# Mechanism Design with Possibilistic Beliefs*

Jing Chen† Silvio Micali‡

August 8, 2013

**Abstract**

We study mechanism design in non-Bayesian settings of incomplete information, when the designer has no information about the players, and the players have arbitrary, heterogeneous, first-order, and *possibilistic* beliefs about their opponents' payoff types.

Using such beliefs, in auctions of a single good, we

- define a revenue benchmark at least as high as the second-highest valuation, and sometimes much higher;
- prove that it is not meaningfully achievable via traditional notions of implementation; and
- prove that it is achievable via a notion of implementation based only on mutual belief of rationality.

**JEL classification:** C72, D80, D82, D44

**Keywords:** incomplete information, single-good auctions, first-order beliefs, conservative beliefs

*We would like to thank Gabriel Carroll, Robert Kleinberg, and Ronald Rivest for discussions that motivated us to prove results stronger than the ones we originally had. We also would like to thank Amos Fiat and Anna Karlin for helping us improve the presentation of our results. Many thanks also to Andrés Perea for helping us understand beliefs in economic settings, to Paul Milgrom for helping us clarify the fragility of implementation at equilibria, to Elchanan Ben-Porath, Sergiu Hart, and Philip Reny for helping us clarify our connections to ex-post equilibria, to Larry Blume and an anonymous referee for many helpful suggestions about the presentation of our results, and to the associate editor for his vision, support and guidance. This work is supported in part by ONR Grant No. N00014-09-1-0597.

†Department of Computer Science, Stony Brook University, jingchen@cs.stonybrook.edu.

‡Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, silvio@csail.mit.edu.

# 1 Introduction

We focus on settings of *incomplete* information. Here, a player $i$ knows precisely $\theta_i$, his own (payoff) type, but not $\theta_{-i}$, the type subprofile of his opponents. Accordingly, he may have all kinds of beliefs (even wrong ones) about $\theta_{-i}$. We refer to such beliefs as $i$'s *external beliefs*, and to $\theta_i$ as his *internal knowledge*.

**Motivation** For achieving a desired outcome, a mechanism designer should in general consider leveraging both the players' internal knowledge and their external beliefs. Mechanisms working in dominant or undominated strategies leverage the former, but not the latter.[1]

Mechanisms using Bayesian Nash equilibrium as their underlying solution concept leverage both, but under a strong assumption: namely, that the type profile $\theta$ is drawn by nature from a distribution $\mathcal{D}$ that is common knowledge to the players (the "common-prior assumption").

A weaker assumption is that the players have *heterogeneous* beliefs: namely, each player $i$ has his own distribution, $D_i$, from which he subjectively believes nature has drawn $\theta$. Although weaker than the common-prior one, this assumption presupposes that, even when the number of players is high and the size of the type space is large, $i$ is sure, for every two type profiles $\theta'$ and $\theta''$, how much more likely is —say— $\theta'$ than $\theta''$.

A player's belief, however, need not be so detailed. For example, in an auction of a house, a player valuing the house for 500,000 dollars may believe (possibly erroneously) that one of his opponents values it for more than one million dollars, without having the vaguest idea about who such a high-valuing player might be, or what the probabilities for her valuation being \$1,000,001, \$1,000,002, etc., might be.

**Goal** In sum, classical mechanisms exploit two extremes: (1) the players have no external beliefs and (2) the players have probabilistic external beliefs. We instead wish to explore what mechanism design can and cannot do when the players hold *possibilistic* beliefs. Specifically we wish to understand which new social choice correspondences we may implement, which solution concepts do not work, and which do.

**Contributions** We let the *conservative belief* of a player $i$ consist of a *set* $\mathcal{B}_i$: the set of all possible candidates for $\theta$ in $i$'s mind. In particular, player $i$ may have no idea about the relative likelihood of any two type profiles in $\mathcal{B}_i$.

In auctions of a single good we use these conservative beliefs to define a revenue benchmark that is always at least as high as the second-highest valuation and sometimes much higher.

We prove that our revenue benchmark cannot be meaningfully achieved under classical non-Bayesian solution concepts, such as implementation in undominated strategies (and thus implementation in dominant strategies), or implementation in ex-post equilibrium. These impossibility results hold even if the designer is allowed to elicit information about the players' beliefs (rather than just their own valuations).

We prove, however, that the players' conservative beliefs can be leveraged, and that our benchmark can be virtually achieved by a simple mechanism without having any a priori information about the players' valuations or beliefs. Although not used before, the solution concept underlying our mechanism is natural and compelling. In particular, it relies on the players' *mutual* —rather than *common*— belief of rationality.

Our mechanism leverages the players' conservative beliefs in a very resilient manner. That is, it virtually achieves our revenue benchmark no matter what additional beliefs the players may have, as long as such beliefs do not contradict the above-mentioned ones.

After presenting our results, in Section 8 we compare them with prior ones.

**Finiteness** While our framework is very general, our theorems focus solely on single-good auctions where all valuations are non-negative integers upperbounded by some value $V$, and all mechanisms provide each player with a finite number of pure strategies.

[1] Whenever such mechanisms exist, they achieve their goals no matter what external beliefs the players may have.

## 2 Conservative Beliefs

In a context of incomplete information, we denote by $N = \{1, \ldots, n\}$ the set of players; by $\Omega$ the finite set of outcomes; by $\Theta = \Theta_1 \times \cdots \times \Theta_n$ the set of all possible (payoff) type profiles; by $u$ the profile of utility functions (where each $u_i$ maps $\Theta_i \times \Omega$ into $\mathbb{R}$); and by $\theta \in \Theta$ the profile of true types. If $t_i \in \Theta_i$ and $\omega$ is a distribution over $\Omega$, then $u_i(t_i, \omega)$ is the expected utility induced by $\omega$. As usual, $N$, $\Omega$, $\Theta$, and $u$ are common knowledge to the players and the mechanism designer, and each player $i$ individually knows $\theta_i$.

In such a context, we model a player $i$'s beliefs as a set, the set $\mathcal{B}_i$ of *all possible candidates for the true type profile in player i's view.* Since $i$ knows his own type, the $i$th component of each element of $\mathcal{B}_i$ coincides with $\theta_i$. More formally:

**Definition 1.** *The **(conservative) belief profile** of a context is a profile $\mathcal{B}$ such that, for each player $i$,*

$$\mathcal{B}_i \subset \Theta_1 \times \cdots \times \Theta_{i-1} \times \{\theta_i\} \times \Theta_{i+1} \times \cdots \times \Theta_n \quad \text{and} \quad i \text{ individually knows } \mathcal{B}_i.$$

*We say that $\mathcal{B}_i$ is **correct** if $\theta \in \mathcal{B}_i$, and that $\mathcal{B}$ is correct if each $\mathcal{B}_i$ is correct.*

Conservative beliefs are deliberately simple. They can specify, as a special case, every context of complete information, but cannot specify even a single non-degenerate Bayesian context. In addition, they do not include the players' higher-order beliefs (i.e., their beliefs about their opponents' beliefs, etc.). Accordingly, $t_i \in \mathcal{B}_i$ is not a full type of player $i$ in the sense of Harsanyi, whether represented as sets or distributions.

Each set $\mathcal{B}_i$, of course, can be described in traditional economic terms. Following Savage and Harsanyi, players have subjective beliefs over the moves of Nature, which include picking full types for the players. For each player $i$ this corresponds to a distribution $D_i$ over his opponents' payoff types, their beliefs, beliefs about beliefs, etc. Thus the set $\mathcal{B}_i$ corresponds to the support of $D_{i|\theta_{-i}}$ (i.e., $i$'s subjective marginal distribution over the payoff types of his opponents). Let us now clarify four points about conservative beliefs.

1. *Arbitrary Additional Beliefs.* Each $\mathcal{B}_i$ consists of all candidates for $\theta$ in player $i$'s mind, but NOT of all beliefs of $i$ about his opponents. The players (or the mechanism designer) may have arbitrary additional knowledge or beliefs, even of a probabilistic nature. In no case, however, can the additional beliefs of a player $i$ contradict $\mathcal{B}_i$. For example, $i$ may additionally believe that another player $j$'s type is $\theta'_j$ with a probability between 1/3 and 2/3, but then there must exist $t \in \mathcal{B}_i$ such that $t_j = \theta'_j$.
2. *Conservative Beliefs Can Be Wrong.* It may even be the case that $\theta \notin \mathcal{B}_i$ for each player $i$.
3. *Leveraging Conservative Beliefs.* A mechanism $M$ leveraging $\mathcal{B}$ should work even when every $\mathcal{B}_i$ is wrong. In addition, to be resilient, $M$ should not assume that $\mathcal{B}$ captures all beliefs of the players, but should achieve its desired outcomes no matter what additional beliefs the players might have.
4. *Conservative Beliefs Always Exist.* Conservative beliefs are more a *model* than an *assumption.* As traditional in mechanism design, each player $i$ is assumed to know his true type $\theta_i$, but we make no requirement about his external beliefs. For instance, $i$ may have no external belief whatsoever. In this case, $\mathcal{B}_i = \Theta_1 \times \cdots \times \Theta_{i-1} \times \{\theta_i\} \times \Theta_{i+1} \times \cdots \times \Theta_n$. On the other extreme, he may have no external uncertainty whatsoever. In this case, $\mathcal{B}_i = \{t\}$ for some type profile $t$ (not necessarily equal to $\theta$).[2]

**Conservative Beliefs and Extended Social Choice Correspondences** We find it natural to let social choice correspondences map *conservative-belief profiles*, rather than *type profiles*, to (distributions over) sets of outcomes. Since "conservative beliefs always exist", each context implicitly has a conservative-belief profile $\mathcal{B}$, and from every such profile one can compute the true type profile $\theta$. Thus, for each traditional correspondence $f$ there exists an extended correspondence $F$ such that "$f(\theta) = F(\mathcal{B})$", but not vice versa.

Extended social choice correspondences legitimately and usefully enlarge the set of possible "targets" in mechanism design. By implementing an extended social choice correspondence that is not expressible in terms of $\theta$ alone, a mechanism designer might be able to produce either outcomes that are more desirable, or, when the "originally desired" ones are impossible to implement (e.g., under a given solution concept), alternative outcomes that are reasonably good.

[2] If the context were one of *complete information*, then necessarily $\mathcal{B}_i = \{\theta\}$ for all $i$.

## 3 The Second-Belief Revenue Benchmark

Our revenue benchmark applies to (finite) single-good auctions with at least two players. In such an auction, a possible type (or *valuation*) is an integer in $\{0, 1, \ldots, V\}$ for some $V$, representing a player's possible value for the good for sale. Accordingly, each conservative belief $\mathcal{B}_i$ is a subset of $\{0, 1, \ldots, V\}^n$.

Intuitively, our benchmark is so described: the highest number $v$ such that there are at least two players believing that there exists a player (whose identity need not be known) valuing the good $v$.

**Definition 2.** *Let $\mathcal{B}$ be the conservative belief profile of a single-good auction. Then relative to $\mathcal{B}$*

$$smv_i \triangleq \min_{t \in \mathcal{B}_i} \max_{j \in N} t_j \quad \text{and} \quad 2^{nd}(\mathcal{B}) \triangleq \text{the second highest value in } \{smv_i : i \in N\}.$$

*We refer to the function $2^{nd}(\cdot)$ as the* ***second-belief benchmark****.*

Naturally, our revenue benchmark defines an extended social choice correspondence which maps $\mathcal{B}$ to the set of outcomes with revenue at least $2^{nd}(\mathcal{B})$.

Let us now reconcile the intuitive description of our benchmark with that of Definition 2. If $t$ were the true valuation profile, then $\max_j t_j$ would be the maximum valuation a player has for the good. Accordingly, since player $i$ believes that $\mathcal{B}_i$ is the set of all possible candidates for the true valuation profile, it follows that $smv_i$, "the sure maximum value according to $i$", is the maximum value which $i$ is *sure some* player (possibly $i$ himself or a player whose identity is not precisely known to $i$) has for the good. Thus the second highest of the $smv_i$'s indeed coincides with the benchmark intuitively described above.

A Simple Example. Consider an auction with three players where $\theta = (100, 80, 60)$ and

$\mathcal{B}_1 = \{(100, x, y) : x \geq 0, y \geq 0\}$, $\mathcal{B}_2 = \{(100, 80, x), (y, 80, 100) : x \geq 0, y \geq 0\}$, and $\mathcal{B}_3 = \{(150, 0, 60)\}$.

Here, the beliefs of players 1 and 2 are correct, but that of player 3 is wrong. Player 1 has no external beliefs: in his eyes, all valuations are possible for his two opponents. Player 2 believes that either player 1 or player 3 has valuation 100, but cannot tell whom. Player 3 has no external uncertainty: in his eyes, $(150, 0, 60)$ is the true valuation profile. According to $\mathcal{B}$, $smv_1 = smv_2 = 100$ and $smv_3 = 150$. Thus $2^{nd}(\mathcal{B}) = 100$, which in this specific case happens to be the highest valuation. □

Our benchmark clearly satisfies the following properties.

1. It is never lower than the second highest valuation. (Indeed $smv_i \geq \theta_i$ for all $i$.)
2. It cannot exceed the highest valuation when the players' beliefs are correct.
3. It may exceed the highest valuation when the beliefs of at least two players are wrong.

## 4 Notation

An auction is a game $G$ consisting of a (conservative-belief) auction context $C$ (describing the players, the outcomes, the players' preferences over outcomes, and the players' beliefs) and a mechanism (describing the strategies available to the players and how strategies lead to outcomes): $G = (C, M)$. All our auctions are (finite and) for a single good.

**Auction Contexts** A *conservative-belief* auction context is a tuple $(n, \Omega, \Theta, u, \theta, \mathcal{B})$ where

- $\Theta = \{0, 1, \ldots, V\}^n$, for some $V$ referred to as the *valuation bound*;
- $\Omega = \{0, 1, \ldots, n\} \times \mathbb{R}^n$;
- For any $t_i \in \Theta_i$, $u_i(t_i, (a, P))$ equals $t_i - P_i$ if $i = a$, and $-P_i$ otherwise; and
- $\mathcal{B}$ is a conservative-belief profile.

If $(a, P) \in \Omega$, then $P$ is the *price profile* and $a$ the *allocation*. If $a = 0$ then the good is unallocated, else $a$ is the player getting the good. If $\omega = (a, P) \in \Omega$, then player $i$'s utility for $\omega$, $u_i(\omega)$, is $u_i(\theta_i, \omega)$; and the *revenue* of $\omega$, $REV(\omega)$, is $\sum_i P_i$.

Notice that such a context $C$ is identified by just $n$, $V$, $\theta$ and $\mathcal{B}$ alone: that is, $C = (n, V, \theta, \mathcal{B})$.

**Auction Mechanisms** A mechanism $M$ for a conservative-belief auction context with $n$ players and outcome set $\Omega$ specifies

- for each player $i$, the set of all pure strategies of $i$, $S_i$, and
- a function from $S = S_1 \times \cdots \times S_n$ to $\Omega$, typically (although a bit ambiguously) also denoted by $M$.

If $s \in S$, then $M(s)$ denotes the outcome (the distribution over outcomes if $M$ is probabilistic) generated by $M$, and $u_i(M(s))$ —or more simply $u_i(s)$ if the underlying mechanism $M$ is clear— the corresponding utility (expected utility if $M$ is probabilistic) of player $i$.

An auction mechanism $M$ is *interim individually rational* (IIR) if, for each player $i$ and possible true type $t_i$ of $i$, there exists a strategy $out_i \in S_i$ such that $u_i(M(out_i, s_{-i})) = 0$ for every strategy subprofile $s_{-i} \in S_{-i}$.

**Domination** Consider an auction (more generally, a game) $G = (C, M)$, and denote as usual by $\Delta(A)$ the set of probabilistic distributions over a set $A$.

A strategy $s_i \in S_i$ is *weakly dominated* by another (possibly mixed) strategy $\sigma_i \in \Delta(S_i)$ if $u_i(\sigma_i, s_{-i}) \geq u_i(s_i, s_{-i})$ for every strategy subprofile $s_{-i}$ of the others, and $u_i(\sigma_i, s'_{-i}) > u_i(s_i, s'_{-i})$ for some strategy subprofile $s'_{-i}$. A strategy $s_i$ is *undominated* if it is not weakly dominated by any strategy. A strategy $s_i$ is *purely undominated* if it is not weakly dominated by any pure strategy.

A strategy $s_i \in S_i$ is *strictly dominant* if for every other strategy $s'_i$, $u_i(s_i, s_{-i}) > u_i(s'_i, s_{-i})$ for every strategy subprofile $s_{-i}$. Strategy $s_i$ is *weakly dominant* if for every other strategy $s'_i$, $s'_i$ is weakly dominated by $s_i$. Strategy $s_i$ is *very weakly dominant* if for every other strategy $s'_i$, $u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$ for every $s_{-i}$.

# 5 The Impossibility of Implementing $2^{nd}(\cdot)$ in Undominated Strategies

Implementation in undominated strategies —which includes implementation in dominant strategies as a special case— is a classical choice for non-Bayesian settings of incomplete information. In this section we prove that the second-belief benchmark cannot be implemented according to this notion. We note that this impossibility result is trivial when the players' beliefs are wrong.[3] Accordingly, we state our result directly for contexts with *correct* conservative beliefs.

We actually prove an impossibility result that is stronger in two ways. First, we prove that no undominated-strategy mechanism $M$ can even *approximately* implement our benchmark. That is, we prove that $M$ cannot guarantee (no matter what the number $n$ of players and the valuation bound $V$ may be) a fraction $\varepsilon$ of our revenue benchmark: specifically, no fraction greater than 1/2 for probabilistic undominated-strategy mechanisms, no fraction greater than 0 for deterministic ones. Second, we prove that this stronger impossibility result applies even if one adopts a *weaker* notion of implementation in undominated strategies.[4]

**Theorem 1.** *For all $\varepsilon \in (\frac{1}{2}, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon - 1/2} \rceil$, and probabilistic IIR mechanisms $M$, there exist (1) an auction context $C$ with $n$ players, valuation bound $V$, and a correct conservative-belief profile $\mathcal{B}$, and (2) a profile $s$ of undominated strategies in the auction $(C, M)$ such that*

$$REV(M(s)) < \varepsilon \cdot 2^{nd}(\mathcal{B}).$$

**Theorem 2.** *For all $\varepsilon \in (0, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon} \rceil$, and deterministic IIR mechanisms $M$, there exist (1) an auction context $C$ with $n$ players, valuation bound $V$, and a correct conservative-belief profile $\mathcal{B}$, and (2) a profile $s$ of purely undominated strategies in the auction $(C, M)$ such that*

$$REV(M(s)) < \varepsilon \cdot 2^{nd}(\mathcal{B}).$$

[3] This is so because, when more than one player's beliefs are not correct, it is trivial to construct contexts for which the second-belief benchmark is much greater than the highest valuation. And no classical notion of implementation can guarantee revenue greater than the highest valuation.

[4] Note that the traditional notion of (full) implementation in undominated strategies —see Jackson [19]— requires not only that every profile of undominated strategies yields an outcome satisfying the desired social choice correspondence, but also that, conversely, for each desired outcome there exists a profile of undominated strategies yielding that outcome. By removing the latter requirement we weaken the notion of implementation and thus strengthen the impossibility result of Theorem 1.

We prove Theorem 1 in Subsection 5.1. The proof of Theorem 2 is not only similar but simpler, and is thus omitted. These theorems have two immediate consequences about implementation in strictly/weakly/very weakly dominant strategies.

**Corollary 1.** *For all $\varepsilon \in (\frac{1}{2}, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon - 1/2} \rceil$, and probabilistic IIR mechanisms $M$, there exists an auction context $C$ with $n$ players, valuation bound $V$, and correct conservative beliefs $\mathcal{B}$ such that either*

- *there is no profile of strictly/weakly/very weakly dominant strategies, or*
- *there is a profile $s$ of strictly/weakly/very weakly dominant strategies such that $REV(M(s)) < \varepsilon \cdot 2^{nd}(\mathcal{B})$.*

**Corollary 2.** *For all $\varepsilon \in (0, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon} \rceil$, and deterministic IIR mechanisms $M$, there exists an auction context $C$ with $n$ players, valuation bound $V$, and correct conservative beliefs $\mathcal{B}$ such that either*

- *there is no profile of strictly/weakly/very weakly dominant strategies, or*
- *there is a profile $s$ of strictly/weakly/very weakly dominant strategies such that $REV(M(s)) < \varepsilon \cdot 2^{nd}(\mathcal{B})$.*

(Of course, in the above corollaries $s$ is unique if composed of strictly or weakly dominant strategies.)

Note that, in the absence of Theorems 1 and 2, these two corollaries would be trivial if the players were restricted to bid valuations only. In such a case, in fact, the second-price mechanism is "the only" (weakly) dominant-strategy mechanism for auctions of a single good. And since the revenue it generates is precisely equal to the second-highest valuation, no other dominant-strategy mechanism can generate second-belief revenue. QED.

We thus wish to emphasize that:

> *All our impossibility results hold without any restrictions on the strategy spaces (in particular when the players are allowed to report their conservative beliefs).*

Implementation in dominant strategies and implementation in undominated strategies ultimately fail to achieve our benchmark because they do not require that the players believe that their opponents are rational. The absence of the latter requirement is a strength, but only if the desired social choice correspondence is implementable. Else it is a "weakness". As we shall see, the solution concept underlying our mechanism relies on mutual belief of rationality (but not on higher-order beliefs of rationality).

## 5.1 Proof of Theorem 1

In the analysis below we solely focus on the case $n = 2$ (the analysis for arbitrary $n > 2$ is very similar and thus omitted). For sake of contradiction, assume that there exist a value $\varepsilon \in (1/2, 1]$, an integer $V > \lceil \frac{1}{\varepsilon - 1/2} \rceil$, and a probabilistic IIR mechanism $M$ such that for all contexts $C$ with 2 players, valuation bound $V$, and a correct conservative-belief profile $\mathcal{B}$, and for all profiles $s$ of undominated strategies in the auction $(C, M)$, we have

$$REV(M(s)) \geq \varepsilon \cdot 2^{nd}(\mathcal{B})$$

(that is, $M$ implements $\varepsilon 2^{nd}$ in undominated strategies for contexts with 2 players, valuation bound $V$, and correct conservative beliefs). To derive the desired contradiction, letting $H$ be an integer such that

$$V \geq H > \frac{1}{\varepsilon - 1/2},$$

we construct two games, $G$ and $G'$, as follows.

1. $G = (C, M)$, where $C = (2, V, \theta, \mathcal{B})$ with $\theta = (H, 0)$ and $\mathcal{B}_1 = \mathcal{B}_2 = \{(H, 0)\}$.
   **Note:** Each belief $\mathcal{B}_i$ is correct, and $2^{nd}(\mathcal{B}) = H$ because $smv_1 = smv_2 = H$.
2. $G' = (C', M)$, where $C' = (2, V, \theta', \mathcal{B}')$ with $\theta' = (1, 0)$ and $\mathcal{B}'_1 = \mathcal{B}'_2 = \{(1, 0)\}$.
   **Note:** Each belief $\mathcal{B}'_i$ is correct and $2^{nd}(\mathcal{B}') = 1$.

After analyzing the (auxiliary) game $G'$, we derive our desired contradiction for $G$. To clarify the game to which a given quantity refers, we shall use the superscripts $G$ and $G'$.

Let $UD^{G'} = UD_1^{G'} \times UD_2^{G'}$, where each $UD_i^{G'}$ is player $i$'s set of undominated strategies in $G'$. Then, by hypothesis:

$$\forall s' \in UD^{G'}, REV(M(s')) \geq \varepsilon 2^{nd}(\mathcal{B}') = \varepsilon. \tag{1}$$

We now prove the following statement:

$$\text{There exists a strategy } \sigma_1' \in \Delta(UD_1^{G'}) \text{ such that } \forall \text{ strategy } s_2 \text{ of player 2, } u_1^{G'}(M(\sigma_1', s_2)) \geq 0. \tag{2}$$

Because $M$ is IIR, player 1 has a strategy $out_1$ such that $u_1^{G'}(M(out_1, s_2)) = 0 \ \forall s_2$. If $out_1 \in UD_1^{G'}$ then Statement 2 follows by taking $\sigma_1' = out_1$. Otherwise, by the finiteness of $M$ there exists $\sigma_1' \in \Delta(UD_1^{G'})$ such that $out_1$ is weakly dominated by $\sigma_1'$, which implies $u_1^{G'}(M(\sigma_1', s_2)) \geq u_1^{G'}(M(out_1, s_2)) = 0 \ \forall s_2$, as desired.

Similarly, we have the following statement:

$$\text{There exists a strategy } \sigma_2' \in \Delta(UD_2^{G'}) \text{ such that } \forall \text{ strategy } s_1 \text{ of player 1, } u_2^{G'}(M(s_1, \sigma_2')) \geq 0. \tag{3}$$

Combining Statements 2 and 3, letting $\omega'$ be the (possibly probabilistic) outcome $M(\sigma_1', \sigma_2')$, and letting $p_i'$ and $EP_i'$ respectively be the probability that player $i$ gets the good and the expected price that $i$ pays according to $\omega'$, we have that

$$u_1^{G'}(\omega') = p_1' - EP_1' \geq 0 \quad \text{and} \quad u_2^{G'}(\omega') = -EP_2' \geq 0. \tag{4}$$

Because of Equation 1, and because $\sigma_i' \in \Delta(UD_i^{G'})$ for each $i$, we have

$$REV(\omega') = EP_1' + EP_2' \geq \varepsilon. \tag{5}$$

Combining Equations 4 and 5, we have

$$p_1' \geq EP_1' \geq \varepsilon - EP_2' \geq \varepsilon. \tag{6}$$

Let us now analyze game $G$. Notice that, under the strategy profile $(\sigma_1', \sigma_2')$, the (possibly probabilistic) outcome of $M$ is still $\omega'$ in game $G$. Accordingly, following Equation 6 we have that

$$u_1^G(M(\sigma_1', \sigma_2')) = u_1^G(\omega') = p_1'H - EP_1' \geq p_1'H - p_1' \geq \varepsilon(H-1),$$

where the second inequality holds further because $H > 1$.

Let $UD^G = UD_1^G \times UD_2^G$, where each $UD_i^G$ is player $i$'s set of undominated strategies in $G$. We now argue that there exists a strategy $\hat{\sigma}_1 \in \Delta(UD_1^G)$ such that

$$u_1^G(M(\hat{\sigma}_1, \sigma_2')) \geq \varepsilon(H-1). \tag{7}$$

To see why Inequality 7 is true, notice that if $\sigma_1' \in \Delta(UD_1^G)$ then we can take $\hat{\sigma}_1 = \sigma_1'$. Otherwise, for each strategy $s_1'$ which is in the support of $\sigma_1'$ but not in $UD_1^G$, there exists $\sigma_1'' \in \Delta(UD_1^G)$ weakly dominating $s_1'$ in game $G$ (again because $M$ is finite). Thus, we can construct $\hat{\sigma}_1$ from $\sigma_1'$ by replacing each such $s_1'$ with the corresponding $\sigma_1''$, and the so-constructed $\hat{\sigma}_1$ satisfies $u_1^G(M(\hat{\sigma}_1, \sigma_2')) \geq u_1^G(M(\sigma_1', \sigma_2')) \geq \varepsilon(H-1)$, as desired.

Because $\theta_2 = \theta_2'$, we have that player 2's set of undominated strategies is the same in $G$ and $G'$, and so is his utility for each possible outcome. That is,

$$UD_2^G = UD_2^{G'} \quad \text{and} \quad u_2^G(\cdot) = u_2^{G'}(\cdot). \tag{8}$$

Equations 3 and 8 directly imply the following statement:

$$\sigma_2' \in \Delta(UD_2^G) \quad \text{and} \quad u_2^G(M(\hat{\sigma}_1, \sigma_2')) \geq 0. \tag{9}$$

Let $\omega = M(\hat{\sigma}_1, \sigma'_2)$, and let $p_i$ and $EP_i$ respectively be the probability that player $i$ gets the good and the expected price that $i$ pays according to $\omega$. Following Equation 7 and the inequality of Statement 9, we have

$$u_1^G(\omega) = p_1 H - EP_1 \geq \varepsilon(H-1) \quad \text{and} \quad u_2^G(\omega) = -EP_2 \geq 0. \tag{10}$$

Combining Equation 10 with the facts that $0 \leq p_1 \leq 1$, $1/2 < \varepsilon \leq 1$, and $H > \frac{1}{\varepsilon - 1/2}$, we have

$$\begin{aligned} REV(\omega) &= EP_1 + EP_2 \leq EP_1 \leq p_1 H - \varepsilon(H-1) = H(p_1 - \varepsilon + \frac{\varepsilon}{H}) \leq H(1 - \varepsilon + \frac{1}{H}) \\ &< H(1 - \varepsilon + \varepsilon - 1/2) = H/2 < \varepsilon H. \end{aligned}$$

Accordingly, there exists a strategy profile $\hat{s}$ such that: (1) $\hat{s}_1$ is in the support of $\hat{\sigma}_1$ and $\hat{s}_2$ is in the support of $\sigma'_2$, which imply that $\hat{s} \in UD^G$; and (2) $REV(M(\hat{s})) \leq REV(\omega) < \varepsilon H = \varepsilon 2^{nd}(\mathcal{B})$. That is, we have finally reached the desired contradiction against the hypothesis that $M$ implements $\varepsilon 2^{nd}$ in undominated strategies for contexts with correct conservative beliefs. Thus Theorem 1 holds. ■

# 6 The Fragility of Implementing $2^{nd}(\cdot)$ at Ex-Post Equilibrium

In this section we analyze the possibility of achieving our revenue benchmark under two other classical notions of implementation: at ex-post and at very weakly dominant equilibrium. (The notions of ex-post and very weakly dominant strategy are almost the same, but do not coincide for some games.) Both implementation notions require only the existence of *one* equilibrium at which our benchmark is achieved. Thus, satisfying this requirement does not contradict the previously stated Corollaries 1 and 2, which only rule out the possibility of guaranteeing second-belief revenue at *all* very weakly dominant equilibria.

Our analysis however shows that both notions are inadequate for implementing our benchmark. The nature of this inadequacy is a bit different than that for implementation in undominated strategies. Indeed, truthfully reporting the conservative beliefs may easily be an ex-post (or very weakly dominant) equilibrium generating the desired revenue. However, an *extremely severe* equilibrium-selection problem arises. Consider the following auction mechanism for 2 players whose possible valuations range between 0 and 100.

**Mechanism** NAIVE. *For each player $i$, the strategy set $S_i$ consists of the set of all possible conservative beliefs of $i$. That is, $S_i = \{X \subseteq \{0, 1, \ldots, 100\}^2 \ : \ t, t' \in X \Rightarrow t_i = t'_i\}$.*

*For a strategy profile $(\mathcal{B}'_1, \mathcal{B}'_2)$, allegedly the true profile $\mathcal{B}$, compute an outcome $(a, P)$ as follows.*

- *For each $i$, let $\theta'_i$ be the $i$th component of a profile in $\mathcal{B}'_i$* (i.e., let $\theta'_i$ be the alleged true type of $i$).
- *Set $w = \text{argmax}_i\, \theta'_i$, breaking ties lexicographically, and $p = \min_{t \in \mathcal{B}'_{-w}} \max_j t_j$.*
- *If $\theta'_w \geq p$, then $a = w$* (i.e., the good is sold to player $w$)*, $P_w = p$ and $P_{-w} = 0$.*
  *Else, $a = 0$ (i.e., the good is unallocated) and $P = (0, 0)$*

It is immediately clear that, in mechanism NAIVE, truthfully announcing one's own conservative beliefs is always an ex-post equilibrium. It is also clear that, when all beliefs are correct, the truthful equilibrium guarantees our revenue benchmark. However, consider the following

**Context $C$:** $\theta = (70, 100)$, $\mathcal{B}_1 = \{(70, x) : x \geq 90\}$, and $\mathcal{B}_2 = \{(x, 100) : x \geq 60\}$.

In this context,

- all beliefs are correct; and
- at the truthful equilibrium $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2)$, $REV(\mathcal{B}) = 90 = 2^{nd}(\mathcal{B})$, as desired.

Let us now illustrate why this equilibrium is far from satisfactory. To begin with, note that

$$\mathcal{B}' \triangleq (\mathcal{B}'_1, \mathcal{B}_2) \triangleq (\{(70, x) : x \geq 0\}, \mathcal{B}_2)$$

is an alternative Nash equilibrium (corresponding to another ex-post equilibrium) whose revenue is only 70.[5]

[5] In fact, any strategy profile $(\mathcal{B}''_1, \mathcal{B}_2)$ with $\mathcal{B}''_1 = \{(70, x) : x \geq b\}$ and $b \leq 70$ is a Nash equilibrium (and corresponding to an ex-post equilibrium) whose revenue is only 70.

The existence of multiple equilibria is always problematic, but NOT necessarily crucial. Indeed one is often able to argue that the players have good reasons to coordinate around the desired equilibrium. But in our case $\mathcal{B}$ and $\mathcal{B}'$ differ only at player 1's strategy. Thus, even if player 1 believes that player 2 will play his truthful strategy, it is *identically* rational for player 1 to play $\mathcal{B}'_1$ instead of $\mathcal{B}_1$. Vice versa, even if player 2 believes that player 1 will play $\mathcal{B}'_1$, it is still rational for player 2 to stick to his own strategy in the truthful equilibrium, because in the above example it coincides with his strategy in the alternative equilibrium.

In sum, whatever reasons player 1 has to play his truthful strategy $\mathcal{B}_1$, he has exactly the same reasons to play his alternative strategy $\mathcal{B}'_1$. And "even more so" for player 2! This being the case,

*Which revenue should we expect from* NAIVE *for the above context $C$?*

The answer is 90 if player 1 feels "generous" towards the seller and 70 otherwise.[6] Dependency on player generosity is of course hardly satisfactory in mechanism design.

Let us now prove that the above extreme "fragility" of the truthful equilibrium in NAIVE applies to
(1) *every* mechanism $M$ that ex-post implements (even in an approximate way) our benchmark, and
(2) *every* ex-post equilibrium at which $M$ implements our benchmark.

### 6.1 Formalization of Fragility and Statement of Our Results

The notions of ex-post equilibrium and implementation at ex-post equilibrium, originally defined for Bayesian settings, readily apply to our setting. Namely:

**Definition 3.** *Let $\mathscr{C}$ be a class of contexts, $M$ a mechanism, and $F$ an extended social choice correspondence.*

- *An **ex-post equilibrium** of a mechanism $M$ for $\mathscr{C}$ is a profile $\boldsymbol{s}$ of functions, where each $\boldsymbol{s}_i$ maps player $i$'s conservative beliefs to his (possibly mixed) strategies in $M$, such that*

$$\boldsymbol{s}(\mathcal{B}) \triangleq (\boldsymbol{s}_1(\mathcal{B}_1), \ldots, \boldsymbol{s}_n(\mathcal{B}_n))$$

  *is a Nash equilibrium of the game $(C, M)$ for all contexts $C \in \mathscr{C}$ with conservative-belief profile $\mathcal{B}$. If the range of each $\boldsymbol{s}_i$ only consists of pure strategies, then $\boldsymbol{s}$ is a pure ex-post equilibrium.*

- *$M$ **implements** $F$ **at ex-post equilibrium** for $\mathscr{C}$ if there exists an ex-post equilibrium $\boldsymbol{s}$ for $\mathscr{C}$ such that for all contexts $C \in \mathscr{C}$ with conservative-belief profile $\mathcal{B}$*

$$M(\boldsymbol{s}(\mathcal{B})) \in F(\mathcal{B}).$$

  *If this is the case, we further say that $M$ implements $F$ at $\boldsymbol{s}$.*

(Note that, if for each player $i$ and each strategy $s_i$ there exists $\mathcal{B}_i$ such that $\mathbf{s}_i(\mathcal{B}_i) = s_i$, then for each $\mathcal{B}$ the strategy profile $\mathbf{s}(\mathcal{B})$ is also a very weakly dominant equilibrium. But otherwise it is not. As already said, ex-post equilibrium and very weakly dominant equilibrium are different notions.)

Let us now formalize the intuitively discussed notion of fragility for implementation at ex-post equilibrium. The corresponding formalization for implementation at very weakly dominant equilibrium is similarly defined.

**Definition 4.** *Let $M$ be a mechanism implementing an extended social choice correspondence $F$ at ex-post equilibrium for a class of contexts $\mathscr{C}$. Then $M$ is **fragile** if, for all ex-post equilibria $\boldsymbol{s}$ at which $M$ implements $F$, there is another ex-post equilibrium $\boldsymbol{s}'$ satisfying the following two properties:*
*(1) There exist a player $i$ and a conservative belief $\mathcal{B}_i$ of $i$ such that $\boldsymbol{s}'$ and $\boldsymbol{s}$ differ only at $\mathcal{B}_i$;[7] and*
*(2) $M(\boldsymbol{s}'(\mathcal{B}')) \notin F(\mathcal{B}')$ for all contexts $C \in \mathscr{C}$ with conservative-belief profile $\mathcal{B}'$ such that $\mathcal{B}'_i = \mathcal{B}_i$.*

[6] Notice that the truthful equilibrium actually specifies a very weakly dominant strategy for each player in each context, and thus illustrates the lack of robustness for implementation at very weakly dominant equilibria as well. Such lack of robustness was already pointed out by Saijo, Sjostrom, and Yamato theoretically [23] and by Casona, Saijo, Sjostrom, and Yamato experimentally [5]. In [23] the authors also propose *secure implementation*: essentially, implementation via mechanisms ensuring that (a) each player has a very weakly dominant strategy, and that (b) the desired property holds at all Nash equilibria (and thus all very weakly dominant ones). As we have discussed, therefore, the second-belief revenue benchmark is not securely implementable.

[7] That is, $\mathbf{s}_i(\mathcal{B}_i) \neq \mathbf{s}'_i(\mathcal{B}_i)$; $\mathbf{s}_i(\mathcal{B}'_i) = \mathbf{s}'_i(\mathcal{B}'_i)$ for all $\mathcal{B}'_i \neq \mathcal{B}_i$; and $\mathbf{s}_j = \mathbf{s}'_j$ for all $j \neq i$.

**Theorem 3.** *For all $\varepsilon \in (\frac{1}{2}, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon - 1/2} \rceil$, and probabilistic interim individually rational mechanisms $M$ implementing $\varepsilon \cdot 2^{nd}(\cdot)$ at ex-post equilibrium for contexts with $n$ players, valuation bound $V$, and correct beliefs,*

$$M \text{ is fragile.}$$

The notion of implementation at pure ex-post equilibrium and the corresponding notion of fragility are similarly defined, and we have the following theorem.

**Theorem 4.** *For all $\varepsilon \in (0, 1]$, $n > 1$, $V > \lceil \frac{1}{\varepsilon} \rceil$, and deterministic interim individually rational mechanisms $M$ implementing $\varepsilon \cdot 2^{nd}(\cdot)$ at pure ex-post equilibrium for contexts with $n$ players, valuation bound $V$, and correct beliefs,*

$$M \text{ is fragile.}$$

These two theorems also hold for implementation at very weakly dominant equilibrium. Since their proof is essentially the same, below we just prove Theorem 3.

### 6.2 Proof of Theorem 3

Similar to the proof of Theorem 1, we focus on the case $n = 2$, as the case $n > 2$ is very similar. Let $\varepsilon$ be a value in $(1/2, 1]$, $V$ an integer greater than $\lceil \frac{1}{\varepsilon - 1/2} \rceil$, and $M$ a probabilistic IIR mechanism implementing $\varepsilon 2^{nd}(\cdot)$ at ex-post equilibrium $\boldsymbol{s}$ for the class $\mathscr{C}$ of contexts with 2 players, valuation bound $V$, and correct beliefs. To prove that $M$ is fragile, let $H$ be an integer such that

$$V \geq H > \frac{1}{\varepsilon - 1/2}.$$

Again we are going to consider different contexts and thus different games, and we use superscripts to clarify the game to which a given quantity refers.

Let $\mathcal{B}_2^* = \{(H, 0)\}$. Notice that there exist some contexts in $\mathscr{C}$ with player 2's conservative belief being $\mathcal{B}_2^*$ —indeed, these are contexts where player 1's true valuation is $H$, player 2's true valuation is 0, and player 2 believes that player 1's true valuation is $H$ (that is, player 1's belief about player 2 is the only undetermined part). Our goal is to show that there exists another ex-post equilibrium $\boldsymbol{s}'$ or $\mathscr{C}$ such that:

(1) $\boldsymbol{s}'$ and $\boldsymbol{s}$ differ only at the conservative belief $\mathcal{B}_2^*$ of player 2; and

(2) for every context $C = (n, V, \theta, \mathcal{B}) \in \mathscr{C}$ with $\mathcal{B}_2 = \mathcal{B}_2^*$, $REV(M(\boldsymbol{s}'(\mathcal{B}))) < \varepsilon 2^{nd}(\mathcal{B})$.

To do so, we analyze two (classes of) related games, $G$ and $G'$, as follows.

| $G = (C, M)$, where $C = (2, V, \theta, \mathcal{B})$ is an arbitrary context in $\mathscr{C}$ with $\mathcal{B}_2 = \mathcal{B}_2^*$. |
|---|

In $G$ we have that $\theta = (H, 0)$, $\theta \in \mathcal{B}_1$, $smv_1 = smv_2 = H$, and thus $2^{nd}(\mathcal{B}) = H$ no matter what $\mathcal{B}_1$ is.

| $G' = (C', M)$, where $C' = (2, V, \theta', \mathcal{B}')$ with<br>$\theta' = (1, 0)$, $\mathcal{B}_1' = \{(1, x) : (H, x) \in \mathcal{B}_1\}$, and $\mathcal{B}_2' = \{(x, 0) : x \geq 1\}$. |
|---|

Notice that $C' \in \mathscr{C}$ and $2^{nd}(\mathcal{B}') = 1$.

Let us now analyze game $G'$. Let $\omega' = M(\boldsymbol{s}(\mathcal{B}'))$, and $p_i'$ and $EP_i'$ respectively be the probability that player $i$ gets the good and the expected price that player $i$ pays according to $\omega'$. Because $M$ is IIR, there exists strategy $out_i'$ for each player $i$ such that $u_i^{G'}(M(out_i', s_{-i})) = 0$ for every $s_{-i}$. Accordingly, and further because $\boldsymbol{s}$ is an ex-post equilibrium at which $M$ implements $\varepsilon 2^{nd}(\cdot)$, we have

$$u_1^{G'}(\omega') = p_1' - EP_1' \geq 0, \quad u_2^{G'}(\omega') = -EP_2' \geq 0, \quad \text{and} \quad EP_1' + EP_2' \geq \varepsilon 2^{nd}(\mathcal{B}') = \varepsilon.$$

Combining these three inequalities, we have

$$p_1' \geq EP_1' \geq \varepsilon - EP_2' \geq \varepsilon. \tag{11}$$

We construct the desired ex-post equilibrium $\boldsymbol{s}'$ as follows:

$$\boldsymbol{s}'_2(\mathcal{B}^*_2) = \boldsymbol{s}_2(\mathcal{B}'_2),$$

and $\boldsymbol{s}'$ coincides with $\boldsymbol{s}$ everywhere else. Note that $\boldsymbol{s}'$ satisfies property (1) of Definition 4 (with $i = 2$). We now prove that the so-constructed $\boldsymbol{s}'$ is indeed an ex-post equilibrium for $\mathscr{C}$, and that it satisfies property (2) of Definition 4.

By construction, for any context $C'' \in \mathscr{C}$ with conservative belief profile $\mathcal{B}''$ such that $\mathcal{B}''_2 \neq \mathcal{B}^*_2$, $\boldsymbol{s}'(\mathcal{B}'') = \boldsymbol{s}(\mathcal{B}'')$, and thus $\boldsymbol{s}'(\mathcal{B}'')$ is a Nash equilibrium of the game $(C'', M)$. Because $C$ is a generic context in $\mathscr{C}$ with player 2's conservative belief being $\mathcal{B}^*_2$, it remains to show that $\boldsymbol{s}'$ satisfies the following properties:
(A) $\boldsymbol{s}'(\mathcal{B})$ is a Nash equilibrium of $G$ (which implies that $\boldsymbol{s}'$ is an ex-post equilibrium for $\mathscr{C}$); and
(B) $REV(M(\boldsymbol{s}'(\mathcal{B}))) < \varepsilon H$.

**Proof of Property A.** We do so by introducing another (auxiliary) game $G''$.

$$\boxed{G'' = (C'', M), \text{ where } C'' = (2, V, \theta'', \mathcal{B}'') \text{ with } \theta'' = \theta,\ \mathcal{B}''_1 = \mathcal{B}_1, \text{ and } \mathcal{B}''_2 = \mathcal{B}'_2.}$$

Notice that $C'' \in \mathscr{C}$, and that $C''$ differs from $C$ only at player 2's belief and from $C'$ only at player 1's true valuation (of course $\mathcal{B}''_1$ has to be consistent with $\theta''_1$ which is $H$, and thus differs from $\mathcal{B}'_1$, but player 1's beliefs about player 2 do not change).

Because $\boldsymbol{s}'_1 = \boldsymbol{s}_1$, $\mathcal{B}_2 = \mathcal{B}^*_2$, $\mathcal{B}_1 = \mathcal{B}''_1$, $\boldsymbol{s}'_2(\mathcal{B}^*_2) = \boldsymbol{s}_2(\mathcal{B}'_2)$, and $\mathcal{B}'_2 = \mathcal{B}''_2$, we have that

$$\boldsymbol{s}'(\mathcal{B}) = (\boldsymbol{s}'_1(\mathcal{B}_1), \boldsymbol{s}'_2(\mathcal{B}_2)) = (\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}'_2(\mathcal{B}^*_2)) = (\boldsymbol{s}_1(\mathcal{B}''_1), \boldsymbol{s}_2(\mathcal{B}'_2)) = (\boldsymbol{s}_1(\mathcal{B}''_1), \boldsymbol{s}_2(\mathcal{B}''_2)) = \boldsymbol{s}(\mathcal{B}'').$$

Because $\boldsymbol{s}(\mathcal{B}'')$ is a Nash equilibrium of $G''$ by the definition of $\boldsymbol{s}$, $\boldsymbol{s}'(\mathcal{B})$ is also a Nash equilibrium of $G''$. Because $G$ and $G''$ have the same true valuation profile, $\boldsymbol{s}'(\mathcal{B})$ is a Nash equilibrium of $G$, and Property A holds. Therefore $\boldsymbol{s}'$ is an ex-post equilibrium for $\mathscr{C}$.

**Proof of Property B.** Notice that in game $G$, the outcome of strategy profile $\boldsymbol{s}(\mathcal{B}')$ is still $\omega'$. Thus

$$u^G_1(M(\boldsymbol{s}(\mathcal{B}'))) = u^G_1(\omega') = p'_1 H - EP'_1 \geq p'_1 H - p'_1 \geq \varepsilon(H-1),$$

where the inequalities hold by Equation 11.

Because $\boldsymbol{s}'(\mathcal{B}) = (\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}'_2(\mathcal{B}^*_2)) = (\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}_2(\mathcal{B}'_2))$ is a Nash equilibrium of $G$, we have that

$$u^G_1(M(\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}_2(\mathcal{B}'_2))) \geq u^G_1(M(\boldsymbol{s}(\mathcal{B}'))) \geq \varepsilon(H-1),$$

and

$$u^G_2(M(\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}_2(\mathcal{B}'_2))) \geq u^G_2(M(\boldsymbol{s}_1(\mathcal{B}_1), out_2)) = 0,$$

where $out_2$ is the strategy of player 2 such that $u^G_2(s_1, out_2) = 0$ for every $s_1$, and the existence of such an $out_2$ is guaranteed by $M$ being IIR.
Let $\omega'' = M(\boldsymbol{s}_1(\mathcal{B}_1), \boldsymbol{s}_2(\mathcal{B}'_2))$, and $p''_i$ and $EP''_i$ respectively be the probability that player $i$ gets the good and the expected price that player $i$ pays according to $\omega''$. Combining with the above two lines of equations, we have

$$u^G_1(\omega'') = p''_1 H - EP''_1 \geq \varepsilon(H-1) \quad \text{and} \quad u^G_2(\omega'') = -EP''_2 \geq 0.$$

Combining with the facts that $0 \leq p''_1 \leq 1$, $1/2 < \varepsilon \leq 1$, and $H > \frac{1}{\varepsilon - 1/2}$, we have

$$\begin{aligned} REV(M(\boldsymbol{s}'(\mathcal{B}))) &= REV(\omega'') = EP''_1 + EP''_2 \leq EP''_1 \leq p''_1 H - \varepsilon(H-1) = H(p''_1 - \varepsilon + \frac{\varepsilon}{H}) \\ &\leq H(1 - \varepsilon + \frac{1}{H}) < H(1 - \varepsilon + \varepsilon - 1/2) = H/2 < \varepsilon H. \end{aligned}$$

Therefore Property B holds, and so does Theorem 3. ■

# 7 The Implementability of $2^{nd}(\cdot)$ in Conservative Strategies

In this section we prove that the second-belief benchmark is virtually implementable under a natural solution concept. We do so in four steps. First we present the underlying solution concept, then we exhibit our mechanism, analyze it, and address three concerns raised about it.

**Notation** Given the set of all strategy profiles $S = S_1 \times \cdots \times S_n$ and the true type profile $\theta$, we denote by $U = U_1 \times \cdots \times U_n$ the set of profiles of strategies that are not *strictly* dominated.

If $T = T_i \times T_{-i}$ is a set of strategy profiles, $t_i \in \Theta_i$, $s_i \in T_i$, and $\sigma_i \in \Delta(T_i)$, then we say that $s_i$ is *strictly dominated by $\sigma_i$ with respect to $t_i$ and $T$*, in symbols $s_i <_T^{t_i} \sigma_i$, if $u_i(t_i, (s_i, s_{-i})) < u_i(t_i, (\sigma_i, s_{-i}))$ for all $s_{-i} \in T_{-i}$. (That is, $s_i$ is strictly dominated by $\sigma_i$ when the set of all strategy profiles is assumed to be $T$ and the true type of $i$ to be $t_i$.)

The set of strategies in $T_i$ that are not strictly dominated with respect to $t_i$ and $T$ is denoted by $U_i(t_i, T_{-i})$.

## 7.1 Conservative Implementation

In an auction $(C, M)$, we assume that every player is rational (i.e., never plays strictly dominated strategies) and believes that his opponents are rational.

Accordingly, $i$ confines his strategy choices to $U_i$. But to which set should he believe his opponents to confine their strategy choices? Although $i$ believes that all players in $-i$ are rational, he cannot compute $U_{-i}$, because he does not know $\theta_{-i}$. However, given his conservative belief $\mathcal{B}_i$, $i$ is sure that $\theta_{-i} \in \bigcup_{t \in \mathcal{B}_i} \{t_{-i}\}$, and that his opponents, for any $t \in \mathcal{B}_i$, only play strategy subprofiles in the set $U_{-i}(t) \triangleq \prod_{j \neq i} U_j(t_j, S_{-j})$.

Therefore, player $i$ can conservatively refine his set of undominated strategies by eliminating every strategy $s_i$ that is strictly dominated, by the *same* strategy $\sigma_i$, in *every* world he considers possible. That is, he "conservatively" eliminates $s_i \in U_i$ if and only if there exists a (possibly mixed) strategy $\sigma_i \in \Delta(U_i)$ such that, for every $t \in \mathcal{B}_i$, $s_i <_{U_{-i}(t)}^{\theta_i} \sigma_i$. In this case, we say that $s_i$ is *conservatively dominated* by $\sigma_i$.

We refer to the strategies surviving the "cautious" elimination procedure above as *conservative strategies.* In sum, under mutual belief of rationality, only profiles of conservative strategies will be played.

**Definition 5.** *In a game $(C, M)$ with conservative belief profile $\mathcal{B}$, the set of* ***conservative strategy*** *profiles is $\mathcal{C} \triangleq \mathcal{C}_1 \times \cdots \times \mathcal{C}_n$, where*

$$\mathcal{C}_i \triangleq U_i \setminus \{s_i : \exists \sigma_i \in \Delta(U_i)\ \forall t \in \mathcal{B}_i\ s_i <_{U_{-i}(t)}^{\theta_i} \sigma_i\}.$$

*A mechanism $M$* ***conservatively implements*** *a social choice correspondence $F$ for a class of contexts $\mathscr{C}$ if, for all contexts $C \in \mathscr{C}$ with belief profile $\mathcal{B}$, and all strategy profiles $s \in \mathcal{C}$,*

$$M(s) \in F(\mathcal{B}).$$

In game theory it is well known that for any player $i$, a strategy $s_i$ is not strictly dominated if and only if it is a best response to some strategy subprofile of others (assuming that the other players may play correlated strategy profiles). Therefore the above definition of conservative strategies can be rephrased as follows:

> Let $BR_i(t_i, S_{-i})$ be the set of best responses of each player $i$, with type $t_i$, to the others' strategy subprofiles in $\Delta(S_{-i})$. Player $i$ then uses his set of possible payoff types for the others, $\mathcal{B}_i$, to construct a restricted set of possible strategy subprofiles for the others,
>
> $$S_{-i}^1(\mathcal{B}_i) = \bigcup_{t \in \mathcal{B}_i} \prod_{j \neq i} BR_j(t_j, S_{-j}),$$
>
> which are the undominated strategies of the payoff types he considers possible. Player $i$ then plays only strategies in $BR_i(\theta_i, S_{-i}^1(\mathcal{B}_i))$. The set of conservative strategy profiles is $\mathcal{C} = \prod_{i \in [n]} BR_i(\theta_i, S_{-i}^1(\mathcal{B}_i))$, i.e., loosely speaking the set of twice subjective-best-responses. □

## 7.2 The Second-Belief Mechanism

For any $\varepsilon \in (0,1]$, $n$, and $V$, the mechanism $\mathcal{M}_{\varepsilon,n,V}$ chooses an outcome $(a,P)$ according to the following steps. Note that the mechanism applies to any context with $n$ players, valuation bound $V$ and correct beliefs, and that the players act simultaneously and only once, in Step **1**. Steps **a** through **d** are just "conceptual steps taken by the mechanism".

**Mechanism $\mathcal{M}_{\varepsilon,n,V}$**

**1**: *Each player $i$, publicly and simultaneously with the others, announces a pair $(e_i, v_i) \in \{0,1\}\times\{0,1,\ldots,V\}$.*

Comment. Allegedly, $v_i = smv_i$, and $e_i$ indicates whether $i$'s announcement is about his internal knowledge (allegedly $e_i = 0$ signifies that $v_i = \theta_i$), or about his external belief.

**a**: *If $v_i = 0$ for every $i$, then set $a$ to be a randomly chosen player, set $P_i = 0$ for each player $i$, and halt.*

**b**: *Order the announced $n$ pairs according to $v_1,\ldots,v_n$ decreasingly, breaking ties in favor of those with $e_i = 0$. If there are still ties among some pairs, then break them according to the corresponding players.*

Comment. It does not matter whether the players are ordered lexicographically (increasingly or decreasingly), or according to some other way.

**c**: *Set $a$ to be the player corresponding to the first pair, and $P_a = \max\{\frac{1}{2}, \max_{j\neq a} v_j\}$.*

**d**: *For each player $i$, $P_i = P_i - \delta_i$, where $\delta_i = \frac{\varepsilon}{4n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right]$.*

Comment. Each player $i$ receives a (positive) reward $\delta_i$.

**Remark**

- Notice that $\mathcal{M}_{\varepsilon,n,V}$ always sells the good.
- *Non-negative Revenue.* Notice that if $\mathcal{M}_{\varepsilon,n,V}$ halts in Step **a** then its revenue is 0. Otherwise, its revenue equals the price charged to player $a$ in Step **c** minus the total rewards given to the players in Step **d**. Because for each player $i$ the reward that $i$ receives in Step **d** is $\delta_i < \frac{\varepsilon}{4n}(1+1) = \frac{\varepsilon}{2n} \leq \frac{1}{2n}$, the total rewards given to the players in Step **d** is at most $\frac{1}{2}$. Because the price charged to player $a$ in Step **c** is at least $\frac{1}{2}$, we have that $\mathcal{M}_{\varepsilon,n,V}$ always has non-negative revenue.
  In fact, Step **a** is needed solely to ensure that the revenue of the mechanism is non-negative. If the seller can withstand a $-\varepsilon$ revenue when all but one $v_i$'s are 0, then we can remove Step **a** and make the mechanism deterministic.
- *Uniform Construction.* As promised, it is clear that $\mathcal{M}_{\varepsilon,n,V}$ is uniformly and efficiently constructible on inputs $\varepsilon$, $n$, and $V$. In addition, it is very simple. It essentially consists of the second-price mechanism together with carefully designed rewards. In light of our impossibility results about implementing $\varepsilon 2^{nd}(\cdot)$ under classical solution concepts, this simplicity suggests that conservative implementation can be quite powerful.
- *From Additive to Multiplicative $\varepsilon$.* Notice that the reward each player gets in Step **d** is at most $\frac{\varepsilon}{2n}$. Thus if a player does not get the good, then his utility is at most $\frac{\varepsilon}{2n}$. This is so because we aim at achieving the second-belief revenue benchmark up to only an additive $\varepsilon$. If we are willing to give up an $\varepsilon$ fraction of the revenue benchmark, then each player could receive a reward proportional to the second highest bid in the mechanism, so that his utility may still be very high even if he does not get the good. For instance, we can use $\delta_i = \frac{\varepsilon \max_{j\neq a} v_j}{4n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right]$.
- *Additional Revenue.* It is of course possible to generate additional revenue by punishing more harshly a player with wrong beliefs. (E.g., when the winner is a player $a$ announcing $(1,v_a)$, without anyone else announcing $(0,v)$ for some $v \geq v_a$, we may ask him to pay $2\cdot v_a$ instead of $\max_{j\neq a} v_j$.) But this does not achieve a benchmark higher than the second-belief one when all beliefs are correct.

### 7.3 Analysis of the Second-Belief Mechanism

Now we prove that conservative implementation succeeds where classical notions fail.

**Theorem 5.** *For any $\varepsilon \in (0,1]$, $n > 1$, and $V > 0$, $\mathcal{M}_{\varepsilon,n,V}$ conservatively implements $2^{nd}(\cdot) - \varepsilon$ for the class of contexts with $n$ players and valuation bound $V$.*

*Proof.* Arbitrarily fix $\varepsilon$, $n$, $V$, $C = (n, V, \theta, \mathcal{B})$, and a strategy profile $s$. Denoting $\mathcal{M}_{\varepsilon,n,V}$ by $\mathcal{M}$ for short, it suffices for us to prove that, if $s$ is a profile of conservative strategies in the game $(C, \mathcal{M})$, then

$$REV(\mathcal{M}(s)) \geq 2^{nd}(\mathcal{B}) - \varepsilon. \tag{12}$$

Letting $s_i \triangleq (e_i, v_i)$ for each $i$, we start by proving three claims.

Claim 1. *$\forall$ player $i$ and $\forall$ type $t_i \in \{0, 1, \ldots, V\}$ of $i$, if $s_i \in U_i(t_i, S_{-i})$ then $v_i \geq t_i$.*

Proof of Claim 1. Assume for sake of contradiction that $s_i \in U_i(t_i, S_{-i})$ and $v_i < t_i$. We shall show that $s_i$ is strictly dominated by $s'_i = (0, t_i)$ with respect to $t_i$ and $S$. By definition, this implies $s_i \notin U_i(t_i, S_{-i})$, a contradiction. For this purpose, letting $s'_{-i}$ be an arbitrary strategy subprofile of $-i$, it suffices to show that

$$u_i(t_i, (s_i, s'_{-i})) < u_i(t_i, (s'_i, s'_{-i})).$$

To do so, let $s'_j = (e'_j, v'_j)$ for each $j \neq i$. Moreover, in the plays of $(s_i, s'_{-i})$ and $(s'_i, s'_{-i})$ respectively, let $(a, P)$ and $(a', P')$ be the outcomes, and $\delta_i$ and $\delta'_i$ the rewards that player $i$ receives in Step **d**.

Because $v_i \geq 0$ by the construction of $\mathcal{M}$ and $v_i < t_i$ by hypothesis, we have that $t_i \geq 1$ and $\mathcal{M}$ does not halt in Step **a** in the play of $(s'_i, s'_{-i})$. Below we shall distinguish two exhaustive cases, according to the play of $(s_i, s'_{-i})$.

*Case 1.* $\mathcal{M}$ halts in Step **a** in the play of $(s_i, s'_{-i})$.

In this case, by the construction of $\mathcal{M}$ we have $v_i = 0$, $v'_j = 0$ for each $j \neq i$, and

$$u_i(t_i, (s_i, s'_{-i})) = \frac{t_i}{n}.$$

Now we consider the play of $(s'_i, s'_{-i})$. Because $t_i \geq 1 > 0 = \max_{j \neq i} v'_j$, we have $a' = i$, $P'_i = \max\{\frac{1}{2}, \max_{j \neq i} v'_j\} - \delta'_i = \frac{1}{2} - \delta'_i$, and $\delta'_i = \frac{\varepsilon}{4n}\left[\frac{t_i}{1+t_i} + \frac{1}{(1+V)^2}\right] > 0$. Accordingly,

$$u_i(t_i, (s'_i, s'_{-i})) = t_i - P'_i = t_i - \frac{1}{2} + \delta'_i > t_i - \frac{1}{2} \geq \frac{t_i}{n},$$

where the second inequality holds because $t_i \geq 1$ and $n \geq 2$. Therefore $u_i(t_i, (s_i, s'_{-i})) < u_i(t_i, (s'_i, s'_{-i}))$ as desired.

*Case 2.* $\mathcal{M}$ does not halt in Step **a** in the play of $(s_i, s'_{-i})$.

In this case, by the construction of $\mathcal{M}$ we have

$$\delta_i = \frac{\varepsilon}{4n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right] \quad \text{and} \quad \delta'_i = \frac{\varepsilon}{4n}\left[\frac{t_i}{1+t_i} + \frac{1}{(1+V)^2}\right].$$

Accordingly,

$$\delta'_i - \delta_i = \frac{\varepsilon}{4n}\left[\frac{t_i}{1+t_i} - \frac{v_i}{1+v_i}\right] + \frac{\varepsilon}{4n}\left[\frac{1-(1-e_i)}{(1+V)^2}\right] = \frac{\varepsilon}{4n}\left[\frac{t_i - v_i}{(1+t_i)(1+v_i)} + \frac{e_i}{(1+V)^2}\right] > 0,$$

where the inequality holds because $v_i < t_i$ by hypothesis and $e_i \geq 0$ by the construction of $\mathcal{M}$. Thus we have

$$\delta'_i > \delta_i.$$

Below we distinguish three exhaustive sub-cases.

*Sub-case 2.1.* $a' \neq i$.

In this sub-case, we also have $a \neq i$, because $v_i < t_i$. Accordingly, $P_i = -\delta_i$ and $P'_i = -\delta'_i$, and thus $u_i(t_i, (s_i, s'_{-i})) = \delta_i$ and $u_i(t_i, (s'_i, s'_{-i})) = \delta'_i$. Therefore $u_i(t_i, (s_i, s'_{-i})) < u_i(t_i, (s'_i, s'_{-i}))$ as desired.

*Sub-case 2.2.* $a' = i$ and $a = i$.

In this sub-case, we have $P'_i = \max\{\frac{1}{2}, \max_{j\neq i} v'_j\} - \delta'_i$ and $P_i = \max\{\frac{1}{2}, \max_{j\neq i} v'_j\} - \delta_i$. Because $\delta'_i > \delta_i$, we further have $P_i > P'_i$, which implies $u_i(t_i, (s_i, s'_{-i})) = t_i - P_i < t_i - P'_i = u_i(t_i, (s'_i, s'_{-i}))$ as desired.

*Sub-case 2.3.* $a' = i$ and $a \neq i$.

In this sub-case, we have $P'_i = \max\{\frac{1}{2}, \max_{j\neq i} v'_j\} - \delta'_i$, $P_i = -\delta_i$, and $t_i \geq \max_{j\neq i} v'_j$. As $t_i \geq 1$ by hypothesis, we further have $t_i \geq \max\{\frac{1}{2}, \max_{j\neq i} v'_j\}$. Accordingly, $u_i(t_i, (s_i, s'_{-i})) = -P_i = \delta_i < \delta'_i \leq (t_i - \max\{\frac{1}{2}, \max_{j\neq i} v'_j\}) + \delta'_i = t_i - P'_i = u_i(t_i, (s'_i, s'_{-i}))$ as desired.

In sum, $u_i(t_i, (s_i, s'_{-i})) < u_i(t_i, (s'_i, s'_{-i}))$ for any $s'_{-i}$, and thus $s_i$ is strictly dominated by $s'_i$ with respect to $t_i$ and $S$, contradicting the fact that $s_i \in U_i(t_i, S_{-i})$. Therefore Claim 1 holds. □

Claim 2. $\forall$ *player* $i$ *and* $\forall$ *type* $t_i \in \{1, 2, \ldots, V\}$ *of* $i$, *if* $s_i = (1, t_i)$ *then* $s_i \notin U_i(t_i, S_{-i})$.

Proof of Claim 2. By definition, it suffices for us to show that $s_i$ is strictly dominated by strategy $s'_i = (0, t_i)$ with respect to $t_i$ and $S$. For this purpose, arbitrarily fixing a strategy subprofile $s'_{-i}$ of $-i$, it suffices to show that

$$u_i(t_i, (s_i, s'_{-i})) < u_i(t_i, (s'_i, s'_{-i})).$$

To do so, first notice that $\mathcal{M}$ does not halt in Step **a** in either the play of $(s_i, s'_{-i})$ or the play of $(s'_i, s'_{-i})$, because $t_i \geq 1$ by hypothesis. The analysis below is very similar to Case 2 of Claim 1. Indeed, in the plays of $(s_i, s'_{-i})$ and $(s'_i, s'_{-i})$ respectively, we denote by $\delta_i$ and $\delta'_i$ the rewards that player $i$ receives in Step **d**, and by $(a, P)$ and $(a', P')$ the final outcomes. Letting $s'_j = (e'_j, v'_j)$ for each player $j \neq i$, we have

$$\delta'_i = \frac{\varepsilon}{4n}\left[\frac{t_i}{1+t_i} + \frac{1}{(1+V)^2}\right] > \frac{\varepsilon}{4n} \cdot \frac{t_i}{1+t_i} = \delta_i,$$

and we distinguish three cases as before:

- If $a' \neq i$, then $a \neq i$ as well, and we have
$$u_i(t_i, (s_i, s'_{-i})) = -P_i = \delta_i < \delta'_i = -P'_i = u_i(t_i, (s'_i, s'_{-i})).$$
- If $a' = i$ and $a = i$, then $P_i = \max\{\frac{1}{2}, \max_{j\neq i} v'_j\} - \delta_i > \max\{\frac{1}{2}, \max_{j\neq i} v'_j\} - \delta'_i = P'_i$, and we have
$$u_i(t_i, (s_i, s'_{-i})) = t_i - P_i < t_i - P'_i = u_i(t_i, (s'_i, s'_{-i})).$$
- Otherwise, we have that $a' = i$ and $a \neq i$, which implies
$$u_i(t_i, (s_i, s'_{-i})) = -P_i = \delta_i < \delta'_i \leq (t_i - \max\{\frac{1}{2}, \max_{j\neq i} v'_j\}) + \delta'_i = t_i - P'_i = u_i(t_i, (s'_i, s'_{-i})).$$

In sum, $s_i$ is strictly dominated by $s'_i$ with respect to $t_i$ and $S$, and Claim 2 holds. □

Claim 3. $\forall$ *player* $i$, *if* $s_i$ *is a conservative strategy in game* $(C, \mathcal{M})$, *then* $v_i \geq smv_i$.

Proof of Claim 3. Assume for sake of contradiction that $s_i$ is a conservative strategy and $v_i < smv_i$. By definition we have $s_i \in U_i$ and $U_i = U_i(\theta_i, S_{-i})$, and thus by Claim 1 we have

$$v_i \geq \theta_i. \tag{13}$$

Let $s'_i = (1, smv_i)$. In order to reach a contradiction it suffices for us to prove the following statement:

$$\forall t \in \mathcal{B}_i, \forall s'_{-i} \in U_{-i}(t), u_i(\theta_i, (s_i, s'_{-i})) < u_i(\theta_i, (s'_i, s'_{-i})). \tag{14}$$

To see why this is sufficient, notice that if $s'_i \in U_i$ then Statement 14 implies that $s_i$ is conservatively dominated by $s'_i$, contradicting the hypothesis that $s_i$ is a conservative strategy. If $s'_i \notin U_i$, then by definition it is strictly dominated with respect to $\theta_i$ and $S$. By well-known properties of strict domination and by the finiteness of $\mathcal{M}$, we have that there exists a strategy $\sigma'_i \in \Delta(U_i)$ such that $s'_i$ is strictly dominated by $\sigma'_i$, that is, the following statement holds:

$$\forall s'_{-i} \in S_{-i}, u_i(\theta_i, (s'_i, s'_{-i})) < u_i(\theta_i, (\sigma'_i, s'_{-i})). \tag{15}$$

Because $U_{-i}(t) \subseteq S_{-i}$ for each $t \in \mathcal{B}_i$, Statements 14 and 15 together imply that

$$\forall t \in \mathcal{B}_i, \forall s'_{-i} \in U_{-i}(t), u_i(\theta_i, (s_i, s'_{-i})) < u_i(\theta_i, (\sigma'_i, s'_{-i})).$$

In turn, this implies that $s_i$ is conservatively dominated by $\sigma'_i$, again contradicting the hypothesis that $s_i$ is a conservative strategy.

Below we shall prove Statement 14. Arbitrarily fixing a type profile $t \in \mathcal{B}_i$ and a strategy subprofile $s'_{-i} \in U_{-i}(t)$, it suffices to show that

$$u_i(\theta_i, (s_i, s'_{-i})) < u_i(\theta_i, (s'_i, s'_{-i})).$$

To do so, let $\star = \operatorname{argmax}_{j\in[n]} t_j$ with ties broken lexicographically. Because $t \in \mathcal{B}_i$ and $smv_i = \min_{t\in\mathcal{B}_i} \max_j t_j$, we have

$$t_\star \geq smv_i.$$

Because $smv_i > v_i \geq \theta_i = t_i$ by hypothesis and by Equation 13, we have $t_\star > t_i$, and thus

$$\star \neq i.$$

Let $s'_j = (e'_j, v'_j)$ for each $j \neq i$. Because $s'_\star \in U_\star(t_\star, S_{-\star}))$, by Claim 1 we have that

$$v'_\star \geq t_\star.$$

In sum, the following sequence of inequalities holds:

$$v'_\star \geq t_\star \geq smv_i > v_i \geq \theta_i \geq 0. \tag{16}$$

By Sequence 16 we have $v'_\star \geq 1$, thus $\mathcal{M}$ does not halt in Step **a** in the play of $(s_i, s'_{-i})$ or in the play of $(s'_i, s'_{-i})$. Below we consider the outcomes of the two plays.

Let $(a, P)$ and $(a', P')$ be the final outcomes of $(s_i, s'_{-i})$ and $(s'_i, s'_{-i})$ respectively. If $v'_\star > smv_i$, then by the construction of $\mathcal{M}$ we have that $(e'_\star, v'_\star)$ is ordered before $(1, smv_i)$, and thus is also ordered before $(e_i, v_i)$. If $v'_\star = smv_i$, then by Sequence 16 we have $v'_\star = t_\star \geq 1$. Thus by Claim 2 we have $e'_\star = 0$, which implies that $(e'_\star, v'_\star)$ is ordered before $(1, smv_i)$, and thus is also ordered before $(e_i, v_i)$. Accordingly, no matter what $v'_\star$ is, we always have

$$a \neq i \quad \text{and} \quad a' \neq i,$$

therefore the utilities of player $i$ only depend on his rewards in Step **d** in both plays.

Let $\delta_i$ and $\delta'_i$ be the rewards that player $i$ receives in Step **d**, in the plays of $(s_i, s'_{-i})$ and $(s'_i, s'_{-i})$ respectively. We have

$$\begin{aligned}
\delta'_i - \delta_i &= \frac{\varepsilon}{4n} \cdot \frac{smv_i}{1 + smv_i} - \frac{\varepsilon}{4n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right] \\
&= \frac{\varepsilon}{4n}\left[\frac{smv_i - v_i}{(1+smv_i)(1+v_i)} - \frac{1-e_i}{(1+V)^2}\right] \geq \frac{\varepsilon}{4n}\left[\frac{1}{(1+smv_i)(1+v_i)} - \frac{1}{(1+V)^2}\right] \\
&> \frac{\varepsilon}{4n}\left[\frac{1}{(1+smv_i)^2} - \frac{1}{(1+V)^2}\right] \geq \frac{\varepsilon}{4n}\left[\frac{1}{(1+V)^2} - \frac{1}{(1+V)^2}\right] = 0,
\end{aligned}$$

where the first inequality holds because $v_i < smv_i$ and $e_i \geq 0$, the second because $v_i < smv_i$, and the last because $smv_i \leq V$. Accordingly we have

$$\delta'_i > \delta_i,$$

which implies

$$u_i(\theta_i, (s_i, s'_{-i})) = \delta_i < \delta'_i = u_i(\theta_i, (s'_i, s'_{-i}))$$

as we wanted to show. Therefore Claim 3 holds. □

Now we are ready to prove that if $s$ is a profile of conservative strategies then Inequality 12 holds, which implies Theorem 5. Because $s$ is a profile of conservative strategies, by Claim 3 we have

$$v_i \geq smv_i \text{ for each } i. \tag{17}$$

If $\mathcal{M}$ halts in Step **a**, then $v_i = 0$ for each $i$, which together with Equation 17 implies that $smv_i = 0$ for each $i$, and thus $2^{nd}(\mathcal{B}) = 0$. Accordingly,

$$REV(\mathcal{M}(s)) = 0 = 2^{nd}(\mathcal{B}) > 2^{nd}(\mathcal{B}) - \varepsilon.$$

Otherwise, by Equation 17 we have that the second highest value in $\{v_1, \ldots, v_n\}$ is greater than or equal to the second highest value in $\{smv_1, \ldots, smv_n\}$, which is precisely $2^{nd}(\mathcal{B})$. By the construction of $\mathcal{M}$ we have that for each reward $\delta_i$ in Step **d**,

$$\delta_i = \frac{\varepsilon}{4n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right] < \frac{\varepsilon}{4n} \cdot (1+1) = \frac{\varepsilon}{2n}.$$

Letting $(a, P)$ be the outcome of $s$, we have: (1) $P_a = \max\{\frac{1}{2}, \max_{j \neq a} v_j\} - \delta_a$; (2) $\forall i \neq a$, $P_i = -\delta_i$; and (3) $\max_{j \neq a} v_j$ is the second highest value in $\{v_1, \ldots, v_n\}$, which implies $\max\{\frac{1}{2}, \max_{j \neq a} v_j\} \geq 2^{nd}(\mathcal{B})$. Accordingly,

$$REV(\mathcal{M}(s)) = P_a + \sum_{i \neq a} P_i \geq 2^{nd}(\mathcal{B}) - \delta_a - \sum_{i \neq a} \delta_i > 2^{nd}(\mathcal{B}) - n \cdot \frac{\varepsilon}{2n} > 2^{nd}(\mathcal{B}) - \varepsilon.$$

Therefore Theorem 5 holds. ■

**Remark.** If a player's belief is not correct, then according to mechanism $\mathcal{M}$ his utility may be negative and he may be "shocked" when seeing the final outcome. But when the game is played he believes that his utility will be non-negative and thus behaves as specified by our solution concept, in particular by Claim 3.

### 7.4 Three Concerns About the Second-Belief Mechanism "in Practice"

A concern raised about the second-belief mechanism is that "$\varepsilon$ rewards" may not be enough motivation for the players to participate. When the relevant players opt to "stay home", the second-belief benchmark cannot be guaranteed, and thus the second-price mechanism might in practice generate higher revenue.

Let us have a closer look. First, it should be appreciated that any rational player prefers a positive utility, no matter how small, to 0 utility, which is the utility he would receive if he opted out of the auction, both in the second-belief and in the second-price mechanism. (Saying otherwise requires an alternative notion of rationality.[8]) Second, as we have already observed, conservative beliefs are implicit in any context, whether

[8]To be sure, such alternative notions exist: in particular, $\varepsilon$-Nash equilibrium. Note however that *any* mechanism which, like ours, achieves a revenue benchmark —at least in some contexts— close to the highest true valuation, *must* rely on the traditional notion of rationality, instead of any $\varepsilon$-alternative. This is so because, when the revenue benchmark equals the highest valuation minus $\varepsilon$, by definition the sum of the players' utilities must be at most $\varepsilon$. Therefore any $\varepsilon$-alternative notion of rationality will make the players indifferent between participating and opting out. And when players opt out, the mechanism cannot guarantee its desired benchmark.

or not a seller tries to leverage them. Accordingly, to compare properly the second-belief and the second-price mechanism, one should consider the same, underlying, conservative belief profile $\mathcal{B}$. Consider a player $i$ who does not believe that his valuation is the highest. Then $i$ concludes that he will receive "$\varepsilon$ utility" under the second-belief mechanism, and 0 utility under the second-price one. Therefore, according to any reasonable (traditional or not) notion of rationality, if $i$ chooses to opt out in the second-belief mechanism, he should also opt out in the second-price mechanism. In neither mechanism, therefore, can player $i$ be relied upon to achieve the corresponding revenue benchmark. Consider now a player $i$ who believes that he might be the one with the highest valuation. Then, in either mechanism, it is dominant for him to participate in the auction. (In particular, in the second-belief mechanism, opting out is strictly dominated by $(0, \theta_i)$, which always has positive utility.) Accordingly, if $i$ chooses to participate in the second-price mechanism, he should also participate in the second-belief one.

Another (related) concern pertains to the case in which the players only have very imprecise external beliefs. In this case, while the revenue generated by the second-price mechanism is equal to the second-highest valuation, denoted by $2^{nd}(\theta)$, the one generated by the second-belief mechanism is "$2^{nd}(\theta) - \varepsilon$". Again, such a concern is based on an unfair comparison. The second-belief mechanism works no matter what beliefs the seller may have about the quality of the players' conservative beliefs, and insists on guaranteeing *strictly positive utilities* to the players (when they play conservatively and not all players have value 0). By contrast, the second-price mechanism only guarantees that the players' utilities are $\geq 0$, and thus cannot guarantee the participation of players who believe that they do not have the highest valuation. Accordingly, for the seller to gain an extra $\varepsilon$ in revenue by adopting the second-price mechanism instead of the second-belief one, it is necessary that he has enough information about the players: namely, *he must be sure that each player believes that he might be the one with the highest valuation.* In absence of this information, to guarantee the participation of all players, the second-price mechanism must be modified so as to provide some form of "$\varepsilon$ rewards" as well, and thus will miss its target revenue in its purest form. To be sure, the second-price mechanism can be perturbed so that all players with non-zero valuations get strictly positive utilities and it is strictly dominant for them to participate. But then the revenue of the seller becomes "$2^{nd}(\theta) - \varepsilon$" as well.

The third concern raised is that the second-belief mechanism may miss its benchmark because its players may prefer decreasing their opponents' utilities to increasing their own ones. Indeed, if (1) the player with the highest valuation is player $i$, (2) $i$ believes that he is the player with the highest valuation, (3) $i$ believes that $\theta_i \geq 2^{nd}(\mathcal{B})$, and (4) $i$ further believes that $2^{nd}(\mathcal{B}) > 2^{nd}(\theta)$, then, when all other players act rationally, by sufficiently underbidding his own valuation —e.g., by bidding $(0, 0)$— player $i$ will cause another player to receive negative utility. However, let us emphasize that, while leveraging the players' external beliefs, we continue to use the *classical utility function* for single-good auctions: namely, the utility of every player equals his true valuation minus the price he pays if he wins the good, and 0 minus the price he pays otherwise. Under such a classical utility function, the second-belief mechanism achieves its benchmark at every rational play. The concern about a player having a different type of preference is therefore out of the model.

# 8 Related Work

In Bayesian settings with a common prior, higher revenue benchmarks can be guaranteed, and, more generally, more social choice correspondences can be implemented, under proper assumptions.[9] These works are not very relevant to ours, since we focus on a non-probabilistic model of incomplete information, and we do not impose any common knowledge assumption about the players' beliefs. Let us instead recall other works, where probabilistic/common-prior assumptions have been substantially relaxed.

[9]For instance, Cremer and McLean [11] show that, for certain valuation distributions, revenue equal to the highest valuation can be achieved in a single-good auction under Bayesian Nash equilibrium or in weakly dominant strategies. Also, Abreu and Matsushima [1] show that, under some technical conditions, any Bayesian incentive compatible social-choice function can be virtually implemented in iteratively undominated strategies.

**Other Models of Incomplete Information** Postlewaite and Schmeidler [22] studied *differential information* settings for exchange economies. They model a player's uncertainty as a partition of the set of all possible states of the world, and assume such partitions to be common knowledge. In our case, we do not assume a player to have any knowledge/beliefs about the knowledge/beliefs of another player, and we certainly do not have any common-knowledge requirements. In addition, they further assume that each player has a probabilistic distribution over the state space, and use Bayesian equilibrium as the key solution concept. Their model actually reduces to Harsanyi's incomplete information model [16] if the state space is finite.

Chung and Ely [10] model a player's belief about the state of the world via a *distribution*, but assume that he prefers one outcome $\omega$ to another $\omega'$ if he locally prefers $\omega$ to $\omega'$ in every state that is possible according to his belief. In this sense, what matters is the support of the distribution, which is possibilistic. The authors show that, even when the players only have very small uncertainty about the state of the world, the set of social choice rules implementable at (essentially) undominated Nash equilibria is highly constrained compared with that in complete-information settings. Their result is less relevant for settings, like ours, where a player has no uncertainty about his own payoff type. In addition, in our purely possibilistic model, we have no requirement on how big a player's uncertainty about his opponents can be. Finally, instead of studying implementation at all equilibria (of a given type), we study the fragility of implementation even at some of them.

Artemov, Kunimoto, and Serrano [2] also model the players' beliefs about each other via distributions. But they assume that each player $i$'s belief about the others' payoff types is from a subset $Q_i$ of the set of all possible distributions, and that the $Q_i$'s are common knowledge among the players. By doing so, they assume that the players have some knowledge about each other's first-order belief. They impose no constraint on the players' higher-order beliefs, and assume that no other player knows player $i$'s true first-order belief. Their model is still different from ours. First of all, in our model a player's belief is possibilistic instead of probabilistic. Second of all, we do not assume that the players have any knowledge about each other. Moreover, their model implicitly assumes that the players' knowledge about each other's first-order belief is *correct* —i.e., player $i$'s true first-order belief is from $Q_i$, while in our model a player can have arbitrary, perhaps totally wrong, beliefs about others. Finally, the social-choice functions studied in [2] are still defined over the players' payoff types rather than their beliefs.

Our model of external information is also related to other notions in decision theory. In particular, Knight [20] and later Bewley [4] have considered players who have incomplete information about *their own* types. Specifically, a Knightian player $i$ does not know his own type $\theta_i$, nor the distribution $D_i$ from which $\theta_i$ has been drawn. Rather, he knows several distributions, one of which is guaranteed to be $D_i$. Recently Knightian players have also been studied in mechanism design, in particular, by Lopomo, Rigotti, and Shannon [21] for games with a single player, and by Chiesa, Micali, and Zhu [9] for auctions with multiple players.

Also, Hyafil and Boutilier [18] study regret-minimizing equilibria in games with multiple players having possibilistic beliefs about each other. But they assume that the players' beliefs come from a common prior, and are always correct. Our model does not make these assumptions.

**"Side Bets"** Eliaz and Spiegler [12] study mechanism design in non-common prior settings, and consider speculative bets between two players with heterogeneous prior beliefs. They characterize the conditions under which unmanipulable bets can be implemented at Bayesian equilibrium, leading to speculative gains to both players. Their framework envisages only two players, two actions for player 2, no actions for player 1, and two states of nature affecting only the utility function of player 2. Player 1 is a pure "speculator". Moreover, each player's belief can be specified by a single parameter, the probability that the first state of nature occurs, and both players' beliefs are drawn *independently* from the *same* distribution. Differently, we only rely on the players' possibilistic beliefs, and do not have assumptions about where such beliefs come from.

It is important to realize that when the players have heterogeneous probabilistic beliefs it is possible to implement speculative trade ("side-bets") leading to gains for both players, but it is not clear whether their result is generalizable to $n$ players, let alone usable to derive ours.

However, our mechanism can be expressed in terms of *degenerate* side-bets. Namely, a player $i$ announcing

$(0, v_i)$ in Step 1 can be interpreted as saying that he is not participating to the "bet" and that his own value is $v_i$; while $i$ announcing $(1, v_i)$ can be interpreted as saying that he "bets" that some other player $j$ will declare $j$'s value to be at least $v_i$.[10]

**Prior-Free Mechanisms** Prior-free mechanisms for auctions have also been investigated (in particular, by Baliga and Vohra [3], Segal [25], and Goldberg, Hartline, Karlin, Saks, and Wright [14]), although not always in auctions of a single good. The term "prior-free" seems to suggest that this approach be relevant to our possibilistic setting, but things are quite different. For instance, all cited prior-free mechanisms work in dominant strategies, and we have proved that no dominant-strategy mechanism can even approximate our revenue benchmark. More generally, as for all mechanisms, prior-free ones must be analyzed based on some underlying solution concept, and as long as they use one of the solution concepts we prove inadequate for our benchmark, they automatically fail to guarantee it.

**Impossibility Results** Several impossibility results have been proved for implementation in dominant strategies: for instance, for many forms of elections (see Gibbard [13] and Satterthwaite [24]), for maximizing social welfare in a budget-balanced way (see Green and Laffont [15] and Hurwicz [17]), and for maximizing revenue in general settings of quasi-linear utilities (see Chen, Hassidim and Micali [6]). As for mechanisms working in undominated strategies, Jackson [19] shows that the set of social choice correspondences (fully) implementable by bounded mechanisms (which include finite ones) is quite constrained. We note, however, that none of these results imply ours for implementing the second-belief benchmark in either dominant or undominated strategies (indeed, our results do not require full implementation).

**Our Own Prior Work** In [7] we studied mechanisms leveraging only (what we now call) *correct external beliefs*, and constructed one such mechanism for truly combinatorial auctions. (This mechanism would also work with incorrect external beliefs, but under a slightly different analysis.) In a later work with Valiant [8], we were able to extend our combinatorial-auction mechanism so as to leverage also, to a moderate extent, the internal knowledge of the players.[11] In neither of these two prior papers we proved any impossibility results: given that no significant revenue guarantees were known for combinatorial auctions, we were satisfied with achieving new, reasonable benchmarks. For instance, in [7] we showed the existence of a very robust mechanism that, in any truly combinatorial auction and without any knowledge about the players' true valuations, generates within a factor of 2 the "maximum revenue that a player could guarantee if he were charged to sell the goods to his competitors by means of take-it-or-leave-it offers."

Perhaps interestingly, our prior mechanisms were of extensive form, and we still do not know whether equivalent, normal-form ones exist.

# 9 Future Directions

We believe that much work can be done in leveraging the players' possibilistic beliefs. Indeed, in a recent and unpublished work with Rafael Pass, we exhibit single-good auction mechanisms guaranteeing even higher revenue benchmarks (based on the players' possibilistic higher-order beliefs), under different solution concepts.

Beyond single-good auctions, we plan to investigate what social choice correspondences can be implemented by leveraging the players' possibilistic beliefs in other strategic settings.

Finally, we should investigate models where some of the players' beliefs are possibilistic, and some are probabilistic, but without assuming the correctness of such beliefs, let alone their being common knowledge.

[10] This alternative language leaves totally open whether it is possible to design an interim individually rational mechanism generating revenue *always* higher than our benchmark by leveraging first-order possibilistic beliefs. The hard case continues to be when all conservative beliefs are correct. Indeed, with possibilistic beliefs, the players cannot compute their "expected utilities" under certain bets, and thus can only bet on events that they are certain about, such as the values $smv_i$.

[11] The emphasis of [8] actually was the possibility of leveraging the internal knowledge of coalitions rather than individual ones.

## References


[1] D. Abreu and H. Matsushima. Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information. Mimeo, 1992.

[2] G. Artemov, T. Kunimoto, and R. Serrano. Robust Virtual Implementation with Incomplete Information: Towards a Reinterpretation of the Wilson Doctrine. Working paper, 2007.

[3] S. Baliga and R. Vohra. Market Research and Market Design. *Advances in Theoretical Economics*, 3(1), Article 5, 2003.

[4] T. F. Bewley. Knightian decision theory Part I. *Decisions in Economics and Finance*, Vol. 25, No. 2, pp. 79-110, 2002.

[5] T. N. Casona, T. Saijob, T. Sjostrom, and T. Yamatoe. Secure Implementation Experiments: Do Strategy-Proof Mechanisms Really Work? *Games and Economic Behavior*, 57(2): 206-235, 2006.

[6] J. Chen, A. Hassidim, and S. Micali. Robust Perfect Revenue from Perfectly Informed Players. *Innovations in Computer Science (ICS)*, pp. 94-105, 2010.

[7] J. Chen and S. Micali. A New Approach to Auctions and Resilient Mechanism Design. *41st Symposium on Theory of Computing (STOC)*, pp. 503-512, 2009.

[8] J. Chen, S. Micali, and P. Valiant. Robustly Leveraging Collusion in Combinatorial Auctions. *Innovations in Computer Science (ICS)*, pp. 81-93, 2010.

[9] A. Chiesa, S. Micali, and Z. A. Zhu. Mechanism Design with Approximate Valuations. *Innovations in Theoretical Computer Science (ITCS)*, pp. 34-38, 2012.

[10] K.S. Chung and J.C. Ely. Implementation with Near-Complete Information. *Econometrica*, 71(3): 857-871, 2003.

[11] J. Cremer and R.P. McLean. Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions. *Econometrica*, Vol.56, No.6, pp. 1247-1257, Nov., 1988.

[12] K. Eliaz and R. Spiegler. A Mechanism-Design Approach to Speculative Trade. *Econometrica*, 75(3): 875-884, 2007.

[13] A. Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4): 587-602, 1973.

[14] A. Goldberg, J. Hartline, A. Karlin, M. Saks, and A. Wright. Competitive Auctions. *Games and Economic Behavior*, 55(2): 242-269, 2006.

[15] J. Green and J. Laffont. Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods. *Econometrica*, 45(2): 427-438, 1977.

[16] J. Harsanyi. Games with Incomplete Information Played by "Bayesian" Players, I-III. Part I. The Basic Model. *Management Science*, 14(3) Theory Series: 159-182, 1967.

[17] L. Hurwicz. On the Existence of Allocation Systems Whose Manipulative Nash Equilibria Are Pareto Optimal. Unpublished. 1975.

[18] N. Hyafil and C. Boutilier. Regret Minimizing Equilibria and Mechanisms for Games with Strict Type Uncertainty. *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 268-277, 2004.

[19] M. Jackson. Implementation in Undominated Strategies: A Look at Bounded Mechanisms. *The Review of Economic Studies*, 59(4): 757-775, 1992.

[20] F. H. Knight. Risk, Uncertainty and Profit. Houghton Mifflin, Boston, MA. 1921.

[21] G. Lopomo, L. Rigotti, and C. Shannon. Uncertainty in Mechanism Design. Revise and resubmit at *Review of Economic Studies*, 2009.

[22] A. Postlewaite and D. Schmeidler. Implementation in Differential Information Economies. *Journal of Economic Theory*, 39(1): 14-33, 1986.

[23] T. Saijo, T. Sjostrom, and T. Yamato. Secure Implementation: Strategy-Proof Mechanisms Reconsidered. Unpublished. 2003.

[24] M. Satterthwaite. Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, 10(2): 187-217, 1975.

[25] I. Segal. Optimal Pricing Mechanisms with Unknown Demand. *American Economic Review*, 93(3): 509-529, 2003.