Mechanism Design with Maxmin Agents: Theory and an Application to Bilateral Trade[*]

Alexander Wolitzky

Stanford University

April 9, 2014

Abstract

This paper studies mechanism design when agents are maxmin expected utility maximizers. The first result gives a general necessary condition for a social choice rule to be implementable. The condition combines an inequality version of the standard envelope characterization of payoffs in quasilinear environments with an approach for relating maxmin agents' subjective expected utilities to their objective expected utilities under any common prior. The condition is then applied to give an exact characterization of when efficient trade is possible in the bilateral trading problem of Myerson and Satterthwaite (1983), under the assumption that agents know each other's expected valuation of the good (which is the information structure that emerges when the agents start with a common prior but are pessimistic about how the other agent might acquire information before participating in the mechanism). Whenever efficient trade is possible, it may be implemented by a relatively simple double auction format.

[*] For helpful comments, I thank Pierpaolo Battigalli, Aaron Bodoh-Creed, Subir Bose, Gabe Carroll, Aytek Erdil, Ben Golub, Fuhito Kojima, Massimo Marinacci, Paul Milgrom, Suresh Mutuswami, Phil Reny, Ilya Segal, Andy Skrzypacz, Joel Sobel, three anonymous referees, and participants in several seminars. I thank Jason Huang for excellent research assistance.

1 Introduction

"Robustness," broadly construed, has been a central concern in game theory and mechanism design since at least the celebrated argument of Wilson (1987). The Wilson doctrine is usually interpreted as calling for mechanisms that perform well in a wide range of environments or for a wide range of agent behaviors. However, there is also growing and complementary interest in robustness concerns on the part of agents instead of (or in addition to) on the part of the designer; that is, in asking what mechanisms are desirable when agents use "robustly optimal" strategies. This paper pursues this question in the case where agents are maxmin expected utility (MMEU) maximizers (Gilboa and Schmeidler, 1989) by developing a general necessary condition for a social choice rule to be implementable, and by applying it to give an exact characterization of when efficient trade is possible in the classical bilateral trade setting of Myerson and Satterthwaite (1983).

The necessary condition for implementation generalizes a well-known necessary condition in the Bayesian independent private values setting, namely that expected social surplus must exceed the expected sum of information rents left to the agents, as given by an envelope theorem. That this condition has any analogue with maxmin agents is rather surprising, for two reasons. First, the usual envelope characterization of payoffs need not hold with maxmin agents. Second, and more importantly, a maxmin agent's belief about the distribution of opposing types depends on her own type. This is also the situation with Bayesian agents and *correlated* types, where results are quite different than with independent types (Crémer and McLean, 1985, 1988; McAfee and Reny, 1992).

The derivation of the necessary condition (Theorem 1) address both of these issues. For the first, I derive an inequality version of the standard envelope condition that does hold with maxmin agents. The key step is replacing the partial derivative of an agent's allocation with respect to her type with the minimum value of this partial derivative over all possible beliefs of the agent. For the second, I note that, by definition, a maxmin agent's expected utility under her own subjective belief is lower than her expected utility under any belief she finds possible. This implies that the sum of agents' subjective expected utilities is lower than the sum of their "objective" expected utilities under any possible common prior, which in turn equals the expected social surplus under that prior (for a budget-balanced mechanism). Hence, a necessary condition for a social choice rule to be implementable is that the resulting expected social surplus exceeds the expected sum of information rents for *any* possible common prior; that is, for any prior with marginals that the agents find possible.

The second part of the paper applies this necessary condition to give an exact characterization of when efficient bilateral trade is implementable, under the assumption that the agents know each other's expected valuation of the good (as well as bounds on the valuations). As explained below, this information structure is the one that emerges when agents have a (unique) common prior on values at an ex ante stage and are maxmin about how the other might acquire information before entering the mechanism. In this setting, the assumption of maxmin behavior may be an appealing alternative to the Bayesian approach of specifying a prior over the set of experiments that the other agent may have access to, especially when this set is large (e.g., consists of all possible experiments) or the agents' interaction is one-shot. Furthermore, the great elegance of Myerson and Satterthwaite's theorem and proof suggests that their setting may be one where relaxing the assumption of a unique common prior is particularly appealing.[1]

The second main result (Theorem 2) shows that the Myerson-Satterthwaite theorem sometimes continues to hold when agents are maxmin about each other's information acquisition technology as described above—but sometimes not. In the simplest bilateral trade setting where the range of possible seller and buyer values is $[0,1]$, the average seller value is c^*, and the average buyer value is v^*, Figure 1 indicates the combination of parameters (c^*, v^*) for which an efficient, maxmin incentive compatible, interim individually rational, and weakly budget balanced mechanism exists. Above the curve—the formula for which is

$$\frac{c^*}{1-c^*}\log\left(1+\frac{1-c^*}{c^*}\right)+\frac{1-v^*}{v^*}\log\left(1+\frac{v^*}{1-v^*}\right)=1$$

[1] This is in line with Gilboa's exhortation in his monograph on decision making under uncertainty to "[consider] the MMEU model when a Bayesian result seems to crucially depend on the existence of a unique, additive prior, which is common to all agents. When you see that, in the course of some proof, things cancel out too neatly, this is the time to wonder whether introducing a little bit of uncertainty may provide more realistic results," (Gilboa, 2009, p.169).

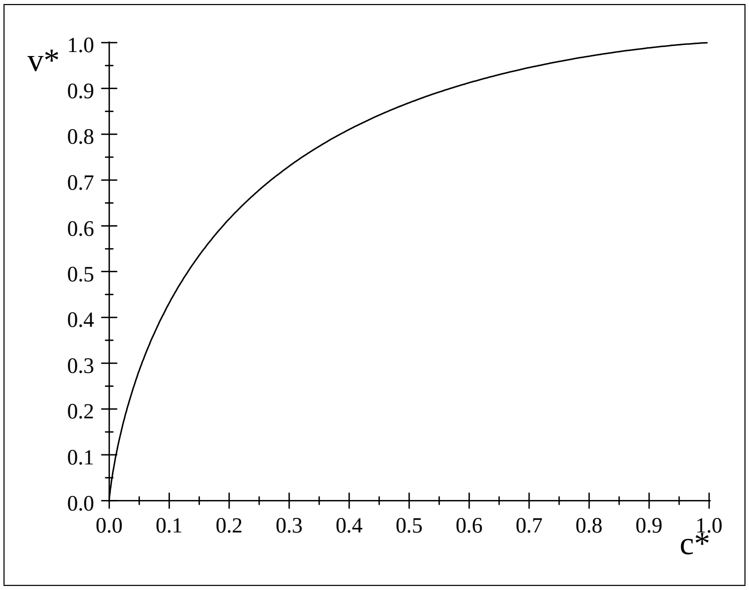

Figure 1: In the bilateral trade setting, efficient trade is possible in the region below the curve and impossible in the region above it.

—the Myerson-Satterthwaite theorem persists, despite the lack of a unique common prior or independent types. Below the curve, the Myerson-Satterthwaite theorem fails.

I call the mechanism that implements efficient trade for all parameters below the curve in Figure 1 the $\alpha_i(\theta_i)$ *double auction*. It is so called because when a type θ_i agent and a type θ_j agent trades, the type θ_i agent receives a share $\alpha_i(\theta_i)$ of the gains from trade that depends only on her own type and not on her opponent's. The $\alpha_i(\theta_i)$ double auction has the property that an agent's worst-case belief is the belief that minimizes the probability that strict gains from trade exist; this may be seen to be the belief that her opponent's type always takes on either the most favorable value for which there are no gains from trade or the most favorable value possible. If an agent misreports her type to try to get a better price, the requirement that her opponent's average value is fixed forces the deviator's worst-case belief to put more weight on the less favorable of these values, which reduces her expected probability of trade. The share $\alpha_i(\theta_i)$ is set so that this first-order cost in terms of probability of trade exactly offsets the first-order benefit in terms of price, which makes the $\alpha_i(\theta_i)$ double auction incentive compatible for maxmin agents. Finally, the $\alpha_i(\theta_i)$ double auction is weakly budget balanced if and only if $\alpha_1(\theta_1) + \alpha_2(\theta_2) \leq 1$ for all θ_1, θ_2; that is, if and only if the shares that must be left to the two agents sum to less than one.

I also consider several extensions and robustness checks in the bilateral trade setting. One observation here is that if the average types of the two agents do not have gains from trade with each other (e.g., if the pair (c^*, v^*) lies below the 45° line in Figure 1), then efficient trade can be implemented with a mechanism even simpler than the $\alpha_i(\theta_i)$ double auction, which I call a *reference rule*. A reference rule works by setting a "reference price" p^* and specifying that trade occurs at price p^* if this is acceptable to both agents, and that otherwise trade occurs (when efficient) at the reservation price of the agent who refuses to trade at p^*. The intuition for why references rules are incentive compatible with maxmin agents if and only if $c^* \geq v^*$ is simple and is given in the text.

This paper joins a growing literature on games and mechanisms with maxmin agents, or with agents who follow "robust" decision rules more generally. In contrast to much of this literature, the current paper shares the following important features of classical Bayesian mechanism design: (i) the implementation concept is (partial) Nash implementation; (ii) the only source of uncertainty in the model concerns exogenous random variables, namely other agents' types; and (iii) for Theorem 1, the model admits the possibility of a unique common prior as a special case. Several recent papers derive permissive implementability results with maxmin agents by relaxing these assumptions, in contrast to the relatively restrictive necessary condition of Theorem 1.

Bose and Daripa (2009), Bose and Mutuswami (2012), and Bose and Renou (2013) relax (i) by considering dynamic mechanisms that exploit the fact that maxmin agents may be time-inconsistent. A central feature of their approach is that agents cannot commit to strategies, so they do not obtain implementation in Nash equilibrium. Their approach also relies on taking a particular position on how maxmin agents update their beliefs, an issue which does not arise here. Di Tillio, Kos, and Messner (2012) and Bose and Renou (2013) relax (ii) by assuming that agents are maxmin over uncertain aspects of the mechanism itself. This lets the designer extract the agents' information by introducing "bait" provisions into the mechanism. The mechanisms considered in these four papers are undoubtedly interesting and may be appealing in particular applications. However, they arguably rely on a more thoroughgoing commitment to maxmin behavior than does the current paper (agents must be time-inconsistent, or must be maxmin over endogenous random variables). Even if one accepts this commitment, it still seems natural to ask what is possible in the more "standard" case where (i) and (ii) are satisfied.

De Castro and Yannelis (2010) relax (iii) by assuming that agents' beliefs are completely unrestricted, and find that efficient social choice rules are then always implementable. This is consistent with Theorem 1, as with completely unrestricted beliefs agents can always expect the worst possible allocation, which implies that the necessary condition of Theorem 1 is vacuously satisfied. For example, in the bilateral trade setting, efficient trade is always implementable, as agents are always certain that they will not trade and are therefore willing to reveal their types.

The model of Lopomo, Rigotti, and Shannon (2009) satisfies (i), (ii), and (iii), but considers agents with incomplete preferences as in Bewley (1986) rather than maxmin preferences. There are two natural versions of incentive compatibility in their model, which bracket maxmin incentive compatibility (and Bayesian incentive compatibility) in terms of strength. Lopomo, Rigotti, and Shannon show that the stronger of their notions of incentive compatibility is often equivalent to ex post incentive compatibility (whereas maxmin incentive compatibility is not), and that full extraction of information rents is generically possible under the weaker of their notions, and is sometimes possible under the stronger one.

Finally, Bodoh-Creed (2012) satisfies (i), (ii), and (iii) and considers maxmin agents. Bodoh-Creed's main result provides conditions for an exact version of the standard envelope characterization of payoffs to apply with maxmin agents, based on Milgrom and Segal's (2002) envelope theorem for saddle point problems. As discussed below, these conditions are not satisfied in my setting. Bodoh-Creed also does not derive a general necessary condition for implementability like Theorem 1, and his treatment of applications (including bilateral trade) focuses not on efficiency but on revenue-maxmization using full-insurance mechanisms, as in Bose, Ozdenoren, and Pape (2006).

The paper proceeds as follows. Section 2 presents the model. Section 3 gives the general necessary conditions for implementation. Section 4 applies these conditions to characterize when efficient bilateral trade is implementable. Section 5 contains additional results in the bilateral trade setting. Section 6 concludes. The appendix contains omitted proofs and auxiliary results.

2 Model

Agents and Preferences: A group N of n agents must make a social choice from a bounded set of alternatives $Y \subseteq \mathbb{R}^n$. Each agent i has a one-dimensional type $\theta_i \in \left[\underline{\theta}_i, \bar{\theta}_i\right] = \Theta_i \subseteq \mathbb{R}$, and agents have quasilinear utility. In particular, if alternative $y = (y_1, \ldots, y_n)$ is selected and a type θ_i agent receives transfer t_i, her payoff is

$$\theta_i y_i + t_i.^2$$

Agent i's type is her private information. In addition, each agent i has a *set of possible beliefs* Φ_{-i} about her opponents' types, where Φ_{-i} is an arbitrary nonempty subset of $\Delta\left(\Theta_{-i}\right)$, the set of Borel measures ϕ_{-i} on Θ_{-i} (throughout, probability measures are denoted by ϕ, and the corresponding cumulative distribution functions are denoted by F). Each agent i evaluates her expected utility with respect to the worst possible distribution of her opponents' types among those distributions in Φ_{-i}; that is, the agents are maxmin optimizers.

Mechanisms: A *direct mechanism* (y, t) consists of a measurable allocation rule $y : \Theta \to Y$ and a measurable and bounded transfer rule $t : \Theta \to \mathbb{R}^n$. Given a mechanism (y, t), let

$$\begin{aligned}
U_i\left(\hat{\theta}_i, \theta_{-i}; \theta_i\right) &= \theta_i y_i\left(\hat{\theta}_i, \theta_{-i}\right) + t_i\left(\hat{\theta}_i, \theta_{-i}\right), \\
U_i\left(\hat{\theta}_i, \phi_{-i}; \theta_i\right) &= E^{\phi_{-i}}\left[U_i\left(\hat{\theta}_i, \theta_{-i}; \theta_i\right)\right], \\
U_i\left(\theta_i\right) &= \inf_{\phi_{-i} \in \Phi_{-i}} U_i\left(\theta_i, \phi_{-i}; \theta_i\right).
\end{aligned}$$

Thus, $U_i\left(\hat{\theta}_i, \theta_{-i}; \theta_i\right)$ is agent i's utility from reporting type $\hat{\theta}_i$ against opposing type profile θ_{-i} given true type θ_i, $U_i\left(\hat{\theta}_i, \phi_{-i}; \theta_i\right)$ is agent i's expected utility from reporting type $\hat{\theta}_i$ against belief ϕ_{-i} given true type θ_i, and $U_i\left(\theta_i\right)$ is agent i's worst-case expected utility from reporting her true type θ_i.[3]

[2] The assumption that utility is multiplicative in θ_i and y_i is for simplicity. One could instead assume that utility equals $v_i\left(y, \theta_i\right) + t_i$ for some absolutely continuous and equidifferentiable family of functions $\left\{v_i\left(y, \cdot\right)\right\}$, as in Milgrom and Segal (2002) or Bodoh-Creed (2012).

[3] The term "worst-case" is only used heuristically in this paper, but the meaning is generally that if $\min_{\phi_{-i} \in \Phi_{-i}} U_i\left(\hat{\theta}_i, \phi_{-i}; \theta_i\right)$ exists, then a minimizer is a worst-case belief; while if the minimum does not

A distinguishing feature of this paper is the notion of incentive compatibility employed, which I call *maxmin incentive compatibility.* A mechanism is maxmin incentive compatible (MMIC) if

$$\theta_i \in \arg\max_{\hat{\theta}_i \in \Theta_i} \inf_{\phi_{-i} \in \Phi_{-i}} U_i\left(\hat{\theta}_i, \phi_{-i}; \theta_i\right) \text{ for all } \theta_i \in \Theta_i, i \in N. \tag{1}$$

I restrict attention to MMIC direct mechanisms throughout the paper. The motivation for doing so is provided in Appendix B, where I prove the appropriate version of the revelation principle. In particular, I show that when agents are maxmin optimizers, any Nash implementable social choice rule is MMIC.[4] The main assumption underlying this result is that an agent cannot "commit to randomize," meaning that mixed strategies are evaluated according to the maxmin criterion realization-by-realization. If instead agents could commit to randomize, then restricting attention to MMIC mechanisms in the sense of (1) would no longer be without loss of generality. Assuming that agents cannot commit to randomize seems as natural as the alternative in most contexts, and is also the more common assumption in the literature on mechanism design with maxmin agents (e.g., Bose, Ozdenoren, and Pape, 2006; de Castro and Yannelis, 2010; Bodoh-Creed, 2012; Di Tillio, Kos, and Messner, 2012).

In addition to MMIC, I also consider the following standard mechanism design criteria.

- *Ex Post Efficiency (EF):* $y(\theta) \in \arg\max_{y \in Y} \sum_i \theta_i y_i$ for all $\theta \in \Theta$.
- *Interim Individual Rationality (IR):* $U_i(\theta_i) \geq 0$ for all $\theta_i \in \Theta_i$.
- *Ex Post Weak Budget Balance (WBB):* $\sum_i t_i(\theta) \leq 0$ for all $\theta \in \Theta$.
- *Ex Post Strong Budget Balance (SBB):* $\sum_i t_i(\theta) = 0$ for all $\theta \in \Theta$.

Efficiency is self-explanatory. Interim individual rationality is imposed with respect to agents' own worst-case beliefs; in addition, all results in the paper continue to hold with ex post individual rationality (i.e., $U_i(\theta_i, \theta_{-i}; \theta_i) \geq 0$ for all $\theta_i \in \Theta_i, \theta_{-i} \in \Theta_{-i}$). The ex post

exist (which is possible, as $U_i\left(\hat{\theta}_i, \phi_{-i}; \theta_i\right)$ may not be continuous in ϕ_{-i}), then a limit point of a sequence $\{\phi_{-i}\}$ that attains the infimum is a worst-case belief.

[4] The solution concept throughout the paper is thus Nash equilibrium (this is made explicit in Appendix B and is captured implicitly by (1) in the text). In particular, there is no strategic uncertainty or "higher order ambiguity" (as in Ahn, 2007).

version of budget balance seems appropriate, since there is no unique common prior. The difference between weak and strong budget balance is that with weak budget balance the mechanism is allowed to run a surplus.

An allocation rule y is *maxmin implementable* if there exists a transfer rule t such that the mechanism (y, t) satisfies MMIC, IR, and WBB.

3 Necessary Conditions for Implementation

My most general result is a necessary condition for maxmin implementation, which generalizes a standard necessary condition for Bayesian implementation with independent private values. In particular, in an independent private values environment with common prior distribution F, it is well-known that an allocation rule y is Bayesian implementable only if the expected social surplus under y exceeds the expected information rents that must be left for the agents in order to satisfy incentive compatibility. It follows from standard arguments that this condition may be written as

$$\sum_i \left(\int_{\theta \in \Theta} \theta_i y_i(\theta)\, d\phi \right) \geq \sum_i \left(\int_{\theta_i \in \Theta_i} (1 - F_i(\theta_i))\, y_i\left(\theta_i, \phi_{-i}\right) d\theta_i \right), \tag{2}$$

where

$$y_i\left(\theta_i, \phi_{-i}\right) = E^{\phi_{-i}}\left[y_i(\theta_i, \theta_{-i}) \right]$$

(recall that ϕ is the measure corresponding to cdf F). I will show that a similar condition is necessary for maxmin implementation, despite the lack of independent types or a unique common prior. Intuitively, the required condition will be that (2) holds *for all distributions F with marginals that the agents find possible*, with the modification that, on the right-hand side of (2), the expected information rents under F are replaced by the expectation under F of type θ_i's minimum possible information rent.

I begin by formalizing these concepts. Given a measure $\phi \in \Delta(\Theta)$, let ϕ_S denote its marginal with respect to Θ_S, for $S \subseteq N$. Let Φ^* be the set of measures $\phi \in \Delta(\Theta)$ such that, for all $i \in N$, the marginals ϕ_i and ϕ_{-i} are pairwise independent and $\phi_{-i} \in \Phi_{-i}$. Some examples may clarify this definition.

- If $n = 2$, then $\Phi^* = \Phi_1 \times \Phi_2$ (where $\Phi_i \equiv \Phi_{-j}$).

- Suppose the set of each agent i's possible beliefs takes the form of a product $\Phi_{-i} = \prod_{j \neq i} \Phi^i_j$ for some sets of measures $\Phi^i_j \subseteq \Delta(\Theta_j)$ (in particular, each agent believes that her opponents' types are independent). Then $\Phi^* = \prod_{j \in N} \left(\bigcap_{i \neq j} \Phi^i_j \right)$.

- If $n > 2$, it is possible that Φ^* is empty. For instance, take the previous example with $\bigcap_{i \neq j} \Phi^i_j = \emptyset$ for some j.

Finally, let

$$\tilde{y}_i(\theta_i) = \inf_{\phi_{-i} \in \Phi_{-i}} y_i(\theta_i, \phi_{-i}).$$

Thus, $\tilde{y}_i(\theta_i)$ is the smallest allocation that type θ_i may expected to receive.

The following result gives the desired necessary condition.

Theorem 1 *If allocation rule y is maxmin implementable, then, for every measure $\phi \in \Phi^*$,*

$$\sum_i \left(\int_{\theta \in \Theta} \theta_i y_i(\theta)\, d\phi \right) \geq \sum_i \left(\int_{\theta_i \in \Theta_i} (1 - F_i(\theta_i))\, \tilde{y}_i(\theta_i)\, d\theta_i \right). \tag{3}$$

Comparing (2) and (3), (2) says that the expected social surplus under F must exceed the expected information rents, whereas (3) says that the expected social surplus must exceed the expectation of the agents' "minimum possible" information rents, reflecting the fact that agents' subjective expected allocations are not derived from F. In addition, (2) must hold only for the "true" distribution F (i.e., the common prior distribution), while (3) must hold for any "candidate" distribution F (i.e., any distribution in Φ^*). Furthermore, (3) is a generalization of (2), as in the case of a unique independent common prior ϕ, it follows that $\Phi_{-i} = \{\phi_{-i}\}$ for all i, $\Phi^* = \{\phi\}$, and $\tilde{y}_i(\theta_i) = y_i(\theta_i, \phi_{-i})$, so (3) reduces to (2). Finally, if y is continuous then (3) also shows that (2) changes continuously as a slight degree of ambiguity aversion is introduced into a Bayesian model.

The differences between (2) and (3) suggest that maxmin implementation is neither easier nor harder than Bayesian implementation in general, and that more generally expanding the sets of possible beliefs Φ_{-i} can make implementation either easier or harder. In particular, expanding the sets Φ_{-i} expands Φ^*, which implies that (3) must hold for a larger set of measures ϕ. On the other hand, expanding Φ_{-i} also reduces $\tilde{y}_i(\theta_i)$, and thus reduces the right-hand side of (3), making (3) easier to satisfy. Indeed, Section 4 shows that efficient bilateral trade is sometimes maxmin implementable in cases where the Myerson-Satterthwaite theorem implies that it is not Bayesian implementable, which demonstrates that maxmin implementation can be easier than Bayesian implementation. But it is also easy to find examples where maxmin implementation is harder than Bayesian implementation. For instance, take a Bayesian bilateral trade setting where all types are certain that gains from trade exist—so that efficient trade is implementable—and expand the set of possible beliefs by adding a less-favorable prior for which the Myerson-Satterthwaite theorem applies. Condition (3) will then imply that efficient trade is not implementable, by exactly the same argument as in Myerson-Satterthwaite.

The first step in proving Theorem 1 is deriving an inequality version of usual envelope characterization of payoffs, Lemma 1. Lemma 1 is related to Theorem 1 of Bodoh-Creed (2012), which gives an exact characterization of payoffs using Milgrom and Segal's (2002) envelope theorem for saddle point problems. The difference comes because the maxmin problem (1) admits a saddle point in Bodoh-Creed but *not* in the present paper; one reason why is that Bodoh-Creed assumes that $y_i\left(\theta_i, \phi_{-i}\right)$ is continuous in θ_i and ϕ_{-i} (his assumption A8), which may not be the case here.[5] For example, efficient allocation rules are not continuous, so Bodoh-Creed's characterization need not apply for efficient mechanisms. I discuss below how Theorem 1 may be strengthened if (1) is assumed to admit a saddle point.[6]

Lemma 1 *In any maxmin incentive compatible mechanism,*

$$U_i(\theta_i) \geq U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \tilde{y}_i(s)\, ds \text{ for all } \theta_i \in \Theta_i. \tag{4}$$

Proof. As Y is bounded, Theorem 2 of Milgrom and Segal (2002) may be applied to

[5] A careful reading of Bodoh-Creed (2012) reveals that some additional assumptions are also required for the existence of a saddle point, such as quasiconcavity assumptions. Bodoh-Creed (2014) provides an alternative derivation of his payoff characterization result using additional continuity assumptions in lieu of assumptions guaranteeing the existence of a saddle point. Neither set of assumptions is satisfied in the current setting.

[6] The present approach of bounding $U_i(\theta_i)$ also bears some resemblence to Segal and Whinston (2002), Carbajal and Ely (2013), or Kos and Messner (2013), with the difference that the difficulty here is not relating $U_i(\theta_i)$ to $\int_{\underline{\theta}_i}^{\theta_i} U_i'(\theta)\, d\theta$, but rather relating $U_i'(\theta)$ to (bounds on) $y\left(\theta, \phi_{-i}\right)$.

the problems $\inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\hat{\theta}_i,\phi_{-i};\theta_i\right)$ and $\max_{\hat{\theta}_i\in\Theta_i}\inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\hat{\theta}_i,\phi_{-i};\theta_i\right)$ to show that $U_i(\theta_i)$ is absolutely continuous. Hence, $U_i(\theta_i)$ is differentiable almost everywhere, and $U_i(\theta_i)=U_i(\underline{\theta}_i)+\int_{\underline{\theta}_i}^{\theta_i}U_i'(s)\,ds$. Given a point of differentiability $\theta_i^0\in\left(\underline{\theta}_i,\bar{\theta}_i\right)$, fix a sequence $\left\{\theta_i^k\right\}$ converging to θ_i^0 from above. Then

$$\begin{aligned}
U_i\left(\theta_i^k\right)-U_i\left(\theta_i^0\right) &\geq \inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\theta_i^0,\phi_{-i};\theta_i^k\right)-\inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\theta_i^0,\phi_{-i};\theta_i^0\right)\\
&= \inf_{\phi_{-i}\in\Phi_{-i}}\left(U_i\left(\theta_i^0,\phi_{-i};\theta_i^0\right)+\left(\theta_i^k-\theta_i^0\right)y\left(\theta_i^0,\phi_{-i}\right)\right)-\inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\theta_i^0,\phi_{-i};\theta_i^0\right)\\
&\geq \inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\theta_i^0,\phi_{-i};\theta_i^0\right)+\inf_{\phi_{-i}\in\Phi_{-i}}\left(\theta_i^k-\theta_i^0\right)y\left(\theta_i^0,\phi_{-i}\right)-\inf_{\phi_{-i}\in\Phi_{-i}} U_i\left(\theta_i^0,\phi_{-i};\theta_i^0\right)\\
&= \inf_{\phi_{-i}\in\Phi_{-i}}\left(\theta_i^k-\theta_i^0\right)y\left(\theta_i^0,\phi_{-i}\right)\\
&\geq \left(\theta_i^k-\theta_i^0\right)\tilde{y}\left(\theta_i^0\right).
\end{aligned}$$

Hence, $\frac{U_i\left(\theta_i^k\right)-U_i\left(\theta_i^0\right)}{\theta_i^k-\theta_i^0}\geq\tilde{y}\left(\theta_i^0\right)$, and therefore $U_i'\left(\theta_i^0\right)\geq\tilde{y}_i\left(\theta_i^0\right)$. The result follows as $U_i(\theta_i)=U_i(\underline{\theta}_i)+\int_{\underline{\theta}_i}^{\theta_i}U_i'(s)\,ds$. ■

The remaining step in the proof of Theorem 1 relates the bound on agents' subjective expected utilities in (4) to the objective social surplus on the left-hand side of (3). The key reason why this is possible is that a maxmin agent's subjective expected utility is a lower bound on her expected utility under any probability distribution she finds possible. Hence, the sum of agents' subject expected utilities is a lower bound on the sum of their objective expected utilities under any measure $\phi\in\Phi^*$, which in turn is a lower bound on the objective expected social surplus under ϕ (if weak budget balance is satisfied). Note that this step relies crucially on the assumption that agents are maxmin optimizers; for example, it would not apply for Bayesian agents with heterogeneous priors.

Proof of Theorem 1. Suppose mechanism (y,t) satisfies MMIC, IR, and WBB. For any measure $\phi_i\in\Delta(\Theta_i)$, integrating (4) by parts yields

$$\int_{\Theta_i}U_i(\theta_i)\,d\phi_i\geq U_i(\underline{\theta}_i)+\int_{\Theta_i}(1-F_i(\theta_i))\,\tilde{y}_i(\theta_i)\,d\theta_i.$$

Recall that

$$U_i(\theta_i)\leq\int_{\Theta_{-i}}(\theta_i y_i(\theta)+t_i(\theta))\,d\phi_{-i}\text{ for all }\phi_{-i}\in\Phi_{-i}.$$

Combining these inequalities implies that, for every measure $\phi = \phi_i \times \phi_{-i} \in \Delta(\Theta_i) \times \Phi_{-i}$,

$$\int_{\Theta_i} \int_{\Theta_{-i}} (\theta_i y_i(\theta) + t_i(\theta))\, d\phi_{-i} d\phi_i \geq U_i(\underline{\theta}_i) + \int_{\Theta_i} (1 - F_i(\theta_i))\, \tilde{y}_i(\theta_i)\, d\theta_i,$$

or

$$\int_{\Theta} (\theta_i y_i(\theta) + t_i(\theta))\, d\phi \geq U_i(\underline{\theta}_i) + \int_{\Theta_i} (1 - F_i(\theta_i))\, \tilde{y}_i(\theta_i)\, d\theta_i. \tag{5}$$

Note that every measure $\phi \in \Phi^*$ is of the form $\phi_i \times \phi_{-i} \in \Delta(\Theta_i) \times \Phi_{-i}$ for each i. Thus, for every $\phi \in \Phi^*$, summing (5) over i yields

$$\sum_i \left(\int_{\Theta} (\theta_i y_i(\theta) + t_i(\theta))\, d\phi \right) \geq \sum_i U_i(\underline{\theta}_i) + \sum_i \left(\int_{\Theta_i} (1 - F_i(\theta_i))\, \tilde{y}_i(\theta_i)\, d\theta_i \right).$$

Finally, $\sum_i U_i(\underline{\theta}_i) \geq 0$ by IR and $\sum_i \int_{\Theta} t_i(\theta)\, d\phi \leq 0$ by WBB, so this inequality implies (3). ■

If the maxmin problem (1) admits a saddle point $\left(\hat{\theta}_i^*(\theta_i), \phi_{-i}^*(\theta_i)\right)$ (where MMIC implies that $\hat{\theta}_i^*(\theta_i) = \theta_i$), then, letting

$$y_i^*(\theta_i) = y_i\left(\theta_i, \phi_{-i}^*(\theta_i)\right)$$

be type θ_i's expected allocation under her worst-case belief $\phi_{-i}^*(\theta_i)$, Theorem 4 of Milgrom and Segal (2002) or Theorem 1 of Bodoh-Creed (2012) implies that (4) may be strengthened to

$$U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} y_i^*(s)\, ds \text{ for all } \theta_i \in \Theta_i. \tag{6}$$

The same argument as in the proof of Theorem 1 then implies that the necessary condition (3) may be strengthened to

$$\sum_i \left(\int_{\theta \in \Theta} \theta_i y_i(\theta)\, d\phi \right) \geq \sum_i \left(\int_{\theta_i \in \Theta_i} (1 - F_i(\theta_i))\, y_i^*(\theta_i)\, d\theta_i \right).$$

Thus, if the maxmin problem admits a saddle point then a necessary condition for maxmin implementation is that, for every measure $\phi \in \Phi^*$, the expected social surplus under ϕ exceeds the expectation under ϕ of the sum of the agents' subjective information rents (i.e., the information rents under the worst-case beliefs $\phi^*_{-i}(\theta_i)$). Unfortunately, I am not aware of sufficient conditions on the allocation rule y alone that ensure the existence of a saddle point. Applying the Debreu-Fan-Glicksberg fixed point theorem to (1), a sufficient condition on the mechanism (y, t) is that y and t are continuous in θ and $U_i\left(\hat{\theta}_i, \theta_{-i}; \theta_i\right)$ is quasiconcave in $\hat{\theta}_i$ and quasiconvex in θ_{-i}.[7]

A well-known necessary condition for an *efficient* allocation rule to be Bayesian implementable is that an individually rational Groves mechanism runs an expected surplus (Makowski and Mezzetti, 1994; Williams, 1999; Krishna and Perry, 2000). This follows because the standard envelope characterization of payoffs implies that the interim expected utility of each type in any efficient and Bayesian incentive compatible mechanism is the same as her interim expected utility in a Groves mechanism. However, this result does *not* go through with maxmin incentive compatibility, even if (1) admits a saddle point. This is because the envelope characterization of payoffs with MMIC, (6), depends on types' expected allocations under their worst-case beliefs $\phi^*_{-i}(\theta_i)$, and these beliefs in turn depend on transfers as well as the allocation rule. In particular, distinct efficient and MMIC mechanisms that give the same interim subjective expected utility to the lowest type of each agent need not give the same interim subjective expected utilities to all types, in contrast to the usual payoff equivalence under Bayesian incentive compatibility.[8] Indeed, I show in Section 4 that, in the context of bilateral trade, the efficient allocation rule may be maxmin implementable even if all IR Groves mechanisms run expected deficits for some measure $\phi \in \Phi^*$.

On the other hand, the condition that an IR Groves mechanism runs an expected surplus is also *sufficient* for an efficient allocation rule to be Bayesian implementable, because, following Arrow (1979) and d'Aspremont and Gérard-Varet (1979), "lump-sum" transfers of the form $h_i(\theta_{-i})$ may be used to balance the budget ex post without affecting incentives. This result also does not carry over with maxmin incentive compatibility, as these transfers can affect agents' worst-case beliefs and thereby affect incentives. This issue makes constructing desirable MMIC mechanisms challenging, and this paper does not contain positive

[7] It should also be noted that if (1) admits a saddle point then the maximizing and minimizing operators in (1) commute, so that maxmin IC coincides with "minmax" IC. In this case, agents may equivalently be viewed as pessimistic Bayesians rather than MMEU maximizers. That is, when (1) admits a saddle point, the approach of this paper does not require that agents are non-Bayesian.

[8] This point was already noted by Bodoh-Creed (2012).

results on maxmin implementation outside of the bilateral trade context—where, however, a full characterization is provided.

4 Application to Bilateral Trade

In this section, I show how Theorem 1 can be applied to obtain a full characterization of when efficient bilateral trade or bilateral public good provision is implementable when agents know each other's expected valuation of the good.

The key assumptions in this section are that $n = 2$ and $Y = \{0 = (0,0), 1 = (1,1)\}$ (e.g., $\{no\ trade, trade\}$, $\{no\ provision, provision\}$), and that in addition every measure $\phi_i \in \Phi_i$ satisfies $E^{\phi_i}[\theta_i] = \theta_i^*$ for some $\theta_i^* \in (\underline{\theta}_i, \bar{\theta}_i)$ (note that, as in the general case, θ_i can be negative).[9] The results in this section actually only require the weaker assumption that $E^{\phi_i}[\theta_i] \geq \theta_i^*$ for all $\phi_i \in \Phi_i$; the intuition is that, with maxmin agents, only bounds on how "bad" an agent's belief can be are binding.[10] However, Section 5.1 shows that the equality assumption $E^{\phi_i}[\theta_i] = \theta_i^*$ is appropriate if agents have a unique common prior with mean (θ_1^*, θ_2^*) and support $[\underline{\theta}_1, \bar{\theta}_1] \times [\underline{\theta}_2, \bar{\theta}_2]$ at an ex ante stage and may acquire additional information prior to entering the mechanism. I therefore adopt the equality assumption for consistency with this interpretation.

The model remains general enough to cover the classical bilateral trade and bilateral public good provision settings, as follows.

Bilateral Trade: Agent 1 is the seller and agent 2 is the buyer. They can trade ($y = 1$) or not ($y = 0$). The seller's type θ_1 is -1 times her value of retaining the object (or equivalently her cost of providing it). The buyer's type θ_2 is his value of acquiring the object. $t_1(\theta_1, \theta_2)$ is the price received by the seller. $t_2(\theta_1, \theta_2)$ is -1 times the price paid by the buyer.

Public Good Provision: Agents 1 and 2 can either share the cost $C \in \mathbb{R}$ of providing a public good ($y = 1$) or not ($y = 0$). An agent's type θ_i is her benefit from the good, net of a benchmark payment of $\frac{C}{2}$. $t_i(\theta_i, \theta_j)$ is -1 times what agent i pays for the good in addition

[9] The assumption that θ_i^* lies in the interior of Θ_i is without loss of generality: if $\theta_i^* \in \{\underline{\theta}_i, \bar{\theta}_i\}$, then there would be no uncertainty about agent i's value, and efficient trade could always be implemented with a Groves mechanism.

[10] Without a bound on how bad beliefs can be, de Castro and Yannelis's (2010) theorem shows that efficient trade is always implementable.

to $\frac{C}{2}$.

The special case of this model where $\underline{\theta}_i = -\bar{\theta}_j$, $i = 1, 2$, will be useful for providing intuition for the results. I call this the *aligned supports* case. With aligned supports, the least favorable type of agent i does not have (strict) gains from trade with any type of agent j, and the most favorable type of agent i has (weak) gains from trade with every type of agent j. For example, "aligned supports" in the bilateral trade setting means that the sets of possible values of the seller and the buyer coincide.

Two special kinds of distributions ϕ_i will play an important role in the analysis. Let $\delta_{\theta_i^*}$ be the Dirac measure on θ_i^*, so that $\delta_{\theta_i^*} \in \Phi_i$ corresponds to the possibility that agent i's value is θ_i^* for sure.[11] Let $\delta_{\theta_i^l,\theta_i^h}$ be the 2-point measure on θ_i^l and θ_i^h satisfying $E^{\delta_{\theta_i^l,\theta_i^h}}[\theta_i] = \theta_i^*$; that is, $\delta_{\theta_i^l,\theta_i^h}$ is given by $\theta_i = \theta_i^l$ with probability $\frac{\theta_i^h - \theta_i^*}{\theta_i^h - \theta_i^l}$ and $\theta_i = \theta_i^h$ with probability $\frac{\theta_i^* - \theta_i^l}{\theta_i^h - \theta_i^l}$. Thus, $\delta_{\theta_i^l,\theta_i^h} \in \Phi_i$ corresponds to the possibility that agent i's value may take on only value θ_i^l or θ_i^h.[12] The results to follow require $\delta_{\theta_i^*} \in \Phi_i$ and $\delta_{\theta_i^l,\theta_i^h} \in \Phi_i$ for certain values of θ_i^l, θ_i^h.

I now turn to the characterization result. A preliminary observation is that if no type of some agent i has gains from trade with the average type of agent j, then efficient trade is always possible. In this case, efficient trade can be implemented by simply giving the entire surplus to agent j: certainty of no-trade is then a worst-case belief for every type of agent i, so truthtelling is trivially optimal for agent i, and truthtelling is optimal for agent j by the usual Vickrey-Clarke-Groves logic.

Proposition 1 *Assume that $\bar{\theta}_i + \theta_j^* \leq 0$ and $\delta_{\theta_j^*} \in \Phi_j$ for some $i \in \{1, 2\}$. Then efficient trade is implementable (with strong budget balance).*

Proof. See Appendix A. ■

In light of Proposition 1, the main result of this section considers the non-trivial case where some type of each agent has gains from trade with the average type of the other agent (i.e., $\bar{\theta}_i + \theta_j^* > 0$ for $i = 1, 2$). I also assume that some type of each agent does *not* have strict gains from trade with the average type of the other agent (i.e., $\underline{\theta}_i + \theta_j^* \leq 0$ for $i = 1, 2$).

[11] In the information acquisition interpretation of Section 5.1, $\delta_{\theta_i^*} \in \Phi_i$ corresponds to the possibility that agent i may acquire no new information about her value before entering the mechanism.

[12] In the information acquisition interpretation of Section 5.1, $\delta_{\theta_i^l,\theta_i^h} \in \Phi_i$ corresponds to the possibility that agent i may observe a binary signal of her value before entering the mechanism, where the "bad" signal lowers her expected value to θ_i^l and the "good" signal raises her expected value to θ_i^h.

This is an analogue of the "overlapping supports" assumption of the Myerson-Satterthwaite theorem. Put together, the assumptions that $\bar{\theta}_i + \theta_j^* > 0$ and $\underline{\theta}_i + \theta_j^* \leq 0$ for $i = 1, 2$ say that the intervals Θ_i and Θ_j are sufficiently wide or sufficiently "well-aligned." For instance, these assumptions hold with aligned supports.

The result is the following.

Theorem 2 *Assume that* $\bar{\theta}_i + \theta_j^* > 0$ *and* $\underline{\theta}_i + \theta_j^* \leq 0$*, and that* $\delta_{\theta_i, \bar{\theta}_i} \in \Phi_i$ *for all* $\theta_i \in [\underline{\theta}_i, \theta_i^*]$*, for* $i = 1, 2$.[13] *Then efficient trade is implementable if and only if*

$$\left(\frac{\bar{\theta}_1 + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}{\bar{\theta}_1 + \bar{\theta}_2}\right)\left(\frac{\bar{\theta}_1 - \theta_1^*}{\theta_1^* + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}\right)\log\left(1 + \frac{\theta_1^* + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}{\bar{\theta}_1 - \theta_1^*}\right)$$
$$+\left(\frac{\bar{\theta}_2 + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}{\bar{\theta}_1 + \bar{\theta}_2}\right)\left(\frac{\bar{\theta}_2 - \theta_2^*}{\theta_2^* + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}\right)\log\left(1 + \frac{\theta_2^* + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}{\bar{\theta}_2 - \theta_2^*}\right) \geq 1. \quad (*)$$

Proof. See Appendix A. ■

Theorem 2 shows that, under mild restrictions, efficient bilateral trade between maxmin agents is possible if and only if Condition (*) holds. In other words, the Myerson-Satterthwaite impossibility result holds with maxmin agents if and only if Condition (*) fails.

As will become clear, Condition (*) says precisely that, in the $\alpha_i(\theta_i)$ double auction, $\alpha_1(\theta_1) + \alpha_2(\theta_2) \leq 1$ for all $\theta_1 \in \Theta_1, \theta_2 \in \Theta_2$. To understand the economic content of Condition (*), I discuss three aspects of the condition. First, what does Condition (*) imply for comparative statics and other economic results? Second, where does Condition (*) come from? And, third, why is Condition (*) necessary and sufficient condition for implementation, while Theorem 1 only gives a necessary condition?

To see the implications of Condition (*), first set aside the first term in each of the products on the left-hand side (i.e., the $\frac{\bar{\theta}_1 + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}{\bar{\theta}_1 + \bar{\theta}_2}$ and $\frac{\bar{\theta}_2 + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}{\bar{\theta}_1 + \bar{\theta}_2}$ terms). These terms disappear with aligned supports, and may be viewed as "adjustments" that are needed if there are some types that are either sure to trade (if $\bar{\theta}_i > -\underline{\theta}_j$) or sure to not trade (if $\bar{\theta}_i < -\underline{\theta}_j$). Next, note that each of the remaining products is of the form $\frac{1}{x}\log(1+x)$, which is decreasing in x. In particular, increasing θ_i^* makes Condition (*) harder to satisfy. A rough intuition for this comparative static is that increasing θ_i^* makes agent j more confident

[13] Note that $\delta_{\theta_i^*, \bar{\theta}_i} = \delta_{\theta_i^*}$, so we have $\delta_{\theta_i^*} \in \Phi_i$.

that he will trade, which makes shading his report to get a better transfer more tempting. For example, as $\theta_i^* \to \underline{\theta}_i$, the bound on how "bad" agents' beliefs can be vanishes, and efficient trade is always implementable as in de Castro and Yannelis (2010); on the other hand, as $\theta_i^* \to \bar{\theta}_i$, agents become certain that they will trade, and the temptation to shade their reports becomes irresistible.

Another observation is that Condition (*) always holds when $\theta_1^* + \theta_2^* \leq 0$; that is, when the average types of each agent do not have strict gains from trade with each other (e.g., this is why the curve in Figure 1 lies above the 45° line). This follows because, using the inequality $\log 1 + x \geq \frac{x}{1+x}$, the left-hand side of Condition (*) is at least

$$\frac{\bar{\theta}_1 - \theta_1^*}{\bar{\theta}_1 + \bar{\theta}_2} + \frac{\bar{\theta}_2 - \theta_2^*}{\bar{\theta}_1 + \bar{\theta}_2} = 1 - \frac{\theta_1^* + \theta_2^*}{\bar{\theta}_1 + \bar{\theta}_2}.$$

This is consistent with Proposition 5 below, which shows that efficient trade is implementable with reference rules when $\theta_1^* + \theta_2^* \leq 0$. In particular, the parameters for which efficient trade is implementable with general mechanisms but not with reference rules are precisely those that satisfy Condition (*) but would violate Condition (*) if the $\log 1 + x$ terms were approximated by $\frac{x}{1+x}$. This gives one measure of how restrictive reference rules are.

To see (heuristically) where Condition (*) comes from, suppose that supports are aligned and that the worst-case belief of a type θ_i agent who reports type $\hat{\theta}_i$ is $\delta_{-\hat{\theta}_i, \bar{\theta}_j}$, which is easily seen to be the belief that minimizes the probability that strict gains from trade exist (i.e., that $\hat{\theta}_i + \theta_j > 0$) among beliefs ϕ_j with $E^{\phi_j}[\theta_j] = \theta_j^*$. Suppose also that the mechanism is ex post individually rational, so that $U_i\left(\hat{\theta}_i, -\hat{\theta}_i; \theta_i\right) = 0$. Then

$$U_i\left(\hat{\theta}_i, \delta_{-\hat{\theta}_i, \bar{\theta}_j}; \theta_i\right) = \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i + t_i\left(\hat{\theta}_i, \bar{\theta}_j\right)\right) + \left(1 - \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\right)(0),$$

where $\frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}$ is the probability of trade (i.e., the probability that $\theta_j = \bar{\theta}_j$ under $\delta_{-\hat{\theta}_i, \bar{\theta}_j}$), and $\theta_i + t_i\left(\hat{\theta}_i, \bar{\theta}_j\right)$ is type θ_i's payoff in the event that trade occurs. Assuming that t_i is differentiable, the first-order condition for truthtelling to be optimal is

$$\frac{\partial}{\partial \theta_i} t_i\left(\theta_i, \bar{\theta}_j\right) = -\frac{\bar{\theta}_j - \theta_j^*}{\left(\bar{\theta}_j + \theta_i\right)\left(\theta_j^* + \theta_i\right)}\left(\theta_i + t_i\left(\theta_i, \bar{\theta}_j\right)\right).$$

This first-order condition captures the tradeoff discussed in the introduction: shading one's report down yields a first-order loss in the probability of trade (i.e., in the probability that $\theta_j = \bar{\theta}_j$), which must be offset by a first-order improvement in the transfer (i.e., in $t_i\left(\theta_i, \bar{\theta}_j\right)$).

Solving this differential equation for $t_i\left(\theta_i, \bar{\theta}_j\right)$ yields

$$t_i\left(\theta_i, \bar{\theta}_j\right) = \frac{\bar{\theta}_j + \theta_i}{\theta_j^* + \theta_i}\left[k - \bar{\theta}_j \frac{\bar{\theta}_j - \theta_j^*}{\bar{\theta}_j + \theta_i} - \left(\bar{\theta}_j - \theta_j^*\right)\log\left(\theta_i + \bar{\theta}_j\right)\right],$$

where k is a constant of integration. The constant that keeps transfers bounded as $\theta_i \to -\theta_j^*$ is $k = \bar{\theta}_j + \left(\bar{\theta}_j - \theta_j^*\right)\log\left(\bar{\theta}_j - \theta_j^*\right)$, which gives

$$t_i\left(\theta_i, \bar{\theta}_j\right) = \alpha_i\left(\theta_i\right)\left(\theta_i + \bar{\theta}_j\right) - \theta_i,$$

where

$$\alpha_i\left(\theta_i\right) = 1 - \frac{\bar{\theta}_j - \theta_j^*}{\theta_j^* + \theta_i}\log\left(1 + \frac{\theta_j^* + \theta_i}{\bar{\theta}_j - \theta_j^*}\right).$$

Now, letting

$$t_i\left(\theta_i, \theta_j\right) = \alpha_i\left(\theta_i\right)\left(\theta_i + \theta_j\right) - \theta_i \text{ for all } \theta_i \in \Theta_i, \theta_j \in \Theta_j,$$

so that the resulting mechanism is an $\alpha_i\left(\theta_i\right)$ double auction as described in the introduction, it may be verified that weak budget balance holds for all $\left(\theta_i, \theta_j\right)$ if and only if it holds for $\left(\bar{\theta}_i, \bar{\theta}_j\right)$ (and it also may be verified that $\delta_{-\hat{\theta}_i, \bar{\theta}_j}$ is indeed a worst-case belief). Therefore, efficient trade is implementable if and only if

$$t_1\left(\bar{\theta}_1, \bar{\theta}_2\right) + t_2\left(\bar{\theta}_1, \bar{\theta}_2\right) \leq 0,$$

or equivalently

$$\alpha_1\left(\bar{\theta}_1\right) + \alpha_2\left(\bar{\theta}_2\right) \leq 1.$$

This is precisely Condition (*) (in the aligned supports case). In other words, Condition (*) says that the shares of the social surplus that must be left to the highest types in the $\alpha_i(\theta_i)$ double auction sum to less than one.

Finally, why does the sufficient condition for implementability that $\alpha_1(\bar{\theta}_1) + \alpha_2(\bar{\theta}_2) \leq 1$ match the necessary condition from Theorem 1? Recall that the necessary condition is that expected social surplus exceeds (a lower bound on) expected information rents (i.e., (3) holds) for any distribution $\phi_1 \times \phi_2 \in \Phi_1 \times \Phi_2$. A first observation is that it suffices to compare the social surplus and information rents under the critical distribution $\delta_{\underline{\theta}_1,\bar{\theta}_1} \times \delta_{\underline{\theta}_2,\bar{\theta}_2}$, as this distribution may be shown to minimize the difference between the left- and right-hand sides of (3). Letting β_i be the probability that $\theta_i = \bar{\theta}_i$ under $\delta_{\underline{\theta}_i,\bar{\theta}_i}$, the expected social surplus under $\delta_{\underline{\theta}_1,\bar{\theta}_1} \times \delta_{\underline{\theta}_2,\bar{\theta}_2}$ in the aligned supports case equals

$$\beta_1\beta_2\left(\bar{\theta}_1 + \bar{\theta}_2\right),$$

as under $\delta_{\underline{\theta}_1,\bar{\theta}_1} \times \delta_{\underline{\theta}_2,\bar{\theta}_2}$ there are strict gains from trade only if $\theta_1 = \bar{\theta}_1$ and $\theta_2 = \bar{\theta}_2$. Next, the expectation of (the lower bound on) agent i's information rent under $\delta_{\underline{\theta}_i,\bar{\theta}_i}$ equals

$$\beta_i \int_{\underline{\theta}_i}^{\bar{\theta}_i} \tilde{y}_i(\theta_i)\, d\theta_i + \underbrace{(1-\beta_i) \int_{\underline{\theta}_i}^{\underline{\theta}_i} \tilde{y}_i(\theta_i)\, d\theta_i}_{=0},$$

which may be shown to equal

$$\begin{aligned} & \beta_i\left(\bar{\theta}_i + \theta_j^*\right)\alpha_i\left(\bar{\theta}_i\right) \\ = \;& \beta_1\beta_2\left(\bar{\theta}_1 + \bar{\theta}_2\right)\alpha_i\left(\bar{\theta}_i\right). \end{aligned}$$

The explanation for the appearance of the $\alpha_i(\bar{\theta}_i)$ term here is that this is the fraction of the social surplus that must be left to type $\bar{\theta}_i$ in an MMIC mechanism when type θ_i's subjective expected allocation is $\tilde{y}_i(\theta_i)$ (in particular, the bound on agent i's subjective information rent given by integrating $\tilde{y}_i(\theta_i)$ is tight in the current setting), and the last equality follows by aligned supports. Combining these observations, the necessary condition from Theorem

1 reduces to

$$\beta_1\beta_2\left(\bar{\theta}_1+\bar{\theta}_2\right)\geq\beta_1\beta_2\left(\bar{\theta}_1+\bar{\theta}_2\right)\left(\alpha_1\left(\bar{\theta}_1\right)+\alpha_2\left(\bar{\theta}_2\right)\right),$$

which is equivalent to $\alpha_1\left(\bar{\theta}_1\right)+\alpha_2\left(\bar{\theta}_2\right)\leq 1$.

The approach taken to constructing the $\alpha_i\left(\theta_i\right)$ double auction is quite different from standard approaches in Bayesian mechanism design. In particular, the approach here is to posit type θ_i's worst-case belief to be $\delta_{-\theta_i,\bar{\theta}_j}$ (the belief the minimizes the probability that strict gains from trade exist); solve a differential equation coming from incentive compatibility for $t_i\left(\theta_i,\bar{\theta}_j\right)$, which gives the formula for $\alpha_i\left(\theta_i\right)$; and then verify that $\delta_{-\theta_i,\bar{\theta}_j}$ is indeed type θ_i's worst-case belief in the $\alpha_i\left(\theta_i\right)$ double auction. In contrast, a standard approach might be to use an AGV mechanism. However, as argued above, standard arguments for why using such mechanisms is without loss of generality do not apply with maxmin agents. Indeed, I close this section by showing that in some cases efficient trade is implementable in an $\alpha_i\left(\theta_i\right)$ double auction, but not in any AGV mechanism.

With maxmin agents, the natural definition of an AGV mechanism is a mechanism where, for all $\theta_i\in\Theta_i,\theta_j\in\Theta_j$,

$$t_i\left(\theta_i,\theta_j\right)=E^{\phi_j^*\left(\theta_i\right)}\left[\theta_j y\left(\theta_i,\theta_j\right)\right]+h_i\left(\theta_j\right)$$

for some worst-case belief $\phi_j^*\left(\theta_i\right)\in\arg\min_{\phi_j\in\Phi_j}U_i\left(\theta_i,\phi_j;\theta_i\right)$ and some lump-sum transfer function $h_i:\Theta_j\rightarrow\mathbb{R}$. Suppose that $\theta_i^*>0$ for $i=1,2$ and Condition (*) and the assumptions of Theorem 2 hold (it may be checked that these assumptions are mutually consistent). Theorem 2 then implies that an $\alpha_i\left(\theta_i\right)$ double auction implements efficient trade, but I claim that efficient trade is not implementable in any AGV mechanism. To see this, first note that individual rationality of type $\underline{\theta}_i$ implies that $h_i\left(\theta_j^*\right)\geq 0$ for $i=1,2$, as otherwise one would have $U_i\left(\underline{\theta}_i\right)\leq U_i\left(\underline{\theta}_i,\delta_{\theta_j^*};\underline{\theta}_i\right)=h_i\left(\theta_j^*\right)<0$. Next, WBB implies that

$$t_1\left(\theta_1^*,\theta_2^*\right)+t_2\left(\theta_1^*,\theta_2^*\right)=E^{\phi_2^*\left(\theta_1^*\right)}\left[\theta_2 y\left(\theta_1^*,\theta_2\right)\right]+E^{\phi_1^*\left(\theta_2^*\right)}\left[\theta_1 y\left(\theta_1,\theta_2^*\right)\right]+h_1\left(\theta_2^*\right)+h_2\left(\theta_1^*\right)\leq 0.$$

On the other hand, EF, $E^{\phi_i}[\theta_i] = \theta_i^*$, and $\theta_i^* > 0$ imply that, for all $\phi_1 \in \Phi_1, \phi_2 \in \Phi_2$,

$$\begin{aligned}
& E^{\phi_2}[\theta_2 y(\theta_1^*, \theta_2)] + E^{\phi_1}[\theta_1 y(\theta_1, \theta_2^*)] \\
= \; & \Pr^{\phi_2}(y(\theta_1^*, \theta_2) = 1) E^{\phi_2}[\theta_2 | y(\theta_1^*, \theta_2) = 1] + \Pr^{\phi_1}(y(\theta_1, \theta_2^*) = 1) E^{\phi_1}[\theta_1 | y(\theta_1, \theta_2^*) = 1] \\
\geq \; & \Pr^{\phi_2}(y(\theta_1^*, \theta_2) = 1) \theta_2^* + \Pr^{\phi_1}(y(\theta_1, \theta_2^*) = 1) \theta_1^* \\
> \; & 0.
\end{aligned}$$

Therefore, $h_1(\theta_2^*) + h_2(\theta_1^*) < 0$, which is inconsistent with $h_i\left(\theta_j^*\right) \geq 0$ for $i = 1, 2$.

A closely related point is that if $\theta_1^* + \theta_2^* > 0$ and Condition (*) and the assumptions of Theorem 2 hold, efficient trade is implementable even though every IR Groves mechanism runs an expected deficit for some measure $\phi \in \Phi^*$, in contrast to the results of Makowski and Mezzetti (1994), Williams (1999), and Krishna and Perry (2000) for Bayesian mechanism design. This follows because IR again implies that $h_i\left(\theta_j^*\right) \geq 0$ for $i = 1, 2$ for a Groves mechanism given by

$$t_i(\theta_i, \theta_j) = \theta_j y(\theta_i, \theta_j) + h_i(\theta_j),$$

so the expected deficit of such a mechanism under the measure $\delta_{\theta_1^*} \times \delta_{\theta_2^*} \in \Phi^*$ is equal to

$$t_1(\theta_1^*, \theta_2^*) + t_2(\theta_1^*, \theta_2^*) = \theta_1^* + \theta_2^* + h_1(\theta_2^*) + h_2(\theta_1^*) > 0.$$

5 Further Results on Bilateral Trade

This sections presents additional results on bilateral trade with maxmin agents. Section 5.1 describes how the assumption that agents know each other's expected valuation may be interpreted in terms of information acquisition. Section 5.2 explores the robustness of Theorem 2 to alternative models of agent preferences. Section 5.3 characterizes when efficient trade is possible with *reference rules*, a particularly simple class of mechanisms.

5.1 Information Acquisition Interpretation

The assumption that agents know the mean and bounds on the support of the distribution of each other's value emerges naturally when agents share a unique common prior at an ex ante stage but are uncertain about the information acquisition technology that one's opponent can access prior to entering the mechanism. This section provides the details of this argument.

Consider the following extension of the model. Each agent i's ex post utility is

$$\tilde{\theta}_i y + t_i,$$

where $\tilde{\theta}_i \in \mathbb{R}$ is her realized ex post value. There is an ex ante stage at which the agents' beliefs about the ex post values $\left(\tilde{\theta}_1, \tilde{\theta}_2\right)$ are given by a (unique) common product measure $\tilde{\phi}$ on $\left[\underline{\theta}_1, \bar{\theta}_1\right] \times \left[\underline{\theta}_2, \bar{\theta}_2\right]$ with mean (θ_1^*, θ_2^*) (the common prior). Before entering the mechanism, each agent i observes the outcome of a signaling function ("experiment") $\Sigma_i : \Theta_i \to M_i$ (where M_i is an arbitrary message set), which is informative of her own ex post value but independent of her opponent's. Agent i's interim value, θ_i—which corresponds to her type in the main model—is then her posterior expectation of $\tilde{\theta}_i$ after observing the outcome of her experiment. That is, after observing outcome m_i, agent i's valuation for the good is given by

$$\theta_i \equiv E^{\tilde{\phi}_i}\left[\tilde{\theta}_i | \Sigma_i\left(\tilde{\theta}_i\right) = m_i\right]. \tag{7}$$

Note that the issue of updating "ambiguous beliefs" does not arise in this model. In particular, each agent "knows" her own signaling function, so the updating in (7) is completely standard. However, the following observation shows that the main model can be interpreted as resulting from each agent's being maxmin about her opponent's signaling function at the interim stage (i.e., after she observes her own signal).

Remark 1 *If a measure ϕ_i is the distribution of $\theta_i = E^{\tilde{\phi}_i}\left[\tilde{\theta}_i | \Sigma_i\left(\tilde{\theta}_i\right) = m_i\right]$ under $\tilde{\phi}_i$ for some experiment Σ_i, then $E^{\phi_i}\left[\theta_i\right] = \theta_i^*$ and* $\operatorname{supp} \phi_i \subseteq \Theta_i$ *(where* $\operatorname{supp} \phi_i$ *denotes the support of ϕ_i).*

The fact that $E^{\phi_i}\left[\theta_i\right] = \theta_i^*$ is the law of iterated expectation. The fact that $\operatorname{supp} \phi_i \subseteq \Theta_i$

follows because $\tilde{\theta}_i \in \left[\underline{\theta}_i, \bar{\theta}_i\right]$ with probability 1 under $\tilde{\phi}_i$. Thus, assuming that agent j finds some particular set of measures ϕ_i satisfying $E^{\phi_i}\left[\theta_i\right] = \theta_i^*$ and $\operatorname{supp}\phi_i \subseteq \Theta_i$ possible amounts to assuming that he finds it possible that agent i may have access to some subset of all possible experiments.[14] With this interpretation, the assumption that $\delta_{\theta_i^*} \in \Phi_i$ means that agent j finds it possible that agent i acquires no information about her value before entering the mechanism (beyond the common prior), while the assumption that $\delta_{\theta_i^l, \theta_i^h} \in \Phi_i$ means that agent j finds it possible that agent i observes a binary signal of her value, where the "bad" realization lowers her expectation of $\tilde{\theta}_i$ to θ_i^l and the "good" realization raises her expectation of $\tilde{\theta}_i$ to θ_i^h.

5.2 Robustness to Different Preferences

Although the MMEU model is perhaps the best-studied model of ambiguity aversion, its emphasis on agents' worst-case beliefs strikes some researchers as "extreme" (see Gilboa (2009) or Gilboa and Marinacci (2011) for discussion). This section therefore considers the robustness of Theorem 2 to less extreme models. As a simple first step, I show that Theorem 2 continues to hold with MMEU when agents do not use weakly dominated strategies. I then consider the robustness of Theorem 2 in the epsilon contamination model axiomatized by Kopylov (2008) and in the variational preferences model axiomatized by Maccheroni, Marinacci, and Rustichini (2006). Both of these models nest MMEU as a special case and allow for a tractable analysis of the robustness of Theorem 2 as one moves away from MMEU. I find that Theorem 2 may fail to be robust to epsilon contamination preferences (depending on the prior that is "contaminated" with ambiguity aversion), but is robust to variational preferences.

[14] More precisely, the set of possible measures ϕ_i is jointly determined by the set of experiments agent i may have access to and the prior. For example, *every* measure ϕ_i such that $E^{\phi_i}\left[\theta_i\right] = \theta_i^*$ and $\operatorname{supp}\phi_i \subseteq \Theta_i$ is the distribution of θ_i for some experiment if and only if the prior puts probability 1 on agent i's ex post value being either $\underline{\theta}_i$ or $\bar{\theta}_i$ (see, for example, Theorem 1 of Shmaya and Yariv (2009) or Proposition 1 of Kamenica and Gentzkow (2011)).

5.2.1 Ruling Out Weakly Dominated Strategies

As constructed in the proof of Theorem 2, the $\alpha_i(\theta_i)$ double auction has the unappealing feature that truthtelling is weakly dominated for some types. In particular, $\alpha_i(\theta_i) = 0$ if $\theta_i \leq -\theta_j^*$, so types $\theta_i \leq -\theta_j^*$ get payoff 0 against every opposing type from truthtelling, but get a positive payoff against some opposing types (and payoff 0 against the others) from shading their reports. However, the specification of $\alpha_i(\theta_i)$ for types $\theta_i \leq -\theta_j^*$ can be altered without affecting the desirable properties of the $\alpha_i(\theta_i)$ double auction, as the following result shows. For instance, in the aligned supports case it suffices to let $\alpha_i(\theta_i) = \frac{1}{2}\left(1 - \alpha_j\left(\bar{\theta}_j\right)\right)$ for $\theta_i \leq -\theta_j^*$, $i = 1, 2$.[15]

Proposition 2 *Theorem 2 continues to hold when the definition of MMIC is strengthened to require that truthtelling is not weakly dominated for any type.*

Proof. See Appendix A. ■

5.2.2 Epsilon Contamination

To model agents with epsilon contamination preferences, assume there is a (unique) common prior $\phi^{CP} = \phi_1^{CP} \times \phi_2^{CP}$ and a constant $\varepsilon \in [0, 1]$ such that the notion of incentive compatibility is

$$\theta_i \in \arg\max_{\hat{\theta}_i \in \Theta_i} (1 - \varepsilon) U_i\left(\hat{\theta}_i, \phi_j^{CP}; \theta_i\right) + \varepsilon \min_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right). \tag{8}$$

Note that $\varepsilon = 0$ corresponds to Bayesian IC and $\varepsilon = 1$ corresponds to maxmin IC.

I show that the conclusion of Theorem 2 may change discontinuously as ε decreases from 1, depending on the prior ϕ^{CP}. The intuition is that this apparently small change in beliefs radically changes the incentives of types for whom certainty of no-trade is a worst-case belief (i.e., types $\theta_i < -\theta_j^*$). In particular, in the MMEU model, incentive compatibility is unusually easy to satisfy for types $\theta_i < -\theta_j^*$, as these types do not expect to trade in the worst-case and are therefore willing to reveal their information. In contrast, in the

[15] An earlier version of this paper shows that the reference rule mechanisms of Section 5.3 can also be modified so as to rule out dependence on weakly dominated strategies.

epsilon contamination model for $\varepsilon < 1$, incentive compatibility is unusually hard to satisfy for $\theta_i < -\theta_j^*$, as these types now condition on the very low probability event that the opponent's type is given by the common prior ϕ^{CP}, and thus behave exactly as if they were Bayesian expected utility maximizers.

One might therefore expect the Myerson-Satterthwaite impossibility result to apply for any $\varepsilon < 1$ whenever a type $\theta_1 < -\theta_2^*$ and a type $\theta_2 < -\theta_1^*$ have gains from trade with each other, which occurs when $\theta_1^* + \theta_2^* < 0$. However, in the epsilon contamination model types $\theta_i > -\theta_j^*$ do not condition only on the opponent's type being given by ϕ^{CP}, and if efficient trade among these types can be implemented with a surplus, this surplus can then be used to subsidize trade among types $\theta_i < -\theta_j^*$. Nonetheless, if ϕ^{CP} puts sufficiently small weight on types $\theta_i > -\theta_j^*$, this subsidy cannot be large enough in expected terms (according to ϕ^{CP}) to allow the $\theta_i < -\theta_j^*$ trades to trade efficiently, at least if the ex post version of individual rationality is imposed.[16]

The following result formalizes this intuition. Maintain the assumptions of Theorem 2, so that if $\theta_1^* + \theta_2^* < 0$ then Theorem 2 implies that efficient trade is possible with maxmin agents. The following result shows that, for some prior ϕ^{CP}, this conclusion fails in the epsilon contamination model for any $\varepsilon < 1$.

Proposition 3 *If $\theta_1^* + \theta_2^* < 0$ and $\bar{\theta}_i + \theta_j^* > 0$ for $i = 1, 2$, then there exists a common prior ϕ^{CP} with positive density on $\left[\underline{\theta}_1, \bar{\theta}_1\right] \times \left[\underline{\theta}_2, \bar{\theta}_2\right]$ such that efficient trade is not implementable with ex post IR in the epsilon contamination model for any $\varepsilon < 1$.*

Proof. See Appendix A. ■

5.2.3 Variational Preferences

When agents have variational preferences, the appropriate notion of incentive compatibility is

$$\theta_i \in \arg\max_{\hat{\theta}_i \in \Theta_i} \min_{\phi_j \in \Delta(\Theta_j)} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) + c_i\left(\phi_j\right) \tag{9}$$

[16] Conversely, I conjecture that if ϕ^{CP} puts sufficiently high weight on types $\theta_i > -\theta_j^*$, Theorem 2 continues to hold for ε close to 1 in (8). Unfortunately, proving this result seems to require a different approach from that in Theorem 2, as the relevant belief of a type θ_i agent is no longer a 2-point distribution.

for some function $c_i : \Delta(\Theta_j) \to \mathbb{R}_+$. Note that variation preferences coincide with MMEU preferences if

$$c_i\left(\phi_j\right) = \left\{ \begin{array}{c} 0 \text{ if } \phi_j \in \Phi_j \\ \infty \text{ if } \phi_j \notin \Phi_j \end{array} \right\}.$$

I show that Theorem 2 holds for a class of variational preferences that extends far beyond the special case of MMEU. In particular, consider the simple class of variational preferences where

$$c_i\left(\phi_j\right) = a\left|E^{\phi_j}\left[\theta_j\right] - \theta_j^*\right| \text{ for some } a > 0.$$

This class admits the model of Section 4 as the limiting case where $a \to \infty$. The following result shows that in fact Theorem 2 holds for variational preferences of this type for any $a \geq 1$. This indicates a significant degree of robustness of Theorem 2 to considering variational preferences that are less extreme than MMEU. The proof proceeds by showing that if $a \geq 1$, an agent's "worst-case" belief with variational preferences (i.e., the minimizing measure ϕ_j in (9)) coincides with her worst-case belief in the MMEU model, so the $\alpha_i(\theta_i)$ double auction remains incentive compatible with variational preferences.

Proposition 4 *Suppose the assumptions of Theorem 2 are satisfied and Condition (*) holds, so that the $\alpha_i(\theta_i)$ double auction implements efficient trade with maxmin agents. Then the $\alpha_i(\theta_i)$ double auction also implements efficient trade when agents have variational preferences with $a \geq 1$.*

Proof. See Appendix A. ■

5.3 Efficient Trade with Reference Rules

A common justification for introducing concerns about robustness into mechanism design is that these concerns may argue for the use of simpler or otherwise more intuitively appealing mechanisms. The $\alpha_i(\theta_i)$ double auction introduced in the previous section is simple in some ways, but it does involve a carefully chosen transfer rule. In this section, I point out that efficient trade can also be implemented in an extremely simple class of mechanisms—which I call *reference rules*—in the case where the average types of the two agents do not have gain from trade with each other (i.e., when $\theta_1^* + \theta_2^* \leq 0$). Reference rules also have the advantage of satisfying strong rather than weak budget balance.[17]

I define a reference rule as follows.

Definition 1 *A mechanism* (y, t) *is a* reference rule *if*

$$y(\theta_i, \theta_j) = \left\{ \begin{array}{ll} 1 & \text{if } \theta_i + \theta_j \geq 0 \\ 0 & \text{if } \theta_i + \theta_j < 0 \end{array} \right\}$$

and there exist transfers $t_i^* \in \mathbb{R}$ *for* $i = 1, 2$ *such that* $t_i^* = -t_j^*$ *and*

$$t_i(\theta_i, \theta_j) = \left\{ \begin{array}{ll} t_i^* & \text{if } \theta_i \geq -t_i^*, \theta_j \geq -t_j^*, \theta_i + \theta_j \geq 0 \\ -\theta_i & \text{if } \theta_i < -t_i^*, \theta_j \geq -t_j^*, \theta_i + \theta_j \geq 0 \\ \theta_j & \text{if } \theta_i \geq -t_i^*, \theta_j < -t_j^*, \theta_i + \theta_j \geq 0 \\ 0 & \text{if } \theta_i + \theta_j < 0 \end{array} \right\}.^{18}$$

With a reference rule, agents trade with "reference transfers" $(t_1^* - t_1^*)$ if they are both willing to do so. Otherwise, the agent who is unwilling to trade at the reference transfers receives her reservation transfer $-\theta_i$, and the other agent receives the full gains from trade. For example, in the classical bilateral trade setting (where c is the seller's cost, v is the buyer's value, and p is the price), a reference rule corresponds to setting a reference price p^*, trading at price p^* if $c \leq p^* \leq v$, and otherwise trading at whichever value is closer to p^* (i.e., price is v if $c \leq v < p^*$; price is c if $p^* < c \leq v$).

Reference rules clearly satisfy EF, (ex post) IR, and SBB, so an MMIC reference rule implements efficient trade. The following result characterizes when MMIC reference rules exist; that is, when efficient trade is implementable with reference rules.

[17] The term "reference rule" is taken from Erdil and Klemperer (2010), who recommend the use of such mechanisms in multi-unit auctions. They highlight that reference rules perform well in terms of agents' "local incentives to deviate," a different criterion from what I consider here. Reference rules also bear some resemblance to the "downward flexible price mechanism" of Börgers and Smith (2012). Their mechanism starts with a fixed price p^* which the seller may then lower to any $p' < p^*$, whereupon the parties decide whether to trade at p'.

[18] Note that this defines $t_i(\theta_i, \theta_j)$ for all θ_i, θ_j, because if $\theta_i < -t_i^*$ and $\theta_j < -t_j^*$ then $\theta_i + \theta_j < -t_i^* - t_j^* = 0$.

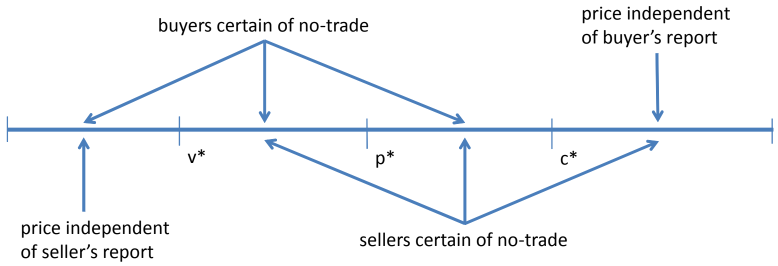
buyers certain of no-trade
price independent of buyer's report
v*
p*
c*
price independent of seller's report
sellers certain of no-trade

Figure 2: Reference rules with $p^* \in [v^*, c^*]$ are incentive compatible.

Proposition 5 *Assume that $\delta_{\theta_i^*} \in \Phi_i$ for $i = 1, 2$. Then efficient trade is implementable with reference rules if and only if either*

1. $\theta_1^* + \theta_2^* \leq 0$ *(the average types do not have strict gains from trade), or*

2. $\underline{\theta}_1 + \underline{\theta}_2 \geq 0$ *(every pair of types has gains from trade).*

Proof. See Appendix A. ■

The intuition for Proposition 5 is particularly easy to see in the classical bilateral trade setting. The intuition for why reference rules are incentive compatible when $c^* \geq v^*$ and $p^* \in [v^*, c^*]$ is captured in Figure 2. Observe that every buyer with value $v \leq c^*$ may be certain that no gains from trade exist, as he may believe that the distribution of seller values is the degenerate distribution on c^*. Hence, certainty of no-trade is a worst-case belief for these buyers, and they are therefore willing to reveal their information. In contrast, buyers with value $v > c^*$ do believe that gains from trade exist with positive probability. But it is optimal for these buyers to reveal their values truthfully as well: misreporting some $\hat{v} > c^*$ does not affect the price regardless of the seller's value (as price equals c if $c > p^*$ and equals p^* if $c \leq p^*$), and misreporting some $\hat{v} \leq c^*$ again gives payoff 0 in the worst-case (as certainty that the seller's value equals c^* would again be a worst-case belief). Therefore, truthtelling is optimal for every buyer type, and the argument for sellers is symmetric.

On the other hand, the intuition for why reference rules are not incentive compatible when $c^* < v^*$ may be seen in Figure 3. Suppose the reference price p^* is greater than c^*. Consider a buyer with value $v \in (c^*, p^*)$. If he reports his value truthfully, then whenever

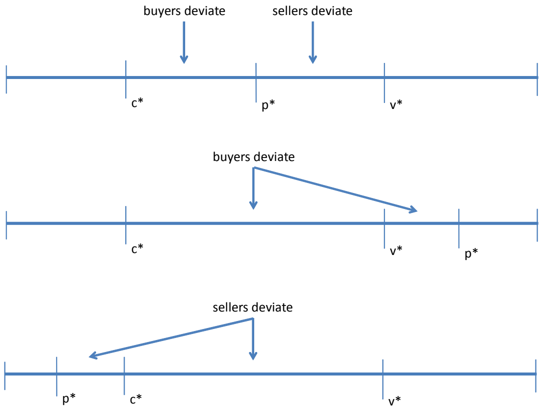

Figure 3: When $c^* < v^*$, no reference rule is incentive compatible.

he trades under the reference rule he does so at price v, which gives him payoff 0. Suppose he instead shades his report down to some $\hat{v} \in (c^*, v)$. Then whenever he trades the price is $\hat{v}$, which gives him a positive payoff, and in addition he expects to trade with positive probability (since $\hat{v} > c^*$). Hence, he will shade down. The same argument shows that in any reference rule a seller with $c \in (p^*, v^*)$ shades up. Figure 3 shows that a consequence of this argument is that a reference rule cannot be MMIC for both agents when $c^* < v^*$, regardless of where the reference price p^* is set.

6 Conclusion

This paper contributes to the study of mechanism design where agents follow "robust" decision rules, in particular where agents are maxmin expected utility maximizers. I establish two main results. First, I give a general necessary condition for the implementability of a social choice rule, which generalizes the well-known condition from Bayesian mechanism design that expected social surplus must exceed expected information rents. This condition involves both a modification of the usual envelope characterization of payoffs and a simple but important conceptual connection between maxmin agents' subjective expected utilities and the "objective" expected social surplus under a common prior. Second, I apply this result to give a complete characterization of when efficient bilateral trade is possible, when agents know little beyond each other's expected valuation of the good (which is the information structure that results when agents are maxmin about the experiment that the other may access before participating in the mechanism). Somewhat surprisingly, the Myerson-Satterthwaite impossible result sometimes continue to hold with maxmin agents, despite the lack of a unique common prior or independent types. When instead efficient trade is possible, it is implementable with a relatively simple double auction format, the $\alpha_i(\theta_i)$ *double auction*; sometimes, it is also implementable with extremely simple *reference rules.*

A clear direction for future work is deriving positive implementation results beyond the bilateral trade context of two agents and two social alternatives. I have argued that standard mechanisms (like the AGV mechanism) may fail to have desirable properties with maxmin agents, and it is not immediately clear how to generalize the mechanisms I construct in this paper (the $\alpha_i(\theta_i)$ double auction and the reference rule) beyond the bilateral trade case. Perhaps the next simplest case to consider is the multilateral public good provision problem with n agents and two alternatives (Mailath and Postlewaite, 1989). In this setting, the same approach used to derive the $\alpha_i(\theta_i)$ double auction can be used to derive a mechanism that sometimes implements efficient trade. However, the attainability of an exact characterization of when efficient trade is possible remains an open question.

More broadly, an important issue is considering models of robust agent behavior beyond the maxmin expected utility model and the related models of ambiguity aversion discussed in Section 5.2. For instance, Linhart and Radner (1989) and Bergemann and Schlag (2008, 2011) consider minmax regret approaches. The integration of models of robust agent behavior in mechanisms and models of robustness concerns on the part of the mechanism designer (Bergemann and Morris, 2005; Chung and Ely, 2007) must also await future research.

Appendix A: Omitted Proofs

Proof of Proposition 1

Consider the mechanism

$$\begin{aligned} y(\theta_i,\theta_j) &= \left\{ \begin{array}{c} 1 \text{ if } \theta_i+\theta_j>0 \\ 0 \text{ if } \theta_i+\theta_j\le 0 \end{array} \right\}, \\ t_i(\theta_i,\theta_j) &= \left\{ \begin{array}{c} -\theta_i \text{ if } \theta_i+\theta_j>0 \\ 0 \text{ if } \theta_i+\theta_j\le 0 \end{array} \right\}, \\ t_j(\theta_i,\theta_j) &= \left\{ \begin{array}{c} \theta_i \text{ if } \theta_i+\theta_j>0 \\ 0 \text{ if } \theta_i+\theta_j\le 0 \end{array} \right\}. \end{aligned}$$

This mechanism clearly satisfies EF, IR, and SBB. For MMIC for agent i, note that $U_i(\theta_i,\theta_j;\theta_i)=0$ for all $\theta_i\in\Theta_i$ and $U_i\left(\hat{\theta}_i,\theta_j^*;\theta_i\right)=0$ for all $\theta_i,\hat{\theta}_i\in\Theta_i$ (as $\bar{\theta}_i+\theta_j^*\le 0$), so $\delta_{\theta_j^*}\in\Phi_j$ implies that $\inf_{\phi_j\in\Phi_j}U_i\left(\theta_i,\phi_j;\theta_i\right)=0\ge\inf_{\phi_j\in\Phi_j}U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)$. For MMIC for agent j, note that $U_j(\theta_j,\theta_i;\theta_j)=\max\{\theta_i+\theta_j,0\}\ge U_j\left(\hat{\theta}_j,\theta_i;\theta_j\right)$ for all $\theta_j,\hat{\theta}_j\in\Theta_j$ and $\theta_i\in\Theta_i$, so $\inf_{\phi_i\in\Phi_i}U_j(\theta_j,\phi_i;\theta_j)\ge\inf_{\phi_i\in\Phi_i}U_j\left(\hat{\theta}_j,\phi_i;\theta_j\right)$.

Proof of Theorem 2: Necessity

By Theorem 1, efficient trade is implementable only if

$$\sum_i\left(\int_{\theta\in\Theta}\theta_i y_i(\theta)\,d\phi\right)-\sum_i\left(\int_{\theta_i\in\Theta_i}(1-F_i(\theta_i))\,\tilde{y}_i(\theta_i)\,d\theta_i\right)\ge 0 \tag{10}$$

for some allocation rule y_i satisfying

$$y_i(\theta)=\left\{ \begin{array}{c} 1 \text{ if } \theta_i+\theta_j>0 \\ 0 \text{ if } \theta_i+\theta_j<0 \end{array} \right\}.$$

Note that, for any such allocation rule y_i,

$$\tilde{y}_i(\theta_i) = \left\{ \begin{array}{ll} 0 & \text{if } \theta_i \le -\theta_j^* \\ \frac{\theta_j^*+\theta_i}{\bar{\theta}_j+\theta_i} & \text{if } \theta_i \in \left(-\theta_j^*, -\underline{\theta}_j\right) \\ 1 & \text{if } \theta_i > -\underline{\theta}_j \end{array} \right\}.$$

This is immediate for the $\theta_i \le -\theta_j^*$ and $\theta_i > -\underline{\theta}_j$ cases, and follows by Chebyshev's inequality in the $\theta_i \in \left(-\theta_j^*, -\underline{\theta}_j\right)$ case.[19,20]

Let $\phi_i = \delta_{\max\left\{\underline{\theta}_i, -\bar{\theta}_j\right\}, \bar{\theta}_i}$ (which is assumed to be an element of Φ_i, as $-\bar{\theta}_j < \theta_i^*$) for $i = 1, 2$. Let $\beta_i = \frac{\theta_i^*+\min\left\{\bar{\theta}_j, -\underline{\theta}_i\right\}}{\bar{\theta}_i+\min\left\{\bar{\theta}_j, -\underline{\theta}_i\right\}}$, which is the probability that $\theta_i = \bar{\theta}_i$ under $\delta_{\max\left\{\underline{\theta}_i, -\bar{\theta}_j\right\}, \bar{\theta}_i}$. Observe that

$$\begin{aligned} \sum_i \left(\int_{\theta \in \Theta} \theta_i y_i(\theta)\, d\phi \right) &= \left(\bar{\theta}_i + \bar{\theta}_j\right) \beta_i \beta_j + \max\left\{\bar{\theta}_i + \underline{\theta}_j, 0\right\} \beta_i \left(1 - \beta_j\right) \\ &\quad + \max\left\{\bar{\theta}_j + \underline{\theta}_i, 0\right\} \beta_j \left(1 - \beta_i\right), \end{aligned}$$

and, using the assumption that $\underline{\theta}_i + \theta_j^* \le 0$,

$$\begin{aligned} \int_{\theta_i \in \Theta_i} \left(1 - F_i(\theta)\right) \tilde{y}_i(\theta_i)\, d\theta_i &= \int_{\max\left\{\underline{\theta}_i, -\theta_j^*\right\}}^{\min\left\{\bar{\theta}_i, -\underline{\theta}_j\right\}} \beta_i \left(\frac{\theta_j^* + \theta_i}{\bar{\theta}_j + \theta_i} \right) d\theta_i + \int_{\min\left\{\bar{\theta}_i, -\underline{\theta}_j\right\}}^{\bar{\theta}_i} \beta_i d\theta_i \\ &= \left(\bar{\theta}_i + \theta_j^*\right) \beta_i \\ &\quad - \left(\bar{\theta}_j - \theta_j^*\right) \beta_i \log\left(1 + \frac{\theta_j^* + \min\left\{\bar{\theta}_i, -\underline{\theta}_j\right\}}{\bar{\theta}_j - \theta_j^*}\right). \end{aligned}$$

Combining these observations and collecting terms, the left-hand side of (10) equals

$$\zeta \left[\begin{array}{c} -1 + \left(\frac{\bar{\theta}_1 + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}{\bar{\theta}_1 + \bar{\theta}_2} \right) \left(\frac{\bar{\theta}_1 - \theta_1^*}{\theta_1^* + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}} \right) \log\left(1 + \frac{\theta_1^* + \min\left\{\bar{\theta}_2, -\underline{\theta}_1\right\}}{\bar{\theta}_1 - \theta_1^*}\right) \\ + \left(\frac{\bar{\theta}_2 + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}{\bar{\theta}_1 + \bar{\theta}_2} \right) \left(\frac{\bar{\theta}_2 - \theta_2^*}{\theta_2^* + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}} \right) \log\left(1 + \frac{\theta_2^* + \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}}{\bar{\theta}_2 - \theta_2^*}\right) \end{array} \right], \tag{11}$$

where $\zeta = \beta_1 \beta_2 \left(\bar{\theta}_1 + \bar{\theta}_2\right) > 0$. The bracketed term in (11) non-negative if and only if

[19] The form of Chebyshev's inequality I use throughout the paper is, for random variable X with mean x^* and upper bound $\bar{x}$, $\Pr(X \ge x) \ge \frac{x^* - x}{\bar{x} - x}$. This follows because $x^* \le \Pr(X \ge x)\, \bar{x} + \Pr(X < x)\, x$. See, for example, p. 319 of Grimmett and Stirzaker (2001).

[20] As will become clear, the value of $\tilde{y}_i\left(-\underline{\theta}_j\right)$ does not matter for the proof.

Condition (*) holds. Hence, Condition (*) is necessary.

Proof of Theorem 2: Sufficiency

The $\alpha_i(\theta_i)$ *double auction* is defined by

$$\begin{aligned} y(\theta_i,\theta_j) &= \left\{ \begin{array}{c} 1 \text{ if } \theta_i+\theta_j>0 \\ 0 \text{ if } \theta_i+\theta_j\leq 0 \end{array} \right\}, \\ t_i(\theta_i,\theta_j) &= \left\{ \begin{array}{cc} \alpha_i(\theta_i)\,\theta_j-(1-\alpha_i(\theta_i))\min\left\{\theta_i,-\underline{\theta}_j\right\} & \text{if } \theta_i+\theta_j>0 \\ 0 & \text{if } \theta_i+\theta_j\leq 0 \end{array} \right\}, \end{aligned}$$

for $i=1,2$, where

$$\alpha_i(\theta_i)=\left\{ \begin{array}{cc} 1-\frac{\bar{\theta}_j-\theta_j^*}{\theta_j^*+\min\left\{\theta_i,-\underline{\theta}_j\right\}}\log\left(1+\frac{\theta_j^*+\min\left\{\theta_i,-\underline{\theta}_j\right\}}{\bar{\theta}_j-\theta_j^*}\right) & \text{if } \theta_i>-\theta_j^* \\ 0 & \text{if } \theta_i\leq-\theta_j^* \end{array} \right\}.$$

This mechanism is clearly efficient. I show that it satisfies IR and MMIC, and that it satisfies WBB if and only if Condition (*) holds.

Claim 1: The $\alpha_i(\theta_i)$ double auction satisfies (ex post) IR.

Proof: If $\theta_i+\theta_j\leq 0$, then $U_i(\theta_i,\theta_j;\theta_i)=0$. If $\theta_i+\theta_j>0$, then

$$\begin{aligned} U_i(\theta_i,\theta_j;\theta_i) &= \theta_i+\alpha_i(\theta_i)\,\theta_j-(1-\alpha_i(\theta_i))\min\left\{\theta_i,-\underline{\theta}_j\right\} \\ &\geq \alpha_i(\theta_i)(\theta_i+\theta_j). \end{aligned}$$

Now $\alpha_i(\theta_i)$ is of the form $1-\frac{1}{x}\log(1+x)$ for $x>0$, and $\frac{1}{x}\log(1+x)\in(0,1)$ for $x>0$, so $\alpha_i(\theta_i)\in(0,1)$ for all θ_i. This yields (ex post) IR.

Claim 2: The $\alpha_i(\theta_i)$ double auction satisfies WBB if and only if Condition (*) holds.

Proof: WBB is trivially satisfied when $\theta_1+\theta_2\leq 0$, so suppose that $\theta_1+\theta_2>0$.

If $\theta_1<-\underline{\theta}_2$ and $\theta_2<-\underline{\theta}_1$,

$$t_1(\theta_1,\theta_2)+t_2(\theta_1,\theta_2)=(\theta_1+\theta_2)(\alpha_1(\theta_1)+\alpha_2(\theta_2)-1).$$

Since $\alpha_i(\theta_i)$ is non-decreasing in θ_i (as $\frac{1}{x}\log(1+x)$ is decreasing in x), this expression is non-positive for all θ_1, θ_2 with $\theta_1 + \theta_2 > 0$ if and only if $\alpha_1(\bar{\theta}_1) + \alpha_2(\bar{\theta}_2) \leq 1$. Condition (*) implies $\alpha_1(\bar{\theta}_1) + \alpha_2(\bar{\theta}_2) \leq 1$ and is equivalent to this inequality when $\bar{\theta}_i \leq -\underline{\theta}_j$ for $i = 1, 2$.

If $\theta_1 \geq -\underline{\theta}_2$ and $\theta_2 \geq -\underline{\theta}_1$,

$$\begin{aligned} t_1(\theta_1,\theta_2) + t_2(\theta_1,\theta_2) &= \alpha_1(\theta_1)\theta_2 + \alpha_2(\theta_2)\theta_1 + (1-\alpha_1(\theta_1))\underline{\theta}_2 + (1-\alpha_2(\theta_2))\underline{\theta}_1 \\ &= \theta_1 + \theta_2 - (1-\alpha_1(\theta_1))(\theta_2 - \underline{\theta}_2) - (1-\alpha_2(\theta_2))(\theta_1 - \underline{\theta}_1). \end{aligned}$$

This expression is non-decreasing in θ_1 and θ_2 (as $\alpha_i(\theta_i) \in (0,1)$ is non-decreasing and $\theta_i \geq \underline{\theta}_i$), so it is non-positive for all θ_1, θ_2 with $\theta_1 + \theta_2 > 0$ if and only if

$$\bar{\theta}_1 + \bar{\theta}_2 - \left(1 - \alpha_1(\bar{\theta}_1)\right)\left(\bar{\theta}_2 - \underline{\theta}_2\right) - \left(1 - \alpha_2(\bar{\theta}_2)\right)\left(\bar{\theta}_1 - \underline{\theta}_1\right) \leq 0.$$

Moving the product terms to the right-hand side and dividing by $\bar{\theta}_1 + \bar{\theta}_2$ (which is positive) shows that this inequality is equivalent to Condition (*) when $\bar{\theta}_i \geq -\underline{\theta}_j$ for $i = 1, 2$ (which is the case under consideration).

Finally, if $\theta_1 < -\underline{\theta}_2$ and $\theta_2 \geq -\underline{\theta}_1$ (which is the hardest case),[21]

$$\begin{aligned} t_1(\theta_1,\theta_2) + t_2(\theta_1,\theta_2) &= (\alpha_1(\theta_1) + \alpha_2(\theta_2) - 1)\theta_1 + \alpha_1(\theta_1)\theta_2 + (1-\alpha_2(\theta_2))\underline{\theta}_1 \\ &= (\theta_1 + \theta_2)\left[\alpha_1(\theta_1) - \frac{\theta_1 - \underline{\theta}_1}{\theta_1 + \theta_2}(1 - \alpha_2(\theta_2))\right]. \end{aligned}$$

This expression is non-positive for all θ_1, θ_2 with $\theta_1 + \theta_2 > 0$ if and only if the bracketed term is non-positive for all such θ_1, θ_2. This term is increasing in θ_2, so it is non-positive for all θ_1, θ_2 with $\theta_1 + \theta_2 > 0$ if and only if

$$\alpha_1(\theta_1) - \frac{\theta_1 - \underline{\theta}_1}{\theta_1 + \bar{\theta}_2}\left(1 - \alpha_2(\bar{\theta}_2)\right) \leq 0 \tag{12}$$

for all θ_1.[22] If $\theta_1 \leq -\theta_2^*$, then $\alpha_1(\theta_1) = 0$ so (12) holds. Suppose toward a contradiction

[21] This assumes $\theta_1 + \theta_2 > 0$. If instead $\theta_1 + \theta_2 = 0$, then $t_1(\theta_1,\theta_2) + t_2(\theta_1,\theta_2) = -(1-\alpha_2(\theta_2))(\theta_1 - \underline{\theta}_1) \leq 0$.

[22] The "only if" part of this statement follows because the hypothesis that $\theta_2 \geq -\underline{\theta}_1$ implies that $\bar{\theta}_2 \geq -\underline{\theta}_1$, which in turn implies that $\theta_1 + \bar{\theta}_2 \geq 0$ for all θ_1.

that (12) fails for some $\theta_1 \in \left[-\theta_2^*, \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}\right]$. Observe first that (12) holds at $\theta_1 = \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}$: this has already been shown if $\bar{\theta}_1 \geq -\underline{\theta}_2$, and if $\bar{\theta}_1 < -\underline{\theta}_2$ it follows by noting that at $\theta_1 = \bar{\theta}_1$ (12) is equivalent to Condition (*) when $\bar{\theta}_1 < -\underline{\theta}_2$ and $\bar{\theta}_2 \geq -\underline{\theta}_1$. Since the left-hand side of (12) is continuous in θ_1 and (12) holds for $\theta_1 = \bar{\theta}_1$ and for all $\theta_1 \geq -\underline{\theta}_2$, (12) fails somewhere on the interval $\left[-\theta_2^*, \min\left\{\bar{\theta}_1, -\underline{\theta}_2\right\}\right]$ if and only if it fails at a local minimum in $(-\theta_2^*, -\underline{\theta}_2)$. Hence, the argument may be completed by showing that no local minimum in $(-\theta_2^*, -\underline{\theta}_2)$ exists. To see this, note that for $\theta_1 \in (-\theta_2^*, -\underline{\theta}_2)$,

$$\alpha_1'(\theta_1) = \frac{1}{\theta_2^* + \theta_1}\left(\frac{\theta_2^* + \theta_1}{\bar{\theta}_2 + \theta_1} - \alpha_1(\theta_1)\right),$$

and therefore the first-order condition for an extremum is

$$\frac{1}{\theta_2^* + \theta_1}\left(\frac{\theta_2^* + \theta_1}{\bar{\theta}_2 + \theta_1} - \alpha_1(\theta_1)\right) = \frac{\bar{\theta}_2 + \underline{\theta}_1}{\left(\bar{\theta}_2 + \theta_1\right)^2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right).$$

In addition, the second derivative of the left-hand side of (12) equals

$$-\frac{\left(\bar{\theta}_2 - \theta_2^*\right)\left(2\left(\bar{\theta}_2 + \theta_1\right) + \theta_2^* + \theta_1\right)}{\left(\theta_2^* + \theta_1\right)^2\left(\bar{\theta}_2 + \theta_1\right)^2} + 2\frac{1 - \alpha_1(\theta_1)}{\left(\theta_2^* + \theta_1\right)^2} - 2\frac{\bar{\theta}_2 + \underline{\theta}_1}{\left(\bar{\theta}_2 + \theta_1\right)^3}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right).$$

At an extremum, using the first-order condition implies that this equals

$$\begin{aligned}
& -\frac{\left(\bar{\theta}_2 - \theta_2^*\right)\left(2\left(\bar{\theta}_2 + \theta_1\right) + \theta_2^* + \theta_1\right)}{\left(\theta_2^* + \theta_1\right)^2\left(\bar{\theta}_2 + \theta_1\right)^2} + 2\frac{1 - \alpha_1(\theta_1)}{\left(\theta_2^* + \theta_1\right)^2} - 2\frac{1}{\left(\bar{\theta}_2 + \theta_1\right)\left(\theta_2^* + \theta_1\right)}\left(\frac{\theta_2^* + \theta_1}{\bar{\theta}_2 + \theta_1} - \alpha_1(\theta_1)\right) \\
= \ & \frac{\bar{\theta}_2 - \theta_2^*}{\left(\theta_2^* + \theta_1\right)^2\left(\bar{\theta}_2 + \theta_1\right)}\left[-\alpha_1(\theta_1) + \left(\frac{\theta_2^* + \theta_1}{\bar{\theta}_2 + \theta_1} - \alpha_1(\theta_1)\right)\right] \\
= \ & \frac{\bar{\theta}_2 - \theta_2^*}{\left(\theta_2^* + \theta_1\right)^2\left(\bar{\theta}_2 + \theta_1\right)}\left[-\alpha_1(\theta_1) + \frac{\left(\theta_2^* + \theta_1\right)\left(\bar{\theta}_2 + \underline{\theta}_1\right)}{\left(\bar{\theta}_2 + \theta_1\right)^2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right)\right].
\end{aligned}$$

Next, observe that

$$\frac{\theta_1 - \underline{\theta}_1}{\theta_1 + \bar{\theta}_2} \geq \frac{\left(\theta_2^* + \theta_1\right)\left(\bar{\theta}_2 + \underline{\theta}_1\right)}{\left(\bar{\theta}_2 + \theta_1\right)^2},$$

which may be seen by cross-multiplying by $\left(\bar{\theta}_2 + \theta_1\right)^2$ and noting that $\theta_1 - \underline{\theta}_1 \geq \theta_2^* + \theta_1$ (as

$\underline{\theta}_1 + \theta_2^* \leq 0$) and $\bar{\theta}_2 + \theta_1 \geq \bar{\theta}_2 + \underline{\theta}_1$. Therefore,

$$-\alpha_1(\theta_1) + \frac{\theta_1 - \underline{\theta}_1}{\theta_1 + \bar{\theta}_2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right) \geq -\alpha_1(\theta_1) + \frac{\left(\bar{\theta}_2 + \underline{\theta}_1\right)\left(\theta_2^* + \theta_1\right)}{\left(\theta_1 + \bar{\theta}_2\right)^2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right),$$

so if (12) fails then the second derivative is non-positive at any local extremum. That is, any local extremum in $(-\theta_2^*, -\underline{\theta}_2)$ must be a local maximum, so no local minimum in $(-\theta_2^*, -\underline{\theta}_2)$ exists, completing the proof. The argument for $\theta_1 \geq -\underline{\theta}_2$ and $\theta_2 < -\underline{\theta}_1$ is symmetric.

Claim 3: The $\alpha_i(\theta_i)$ double auction satisfies MMIC.

Proof: Suppose $\theta_i \leq -\theta_j^*$. By IR, $U_i(\theta_i) \geq 0$. By ex post IR and WBB, $t_i\left(\hat{\theta}_i, \theta_j\right) \leq \theta_j$ for all $\hat{\theta}_i, \theta_j$, and therefore $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) \leq \max\left\{\theta_i + \theta_j^*, 0\right\} \leq 0$ for all $\hat{\theta}_i$. Hence, $\delta_{\theta_j^*} \in \Phi_j$ implies that $U_i(\theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) \geq \min_{\phi_j \in \Delta(\Theta_j)} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$, which yields MMIC.

For the remainder of the proof, suppose $\theta_i > -\theta_j^*$. I show that no misreport $\hat{\theta}_i$ can be profitable in each of the following four cases: (i) $\hat{\theta}_i > \theta_i$, (ii) $\hat{\theta}_i \leq -\theta_j^*$, (iii) $\hat{\theta}_i \in \left(-\underline{\theta}_j, \theta_i\right)$, (iv) $\hat{\theta}_i \in \left(-\theta_j^*, \min\left\{\theta_i, -\underline{\theta}_j\right\}\right]$. These cases cover all possible misreports, so the $\alpha_i(\theta_i)$ double auction satisfies MMIC.

Case (i): $\hat{\theta}_i > \theta_i$.

In this case, I claim that $U_i(\theta_i, \theta_j; \theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$ for all θ_j. The key step is the following observation.

Lemma 2 *In the $\alpha_i(\theta_i)$ double auction, $t_i(\theta_i, \theta_j)$ is non-increasing in θ_i in the region where $\theta_i + \theta_j \geq 0$.*

Proof. See below. ■

Now, if $\theta_i + \theta_j \leq 0$, then $U_i(\theta_i, \theta_j; \theta_i) = 0$, while ex post IR and WBB imply that $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) \leq \max\{\theta_i + \theta_j, 0\} \leq 0$. If instead $\theta_i + \theta_j > 0$, then EF and Lemma 2 imply that $U_i(\theta_i, \theta_j; \theta_i) \geq U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$. The claim follows, and therefore $U_i(\theta_i) \geq \inf_{\phi_j \in \Delta(\Theta_j)} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$.

Case (ii): $\hat{\theta}_i \leq -\theta_j^*$.

Here, $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) = 0$. Hence, $\delta_{\theta_j^*} \in \Phi_j$ and IR imply that $U_i(\theta_i) \geq \inf_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$.

Case (iii): $\hat{\theta}_i \in \left(-\underline{\theta}_j, \theta_i\right)$.

Note that $\alpha_i\left(\hat{\theta}_i\right) = \alpha_i\left(-\underline{\theta}_j\right)$, so $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) = \theta_i + \alpha_i\left(-\underline{\theta}_j\right)\theta_j - \left(1 - \alpha_i\left(-\underline{\theta}_j\right)\right)\underline{\theta}_j$ for all

θ_j. Therefore, $U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)=\theta_i+\alpha_i\left(-\underline{\theta}_j\right)\theta_j^*-\left(1-\alpha_i\left(-\underline{\theta}_j\right)\right)\underline{\theta}_j$ for all $\phi_j\in\Phi_j$. Similarly, $U_i\left(\theta_i,\phi_j;\theta_i\right)=\theta_i+\alpha_i\left(-\underline{\theta}_j\right)\theta_j^*-\left(1-\alpha_i\left(-\underline{\theta}_j\right)\right)\underline{\theta}_j$, so $U_i\left(\theta_i\right)=\inf_{\phi_j\in\Delta(\Theta_j)}U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)$.

Case (iv): $\hat{\theta}_i\in\left(-\theta_j^*,\min\left\{\theta_i,-\underline{\theta}_j\right\}\right]$.

In this case, I claim that

$$\inf_{\phi_j\in\Delta(\Theta_j)}U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)=\frac{\theta_j^*+\hat{\theta}_i}{\bar{\theta}_j+\hat{\theta}_i}\left(\theta_i+t_i\left(\hat{\theta}_i,\bar{\theta}_j\right)\right).{}^{23} \tag{13}$$

To see that (13) is an upper bound on $\inf_{\phi_j\in\Delta(\Theta_j)}U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)$, observe that $\delta_{-\hat{\theta}_i,\bar{\theta}_j}\in\Phi_j$ and that $U_i\left(\hat{\theta}_i,\delta_{-\hat{\theta}_i,\bar{\theta}_j};\theta_i\right)$ equals (13). To see that (13) is a lower bound on $\inf_{\phi_j\in\Delta(\Theta_j)}U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)$, note that

$$\begin{aligned}
& U_i\left(\hat{\theta}_i,\phi_j;\theta_i\right)\\
= \; & \Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)E^{\phi_j}\left[\theta_i+t_i\left(\hat{\theta}_i,\theta_j\right)|\theta_j>-\hat{\theta}_i\right]+\Pr^{\phi_j}\left(\theta_j\leq-\hat{\theta}_i\right)(0)\\
= \; & \Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)E^{\phi_j}\left[\theta_i+\alpha_i\left(\hat{\theta}_i\right)\theta_j-\left(1-\alpha_i\left(\hat{\theta}_i\right)\right)\hat{\theta}_i|\theta_j>-\hat{\theta}_i\right]\\
= \; & \Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)\left(\theta_i-\hat{\theta}_i\right)\\
& +\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)\alpha_i\left(\hat{\theta}_i\right)\left(E^{\phi_j}\left[\theta_j|\theta_j>-\hat{\theta}_i\right]+\hat{\theta}_i\right).
\end{aligned} \tag{14}$$

I show that (13) is a lower bound on (14) for all ϕ_j with expectation θ_j^*, and hence for all $\phi_j\in\Phi_j$. To see this, consider the problem of minimizing (14) over ϕ_j with expectation θ_j^* in two steps: first minimize over ϕ_j with a given value of $\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)$, and then minimize over $\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)$. For a given value of $\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)$, (14) is minimized by minimizing $E^{\phi_j}\left[\theta_j|\theta_j>-\hat{\theta}_i\right]$ over ϕ_j with expectation θ_j^*. Observe that

$$\left(1-\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)\right)E^{\phi_j}\left[\theta_j|\theta_j\leq-\hat{\theta}_i\right]+\Pr^{\phi_j}\left(\theta_j>-\hat{\theta}_i\right)E^{\phi_j}\left[\theta_j|\theta_j>-\hat{\theta}_i\right]=\theta_j^*$$

[23] Intuitively, the claim is that $\delta_{\max\left\{-\hat{\theta}_i,\underline{\theta}_j\right\},\bar{\theta}_j}$ is a worst-case belief for an agent of type θ_i who misreports as type $\hat{\theta}_i\in\left(-\theta_j^*,\theta_i\right]$.

for all ϕ_j with expectation θ_j^*. Noting that $E^{\phi_j}\left[\theta_j|\theta_j \leq -\hat{\theta}_i\right] \leq -\hat{\theta}_i$ and rearranging yields

$$E^{\phi_j}\left[\theta_j|\theta_j > -\hat{\theta}_i\right] \geq \frac{1}{\Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)}\left(\theta_j^* + \left(1 - \Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)\right)\hat{\theta}_i\right).$$

Hence, the minimum of (14) over ϕ_j with expectation θ_j^* and a given value of $\Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)$ equals

$$\begin{aligned}
& \Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)\left(\theta_i - \hat{\theta}_i\right) \\
& + \Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)\alpha_i\left(\hat{\theta}_i\right)\left(\frac{1}{\Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)}\left(\theta_j^* + \left(1 - \Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)\right)\hat{\theta}_i\right) + \hat{\theta}_i\right) \\
= \;& \Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)\left(\theta_i - \hat{\theta}_i\right) + \alpha_i\left(\hat{\theta}_i\right)\left(\theta_j^* + \hat{\theta}_i\right).
\end{aligned}$$

As $\hat{\theta}_i \leq \theta_i$, (14) is minimized over ϕ_j with expectation θ_j^* by minimizing $\Pr^{\phi_j}\left(\theta_j > -\hat{\theta}_i\right)$, which by Chebyshev's inequality yields

$$\begin{aligned}
\frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i - \hat{\theta}_i\right) + \alpha_i\left(\hat{\theta}_i\right)\left(\theta_j^* + \hat{\theta}_i\right) &= \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i - \hat{\theta}_i + \alpha_i\left(\hat{\theta}_i\right)\left(\bar{\theta}_j + \hat{\theta}_i\right)\right) \\
&= \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i + t_i\left(\hat{\theta}_i, \bar{\theta}_j\right)\right).
\end{aligned}$$

This gives (13), proving the claim.

Therefore,

$$\sup_{\hat{\theta}_i \in \left(-\theta_j^*, \min\left\{\theta_i, -\underline{\theta}_j\right\}\right]} \inf_{\phi_j \in \Delta(\Theta_j)} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) = \sup_{\hat{\theta}_i \in \left(-\theta_j^*, \min\left\{\theta_i, -\underline{\theta}_j\right\}\right]} \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i + t_i\left(\hat{\theta}_i, \bar{\theta}_j\right)\right).$$

To complete the proof, it suffices to show that the objective is non-decreasing in $\hat{\theta}_i$ over $\left(-\theta_j^*, \min\left\{\theta_i, -\underline{\theta}_j\right\}\right]$. This holds because

$$\begin{aligned}
\alpha_i'\left(\hat{\theta}_i\right) &= \frac{1}{\bar{\theta}_j + \hat{\theta}_i} - \frac{1}{\theta_j^* + \hat{\theta}_i}\alpha_i\left(\hat{\theta}_i\right), \\
\frac{\partial}{\partial \hat{\theta}_i} t_i\left(\hat{\theta}_i, \bar{\theta}_j\right) &= -\frac{\bar{\theta}_j - \theta_j^*}{\theta_j^* + \hat{\theta}_i}\alpha_i\left(\hat{\theta}_i\right),
\end{aligned}$$

and

$$\frac{\partial}{\partial \hat{\theta}_i}\left[\frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}\left(\theta_i + t_i\left(\hat{\theta}_i, \bar{\theta}_j\right)\right)\right] = \frac{\left(\bar{\theta}_j - \theta_j^*\right)\left(\theta_i - \hat{\theta}_i\right)}{\left(\bar{\theta}_j + \hat{\theta}_i\right)^2} \geq 0,$$

where the last inequality follows because $\theta_i \geq \hat{\theta}_i \geq -\theta_j^* \geq -\bar{\theta}_j$.

Proof of Lemma 2

The result is immediate when $\theta_i \leq -\theta_j^*$ or $\theta_i \geq -\underline{\theta}_j$, as in both cases $\alpha_i'(\theta_i) = 0$, which immediately implies that $t_i(\theta_i, \theta_j)$ is non-increasing in θ_i.

If $\theta_i \in \left(-\theta_j^*, -\underline{\theta}_j\right)$ then $t_i(\theta_i, \theta_j) = \alpha_i(\theta_i)\theta_j - (1 - \alpha_i(\theta_i))\theta_i$, and therefore

$$\frac{\partial}{\partial \theta_i} t_i(\theta_i, \theta_j) = -(1 - \alpha_i(\theta_i)) + \alpha_i'(\theta_i)(\theta_j + \theta_i).$$

In addition,

$$\alpha_i'(\theta_i) = \frac{1}{\bar{\theta}_j + \theta_i} - \frac{1}{\theta_j^* + \theta_i}\alpha_i(\theta_i),$$

and therefore

$$\begin{aligned}\frac{\partial}{\partial \theta_i} t_i(\theta_i, \theta_j) &= \frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i}(1 - \alpha_i(\theta_i)) - \frac{\bar{\theta}_j - \theta_j^*}{\theta_j^* + \theta_i}\frac{\theta_j + \theta_i}{\bar{\theta}_j + \theta_i} \\ &= \frac{\bar{\theta}_j - \theta_j^*}{\theta_j^* + \theta_i}\left[\frac{\theta_j - \theta_j^*}{\theta_j^* + \theta_i}\log\left(1 + \frac{\theta_j^* + \theta_i}{\bar{\theta}_j - \theta_j^*}\right) - \frac{\theta_j + \theta_i}{\bar{\theta}_j + \theta_i}\right].\end{aligned}$$

Since $\theta_i > -\theta_j^*$, the sign of $\frac{\partial}{\partial \theta_i} t_i(\theta_i, \theta_j)$ equals the sign of the term in brackets. Using the fact that $\frac{1}{x}\log(1 + x) < 1$ for all $x > 0$, this term is less than

$$\frac{\theta_j - \theta_j^*}{\bar{\theta}_j - \theta_j^*} - \frac{\theta_j + \theta_i}{\bar{\theta}_j + \theta_i} = -\frac{\left(\bar{\theta}_j - \theta_j\right)\left(\theta_j^* + \theta_i\right)}{\left(\bar{\theta}_j - \theta_j^*\right)\left(\bar{\theta}_j + \theta_i\right)} \leq 0,$$

where the last inequality again uses $\theta_i > -\theta_j^*$. Hence, $t_i(\theta_i, \theta_j)$ is non-increasing in θ_i.

Proof of Proposition 2

The proof of necessity is unchanged. For sufficiency, modify the $\alpha_i(\theta_i)$ double auction constructed in the proof of Theorem 2 by letting $\alpha_i(\theta_i) = \min\left\{\frac{1}{2}, \frac{\theta_i - \underline{\theta}_i}{\bar{\theta}_j - \theta_j^*}\right\}\left(1 - \alpha_j\left(\bar{\theta}_j\right)\right)$ for all $\theta_i \le -\theta_j^*$, $i = 1, 2$ (rather than $\alpha_i(\theta_i) = 0$ for $\theta_i \le -\theta_j^*$). The modified mechanism satisfies EF and (ex post) IR as in the proof of Theorem 2. In addition, $t_i(\theta_i, \theta_j)$ is unchanged for all $\theta_i > -\theta_j^*$, and reporting $\hat{\theta}_i \le -\theta_j^*$ continues to give payoff 0 in the worst case, so MMIC also follows as in the proof of Theorem 2.

Next, as $\alpha_i(\theta_i) > 0$ for all $\theta_i > \underline{\theta}_i$ in the modified mechanism, truthtelling is not weakly dominated for any type. In particular, truthtelling was weakly dominated only for types $\theta_i \le -\theta_j^*$ in the unmodified $\alpha_i(\theta_i)$ double auction, and in the modified mechanism such a type does strictly better from truthtelling than from misreporting as type $\hat{\theta}_i < \theta_i$ against any opposing type $\theta_j \in \left(-\theta_i, -\hat{\theta}_i\right)$ (and $t_i(\theta_i, \theta_j)$ is unchanged for all $\theta_i > -\theta_j^*$, so truthtelling does not become weakly dominated for any of these types).

Finally, the argument for WBB is also similar to the proof of Theorem 2. More specifically, as in that proof, consider three cases:

If $\theta_1 < -\underline{\theta}_2$ and $\theta_2 < -\underline{\theta}_1$, then as in the proof of Theorem 2 WBB holds if and only if $\alpha_1(\theta_1) + \alpha_2(\theta_2) \le 1$. If $\theta_i > -\theta_j^*$ for $i = 1, 2$, this holds because $\alpha_1\left(\bar{\theta}_1\right) + \alpha_2\left(\bar{\theta}_2\right) \le 1$ (recalling that $\alpha_i(\theta_i)$ is non-decreasing in the unmodified $\alpha_i(\theta_i)$ double auction). If $\theta_1 > -\theta_2^*$ and $\theta_2 \le -\theta_1^*$, it holds because

$$\alpha_1(\theta_1) + \alpha_2(\theta_2) \le \frac{1}{2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right) + \alpha_2(\theta_2) \le \frac{1}{2} + \frac{1}{2}\alpha_2\left(\bar{\theta}_2\right) < 1.$$

And if $\theta_i \le -\theta_j^*$ for $i = 1, 2$, it holds because

$$\alpha_1(\theta_1) + \alpha_2(\theta_2) \le \frac{1}{2}\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right) + \frac{1}{2}\left(1 - \alpha_1\left(\bar{\theta}_1\right)\right) < 1.$$

If $\theta_1 \ge -\underline{\theta}_2$ and $\theta_2 > -\underline{\theta}_1$, then a fortiori $\theta_1 > -\theta_2^*$ and $\theta_2 > -\theta_1^*$, so the argument is exactly as in the proof of Theorem 2.

Lastly, if $\theta_1 < -\underline{\theta}_2$ and $\theta_2 \ge -\underline{\theta}_1$, then, as in the proof of Theorem 2, WBB reduces to (12). If $\theta_1 > -\theta_2^*$ then the argument is exactly as in the proof of Theorem 2. If instead

$\theta_1 \leq -\theta_2^*$ then (12) becomes

$$\left[\min\left\{\frac{1}{2}, \frac{\theta_1 - \underline{\theta}_1}{\bar{\theta}_2 - \theta_2^*}\right\} - \frac{\theta_1 - \underline{\theta}_1}{\bar{\theta}_2 + \theta_1}\right]\left(1 - \alpha_2\left(\bar{\theta}_2\right)\right) \leq 0,$$

which holds as the term in brackets is non-positive.

Proof of Proposition 3

If $\theta_i \leq -\theta_j^*$, then ex post IR and $\delta_{\theta_j^*} \in \Phi_j$ imply that

$$\min_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) \leq 0,$$

with equality if $\hat{\theta}_i \leq \theta_i$. Thus, if $\varepsilon < 1$ a necessary condition for incentive compatibility (8) is

$$\theta_i \in \arg\max_{\hat{\theta}_i \leq \theta_i} U_i\left(\hat{\theta}_i, \phi_j^{CP}; \theta_i\right) \text{ for all } \theta_i \leq -\theta_j^*.$$

The standard envelope argument now implies that

$$U_i\left(\theta_i, \phi_j^{CP}; \theta_i\right) \geq U_i\left(\underline{\theta}_i, \phi_j^{CP}; \underline{\theta}_i\right) + \int_{\underline{\theta}_i}^{\theta_i} y\left(s, \phi_j^{CP}\right) ds \text{ for all } \theta_i \leq -\theta_j^*.$$

Rearranging and integrating over $\theta_i \leq -\theta_j^*$ (which is possible as $-\theta_j^* < \bar{\theta}_i$) yields

$$\begin{aligned} U_i\left(\underline{\theta}_i, \phi_j^{CP}; \underline{\theta}_i\right) &\leq \int_{\underline{\theta}_i}^{-\theta_j^*}\left[\theta_i y\left(\theta_i, \phi_j^{CP}\right) + t_i\left(\theta_i, \phi_j^{CP}\right) - \left(\int_{\underline{\theta}_i}^{\theta_i} y\left(s, \phi_j^{CP}\right) ds\right)\right] f_i^{CP}\left(\theta_i\right) d\theta_i \\ &= \int_{\underline{\theta}_i}^{-\theta_j^*}\left(\theta_i - \frac{F_i^{CP}\left(-\theta_j^*\right) - F_i^{CP}\left(\theta_i\right)}{f_i^{CP}\left(\theta_i\right)}\right) y\left(\theta_i, \phi_j^{CP}\right) f_i^{CP}\left(\theta_i\right) d\theta_i \\ &\quad + \int_{\underline{\theta}_i}^{-\theta_j^*} t_i\left(\theta_i, \phi_j^{CP}\right) f_i^{CP}\left(\theta_i\right) d\theta_i. \end{aligned} \tag{15}$$

By ex post IR, $t_i\left(\theta_i, \phi_j^{CP}\right)$ may be bounded by

$$\begin{array}{rcl} t_i\left(\theta_i, \phi_j^{CP}\right) & = & \displaystyle\int_{\underline{\theta}_j}^{\bar{\theta}_j} t_i\left(\theta_i, \theta_j\right) f_j^{CP}\left(\theta_j\right) d\theta_j \\ & \leq & \displaystyle\int_{\underline{\theta}_j}^{-\theta_i^*} t_i\left(\theta_i, \theta_j\right) f_j^{CP}\left(\theta_j\right) d\theta_j + \left(1 - F_j\left(-\theta_i^*\right)\right) \bar{\theta}_j. \end{array}$$

Thus, summing (15) over $i = 1, 2$ and using EF and WBB yields

$$\begin{array}{rl} & U_1\left(\underline{\theta}_1, \phi_2^{CP}; \underline{\theta}_1\right) + U_2\left(\underline{\theta}_2, \phi_1^{CP}; \underline{\theta}_2\right) \\ \leq & \displaystyle\int_{\underline{\theta}_1}^{-\theta_2^*} \int_{-\theta_1}^{-\theta_1^*} \left[\theta_1 + \theta_2 - \frac{F_1^{CP}\left(-\theta_2^*\right) - F_1^{CP}\left(\theta_1\right)}{f_1^{CP}\left(\theta_1\right)} - \frac{F_2^{CP}\left(-\theta_1^*\right) - F_2^{CP}\left(\theta_2\right)}{f_2^{CP}\left(\theta_2\right)}\right] f_2^{CP}\left(\theta_2\right) f_1^{CP}\left(\theta_1\right) d\theta_2 d\theta_1 \\ & \displaystyle + \int_{\underline{\theta}_1}^{-\theta_2^*} \int_{\underline{\theta}_2}^{-\theta_1^*} \left[t_1\left(\theta_1, \theta_2\right) + t_2\left(\theta_1, \theta_2\right)\right] f_2^{CP}\left(\theta_2\right) f_1^{CP}\left(\theta_1\right) d\theta_2 d\theta_1 \\ & + \left(1 - F_1^{CP}\left(-\theta_2^*\right)\right) \bar{\theta}_1 + \left(1 - F_2^{CP}\left(-\theta_1^*\right)\right) \bar{\theta}_2 \\ \leq & \displaystyle - \int_{\underline{\theta}_1}^{-\theta_2^*} \left[\left(\theta_1 + \theta_2 - \frac{F_1^{CP}\left(-\theta_2^*\right) - F_1^{CP}\left(\theta_1\right)}{f_1^{CP}\left(\theta_1\right)}\right)\left(F_2^{CP}\left(-\theta_1^*\right) - F_2^{CP}\left(\theta_2\right)\right)\right]_{-\theta_1}^{-\theta_1^*} f_1^{CP}\left(\theta_1\right) d\theta_1 \\ & + \left(1 - F_1^{CP}\left(-\theta_2^*\right)\right) \bar{\theta}_1 + \left(1 - F_2^{CP}\left(-\theta_1^*\right)\right) \bar{\theta}_2 \\ = & \displaystyle - \int_{\underline{\theta}_1}^{-\theta_2^*} \left(F_1^{CP}\left(-\theta_2^*\right) - F_1^{CP}\left(\theta_1\right)\right)\left(F_2^{CP}\left(-\theta_1^*\right) - F_2^{CP}\left(-\theta_1\right)\right) d\theta_1 \\ & + \left(1 - F_1^{CP}\left(-\theta_2^*\right)\right) \bar{\theta}_1 + \left(1 - F_2^{CP}\left(-\theta_1^*\right)\right) \bar{\theta}_2. \end{array}$$

As $\theta_1^* + \theta_2^* < 0$ and $\bar{\theta}_i + \theta_j^* > 0$ for $i = 1, 2$, for every $\eta > 0$ there exists a common prior ϕ^{CP} with positive density on $\left[\underline{\theta}_1, \bar{\theta}_1\right] \times \left[\underline{\theta}_2, \bar{\theta}_2\right]$ such that $F_i^{CP}\left(\frac{2}{3}\left(-\theta_j^*\right) + \frac{1}{3}\left(\theta_i^*\right)\right) - F_i^{CP}\left(\frac{1}{3}\left(-\theta_j^*\right) + \frac{2}{3}\left(\theta_i^*\right)\right) > 1 - \eta$ for $i = 1, 2$. With such a prior, the preceding sum is at most

$$\frac{1}{3}\left(1 - \eta\right)^2\left(\theta_1^* + \theta_2^*\right) + \eta\left(\bar{\theta}_1 + \bar{\theta}_2\right).$$

This is negative for sufficiently small η. Hence, if $\varepsilon < 1$ then a mechanism satisfying (8), EF, and WBB must violate IR for type $\underline{\theta}_1$ or $\underline{\theta}_2$.

Proof of Proposition 4

It suffices to show that the $\alpha_i(\theta_i)$ double auction satisfies (9) when $a \geq 1$. Note that the $\alpha_i(\theta_i)$ double auction satisfies MMIC, and that, for any report she makes, an agent is made weakly worse-off by nature's ability to choose a distribution outside of Φ_j; that is,

$$\min_{\phi_j \in \Delta(\Theta_j)} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) + a\left|E^{\phi_j}[\theta_j] - \theta_j^*\right| \leq \min_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) \text{ for all } \theta_i, \hat{\theta}_i \in \Theta_i. \tag{16}$$

Hence, to show that truthtelling remains optimal with variational preferences, it is enough to show that (16) holds with equality when the agent is truthful. If $\theta_i \leq -\theta_j^*$ then an agent's expected utility from truthtelling remains 0 under variational preferences, so the non-trivial case is where $\theta_i > -\theta_j^*$. In this case, nature's minimization problem

$$\min_{\phi_j \in \Delta(\Theta_j)} U_i\left(\theta_i, \phi_j; \theta_i\right) + a\left|E^{\phi_j}[\theta_j] - \theta_j^*\right|$$

may be decomposed into two steps: first minimizing over distributions ϕ_j with a given expectation $E^{\phi_j}[\theta_j]$, and then minimizing over expectations. Consider the $\theta_i \geq -\underline{\theta}_j$ and $\theta_i < -\underline{\theta}_j$ cases separately.

$\theta_i \geq -\underline{\theta}_j$ case:

For any belief ϕ_j with expectation θ_j^{**}, one has $U_i\left(\theta_i, \phi_j; \theta_i\right) = \theta_i + \alpha_i(\theta_i)\theta_j^{**} + (1 - \alpha_i(\theta_i))\underline{\theta}_j$. Thus, nature's problem becomes

$$\min_{\theta_j^{**}} \theta_i + \alpha_i(\theta_i)\theta_j^{**} + (1 - \alpha_i(\theta_i))\underline{\theta}_j + a\left|\theta_j^{**} - \theta_j^*\right|.$$

As $\alpha_i(\theta_i) \in (0,1)$ and $a \geq 1$, the solution is $\theta_j^{**} = \theta_j^*$. Hence, (16) holds with equality.

$\theta_i < -\underline{\theta}_j$ case:

The proof of Theorem 2 shows that the worst-case belief with expectation θ_j^{**} is the 2-point distribution on $\left\{-\theta_i, \bar{\theta}_j\right\}$, which yields expected utility

$$\begin{aligned}\frac{\theta_j^{**} + \theta_i}{\bar{\theta}_j + \theta_i} U_i\left(\theta_i, \bar{\theta}_j; \theta_i\right) &= \frac{\theta_j^{**} + \theta_i}{\bar{\theta}_j + \theta_i} \alpha_i(\theta_i)\left(\bar{\theta}_j + \theta_i\right) \\ &= \left(\theta_j^{**} + \theta_i\right)\alpha_i(\theta_i).\end{aligned}$$

Thus, nature's problem becomes

$$\min_{\theta_j^{**}} \left(\theta_j^{**} + \theta_i\right) \alpha_i \left(\theta_i\right) + a \left|\theta_j^{**} - \theta_j^*\right| .$$

Again, the solution is $\theta_j^{**} = \theta_j^*$, so (16) holds with equality.

Proof of Proposition 5

Sufficiency:

When $\underline{\theta}_1 + \underline{\theta}_2 \geq 0$, any reference rule with $t_i^* \in \left[-\underline{\theta}_i, \underline{\theta}_j\right]$ satisfies MMIC.

When $\theta_1^* + \theta_2^* \leq 0$, I show that any reference rule with $t_i^* \in \left[\theta_j^*, -\theta_i^*\right]$ satisfies MMIC.

First, suppose that $\theta_i < -\theta_j^*$. Observe that $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) \leq 0$ for all $\hat{\theta}_i$. This follows because if $\hat{\theta}_i + \theta_j^* < 0$ then $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) = 0$, while if $\hat{\theta}_i + \theta_j^* \geq 0$ then $U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right) = \theta_i + \theta_j^* < 0$. Hence, for all $\hat{\theta}_i$, IR and $\delta_{\theta_j^*} \in \Phi_j$ imply that $U_i\left(\theta_i\right) \geq 0 \geq \inf_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right)$, which yields MMIC.

Next, suppose that $\theta_i \geq -\theta_j^*$. First, note that misreports of $\hat{\theta}_i \leq -\theta_j^*$ cannot be profitable because $U_i\left(\theta_i\right) \geq 0 = U_i\left(\hat{\theta}_i, \theta_j^*; \theta_i\right)$ and $\delta_{\theta_j^*} \in \Phi_j$. Next, consider misreports of $\hat{\theta}_i > -\theta_j^*$. If $y\left(\theta_i, \theta_j\right) = y\left(\hat{\theta}_i, \theta_j\right)$, then $t_i\left(\theta_i, \theta_j\right) = t_i\left(\hat{\theta}_i, \theta_j\right)$ (as $\theta_i, \hat{\theta}_i \geq -\theta_j^* \geq -t_i^*$), and hence $U_i\left(\theta_i, \theta_j; \theta_i\right) = U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$. In addition, if $y\left(\theta_i, \theta_j\right) = 1$ and $y\left(\hat{\theta}_i, \theta_j\right) = 0$ then $U_i\left(\theta_i, \theta_j; \theta_i\right) \geq 0 = U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$ by ex post IR. Finally, if $y\left(\theta_i, \theta_j\right) = 0$ and $y\left(\hat{\theta}_i, \theta_j\right) = 1$ then $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) \leq \theta_i + \theta_j < 0 = U_i\left(\theta_i, \theta_j; \theta_i\right)$. Hence, $U_i\left(\theta_i, \theta_j; \theta_i\right) \geq U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right)$ for all θ_j, so misreports of $\hat{\theta}_i > -\theta_j^*$ cannot be profitable, either. This yields MMIC.

Necessity:

Since $\theta_1^* + \theta_2^* > 0$, every reference rule satisfies either $t_1^* < \theta_2^*$ or $t_1^* > -\theta_1^*$, and hence $t_i^* < \theta_j^*$ for some $i \in \{1, 2\}$. Fix this choice of i.

First, suppose that $\underline{\theta}_i \leq -t_i^*$. Then there exists a type $\theta_i \in \left(-\theta_j^*, -t_i^*\right] \cap \Theta_i$. Note that $U_i\left(\theta_i, \theta_j; \theta_i\right) = 0$ for all θ_j. However, $U_i\left(\hat{\theta}_i, \theta_j; \theta_i\right) = \theta_i - \hat{\theta}_i > 0$ whenever $\hat{\theta}_i < \theta_i$ and $\hat{\theta}_i + \theta_j \geq 0$. In addition, for all $\hat{\theta}_i \in \left(-\theta_j^*, \theta_i\right)$ and all $\phi_j \in \Phi_j$, Chebyshev's inequality yields $\Pr^{\phi_j}\left(\hat{\theta}_i + \theta_j \geq 0\right) \geq \frac{\theta_j^* + \hat{\theta}_i}{\bar{\theta}_j + \hat{\theta}_i}$, and therefore $U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) \geq \frac{\left(\theta_j^* + \hat{\theta}_i\right)\left(\theta_i - \hat{\theta}_i\right)}{\bar{\theta}_j + \hat{\theta}_i} > 0$. Hence, $\inf_{\phi_j \in \Phi_j} U_i\left(\hat{\theta}_i, \phi_j; \theta_i\right) \geq \frac{\left(\theta_j^* + \hat{\theta}_i\right)\left(\theta_i - \hat{\theta}_i\right)}{\bar{\theta}_j + \hat{\theta}_i} > 0 = \inf_{\phi_j \in \Phi_j} U_i\left(\theta_i, \phi_j; \theta_i\right)$, so MMIC fails.

Next, suppose that $\underline{\theta}_i > -t_i^*$. Then there exists a type $\theta_j \in \left(-\underline{\theta}_i, -t_j^*\right) \cap \Theta_j$ (as

$\underline{\theta}_j < -\underline{\theta}_i < t_i^* = -t_j^* < \theta_j^* < \bar{\theta}_j$). Note that $U_j(\theta_j, \theta_i; \theta_j) = 0$ for all θ_i. However, $U_j\left(\hat{\theta}_j, \theta_i; \theta_j\right) = \theta_j - \hat{\theta}_j > 0$ whenever $\hat{\theta}_j < \theta_j$ and $\theta_i + \hat{\theta}_j \geq 0$. Now for all $\hat{\theta}_j \in (-\underline{\theta}_i, \theta_j)$ and all $\phi_i \in \Phi_i$, one has $\Pr^{\phi_i}\left(\theta_i + \hat{\theta}_j \geq 0\right) = 1$ and therefore $U_j\left(\hat{\theta}_j, \phi_i; \theta_j\right) = \theta_j - \hat{\theta}_j$. Hence, $\inf_{\phi_i \in \Phi_i} U_j\left(\hat{\theta}_j, \phi_i; \theta_j\right) = \theta_j - \hat{\theta}_j > 0 = \inf_{\phi_i \in \Phi_i} U_j(\theta_j, \phi_i; \theta_j)$, so MMIC fails.

Appendix B: Revelation Principle

In this appendix, I consider the more general setting with an arbitrary set of alternatives Y and payoff functions $u_i(y, \theta_i)$, to emphasize that the revelation principle does not depend on quasilinear utility.

Consider an arbitrary mechanism (game form) $G = (S, g)$, with strategy set $S = (S_i)_{i \in N}$ and outcome function $g : S \to Y$. A *mixed strategy profile* is a function $\sigma : \Theta \to \prod_i \Delta(S_i)$. Type θ_i's *interim maxmin expected utility* at mixed strategy profile σ is

$$E^{\sigma_i(\theta_i)}\left[\inf_{\phi_{-i} \in \Phi_{-i}} E^{\phi_{-i}}\left[E^{\sigma_{-i}(\theta_{-i})}\left[u_i\left(g\left(s_i, s_{-i}\right), \theta_i\right)\right]\right]\right],$$

where the outer expectation is over s_i, the middle expectation is over θ_{-i}, and the inner expectation is over s_{-i}. This definition implicitly assumes that agents cannot commit to randomize, in that the maxmin criterion is applied to $\sigma_i(\theta_i)$ realization-by-realization. A mixed strategy profile σ^* is a *maxmin Nash equilibrium* if, for all $\theta_i \in \Theta_i$,

$$\sigma_i^*(\theta_i) \in \arg\max_{\sigma_i(\theta_i) \in \Delta(S_i)} E^{\sigma_i(\theta_i)}\left[\inf_{\phi_{-i} \in \Phi_{-i}} E^{\phi_{-i}}\left[E^{\sigma_{-i}(\theta_{-i})}\left[u_i\left(g\left(s_i, s_{-i}\right), \theta_i\right)\right]\right]\right].$$

A social choice rule $f : \Theta \to Y$ is *maxmin Nash implementable* if there exists a mechanism $G = (S, g)$ and a pure strategy maxmin Nash equilibrium s^* of G such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$. A social choice rule f is maxmin incentive compatible if it is MMIC when viewed as a direct mechanism. The relevant version of the revelation principle is as follows.

Proposition 6 (Revelation Principle) *If a social choice rule f is maxmin Nash implementable, then it is maxmin incentive compatible.*

Proof. If f is maxmin Nash implementable, then there exists a mechanism (S, g) and a pure strategy profile $s^* \in S$ such that $g\left(s^*\left(\theta\right)\right) = f\left(\theta\right)$ for all $\theta \in \Theta$ and

$$s_i^*\left(\theta_i\right) \in \arg \max_{\sigma_i(\theta_i) \in \Delta(S_i)} E^{\sigma_i(\theta_i)}\left[\inf_{\phi_{-i} \in \Phi_{-i}} E^{\phi_{-i}}\left[u_i\left(g\left(s_i, s_{-i}^*\left(\theta_{-i}\right)\right), \theta_i\right)\right]\right] \text{ for all } \theta_i \in \Theta_i.$$

In particular,

$$s_i^*\left(\theta_i\right) \in \arg \max_{\hat{\theta}_i \in \Theta_i} \inf_{\phi_{-i} \in \Phi_{-i}} E^{\phi_{-i}}\left[u_i\left(g\left(s_i^*\left(\hat{\theta}_i\right), s_{-i}^*\left(\theta_{-i}\right)\right), \theta_i\right)\right] \text{ for all } \theta_i \in \Theta_i.$$

In other words, the direct mechanism $g \circ s^* = f$ is maxmin incentive compatible. ■

The notion of implementability considered here is more general than that in the text, as here quasilinearity, individual rationality, and budget balance are not imposed. The same argument applies under these additional requirements.

References

[1] Ahn, D.S. (2007), “Hierarchies of Ambiguous Beliefs,” *Journal of Economic Theory*, 136, 286-301.

[2] Arrow, K. (1979), “The Property Rights Doctrine and Demand Revelation under Incomplete Information,” in *Economics and Human Welfare*, M. Boskin, ed., Academic Press: New York.

[3] Bergemann, D. and S. Morris (2005), “Robust Mechanism Design,” *Econometrica*, 73, 1771-1813.

[4] Bergemann, D. and K.H. Schlag (2008), “Pricing without Priors,” *Journal of the European Economic Association*, 6, 560-569.

[5] Bergemann, D. and K. Schlag (2011), “Robust Monopoly Pricing,” *Journal of Economic Theory*, 146, 2527-2543.

[6] Bewley, T. (1986), “Knightian Decision Theory: Part I,” *mimeo.*

[7] Bodoh-Creed, A. (2012), “Ambiguous Beliefs and Mechanism Design,” *Games and Economic Behavior*, 75, 518-537.

[8] Bodoh-Creed, A. (2014), “Correction to Ambiguous Beliefs and Mechanism Design,” *mimeo.*

[9] Börgers, T. and D. Smith (2012), “Robustly Ranking Mechanisms,” *American Economic Review (Papers and Proceedings)*, 102, 325-329.

[10] Bose, S. and A. Daripa (2009), "A Dynamic Mechanism and Surplus Extraction Under Ambiguity," *Journal of Economic Theory*, 144, 2084-2115.

[11] Bose, S. and S. Mutuswami (2012), "Bilateral Bargaining in an Ambiguous Environment," *mimeo.*

[12] Bose, S., E. Ozdenoren, and A. Pape (2006), "Optimal Auctions with Ambiguity," *Theoretical Economics*, 1, 411-438.

[13] Bose, S. and L. Renou (2013), "Mechanism Design with Ambiguous Communication Devices," *mimeo.*

[14] Carbajal, J.C. and J.C. Ely (2013), "Mechanism Design without Revenue Equivalence," *Journal of Economic Theory*, 148, 104-133.

[15] Chung, K.-S. and J.C. Ely. (2007), "Foundations of Dominant-Strategy Mechanisms," *Review of Economic Studies*, 74, 447-476.

[16] Crémer, J. and R.P. McLean (1985), "Optimal Selling Strategies under Uncertainty for a Discriminating Monopolist when Demands are Interdependent," *Econometrica*, 53, 345-361.

[17] Crémer, J. and R.P. McLean (1988), "Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions," *Econometrica*, 56, 1247-1257.

[18] d'Aspremont, C. and L.-A. Gérard-Varet (1979), "Incentives and Incomplete Information," *Journal of Public Economics*, 11, 25-45.

[19] De Castro, L. and N. Yannelis (2010), "Ambiguity Aversion Solves the Conflict Between Efficiency and Incentive Compatibility," *mimeo.*

[20] Di Tillio, A., N. Kos, and M. Messner (2012), "Designing Ambiguous Mechanisms," *mimeo.*

[21] Erdil, A. and P. Klemperer (2010), "A New Payment Rule for Core-Selecting Package Auctions," *Journal of the European Economic Association*, 8, 537-547.

[22] Gilboa, I. (2009), *Theory of Decision under Uncertainty*, Cambridge University Press: Cambridge.

[23] Gilboa, I. and M. Marinacci (2011), "Ambiguity and the Bayesian Paradigm," in *Advances in Economics and Econometrics: Tenth World Congress*, D. Acemoglu, M. Arellano, E. Dekel, eds., Cambridge University Press: Cambridge.

[24] Gilboa, I. and D. Schmeidler (1989), "Maxmin Expected Utility with Non-Unique Prior," *Journal of Mathematical Economics*, 18, 141-153.

[25] Grimmett, G. and D. Stirzaker (2001), *Probability and Random Processes*, Oxford University Press: Oxford.

[26] Kamenica, E. and M. Gentzkow (2011), "Bayesian Persuasion," *American Economic Review*, 101, 2590-2615.

[27] Kopylov, I. (2008), "Subjective Choice and Confidence," *mimeo.*

[28] Kos, N. and M. Messner (2013), "Extremal Incentive Compatible Transfers," *Journal of Economic Theory*, 148, 134-164.

[29] Krishna, V. and M. Perry (2000), "Efficient Mechanism Design," *mimeo.*

[30] Linhart, P.B., and R. Radner (1989), "Minimax-Regret Strategies for Bargaining over Several Variables," *Journal of Economic Theory*, 48, 152-178.

[31] Lopomo, G., L. Rigotti, and C. Shannon (2009), "Uncertainty in Mechanism Design," *mimeo.*

[32] Maccheroni, F., M. Marinacci, and A. Rustichini (2006), "Ambiguity Aversion, Robustness, and the Variational Representation of Preferences," *Econometrica*, 74, 1447-1498.

[33] Mailath, G. and A. Postlewaite (1990), "Asymmetric Information Bargaining Problems with Many Agents," *Review of Economic Studies*, 57, 351-367.

[34] Makowski, L. and C. Mezzetti (1994), "Bayesian and Weakly Robust First Best Mechanisms: Characterizations," *Journal of Economic Theory*, 64, 500-519.

[35] McAfee, R.P. and P.J. Reny (1992), "Correlated Information and Mechanism Design," *Econometrica*, 60, 395-421.

[36] Milgrom, P. and I. Segal (2002), "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70, 583-601.

[37] Myerson, R. and M. Satterthwaite (1983), "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory*, 29, 265-281.

[38] Segal, I. and M.D. Whinston (2002), "The Mirrlees Approach to Mechanism Design with Renegotiation (with Applications to Hold-Up and Risk-Sharing)," *Econometrica*, 70, 1-45.

[39] Shmaya, E. and L. Yariv (2009), "Foundations for Bayesian Updating," *mimeo.*

[40] Williams, S.R. (1999), "A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms," *Economic Theory*, 14, 155-180.

[41] Wilson, R. (1987), "Game Theoretic Analysis of Trading Processes," in *Advances in Economic Theory*, T. Bewley, ed., Cambridge University Press: Cambridge.