A Noncooperative Foundation of the Neutral Bargaining Solution*

Jin Yeub Kim†

June 16, 2020

Abstract

This paper studies Myerson's neutral bargaining solution for a class of Bayesian bargaining problems in which the solution is unique. For this class of examples, I consider a noncooperative mechanism-selection game. I find that all of the interim incentive efficient mechanisms can be supported as sequential equilibria. Further, standard refinement concepts and selection criteria do not restrict the large set of interim Pareto-undominated sequential equilibria. I provide a noncooperative foundation of the neutral bargaining solution by characterizing the solution as a unique coherent equilibrium allocation.

JEL Classification: C71, C72, C78, D82,
Keywords: Neutral bargaining solution, mechanism-selection game, equilibrium refinement, equilibrium selection criterion, credibility, coherent plan

*I am deeply grateful to Roger Myerson, Lars Stole, and Ethan Bueno de Mesquita for their valuable advice, helpful conversations, and continuous support. I have benefited greatly from several discussions with Tiberiu Dragu, Alex Frankel, Johannes Hörner, Heung Jin Kwon, Kristof Madarasz, Adam Meirowitz, Roberto Serrano, Doron Ravid, and Richard van Weelden. I also appreciate the feedback from the seminar participants at various universities and conferences.

†School of Economics, Yonsei University, 50 Yonsei-ro Seodaemun-gu, Seoul 03722, Republic of Korea (email: shiningjin@gmail.com).

1. Introduction

In two seminal papers, Myerson (1983, 1984) axiomatically derived a concept of neutral bargaining solution for two-person bargaining problems with incomplete information. The general conditions for a neutral solution are well determined, but there is no general uniqueness theorem. Further, because the characterization of the solution concept takes place in an abstract context, it is hard to see what the concept entails for concrete non-cooperative procedures that could implement the neutral bargaining solution. The axioms that define the concept per se may justify the neutral bargaining solution, but one might still wonder in what exact conduit informational or strategic concerns lead to the solution. For these reasons, the neutral bargaining solution has been seldom used despite its analytical power as an interim solution concept in bargaining.

My goal in this paper is to provide an exact noncooperative foundation of the neutral bargaining solution. This is accomplished by modeling the mechanism-selection process as a suitably defined noncooperative game with a negotiation structure that yields the neutral bargaining solution as a unique equilibrium outcome. By doing so, we can identify the assumptions to make about the individualistic bargaining behavior, and clarify certain interpretive ambiguities in Myerson's axiomatic approach. This paper thus contributes to a clear understanding of the neutral bargaining solution concept proposed for cooperative games and provides a more solid justification for applications of the neutral bargaining solution in interim bargaining situations.

In this paper, I focus on a class of symmetric bargaining problems in which the concept of neutral bargaining solution uniquely selects the ex ante worst mechanism among all of the interim incentive efficient mechanisms. With two symmetric players, I can restrict attention to the informed principal's selection problem without loss of generality. Hence, I consider the following informed principal's mechanism-selection game: after the players learn their own types, the principal selects and announces

a mechanism; then the agent makes some inferences about the principal's type; and finally the mechanism is implemented with each player using some best response strategy given his information.

Using the concept of *expectational equilibria* (Myerson, 1983), I find that all of the interim incentive efficient mechanisms within the class of examples can be supported as sequential equilibria of the mechanism-selection game. In an attempt to obtain a sharp description of noncooperative behavior, I first investigate several equilibrium refinement concepts developed by Cho and Kreps (1987) and Banks and Sobel (1987): the Intuitive Criterion, D1, D2, Divinity, Universal Divinity. I find that those refinements do not restrict the large set of interim Pareto-undominated sequential equilibria. In general, the noncooperative analysis is limited by the fact that many games admit multiple equilibria. This issue remains in my game even after adding interim Pareto efficiency as well as imposing standard refinements to reduce the set of sequential equilibria.

Due the inability of these refinements adequately to eliminate equilibrium outcomes, I examine a criterion for selecting among equilibria based on the assumptions that players share a rich language for communication and that the principal's negotiation statements permit exogenous literal meanings. In particular, I employ Myerson's (1989) notion of *coherence* based on some credibility criterion. An announcement is said to be credible with respect to a reference allocation iff the announcement gives weakly higher expected payoffs than the payoffs from the reference allocation for all types of the principal. A reference allocation is then defined to be *attractive* iff it is the limit of a sequence of payoff allocations that admit essentially no credible announcements. The essential equilibrium-selection criterion that is to be applied is that an equilibrium that supports mechanism μ is selected among sequential equilibria if μ is credible with respect to some attractive reference allocation. I find that there is a unique sequential equilibrium that is selected by this criterion, and the selected equilibrium supports the neutral bargaining solution.

The coherence concept of restriction on equilibria together with the specification of the mechanism-selection game offers a precise non-cooperative characterization of players' bargaining behavior in implementing the neutral bargaining solution. In short, the standard of credibility in the concept of coherence incorporates the following idea of bargaining behavior: When a player deliberates which mechanism to negotiate for, he judges any deviation from an equilibrium with respect to some payoff allocations against which no mechanism could be credibly negotiated for; and so, a player must bargain for a mechanism that is "sufficiently" credible. This standard of credibility is stronger than the related notions of credible deviations proposed by Esö and Schummer (2009), Farrell (1993), and Grossman and Perry (1986). These three papers as well as other standard refinements specify the reference payoffs as expected equilibrium payoffs, which themselves may admit credible deviations.

The paper is organized as follows: Section 2 defines a class of Bayesian bargaining problems and characterizes the neutral bargaining solution. Section 3 presents a noncooperative mechanism-selection game and characterizes sequential equilibria. Section 4 concerns the standard equilibrium refinements and selection criteria, which fail to yield the neutral bargaining solution as a unique equilibrium outcome. Section 5 employs the concept of coherence formalized by a standard of credibility that uniquely delivers the neutral bargaining solution, and discusses a noncooperative foundation of the neutral bargaining solution. Section 6 offers concluding remarks. All of the proofs are in the Appendix.

2. Neutral Bargaining Solution

2.1. A Class of Bayesian 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 by Myerson (1984). A two-person Bayesian bargaining problem of my interest is

characterized by the following structures:

$$\Gamma = (D, d_0, T_1, T_2, u_1, u_2, \bar{p}_1, \bar{p}_2)\,,$$

whose components are interpreted as follows. $D = \{d_0, d_1\}$ is the set of feasible bargaining outcomes available to the two players where d_0 is the conflict outcome and d_1 is the agreement outcome. For each player i, $T_i = \{s, w\}$ is the set of possible types t_i. Let $T = T_1 \times T_2$ denote the set of all possible type combinations $t = (t_1, t_2)$. For each i, 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 chosen 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. I adopt a fully symmetric model in the sense that $T_1 = T_2$ and the payoffs do not depend on the identities of the players. Also the payoffs $(u_i)_{i\in\{1,2\}}$ satisfy the following assumptions: (i) $\sum_i u_i(d_1, t) > 0$, $\forall t$; and (ii) $u_i(d_1, t) < 0$ when $t_i = s$ and $t_{-i} = w$, $\forall i$. Lastly, I assume that the types are independently distributed with the probability distribution of i's type denoted by $\bar{p}_i$ 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.

Because I assume a symmetric model, I will often use p as the marginal probability of type s, which is common for both players. Further, I can simplify the notation by letting $\upsilon_{ss} \equiv u_1(d_1, s, s) = u_2(d_1, s, s)$, $\upsilon_{sw} \equiv u_1(d_1, s, w) = u_2(d_1, w, s)$, $\upsilon_{ws} \equiv u_1(d_1, w, s) = u_2(d_1, s, w)$, and $\upsilon_{ww} \equiv u_1(d_1, w, w) = u_2(d_1, w, w)$. Then these parameters satisfy $\upsilon_{ss} > 0$, $\upsilon_{ww} > 0$, $\upsilon_{ws} > 0$, $\upsilon_{sw} < 0$, and $\upsilon_{sw} + \upsilon_{ws} > 0$.

2.2. Interim Incentive Efficient Mechanisms

In the Bayesian bargaining problem Γ, the players can agree on some mechanism that specifies how the choice $d \in D$ should depend on the players' types. Formally, a 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 were the combination of the players' types. By restricting my attention to the symmetric mechanisms, I can simplify the notation by letting $q_S = \mu(d_0|s,s)$, $q_M = \mu(d_0|s,w) = \mu(d_0|w,s)$, and $q_W = \mu(d_0|w,w)$. By the revelation principle (Myerson, 1979), there is no loss of generality in focusing on direct-revelation mechanisms that are incentive compatible and individually rational in the sense of the following conditions:

$$\begin{aligned}
&p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw} \geq p(1-q_M)\upsilon_{ss} + (1-p)(1-q_W)\upsilon_{sw},\\
&p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww} \geq p(1-q_S)\upsilon_{ws} + (1-p)(1-q_M)\upsilon_{ww};\\
&p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw} \geq 0,\\
&p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww} \geq 0.
\end{aligned} \tag{2.1}$$

The concept of interim incentive efficiency can be applied to identify a set of "optimal" mechanisms among which the players should reasonably choose from in a setting in which each player already knows only his own type at the initial decision-making stage. Proposition 2 of Kim (2017) gives a complete characterization of the symmetric interim incentive efficient (IIE) mechanisms for the same class of bargaining problems considered in this paper. Hence, I omit the precise characterization formula, but let $S(\Gamma)$ denote the set of all IIE mechanisms for Γ. What is essential here is that, for a reasonably wide range of parameter values (i.e. $p/(1-p) < -\upsilon_{sw}/\upsilon_{ss} + \upsilon_{ww}/\upsilon_{ws}$), there is a continuum of IIE mechanisms in $S(\Gamma)$.[1] These IIE mechanisms essentially differ in the value of parameter q_M, while q_W is zero and q_S is determined by q_M for any IIE mechanism.

2.3. Neutral Bargaining Solution

To delimit further the set of mechanisms that the players could reasonably consider, the concept of *neutral bargaining solution* (Myerson, 1984) can be applied. The neu-

[1] When $p/(1-p) \geq -\upsilon_{sw}/\upsilon_{ss} + \upsilon_{ww}/\upsilon_{ws}$, there is a unique IIE mechanism.

tral bargaining solution is a generalization of the Nash bargaining solution for two-person bargaining problems with incomplete information. The neutral bargaining solutions form the smallest set satisfying three axioms: a probability-invariance axiom, an extension axiom, and a random-dictatorship axiom. The detailed expositions of these axioms can be found in Myerson (1984).

More importantly, Kim (2017) proves that for the class of environments considered here, the neutral bargaining solution concept gives a unique prediction of which mechanism would reasonably be selected by the two privately-informed players, which I call the *NBS mechanism*. This mechanism is associated with the highest value of q_M among all IIE mechanisms. Further, the NBS mechanism exhibits an interesting feature according to the interim welfare criterion, and can be compared to other IIE mechanisms in terms of the ex ante expected gains of the players. In particular, the NBS mechanism gives the highest interim expected utility for the strong type and the lowest interim expected utility for the weak type, among all of the IIE mechanisms, and is ex ante Pareto inferior to any other IIE mechanism.

Because the two players are symmetric in Γ, the solution that emerges when player 1 has all of the bargaining ability is the same as the solution that emerges when player 2 has all of the bargaining ability. This common solution to the informed principal's selection problem in the sense of *neutral optimum* (Myerson, 1983) is the NBS mechanism that is selected by the two players with equal bargaining ability. Hence, I can restrict attention to the informed principal's selection problem without losing the important features of the two-person bargaining problem Γ that essentially drive the uniqueness of the neutral bargaining solution.

3. Noncooperative Mechanism-Selection Game

To provide a noncooperative, individualistic foundation for the neutral bargaining solution obtained in the bargaining problem Γ, the first step is to model the informed principal's selection of a mechanism as part of a noncooperative game.

For any mechanism-selection game, the *inscrutability principle* (Myerson, 1983) implies that there is no loss of generality in restricting our attention to equilibria in which all types of any player should choose the same mechanism. If a player who selects an incentive feasible mechanism can "negotiate for both the inscrutable equilibrium of the mechanism-selection game and the honest participation equilibrium in the implementation of the mechanism itself" (Myerson, 1991, 511), then it seems reasonable to believe that the players will be able to implement an equilibrium that is Pareto-undominated within the set of equilibria. Hence, I assume that at the start of the interaction of any mechanism-selection game, the two players are given the set of IIE mechanisms, $S(\Gamma)$, from which a mechanism can be selected.

I consider a variation of the noncooperative mechanism-selection game introduced in Myerson (1983) to fit the class of environments in this paper. Because the players are symmetric, I refer to player 1 as the principal and player 2 as the agent without loss of generality. In the selection game, after the players learn their own types, the principal selects and announces a mechanism in $S(\Gamma)$.[2] Then the agent makes some inferences about the principal's type, based on this announcement; and finally the mechanism is implemented, with each player using some participation strategy that is rational for him given his information.

I closely follow the formulation of the idea of *expectational equilibria* that is developed in Myerson (1983), but the notations and definitions are modified. For each type t_1 of the principal, let $r(t_1)$ denote the conditional probability of the principal selecting mechanism $\delta \in S(\Gamma)$ when his type is t_1. A mechanism δ might have zero likelihood of being selected by each type of the principal so that $r = \mathbf{0}$. Hence for any vector $r = (r(s), r(w)) \in \mathbb{R}^{T_1}$, let Q be the normalized-likelihood vector corresponding to r, defined by $Q(t_1)\left(\sum_{s_1 \in T_1} r(s_1)\right) = r(t_1),\ \forall t_1 \in T_1$, such that $\sum_{s_1 \in T_1} Q(s_1) = 1$

[2] In Myerson (1983), the principal's selection is not restricted to a direct revelation mechanism on the interim Pareto-undominated frontier of the incentive feasible set. But this restriction does not drive the result.

and $Q(t_1) \geq 0$ for all $t_1 \in T_1$. Then for such vector Q, let

$$p_2^*(s|Q) = \frac{pQ(s)}{pQ(s) + (1-p)Q(w)} \text{ and } p_2^*(w|Q) = \frac{(1-p)Q(w)}{pQ(s) + (1-p)Q(w)}$$

be the posterior probabilities that the agent assigns to the principal's type being s and w respectively if the principal selected mechanism δ; and let

$$p_1^*(s|Q) = p \text{ and } p_1^*(w|Q) = 1 - p$$

be the posterior probabilities that the principal assigns to the agent's type being s and w respectively when the principal selects δ.

When the IIE mechanism δ is implemented, each player i will determine his reports and "actions" according to some participation strategy. In the Bayesian bargaining problem of the form Γ, the players are given two feasible bargaining outcomes – the agreement outcome and the conflict outcome; hence, each player's set of private "actions" is simply {"accept", "reject"}, where "accept" is interpreted as following the recommendation of mechanism and "reject" is interpreted as resorting to conflict. Then for any player i and any type t_i, I let $\psi_i(\hat{t}_i; t_i)$ denote the probability that i will resort to conflict if t_i is his true type but reported $\hat{t}_i$ and then received the recommendation of d_1 in the implementation of δ. For any $\hat{t}_i$ in T_i, I let $\tau_i(\hat{t}_i|t_i)$ denote the probability that i will report $\hat{t}_i$ when δ is implemented if his type is t_i. From these definitions, the quantities (ψ, τ) must be nonnegative and must satisfy $\sum_{\hat{t}_i \in T_i} \tau_i(\hat{t}_i|t_i) = 1$, $\forall i$, $\forall t_i \in T_i$, and $\psi_i(\hat{t}_i; t_i) \leq 1$, $\forall i$, $\forall \hat{t}_i \in T_i$, $\forall t_i \in T_i$. I denote player i's participation strategy by a pair (ψ_i, τ_i).

Let $W_i(\delta, \psi, \tau|t_i, Q)$ denote the expected utility for player i in the mechanism δ if his type is t_i, his posterior distribution given the principal's selection of δ is characterized by the normalized-likelihood vector Q, and $\psi = (\psi_1, \psi_2)$ and $\tau = (\tau_1, \tau_2)$

characterize the participation strategies of the two players. That is,

$$W_i(\delta, \psi, \tau | t_i, Q) = \sum_{t_{-i} \in T_{-i}} \sum_{\hat{t} \in T} p_i^*(t_{-i}|Q)\tau(\hat{t}|t)\delta(d_1|\hat{t})\pi(\hat{t};t)u_i(d_1,t)$$

where $\tau(\hat{t}|t) = \tau_1(\hat{t}_1|t_1)\tau_2(\hat{t}_2|t_2)$ and $\pi(\hat{t};t) = \big(1-\psi_1(\hat{t}_1;t_1)\big)\big(1-\psi_2(\hat{t}_2;t_2)\big)$. Then the participation strategies (ψ, τ) which the players would use in the implementation of δ form a Nash equilibrium for δ given Q if and only if $W_i(\delta, \psi, \tau|t_i, Q) \geq W_i(\delta, (\psi_{-i}, \psi_i'), (\tau_{-i}, \tau_i')|t_i, Q)$, $\forall i$, $\forall t_i \in T_i$, $\forall(\psi_i', \tau_i')$ satisfying the probability constraints. That is, every player's participation strategy maximizes the expected utility for each type given the other player's strategy.

Given the set of IIE mechanisms $S(\Gamma)$, a mechanism μ is said to be an *expectational equilibrium* iff $\mu \in S(\Gamma)$ and, for every mechanism $\gamma \in S(\Gamma)$, there exist Q, ψ, and τ satisfying the relevant probability constraints such that (ψ, τ) is a Nash equilibrium for δ given Q and $U_1(\mu|t_1) \geq W_1(\delta, \psi, \tau|t_1, Q)$, $\forall t_1 \in T_1$, where $U_1(\mu|s) = p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw}$ and $U_1(\mu|w) = p(1-q_M)\upsilon_{ws} + (1-p)(1-q_W)\upsilon_{ww}$. If μ is an expectational equilibrium, then it can be supported as a sequential equilibrium of the mechanism-selection game in which all types of the principal announce μ followed by truthful and obedient participation in μ by both players. The following proposition characterizes the set of sequential equilibria in this game.

Proposition 1. *For a given Bayesian bargaining problem Γ, every IIE mechanism in $S(\Gamma)$ can be supported as a sequential equilibrium of the informed principal's mechanism-selection game.*

Proposition 1 asserts that for every IIE mechanism, we can show a sequential equilibrium in which the principal inscrutably selects the mechanism followed by truthful revelation (and obedience) in the mechanism. Hence, the set of *interim Pareto-undominated sequential equilibria* (henceforth given the shorter name of sequential equilibria) in the mechanism-selection game is quite large for the class of Bayesian bargaining problems considered in this paper. Adopting different game forms and

equilibrium concepts that are considered in Cramton and Palfrey (1995), Holmström and Myerson (1983), and Maskin and Tirole (1992) also admits large numbers of equilibria.

This result is unfortunate in the sense that the behavior of the players that implicitly drives the unique selection of the neutral bargaining solution in the cooperative game cannot be pinned down only by analyzing a sequential equilibrium that supports the NBS mechanism in the noncooperative game. Natural questions that follow are whether there is a valid off-equilibrium belief refinement or equilibrium-selection criterion that only delivers the neutral bargaining solution out of the set of sequential equilibria; and if so, what kind of noncooperative restrictions would correspond to the cooperative mathematics in the deliverance of the unique neutral bargaining solution. Answering these questions will enable us to clarify a noncooperative, individualistic foundation of the cooperative concept of the neutral bargaining solution in relation to the exact refinement or selection criterion that uniquely delivers it.

4. Standard Restrictions on Equilibria

4.1. Equilibrium Refinements

I first attempt to delimit the set of sequential equilibria by investigating several equilibrium refinements. Cho and Kreps (1987) and Banks and Sobel (1987) develop a number of solution concepts that refine the set of sequential equilibria in many examples of signaling games. It is not my intention to repeat their development of formal definitions in detail, but to make this paper close to self-contained, I provide below a brief review of the alternative refinements.

The *Intuitive Criterion* (Cho and Kreps, 1987) requires that the agent's off-path-beliefs assign zero probability to those types of principal that cannot possibly gain by deviating, regardless of how the agent responds. An equilibrium fails the Intuitive Criterion if there exists some type of the principal who prefers to deviate to

announcing another mechanism which provides him with a higher payoff than his equilibrium announcement in every continuation for which the agent responds with a strategy that is optimal based on a belief that assigns zero probability to those types of principal that cannot gain from the deviation.

The *D1-Criterion* (Cho and Kreps, 1987) requires that zero weight be put on the type t when an off-the-equilibrium announcement is made if there exists another type t' such that t' always strictly benefits from the deviation whenever t benefits from deviation. The *D2-Criterion* (Cho and Kreps, 1987) requires that zero weight be put on the type t when an off-the-equilibrium announcement is made if, for every best response of the agent that causes t to deviate, there exists a t' (which may be different among responses) that wishes to strictly do so.

Closely related to Cho and Kreps's (1987) D1 and D2 restrictions on equilibria are Banks and Sobel's (1987) concepts of *divinity* and *universal divinity*. Divinity is roughly a weakening of D1, which requires that, rather than putting zero weight on types t satisfying D1, out-of-equilibrium beliefs to place relatively more weight on types that gain more from deviating from a fixed equilibrium. Universal divinity is an iterated application of D2, for which the updated beliefs do not depend on the prior.

For a given refinement criterion, roughly speaking, with the agent's off-path beliefs concentrated on some types that are not pruned, if those types gain by the deviation from an equilibrium, then the equilibrium is said to fail the refinement criterion. On the other hand, if an equilibrium can be supported by beliefs concentrated on types that survive the relevant restriction or if no type survives the restriction, then the equilibrium is said to survive the refinement criterion. Applying alternative refinements to sequential equilibria of the informed principal's mechanism selection game produces the following proposition.

Proposition 2. *Any sequential equilibrium of the mechanism-selection game survives the Intuitive, D1, and D2 criteria as well as the tests of divinity and universal divinity.*

The cooperative concept of neutral bargaining solution delineates other mechanisms as unreasonable equilibria, but those equilibria are not eliminated by the alternative equilibrium refinements considered above beyond what the concept of sequential equilibria characterizes. That is, these standard refinements have no analytical power in examining the noncooperative foundation of the neutral bargaining solution.

4.2. Equilibrium Selection Criterion

The standard equilibrium refinements have no bite, so a natural next step is to consider equilibrium selection criteria. In particular, I consider several notions of "credible deviations" proposed by Esö and Schummer (2009), Farrell (1993), and Grossman and Perry (1986).[3] Each of these work formalizes the concept of credibility that is a restriction of sequential equilibrium (or perfect Bayesian equilibrium) on the updating rule relative to a proposed equilibrium play of the game.

Briefly summarized, under Grossman and Perry's (1986) *perfect sequential equilibrium* (PSE), a set of types C (of the principal) breaks an equilibrium with an out-of-equilibrium announcement δ if all types in C improve their payoff by that announcement as long as the agent believes that all (and only) the types in C would always deviate and announce δ. Analogous to PSE, an equilibrium fails *neologism-proofness* (Farrell, 1993) if the types in C are precisely the ones who gain when, in response to the announcement, the agent's beliefs are a Bayesian update of his prior beliefs on C. The concept of neologism-proofs differs from PSE because it requires

[3] Although the concepts of Esö and Schummer (2009), Farrell (1993), and Grossman and Perry (1986) are commonly classified as refinement criteria in the literature, Myerson (1991, 241) emphasizes that we should draw a distinction between equilibrium refinements and criteria for selection among the set of equilibria. "A refinement $[\cdots]$ is a solution concept that is intended to offer a more accurate characterization of rational intelligent behavior in games. $[\cdots]$ A selection criterion is then any objective standard, defined in terms of the given structure of the mathematical game, that can be used to determine the focal equilibrium that everyone expects" (Myerson, 1991, 241). He then points out that the solution concepts by Farrell (1993) and Myerson (1989), which are based on the assumption that players share a rich natural language for communication, should be considered as equilibrium-selection criteria, rather than as equilibrium refinements.

an equilibrium to be supported by all, rather than one, credible updating rules for rationalizing observations off the equilibrium path. Related to PSE and neologism-proofness is Esö and Schummer's (2009) concept of *credible deviations*. Under their definition, an equilibrium is vulnerable to a credible deviation if, for some out-of-equilibrium announcement δ, there is a unique set C such that C is precisely the set of principal's types that would benefit from deviating to δ, whenever the agent plays any best response to δ with beliefs restricted to C.

Proposition 3. *All of the sequential equilibria in the mechanism-selection game are perfect sequential equilibria (Grossman and Perry, 1986), neologism-proof (Farrell, 1993), and immune to credible deviations in the sense of Esö and Schummer (2009).*

The uniqueness of the neutral bargaining solution is still not captured by those more "restrictive" equilibrium concepts, whether being considered as refinements or selection-criteria. Although the restrictions considered in this section do not sufficiently clarify the derived behavior underlying the neutral bargaining solution, they provide, at the very least, a necessary characterization of the individual behavior in the noncooperative implementation of the cooperative solution concept. In particular, by identifying what leads to the inability of the existing refinements adequately to refine the set of equilibria, I can determine what should *not* be included in the salient features of behavior, implied by the refinement criteria, that define the unique neutral bargaining solution. I relegate the relevant discussion to Section 5.

5. A Noncooperative Foundation

In this section, I invoke an equilibrium-selection criterion proposed by Myerson (1989), which uniquely selects the equilibrium that implements the NBS mechanism among many equilibria. I then investigate the noncooperative procedure that yields the neutral bargaining solution as its (focal) equilibrium outcome.

5.1. Coherent Equilibrium

Myerson's (1989) notion of *coherence* determined by some standard of credibility is in a similar spirit to Farrell's (1993) neologism proofness and Grossman and Perry's (1986) notion of PSE. The key difference lies on how the reference payoffs are determined. In particular, both of the latter papers implicitly equate reference payoffs with the equilibrium payoffs. On the other hand, Myerson (1989) defines the reference payoffs such that there is no credible offer against allocations that are arbitrarily close to that reference payoffs. I do not repeat the mathematical structures of a generalized model nor the details of the formal development of the concept introduced in Myerson (1989). For the sake of completeness, however, I provide below a review of the essential definitions that are adapted to my setting.

Let $w = (w(t_1))_{t_1 \in T_1}$ in $\mathbb{R}^{T_1}$ denote a reference payoff allocation. An announcement μ is *credible with respect to* w if and only if μ is incentive compatible and individually rational, μ is announced by at least one $t_1 \in T_1$, and $U_1(\mu|t_1) \geq w(t_1)$ for all t_1 that announce μ. An allocation w is *strongly attractive* if and only if there are no credible announcements with respect to w. An allocation w is *attractive* if and only if it is the limit of a sequence of strongly attractive allocations. The concept of coherence determines an announcement that satisfies the above credibility standard with respect to some attractive reference allocation. Formally, the equilibrium-selection criterion can be stated as follows.

Definition 1. An equilibrium that supports μ is selected among sequential equilibria of the informed principal's mechanism-selection game if and only if μ is incentive feasible and there exists some attractive reference allocation w such that $U_1(\mu|t_1) \geq w(t_1)$ for all $t_1 \in T_1$.

I call a sequential equilibrium that is selected by this criterion a *coherent* equilibrium, based on the development by Myerson (1989).[4]

[4]Myerson (1989) assumes that the players have an ability to communicate and the negotiation

Proposition 4. *There is a unique coherent equilibrium, which delivers the neutral bargaining solution, among all sequential equilibria of the mechanism-selection game.*

The equilibrium-selection criterion in the sense of Myerson (1989) selects a sequential equilibrium that supports the NBS mechanism as a focal equilibrium among sequential equilibria of the mechanism-selection game. Under the concept of coherence, the credibility criterion is formalized by the reference payoffs that are determined in a way that narrows the principal's range of credible announcements, in order to compel all types of the principal to announce the same mechanism. Hence, the NBS mechanism is the only mechanism that is credible with respect to a standard of credibility that admits essentially no other credible announcements.

An immediate observation of this criterion in relation to other notions of credibility is that the specification of reference allocation plays a crucial role in selecting among equilibria. By viewing the coherence concept as the credibility test for deviations, an announcement $\delta \in S(\Gamma) \setminus \{\mu\}$ is considered a credible deviation from an equilibrium outcome $\mu \in S(\Gamma)$ if and only if δ is credible with respect to some attractive allocation w for which any announcement is not credible with respect to allocations that are arbitrarily close to that reference payoffs. In contrast, according to Farrell (1993) and Grossman and Perry (1986), an announcement δ is a credible deviation from an equilibrium outcome μ if and only if δ is credible with respect to the expected equilibrium payoffs from μ given that the agent's beliefs are concentrated on those and only those types that gain from this deviation. If there exists no credible deviation in either sense, then the equilibrium outcome μ passes the credibility test.

In some sense, the coherence selection criterion may seem rather a strong restriction compared to other refinements or criteria widely used in practice (e.g. Intuitive

statements have literal meanings that every player can understand. Hence, negotiation is defined as any process of preplay communication between players that serves to influence the selection of a focal equilibrium that they will play thereafter in some game. With this "cooperative" nature, a statement (or announcement) μ is defined as a coherent plan iff every type of the negotiator surely makes that statement and there exists some attractive reference allocation w such that μ is credible with respect to w. Further, the coherent plans must be interim incentive efficient.

Criterion, D1, neologism-proofness, etc.). However, its ability to predict one mechanism that the principal should be expected to negotiate, no matter what his type is, leads us to obtain more precise description of the players' bargaining behavior in a noncooperative implementation of the neutral bargaining solution, thus providing the individualistic foundation for the cooperative solution concept.

5.2. Discussion

The noncooperative mechanism-selection game described in Section 3 and the concept of coherent equilibrium characterize a noncooperative foundation of the unique neutral bargaining solution for the class of bargaining problems. In particular, based on the specification of the mechanism-selection game and certain restrictions underlying the concept of coherence, I can identify the assumptions to make about the noncooperative procedure and players' bargaining behavior that drive the unique selection of the NBS mechanism.

The noncooperative implementation of the neutral bargaining solution embodies the negotiation structure based on the assumptions that the principal has an ability to make himself understood by the agent and the agent accepts the literal meaning of the principal's negotiation statements. Further, it incorporates the idea that the principal considers which negotiation statement in preplay communication should be credible. The credibility of a negotiation statement will determine the effective equilibrium that is actually played.

Farrell (1993) and Grossman and Perry (1986) also articulate the focal role of credible statements and signals, which have literal meanings, in selecting among equilibria. In contrast to what is implied by the concept of coherent equilibrium, however, a noncooperative implementation of the neutral bargaining solution is not captured by the solution concepts of those two papers as well as standard concepts in the refinements literature. This inability arises from the assumption that equates reference payoffs with expected equilibrium payoffs.

In my class of problems, and in general, each type of the principal has different preferences over the set of incentive compatible mechanisms. Hence, when any type is tempted to advocate out-of-equilibrium mechanism, there is a "plausible" inference that the agent could make about the principal's type after the announcement of any other mechanism that is incentive compatible for the principal, such that that mechanism would no longer be incentive compatible for the agent. The restrictions on the plausibility of posterior beliefs after deviations are determined by the criteria for assessing "deviations from a sequential equilibrium." In all of the previously examined equilibrium concepts other than the concept of coherence, expected payoffs from deviations are evaluated against expected payoffs from the equilibrium mechanism; whereas in Myerson (1989), deviating payoffs are evaluated against expected payoffs from the reference allocation that accepts one negotiation statement (that all types can make with likelihood one) but rejects all other statements that any type might prefer over it.

This distinction is the key to clarifying the characterization of the individual behavior in the noncooperative implementation of the neutral bargaining solution. Intuitively, when a player deliberates which mechanism to negotiate for in a mechanism-selection process, he should take into account whether his selection could be considered "sufficiently" credible in the sense that it is more profitable than some limit of allocations against which no mechanism could be credibly negotiated for. As a criterion for evaluating the credibility of possible negotiation statements, the concept of coherence provides one behavioral motivation for uniquely selecting the neutral bargaining solution in my class of examples.

Table 1 illustrates a comparison between the strategic incentives of the principal in the unique coherent equilibrium that implements the NBS mechanism and the inscrutable intertype compromise inherent in the cooperative concept of NBS. Under the cooperative approach, intertype compromise is considered in a virtual bargaining problem in which transferable virtual utilities, defined by the Lagrange multipliers, are

allocated equally among the players. The interim expected payoffs from the neutral bargaining solution should then be at least the virtually equitable utility allocations. Corresponding to the idea of intertype compromise is the credibility criterion imposed in the noncooperative game in the sense that the equilibrium payoffs should be at least some attractive reference payoffs.

Table 1: Connecting the intuitions

Cooperative concepts	Noncooperative foundation
Inscrutability principle	Pooling (sequential) equilibrium
Efficiency	Pareto-undominated frontier
Equity or intertype compromise	Credibility
↓	↓
Neutral bargaining solution	Coherent equilibrium

In the mechanism-selection game, the principal can send a signal that his is of a certain type indirectly through his strategy of proposing a particular mechanism. In the sequential equilibrium that supports the NBS mechanism, the rational behavior of the agent forces all types of the principal to propose the NBS mechanism. If the principal were to announce some other mechanism, then the agent may infer that the principal's type is weak. With this posterior expectation about such a zero-probability event, the agent uses his equilibrium participation strategy of "resorting to conflict" with probability one if he is strong and "reporting his type honestly and following the recommendations" if he is weak. This strategy leaves the principal no better off than when the NBS mechanism is implemented, no matter what her type may be. So both the strong and weak types of the principal would prefer to select the NBS mechanism than other alternatives.

However, the same reasoning applies to the sequential equilibrium that supports any other IIE mechanism. If the principal were to announce the NBS mechanism, for example, when the predicted equilibrium play is the mechanism that gives the highest interim payoff to the weak type, then the agent may infer that the principal's type is

strong. With these posterior beliefs, the agent would then lie in the implementation of the NBS mechanism, destroying the incentive compatibility of the NBS mechanism and dissuading the principal to make that deviation.

Beyond what the sequential equilibrium delineates about the individual behavior, the noncooperative foundation of the neutral bargaining solution relies on a stronger assumption about the derivation of behavior that is only underlying the focal equilibrium that implements the NBS mechanism. That is, the cooperative idea of inscrutable intertype compromise incorporates the noncooperative component of *credible signaling*: Every players "pretends" to be strong and inscrutably picks in a cooperative sense the mechanism that is best for the strong type, effectively pooling in a way such that no information is revealed. The nature of this implicit pooling towards the NBS mechanism is reminiscent of both types of the principal sending the same message by credibly announcing the NBS mechanism. In my setting, only the NBS mechanism can be considered a credible signal with respect to the attractive reference allocation. Thus, it is as if the players in the cooperative bargaining problem wanted to always credibly "signal" that they were the strong type. The combination of the inscrutability principle and the requirements for intertype compromise generates this kind of built-in credible signaling distortion in the mathematics of the cooperative concept.

6. Concluding Remarks

In this paper, I establish a noncooperative foundation of the neutral bargaining solution by selecting a unique equilibrium that supports the neutral bargaining solution mechanism in a well-defined noncooperative mechanism-selection game. A stronger version of credibility restrictions based on criteria proposed by Grossman and Perry (1986) and Farrell (1993) deliver the unique equilibrium, referred to as *coherent* equilibrium. By the equivalence between the unique neutral bargaining solution and the mechanism supported in the unique coherent equilibrium, I can justify the neutral

bargaining solution as a powerful interim bargaining solution concept without having to model the details of mechanism-selection games. Further, the equivalence enables us to carry the strategic intuitions behind the noncooperative equilibrium concept of mechanism-selection over the cooperative idea of inscrutable intertype compromise that arises in the deliverance of the neutral bargaining solution. For future research, a fruitful analysis would be to identify the exact counterparts of noncooperative concepts that correspond to the axioms that characterize the neutral bargaining solution. In any case, it is my opinion that the equivalence result should be taken as only a first step in an attempt to build a broader bridge between cooperative and noncooperative game theories in the context of bargaining with incomplete information.

It is worth comparing this paper's relationship to other papers on noncooperative bargaining models of "implementable" mechanisms. 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 all of these works 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 implementable. Imposing the concept of durability or security (non-ratifiability) does not rule out any IIE mechanism for the benchmark class of examples that I study in this paper. Thus those concepts would not give any stronger characterization of noncooperative implementation beyond the notion of interim incentive efficiency in my framework.

Appendix. Proofs

Proof of Proposition 1. Theorem 3 in Myerson (1983) proves the existence of at least one expectational equilibrium for any Bayesian incentive problem, and his Theorem 5 further asserts that any neutral optimum is an expectational equilibrium. Hence, the NBS mechanism in Γ is an expectational equilibrium; and this expectational equilibrium can be supported as a sequential equilibrium of the mechanism-selection game. Note that the NBS mechanism has the highest q_M among all IIE mechanisms given the primitives; that is, $q_M = 1$. To prove that any other IIE mechanism μ_q with $q = (q_S, q_M, q_W)$ where $q_M < 1$ is an expectational equilibrium, first suppose that the principal were trying to implement some other mechanism $\mu_{q'}$ where $q'_M < q_M$. With the posterior expectation characterized by $Q = (Q(s), Q(w)) = (0, 1)$ such that $p_2^*(s|Q) = 0$ and $p_2^*(w|Q) = 1$, the participation strategies $\psi_2(s; s) = \psi_2(w; s) = 1$, $\tau_2(s|s) = 1$, $\psi_2(w; w) = \psi_2(s; w) = 0$, and $\tau_2(w|w) = 1$ for the agent along with the honest and obedient participation for the principal form a Nash equilibrium for $\mu_{q'}$ given $(Q(s), Q(w)) = (0, 1)$. With these participation strategies (ψ, τ), we have $U_1(\mu_q|s) = p(1-q_S)\upsilon_{ss} + (1-p)(1-q_M)\upsilon_{sw} > W_1(\mu_{q'}, \psi, \tau|s, Q) = (1-p)(1-q'_M)\upsilon_{sw}$ where $\upsilon_{ss} > 0$, $\upsilon_{sw} < 0$, and $q'_M < q_M$. Also, $U_1(\mu_q|w) = p(1 - q_M)\upsilon_{ws} + (1 - p)(1 - q_W)\upsilon_{ww} > W_1(\mu_{q'}, \psi, \tau|w, Q) = (1 - p)(1 - q'_W)\upsilon_{ww}$ where $\upsilon_{ws} > 0$, $\upsilon_{ww} > 0$, and $q_W = q'_W = 0$ for any IIE mechanism. Now suppose that the principal were trying to implement some mechanism $\mu_{q'}$ other than μ_q where $q'_M > q_M$. Then the participation strategies $\psi_i(\hat{t}_i; s) = 1$, $\forall i$, $\forall \hat{t}_i \in T_i$, $\psi_i(\hat{t}_i, w) = 0$, $\forall i$, $\forall \hat{t}_i \in T_i$; and $\tau_1(s|s) = \tau_1(s|w) = \tau_2(s|s) = \tau_2(s|w) = 1$ constitute a Nash equilibrium for $\mu_{q'}$ given $Q = (Q(s), Q(w)) = (1, 0)$. Then, $W_1(\mu_{q'}, \psi, \tau|s, Q) = 0$ and $W_1(\mu_{q'}, \psi, \tau|w, Q) = (1 - p)\upsilon_{ww}$. Hence, $U_1(\mu_q|s) = p(1 - q_S)\upsilon_{ss} + (1 - p)(1 - q_M)\upsilon_{ws} \geq W_1(\mu_{q'}, \psi, \tau|s, Q)$ and $U_1(\mu_q|w) = p(1-q_M)\upsilon_{ws} + (1-p)\upsilon_{ww} > W_1(\mu_{q'}, \psi, \tau|w, Q)$ for any IIE mechanism $\mu_{q'}$ with $q'_M > q_M$ where $q_M < 1$. Therefore, given an IIE mechanism μ_q, for every IIE mechanism $\mu_{q'} \in S(\Gamma)$, we can find Q, ψ, and τ satisfying the probability constraints such that

(ψ, τ) is a Nash equilibrium for $\mu_{q'}$ given Q and $U_1(\mu_q|t_1) \geq W_1(\mu_{q'}, \psi, \tau|t_1, Q)$ for all $t_1 \in T_1$. By definition, an IIE mechanism μ_q is incentive compatible and individually rational. Therefore any IIE mechanism $\mu \in S(\Gamma)$ is an expectational equilibrium. $\square$

Proof of Proposition 2. I begin by checking that all sequential equilibria pass the Intuitive Criterion. I apply the Intuitive Criterion in two steps: First ask which types of the principal could benefit by deviating from the equilibrium announcement; then ask, if deviations can only come from the types identified in the first step, is the lowest payoff from deviating higher than the equilibrium payoff; if yes, then the equilibrium fails the Intuitive Criterion, otherwise survives. Formally, fix an equilibrium in which the principal announces $\mu \in S(\Gamma)$ and obtains utility $u^*(s) = U_1(\mu|s)$ if his type is s and $u^*(w) = U_1(\mu|w)$ if his type is w. For all $t_1 \in \{s, w\}$, the probability that t_1 announces any $\delta \in S(\Gamma) \setminus \{\mu\}$ is zero. For each out-of-equilibrium announcement δ, form the set $T'(\delta)$ consisting of all types t_1 such that $u^*(t_1) > \max_{(\psi,\tau)\in BR(T(\delta),\delta)} W_1(\delta, \psi, \tau|t_1, Q)$, where $BR(T(\delta), \delta)$ is the set of best response strategies that the players would use in the implementation of δ given the probability assessments concentrated on the set $T(\delta)$ of types of the principal who might have sent that announcement. Here, Q is such that $Q(t_1) > 0$ for all $t_1 \in T(\delta)$. If for any one announcement δ, there is some type $t' \in T_1$ such that $u^*(t') < \min_{(\psi,\tau)\in BR(T(\delta)\setminus T'(\delta),\delta)} W_1(\delta, \psi, \tau|t', Q)$, then the equilibrium is said to fail the Intuitive Criterion. In my setting, for any fixed equilibrium, if the principal announces some other δ associated with a higher interim expected payoff for the strong type, then this announcement is equilibrium dominated for type w, so $T'(\delta) = \{w\}$; but if the agent's beliefs are restricted to $T(\delta) \setminus T'(\delta) = \{s\}$ after the announcement δ, then the weak-type agent's best response would be to lie in the implementation of δ, hence type s cannot possibly gain by the deviation. If the principal announces some other δ associated with a higher interim expected payoff for the weak type, then this announcement is equilibrium dominated for type s, so the agent's beliefs would be concentrated on type w after the announcement; then the strong-type agent's best response would be resort to conflict, destroying any benefit

type w could have gotten from the deviation. Essentially, no type of the principal can ever gain by deviating when the agent response with a strategy that is optimal based on the beliefs concentrated on only that type. For the D1-criterion, the first step of its application asks which types of the principal are more likely to deviate from the equilibrium announcement (and the second step coincides with that of the Intuitive Criterion). In general, the types of the principal who benefit from deviating according to the D1-criterion are a subset of those who benefit from deviating according to the Intuitive Criterion. In my setting, for any deviation, there is at most one type t of the principal who could possibly gain from deviation, whereas the other type t' loses. That is, the set of best responses that would cause t' to deviate from the equilibrium (or to be indifferent) is empty. Hence by pruning the type t', with the agent's beliefs concentrated on t, the lowest payoff for type t from deviating is lower than the equilibrium payoff. Therefore, the Intuitive and D1- criteria coincide. Further, D2 differs from D1 in that some type that strictly wishes to defect whenever type t weakly wishes to defect may change with the response that causes t to defect. But with two types of the principal, whenever t deviates, regardless of any response that causes t to deviate, there is only one other type t' that could possibly strictly gain from the deviation. Therefore, D2 is equivalent to D1 in my game. In the game I study with only two types of the principal, the Intuitive, D1-, and D2- criteria coincide; further, universal divinity is essentially equivalent to the D2 refinement. Hence, the sequential equilibria that survive any concept of these refinements coincide. □

Proof of Proposition 3. First, I prove that all sequential equilibria survive to be perfect sequential equilibria. Fix a sequential equilibrium outcome μ. If the principal announces some other mechanism δ, then the agent will try to rationalize the deviation by identifying a (new) deviation belief that is consistent with the principal's incentive to deviate. Define the deviation set $C \subseteq T_1$ as those types that deviate with positive probability: $C = \{t_1 \in T_1 | Q(t_1) > 0\}$, where $Q(t_1)$ can be interpreted as the conditional probability of the principal deviating when his type is t_1.

The set C is required to be nonempty. A probability distribution $(p_2^*(s|Q), p_2^*(w|Q))$ on T_1 is a credible deviation belief about the principal relative to μ and δ if there exists a continuation equilibrium (ψ, τ) and deviating probabilities $Q(\cdot)$ such that $(p_2^*(s|Q), p_2^*(w|Q))$, (ψ, τ), and $Q(\cdot)$ together satisfy: (i) $Q(t_1) > 0$ for some $t_1 \in T_1$, (ii) $Q(t_1) = 1$ for all $t_1 \in T_1$ such that $U_1(\mu|t_1) < W_1(\delta, \psi, \tau|t_1, Q)$, (iii) $Q(t_1) = 0$ for all $t_1 \in T_1$ such that $U_1(\mu|t_1) > W_1(\delta, \psi, \tau|t_1, Q)$, and (iv) $(p_2^*(s|Q), p_2^*(w|Q))$ satisfies Bayes' rule, given the agent's priors $(p_2(s), p_2(w))$ and the principal's deviating probabilities $(Q(s), Q(w))$:

$$p_2^*(t_1|Q) = \begin{cases} \frac{p_2(t_1)Q(t_1)}{\sum_{t_1 \in C} p_2(t_1)Q(t_1)} & \text{for } t_1 \in C \\ 0 & \text{for } t_1 \notin C. \end{cases}$$

The set of types C that deviate with positive probability in $(p_2^*(s|Q), p_2^*(w|Q))$ is a credible deviation set. Then a sequential equilibrium that implements an IIE mechanism μ is perfect sequential equilibrium if for any deviation δ, (1) there does not exist a credible deviation belief, or (2) there exists a credible deviation belief $(p_2^*(s|Q), p_2^*(w|Q))$ and a corresponding equilibrium (ψ, τ) in the implementation of δ under beliefs $(p_2^*(s|Q), p_2^*(w|Q))$ such that $U_1(\mu|t_1) = W_1(\delta, \psi, \tau|t_1, Q)$ for all $t_1 \in C$. In my setting, the logic of the proof here is similar to that in the proof of Proposition 2. Briefly, fix an equilibrium outcome $\mu \in S(\Gamma)$. For any out-of-equilibrium announcement δ that gives a higher interim expected utility to the strong type of the principal, the deviation set is $C = \{s\}$. Then by the condition (iv), $p_2^*(s|Q) = 1$ and $p_2^*(w|Q) = 0$. These posterior beliefs of the agent in turn dissuade the strong type of the principal from deviating, because there is no continuation equilibrium in δ that would actually make the strong type principal strictly gain from the deviation; a contradiction to $Q(s) > 0$. A similar argument holds for any out-of-equilibrium announcement δ that gives a lower interim expected utility to the weak type of the principal with the deviation set being $C = \{w\}$. Hence, for any sequential equilib-

rium, there exist no credible veto belief for any deviation (or at the very least, the deviator-types are indifferent between deviating and not); so all sequential equilibria in my game are perfect sequential equilibria. For neologism-proof, the condition (1) remains the same for a sequential equilibrium to be neologism-proof, but the condition (2) changes to (2'): for every credible deviation belief $(p_2^*(s|Q), p_2^*(w|Q))$, $U_1(\mu|t_1) = W_1(\delta, \psi, \tau|t_1, Q)$ for all $t_1 \in C$. The two definitions of PSE and neologism-proofness coincide if there is at most one credible deviation belief for the principal, which trivially is the case in the class of examples studied in this paper. Therefore, all sequential equilibria are neologism-proof. Further, it is easy to see that the requirements for D1-Criterion and immunity to credible deviations are essentially the same in my setting, because there is always (and only) a single type (that does not vary according to responses) that could possibly benefit from deviating given the priors but would be dissuaded from deviating whenever the agent plays any best response with beliefs restricted to that type. Thus, all of the sequential equilibria are immune to credible deviations in the sense of Esö and Schummer (2009). □

Proof of Proposition 4. Myerson (1989) shows that there exists a sequence of warranted claims that satisfy the conditions that characterize the principal's neutral optima for an incentive compatible mechanism μ if and only if a sequence of strongly attractive allocations can converge to an attractive reference allocation that supports μ as a coherent plan. For the class of problems that I consider, the NBS mechanism is the unique neutral optimum for the principal. Hence for any mechanism δ in $S(\Gamma)$ other than the neutral optimum, there exists no sequence of warranted claims that satisfy the conditions for the characterization of neutral optima; this statement is equivalent to saying that a sequence of strongly attractive allocations cannot converge to an attractive reference allocation that supports δ as a coherent plan. That is, for any reference allocation w with respect to which δ is credible, there exist no sequence of strongly attractive vectors that converge to w. By definition of coherent plan (Myerson, 1989), any $\delta \in S(\Gamma)$ that is not the neutral op-

timum is not a coherent plan, because there exist no attractive reference allocation w such that δ is credible with respect to w. In the terminology of my formulation, only for the neutral optimum, there exists a sequence of warranted claims that satisfy the characterization conditions. Hence, the only attractive reference allocation is $w = (w(s), w(w)) = (U_1(\mu|s), U_1(\mu|w))$, where μ is the neutral optimum. For any equilibrium announcement $\delta \in S(\Gamma) \setminus \{\mu\}$, $U_1(\delta|s) < w(s)$; whereas for an equilibrium announcement μ, the conditions in the selection criterion are satisfied. Therefore, the coherence-credibility selection criterion uniquely selects the sequential equilibrium that supports the neutral optimum. □

References

Banks, Jeffrey S. and Joel Sobel. 1987. "Equilibrium Selection in Signaling Games." *Econometrica* 55(3):647–661.

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

Cho, In-Koo and David M. Kreps. 1987. "Signaling Games and Stable Equilibria." *Quarterly Journal of Economics* 102(2):179–222.

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

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

Farrell, Joseph. 1993. "Meaning and Credibility in Cheap-Talk Games." *Games and Economic Behavior* 5:514–531.

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

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

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* 102:645–665.

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. 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. 1989. "Credible Negotiation Statements and Coherent Plans." *Journal of Economic Theory* 48:264–303.

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