a mechanism μ_q incentive

feasible in Γ if and only if it satisfies the following two informational incentive constraints and two participational incentive constraints:

$$\begin{aligned} p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw} &\geq p(1-q_M)\upsilon_{ss} + (1-p)(1-q_W)\upsilon_{sw}, \\ p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww} &\geq p(1-q_S)\upsilon_{ws} + (1-p)(1-q_M)\upsilon_{ww}; \\ p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw} &\geq 0, \\ p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww} &\geq 0. \end{aligned} \tag{10}$$

Which mechanism should be selected depends crucially on when the selection decision is made. I characterize the choice of a mechanism in the two benchmark cases of bargaining stage: the ex ante stage ($\varepsilon = 0$) and the interim stage ($\varepsilon = 1$).

4.1 Mechanism Selection in Two Benchmarks

If the selection is made *ex ante*, before any player's type is specified, then the players should bargain for mechanisms that respect their ex ante preferences as well as their incentive constraints. The ex ante evaluation of a mechanism μ_q by player i in Γ is given by:

$$U_i(\mu_q) = p^2(1-q_S)\upsilon_{ss} + p(1-p)(1-q_M)(\upsilon_{sw} + \upsilon_{ws}) + (1-p)^2(1-q_W)\upsilon_{ww}. \tag{11}$$

The reason for taking incentive constraints into account is because players evaluate possible mechanisms given their rational expectations of the effect of any incentive they might create at the interim stage of implementation. These considerations are captured by the concept of ex ante incentive-efficiency due to Holmström and Myerson (1983). Formally, a mechanism μ_q is ex ante incentive-efficient if and only if it is an optimal solution to the following problem:

$$\max_{\mu_q} \Big[p\big[p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw}\big] + (1-p)\big[p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww}\big] \Big] \tag{12}$$

subject to the constraints (10).[4]

Proposition 1. *For any two-person bargaining problem* Γ*, when bargaining over mechanisms takes place at the ex ante stage, the two players choose a unique ex ante incentive-efficient mechanism.*

Proof. The objective function in (12) is the ex ante expected utility in implementing μ_q for any player, which can be rewritten as:

$$U(\mu_q) = p^2(1-q_S)\upsilon_{ss} + p(1-p)(1-q_M)(\upsilon_{sw}+\upsilon_{ws}) + (1-p)^2(1-q_W)\upsilon_{ww}.$$

The $U(\mu_q)$ is linear is $\upsilon_{ss} > 0$, $\upsilon_{sw} + \upsilon_{ws} > 0$, and $\upsilon_{ww} > 0$. Setting $q_W = 0$ maximizes the value of the objective function only to relax (or not affect) the incentive constraints. Given $q_W = 0$, setting both q_S and q_M as low as possible maximizes the value of the objective function while q_S and q_M must together satisfy the relevant binding constraints. (In particular, q_S is either zero or increasing with q_M for any symmetric interim incentive-efficient mechanism, which is proved in Kim (2017).) Consequently, a higher q_M decreases $U(\mu_q)$, and $U(\mu_q)$ is maximized at the lowest q_M. Hence, players should be able to agree on this unique mechanism whose ultimate effect will be an ex ante incentive-efficient allocation, when they bargain over mechanisms given no private information. □

For this selection of mechanism at the ex ante stage to be applied to the real situations of disputes or bargaining problems, the disputants must be assumed to commit themselves to a mechanism that is to be selected before either has any private information. Alternatively, the ex ante efficient selection can be achieved if there is some enforceable rules for bargaining over mechanisms that prevent the disputants from using their private information in any way during the process of mechanism selection. But these requirements may be hard to justify because it must be absolutely common knowledge among the players that they are

[4]Due to symmetry, the utility weight is the same for both players, so I can focus only on maximizing the ex ante expected utility of one player.

not informed of any private information.

The disputants often seek the assistant of a mutually agreed mechanism to help reduce conflicts that arise precisely because of information asymmetries. Then the more natural assumption is that the disputants already have their private information at the time when they make a decision about which mechanism to implement. One way to provide a more solid grounding for the relevance of mechanism-selection at the interim stage to the real bargaining problem is to conversely show non-robustness of the ex ante selection. To do so, I assume that players at the stage of bargaining over mechanisms are not absolutely sure that everyone is uninformed.

At the almost ex-ante stage of bargaining over mechanisms, the players should choose among the set of almost ex-ante neutral mechanisms. Due to Theorem 3, it suffices to characterize the set of all interim neutral mechanisms for Γ that are selected at the *interim* stage. The interim evaluations of a mechanism μ_q by player i of the strong type and the weak type respectively in Γ are:

$$U_i(\mu_q|s) = p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw}, \tag{13}$$

$$U_i(\mu_q|w) = p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww}. \tag{14}$$

Then an incentive feasible mechanism μ_q is interim neutral mechanism if and only if (i) it is an optimal solution to the following problem:

$$\max_{\mu_q} \Big[\lambda(s)[p(1-q_S)\upsilon_{ss}+(1-p)(1-q_M)\upsilon_{sw}]+\lambda(w)[p(1-q_M)\upsilon_{ws}+(1-p)(1-q_W)\upsilon_{ww}]\Big] \tag{15}$$

subject to the constraints in (10) where the utility weights $\lambda(s)$ and $\lambda(w)$ are strictly positive numbers for the strong and the weak type respectively;[5] and (ii) for each positive number ϵ, there exist vectors $\alpha = (\alpha(w|s), \alpha(s|w)$ and $\varpi = (\varpi(s), \varpi(w))$ together with $\lambda = (\lambda(s), \lambda(w))$

[5]Due to symmetry, I need not distinguish between the two players but restrict to the λ-weighted sum of the expected utilities of all types of one player.

such that $\alpha(w|s) \geq 0$, $\alpha(s|w) \geq 0$,

$$\begin{aligned}
&\big[(\lambda(s) + \alpha(w|s))\varpi(s) - \alpha(s|w)\varpi(w)\big] = \max\{V_{ss}, 0\} + \max\{(V_{sw} + V_{ws})/2, 0\},\\
&\big[(\lambda(w) + \alpha(s|w))\varpi(w) - \alpha(w|s)\varpi(s)\big] = \max\{(V_{ws} + V_{sw})/2, 0\} + \max\{V_{ww}, 0\},\\
&p(1 - q_S)\upsilon_{ss} + (1 - p)(1 - q_M)\upsilon_{sw} \geq \varpi(s) - \epsilon\\
&p(1 - q_M)\upsilon_{ws} + (1 - p)(1 - q_W)\upsilon_{ww} \geq \varpi(w) - \epsilon,
\end{aligned}$$

where $\alpha(w|s)$ and $\alpha(s|w)$ are the dual variables for the IC constraints, and

$$\begin{aligned}
V_{ss} &= p\big[(\lambda(s) + \alpha(w|s))\upsilon_{ss} - \alpha(s|w)\upsilon_{ws}\big],\\
V_{sw} &= (1 - p)\big[(\lambda(s) + \alpha(w|s))\upsilon_{sw} - \alpha(s|w)\upsilon_{ww}\big],\\
V_{ws} &= p\big[(\lambda(w) + \alpha(s|w))\upsilon_{ws} - \alpha(w|s)\upsilon_{ss}\big],\\
V_{ww} &= (1 - p)\big[(\lambda(w) + \alpha(s|w))\upsilon_{ww} - \alpha(w|s)\upsilon_{sw}\big].
\end{aligned}$$

Then for any two-person bargaining problem Γ, there is a unique interim neutral mechanism. The following proposition summarizes my main result.

Proposition 2. *For any two-person bargaining problem Γ, when bargaining over mechanisms takes place at the almost ex-ante stage, the two players choose a unique almost ex-ante neutral mechanism that is best for the strong type but worst for the weak type among a continuum of symmetric almost ex-ante (or interim) incentive-efficient mechanisms. This selection is ex ante Pareto inferior to any other almost ex-ante incentive-efficient mechanisms.*

Proof. The proof follows from Theorem 1 in Kim (2017) and Theorem 3 in this paper. □

The reasoning behind Proposition 2 is as follows. If the selection is made at the almost ex-ante stage as described earlier, each player has uncertainty over whether the other player possesses private information. In such case, a player insisting heavily on the ex ante incentive-efficient mechanism could be taken as a signal of being the weak type, regardless of whether the player was the informed weak type or was actually uninformed. The informed

strong-type player will then be convinced to force the conflict outcome. Therefore, each player—whether strong, weak, or uninformed—would not want the other player to infer via his mechanism choice that he is weak. In some sense, both the informed weak-type player and the uninformed player would have incentives to conceal their "types." Accordingly, these players would mimic the strong type by choosing whatever the informed strong-type player would have chosen. Even if ε is very small so that there is only an infinitesimal probability εp that a player already knows that she is the strong type, the effect created by the early informed strong-type player who wants to break off from the ex ante incentive-efficient mechanism is influential on the players' behavior in bargaining over mechanisms. Thus each player would bargain for the mechanism that is most favorable to the strong type.

The ex ante efficient choice of mechanism is worst for the strong-type player but best for both the weak-type player and the uninformed player among all almost ex-ante incentive-efficient mechanisms. On the other hand, the almost ex-ante choice of mechanism characterized in Proposition 2 is ex ante worst for the players. But the proper welfare criterion to evaluate the selected mechanism depends on when bargaining over mechanisms takes place. Hence, the fact that players, if bargaining almost ex-ante, are not going to do as well as when bargaining ex ante in terms of the ex ante welfare criterion is not surprising by itself. Nor is the property of three types' preferences toward the ex ante choice interesting. However, the result that there is a difference between ex-ante versus almost ex-ante bargaining over mechanisms, not in terms of welfare evaluations per se, but in terms of the bargaining outcomes, is notable. An implication is that once there is even a very small possibility that some player may be informed of his type, the players will not choose the ex ante incentive-efficient mechanism. Hence, the ex-ante selected optimal mechanisms are not robust with respect to a small perturbation of the information structure at the bargaining stage.

5 Concluding Remarks

If there is even a small chance that the players might have learned their types at the stage of bargaining over mechanisms, then the selected mechanism might not be ex ante incentive-efficient. The analysis of ex-ante mechanism-selection in bargaining problems with incomplete information crucially depends on there being no doubt at all that nobody knows his or her type. If that doubt is there, the players may play on each other's doubt. Hence, the result under the assumption of ex-ante bargaining stage is not robust to a small perturbation of the information structure. This implies that ex ante efficiency can be seriously misleading as a solution concept for a theory of bargaining or as a welfare measure for evaluating mechanisms. Further, this paper reinforces the relevance of the interim bargaining solution suggested by Myerson (1984) to models of the process of agreeing on a mechanism.

References

Balkenborg, Dieter and Miltiadis Makris. 2015. "An Undominated Mechanism for a Class of Informed Principal Problems with Common Values." *Journal of Economic Theory* 157:918–958.

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.

de Clippel, Geoffroy and Enrico Minelli. 2004. "Two-Person Bargaining with Verifiable Information." *Journal of Mathematical Economics* 40:799–813.

Harsanyi, John C. 1967-8. "Games with Incomplete Information Played by 'Bayesian' Players." *Management Science* 14:159–189, 320–334, 348–502.

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.

Kim, Jin Yeub. 2017. "Interim Third-Party Selection in Bargaining." *Games and Economic Behavior* forthcoming.

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.

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

Maskin, Eric and Jean Tirole. 1992. "The Principal-Agent Relationship with an Informed Principal, II: Common Values." *Econometrica* 60(1):1–42.

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. "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.