Mechanism Design with Set-Theoretic Beliefs

Jing Chen
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Silvio Micali
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

October 22, 2011

This material is based on work supported by the U.S. Office of Naval Research, Grant No. N00014-09-1-0597. Any opinions, findings, conclusions or recommendations therein are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

2011 IEEE 52nd Annual Symposium on Foundations of Computer Science

Mechanism Design with Set-Theoretic Beliefs*

Jing Chen, Silvio Micali
CSAIL, MIT
Cambridge, MA 02139, USA
$\{$*jingchen, silvio*$\}$*@csail.mit.edu*

Abstract— In settings of incomplete information, we put forward

(1) a very conservative —indeed, purely set-theoretic— model of the beliefs (including totally wrong ones) that each player may have about the payoff types of his opponents, and
(2) a new and robust solution concept, based on *mutual* belief of rationality, capable of leveraging such conservative beliefs.

We exemplify the applicability of our new approach for single-good auctions, by showing that, under our solution concept, a normal-form, simple, and deterministic mechanism guarantees —up to an arbitrarily small, additive constant— a revenue benchmark that is always greater than or equal to the second-highest valuation, and sometimes much greater. By contrast, we also prove that the same benchmark cannot even be approximated within any positive factor, under classical solution concepts.

Keywords-single-good auctions; beliefs; revenue

1. Introduction

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

For achieving a desired goal, 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] Bayesian mechanisms leverage both, under the assumption that the players' beliefs consist of probability distributions.[2] Such an assumption is quite natural because uncertainty is traditionally modeled by probability distributions, but is an assumption nonetheless. Independent of whatever additional assumptions may be required by specific mechanisms (e.g., that the distribution from which θ is drawn is known to the designer or is common knowledge to the players), it imposes significant structural constraints on the players' external beliefs. For instance, consider a player i who, in a single-good auction, values the item for sale 50 and believes that one of his opponents values for more than 100. Such a belief is not a distribution —i may not know whom such a high-valuing player might be, nor what the probabilities for his valuation being 101, 102, etc. might be— and is not leverageable by Bayesian mechanisms.

In sum, classical mechanisms exploit two extremes — namely, (1) the players have no external beliefs and (2) their external beliefs consist of probability distributions— but not the vast ground in between. Personally, we consider the first extreme as too pessimistic and the second as too optimistic, and wish to explore a "middle road" to mechanism design.

Our Focus While our belief model and solution concept are 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.

1.1. The Conservative-Belief Model

Definition 1. *A* ***conservative context*** C *consists of a tuple* $(n, \Omega, \Theta, u, \theta, \mathscr{B})$*, where*

- $(n, \Omega, \Theta, u, \theta)$ *is a traditional context of incomplete information,*[3] *and*
- $\mathscr{B}$ *is a profile such that, for each player* i*, (1)* $\mathscr{B}_i \subseteq \Theta$ *and (2)* $t_i = \theta_i$ *for all* $t \in \mathscr{B}_i$*.*

We refer to $\mathscr{B}$ *as the* ***conservative belief profile****, and say that* $\mathscr{B}_i$ *is* ***correct*** *if* $\theta \in \mathscr{B}_i$*.*

In a conservative context, $\mathscr{B}_i$ represents all possible candidates for the true type profile in player i's view. (We do not include the players' higher-level beliefs in our contexts because our solution concept prevents such beliefs from affecting a rational play of our mechanism.)

Knowledge and Beliefs Components n, Ω, Θ, and u are common knowledge to everyone. Each player i individually knows θ_i and $\mathscr{B}_i$, is rational, and believes that his opponents are rational. (Any unspecified knowledge and belief of players or mechanism designers can be chosen arbitrarily.)

*Supported in part by ONR Grant No. N00014-09-1-0597.

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

[2] For instance, when θ is assumed to be drawn from a distribution $\mathcal{D}$ that is common knowledge to the players, and the underlying solution concept is Bayesian equilibrium, the external belief of a player i can be taken to be $\mathcal{D}|\theta_i$, that is, "the distribution of θ_{-i} obtained by conditioning $\mathcal{D}$ on θ_i."

[3] That is, $\{1, \ldots, n\}$ is the set of players; Ω the set of outcomes; $\Theta = \Theta_1 \times \cdots \times \Theta_n$ the set of all possible (payoff) type profiles; u the profile of utility functions, each u_i mapping $\Theta_i \times \Omega$ to $\mathbb{R}$, the set of reals; and $\theta \in \Theta$ the profile of true types. If $t_i \in \Theta_i$ and ω a distribution over Ω, then $u_i(t_i, \omega)$ is the expectation induced by ω.

0272-5428/11 $26.00 © 2011 IEEE
DOI 10.1109/FOCS.2011.11

Conservative Single-Good Auction Contexts A *conservative single-good auction context* is a conservative context $(n, \Omega, \Theta, u, \theta, \mathscr{B})$ where: $\Theta = \{0, 1, \dots, V\}^n$ for some positive integer V referred to as the *valuation bound*; $\Omega = \{0, 1, \dots, n\} \times \mathbb{R}^n$;[4] and each utility function u_i is so defined: $u_i(t_i, (a, P))$ equals $t_i - P_i$ if $i = a$, and $-P_i$ otherwise.

If $\omega = (a, P) \in \Omega$, then player i's utility for ω, $u_i(\omega)$, is $u_i(\theta_i, \omega)$; and the *revenue* of ω, $REV(\omega)$, is $\sum_i P_i$. We denote by $\mathscr{C}_n^V$ the set of all conservative single-good auction contexts with n players and valuation bound V, and by $\mathscr{D}_n^V$ the set of all contexts in $\mathscr{C}_n^V$ where the conservative belief of every player is correct.

Remarks

- Working solely in our model, we may drop the term "conservative" or use it for emphasis/clarity only. Further, since all auctions we consider are single-good, we may also omit the term "single-good."
- Note that an auction context C is identified by n, V, θ and $\mathscr{B}$ alone: that is, $C = (n, V, \theta, \mathscr{B})$.
- In an auction context, a player i's true type θ_i —also called i's true *valuation*— represents i's value for the good for sale, and i's conservative belief $\mathscr{B}_i$ is a set of non-negative integer profiles.
- Players' beliefs can be wrong. Indeed it may even be the case that $\theta \notin \mathscr{B}_i$ for each player i.
- The profile $\mathscr{B}$ is compatible with the players having *additional* beliefs, even of a probabilistic nature. In no case, however, can these additional beliefs contradict $\mathscr{B}$. For instance, if a player i believes that the true type profile has been drawn from some distribution $\mathcal{D}$, then $\mathscr{B}_i$ should coincide with $\mathcal{D}$'s support.
- Conservative belief is a *model* rather than an *assumption*. As usual, a player i knows θ_i, but we make no requirement about his external belief. For instance, he may have no external belief whatsoever. In this case, $\mathscr{B}_i = \Theta_1 \times \cdots \times \Theta_{i-1} \times \{\theta_i\} \times \Theta_{i+1} \times \cdots \times \Theta_n$. On the other extreme, he may have no external uncertainty whatsoever. In this case, $\mathscr{B}_i = \{t\}$ for some type profile t (not necessarily equal to θ).[5]
- Aiming at robustness, we are very conservative when modeling what we may exploit, but very liberal when modeling what may hurt us. That is, our mechanisms only leverage $\mathscr{B}$ in order to achieve their goals, but must work no matter what additional beliefs (compatible with $\mathscr{B}$) the players might have.
- Relative to $\mathscr{B}$, the *external belief* of a player i, $\mathscr{E}_i$, is formally defined to be the set $\{t_{-i} : (\theta_i, t_{-i}) \in \mathscr{B}_i\}$.
- As a player i's type is a comprehensive description of i in the strategic situation at hand, we are essentially separating i's *payoff* type, θ_i, from his *external-belief* type, $\mathscr{E}_i$.

1.2. Conservative-Belief Social Choice Correspondences and Their Advantages

Traditionally, social choice correspondences map *type profiles* to sets of (distributions over) outcomes, but can be naturally extended to map *conservative-belief profiles* to sets of outcomes. This extension strictly enriches the set of "targets" for mechanism design. As noted, each context C implicitly has a conservative-belief profile $\mathscr{B}$, from which the true type profile θ could be easily computed. Thus, for each traditional correspondence f there exists an extended one F such that $f(\theta) = F(\mathscr{B})$, but not vice versa.

The advantage of a meaningful and enlarged "target space" is pretty clear. Very often we do not know how to design mechanisms implementing a given, traditional, social choice correspondence f. Sometimes we can actually prove that designing such mechanisms is impossible (at least for some type of implementation —e.g., in dominant strategies). In these cases, while one can always shop around for new, meaningful, and achievable targets among traditional social choice correspondences, extended social choice correspondences provide access to *additional* ones, i.e., targets that are not even expressible in terms of θ alone, more tractable, and yet reasonable. For instance, in [4] we prove 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."

In this paper, rather than replacing classical social choice correspondences with "tamer" ones, we use conservative beliefs in order to define and then achieve a new correspondence "tougher" than classical ones.

1.3. The Second-Belief Revenue Benchmark

In auction contexts, a *revenue benchmark* F is a function mapping each conservative belief profile $\mathscr{B}$ to a real number. Thus, *de facto*, F is a social choice correspondence: the one mapping each $\mathscr{B}$ to the set of outcomes whose revenue is at least $F(\mathscr{B})$. Let us now define a revenue benchmark for single-good auctions.

Definition 2. *The* ***second-belief*** *benchmark, denoted by* 2^{nd}*, is the revenue benchmark so defined. For a belief profile* $\mathscr{B}$*, let* $smp_i = \min_{t \in B_i} \max_j t_j$ *for each player* i*. Then,* $2^{nd}(\mathscr{B})$ *is the second highest value in* $\{smp_1, \dots, smp_n\}$*.*

If t were the true valuation profile, then $\max_j t_j$ would be the maximum price that a player is willing to pay for the good. Thus, relative to $\mathscr{B}_i$, smp_i is the maximum price

[4] In an outcome (a, P), a denotes the player getting the good if > 0, or that the good is unallocated if $= 0$; and P denotes the price profile.

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

for which player i is *sure* that *some* player (possibly i or a player whose identity is not precisely known to i) is willing to pay for the good.

A Simple Example Consider an auction with three players where

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

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 $\mathscr{B}$, $smp_1 = smp_2 = 100$, $smp_3 = 150$, and thus $2^{nd}(\mathscr{B}) = 100$, which happens to be the highest valuation.

Remark Sometimes $2^{nd}(\mathscr{B})$ can be greater than the highest valuation, but never when all beliefs are correct. However, since $smp_i \geq \theta_i$ for every player i, it is always the case that "$2^{nd}(\mathscr{B}) \geq 2^{nd}(\theta)$": that is, our benchmark is always greater than or equal to the second highest true valuation. Accordingly, a mechanism designer concerned with generating revenue should try to achieve the second-belief benchmark instead of using the second-price mechanism to generate revenue equal to the second-highest valuation. If he succeeds, *the seller may have something (possibly a lot) to gain and nothing to lose.*

As we prove, however, this more demanding benchmark cannot be achieved via classical solution concepts.

1.4. The Impossibility of Classically Implementing the Second-Belief Benchmark

Recall that a mechanism M provides each player i with a set of pure strategies, consistently denoted by S_i in this paper, and maps each strategy profile σ to an outcome (or a distribution over outcomes, if M is probabilistic or σ a mixed-strategy profile) denoted by $M(\sigma)$. Also recall that a mechanism is finite if each S_i is finite, and that a game G consists of a context C and a mechanism M: $G = (C, M)$. Finally, when the mechanism M is clear, for any strategy profile σ, we may denote $u_i(M(\sigma))$ by $u_i(\sigma)$ for short.

For our impossibility results, we consider mechanisms that allow the players to "stay home", that is, to opt out of the auction. Otherwise, one could trivially and meaninglessly generate high revenue by forcing the players to participate in a mechanism always giving them very negative utility.

Definition 3. *A mechanism M is* ***reasonable*** *if it is finite and satisfies the following* ***opt-out condition****: $\forall$ player i $\exists out_i \in S_i$ such that for (any possible true type θ_i and) any strategy subprofile $s_{-i} \in S_{-i}$,*

$$u_i(M(out_i, s_{-i})) = 0.$$

Remarks

- Having the opt-out condition requiring i's utility to be 0 in expectation, rather than for every outcome in the support of $M(out_i, s_{-i})$, can only make our impossibility results stronger.
- Our impossibility results already hold for auctions with just two players, and when all beliefs are correct. Actually, when the players' beliefs are not correct these results become trivial.[6] Accordingly, we state our impossibility results in terms of $\mathscr{D}_n^V$ instead of $\mathscr{C}_n^V$.
- In our impossibility results we never assume any restrictions on the strategy spaces. In particular, our results also apply to normal-form mechanisms that let the players report their (alleged) conservative beliefs, as it is fair to do so when trying to leverage them.

1.4.1. Impossibility of Implementation in Undominated Strategies

Implementation in undominated strategies is a classical notion for settings of incomplete information.[7] We strengthen our first impossibility result by adopting a *weaker* notion of such implementation.[8] Notice that this weaker notion is already *sufficient* from a mechanism designer's point of view.

Definition 4. *A mechanism M* ***sufficiently*** *implements a revenue benchmark F for a class $\mathscr{C}$ of auction contexts in undominated strategies if, $\forall$ contexts $C \in \mathscr{C}$ and $\forall$ profiles s of undominated strategies in the game (C, M), denoting by $\mathscr{B}$ the belief profile of C, we have that*

$$REV(M(s)) \geq F(\mathscr{B}).$$

Theorem 1. *$\forall \epsilon \in (1/2, 1]$ and $\forall$ $V > \lceil \frac{1}{\epsilon - 1/2} \rceil$, no reasonable mechanism sufficiently implements $\epsilon 2^{nd}$ for $\mathscr{D}_2^V$ in undominated strategies.*

For deterministic mechanisms and purely undominated strategies, our impossibility result holds for arbitrary approximation factors.

[6] 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.

[7] Given a game $G = (C, M)$, a strategy s_i of player i is *weakly dominated* by another (possibly mixed) strategy σ_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. Thus, to compute his own undominated strategies in a game, a player needs not have any information about his opponents' (payoff) types.

[8] Note that the traditional notion of (full) implementation in undominated strategies —see Jackson [13]— 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.

Theorem 2. *$\forall \epsilon \in (0, 1]$ and $\forall\ V > \lceil 1/\epsilon \rceil$, no reasonable deterministic mechanism sufficiently implements $\epsilon 2^{nd}$ for $\mathscr{D}_2^V$ in purely undominated strategies.*

The proof of Theorem 1 is provided in the full version of this paper [5]. The proof of Theorem 2 is very similar and thus omitted.

1.4.2. Impossibility of Implementation in Dominant Strategies

Theorems 1 and 2 immediately yield the following about strictly/weakly/very weakly dominant strategies.[9]

Corollary 1. *$\forall \epsilon \in (1/2, 1]$ and $\forall\ V > \lceil \frac{1}{\epsilon - 1/2} \rceil$, no reasonable mechanism implements $\epsilon 2^{nd}$ for $\mathscr{D}_2^V$ in strictly/weakly dominant strategies or in (**all**) very weakly dominant strategies.*

Corollary 2. *$\forall \epsilon \in (0, 1]$ and $\forall\ V > \lceil 1/\epsilon \rceil$, no reasonable deterministic mechanism implements $\epsilon 2^{nd}$ for $\mathscr{D}_2^V$ in strictly/weakly dominant strategies or in (**all**) very weakly dominant strategies.*

A Crucial Clarification Note that, Theorems 1 and 2 not withstanding, the above 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 again that all our impossibility results hold without any restrictions on strategy spaces, and in particular that a mechanism asking the players to announce conservative beliefs cannot be "simulated" by one asking them to announce only valuations.

1.4.3. Extra Fragility of Implementation at Some Ex-Post/Very Weakly Dominant Equilibria

A mechanism guaranteeing a given property at *some* equilibria of a given type is certainly more fragile than one guaranteeing it at *all* equilibria of that type. Indeed, one has no control over the equilibrium ultimately selected by the players. But mechanisms implementing $\epsilon 2^{nd}$ at some ex-post or very weakly dominant equilibria have some **extra fragility**. Consider the following mechanism for $\mathscr{C}_2^{100}$.

> **Mechanism** NAIVE. *A strategy of player i has two components: an integer a_i and a set $b_i \subseteq \{0, 1, \ldots, 100\}$. (Allegedly, a_i is player i's true valuation, and b_i his true external belief.) The winner and prices are decided as follows. Let $w = \operatorname{argmax}_i a_i$ (ties broken lexicographically), and let $P = \min_{t \in \mathscr{B}'_{-w}} \max_j t_j$ where $\mathscr{B}'_{-w} = \{a_{-w}\} \times b_{-w}$. If $a_w \geq P$, then the good is sold to player w, w pays P, and his opponent pays 0. Else, the good is unsold and both players pay 0.*

According to NAIVE, it is clear that every player announcing his true valuation and true external belief in every context is an ex-post equilibrium. When the players' beliefs are correct, this equilibrium guarantees second-belief revenue. However, consider the context C where

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

In this context, all beliefs are correct, $2^{nd}(\mathscr{B}) = 90$, the truthful ex-post equilibrium yields the strategy profile $((70, \{x : x \geq 90\}), (100, \{x : x \geq 60\}))$, and it generates revenue 90 as desired. However, it is also clear that $((70, \{x : x \geq 0\}), (100, \{x : x \geq 60\}))$ is an *alternative* Nash equilibrium —corresponding to another ex-post equilibrium— whose revenue is only 70.

In principle —e.g., when two Nash equilibria differ at multiple players, one can argue that a player may be able to establish some belief about which equilibrium is going to be played out by the others, and best respond to his belief. But in the above example, the "truthful" and the "alternative" equilibria differ only at player 1's strategy. Thus, even if player 1 believed that player 2 will play his truthful strategy, it would also be perfectly rational for player 1 to play his own alternative strategy. Viceversa, even if player 2 believed that player 1 will play his alternative strategy, it would also be perfectly rational for player 2 to stick to his own truthful strategy (which coincides with his alternative one in the above example).

Accordingly, *which revenue should we expect from* NAIVE *at context C?* The answer is 90 if player 1 is "generous" towards the seller and 70 otherwise.[10] In the full version of this paper [5], we formalize this phenomenon and prove that such extra fragility is actually *unavoidable* for *any* mechanism implementing (or even approximating) the second-belief benchmark at some ex-post or very weakly dominant equilibria.

1.5. Conservative Strict Implementation: Our New Solution Concept

The inability of achieving the second-belief benchmark via classical notions of implementation encourages us to

[9] A strategy s_i of player 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, $u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$ for every s_{-i}, and the inequality is strict for some 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}.

[10] Notice that the truthful ex-post equilibrium actually specifies a very weakly dominant strategy for each player in each context, and thus illustrates the lack of robustness for implementation at some very weakly dominant equilibria as well. Such lack of robustness was already pointed out by Saijo, Sjostrom, and Yamato theoretically [15] and by Casona, Saijo, Sjostrom, and Yamato experimentally [2]. In [15] 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.

develop a new one. Intuitively, but erroneously, our notion can be taken to consist of *"two-round elimination of strictly dominated strategies"* (hardly a new solution concept!). The problem is that such elimination is not well defined in a setting of *incomplete* information: without knowing his opponents' payoff types, a player is not capable of figuring out what strategies are left for them after the first round, and thus is not capable of figuring out which of his own strategies are dominated in the second round. Therefore we must be more careful. Our notion is formally defined in Section 3, but can be summarized as follows.

Sketch of Our Notion We say that a normal-form mechanism M *conservatively strictly implements* a social choice correspondence F for a class of contexts $\mathscr{C}$ if, for any context $C \in \mathscr{C}$, denoting by $\mathscr{B}$ the belief profile of C, we have $M(s) \in F(\mathscr{B})$ for any strategy profile s surviving the following two-step elimination procedure:

1. Each player eliminates all of his strictly dominated strategies;
2. Based on his conservative belief $\mathscr{B}_i$, and assuming that everyone completes Step 1, each player i eliminates all his remaining strategies that are *dominated relative to* $\mathscr{B}_i$.

The real novelty of our notion, and the key for meaningfully leveraging set-theoretic beliefs, lie with properly defining "domination relative to $\mathscr{B}_i$" in Step 2. As usual, after Step 1, to determine which of his remaining strategies are dominated, i should know what are the currently surviving strategies of the other players. However, to figure this out, player i must also know what are the true types of the other players —which is precisely a piece of information that he does not have in a setting of incomplete information. We address this concern by breaking down Step 2 into two conceptual sub-steps as follows.

2.1 Each player i, for each type profile t in $\mathscr{B}_i$, computes the profile $S(t)$, where each $S(t)_j$ represents the set of surviving strategies for player j after Step 1, if t were the true type profile.

2.2 Each player i eliminates a Step-1 surviving strategy s_i if and only if there exists another (possibly mixed) Step-1 surviving strategy σ_i that (classically) strictly dominates s_i with respect to $S(t)$ for *each* $t \in \mathscr{B}_i$.

Remark Let us emphasize a subtle point hidden in Step 2.2. Consider the following two ways of defining s_i to be "dominated relative to $\mathscr{B}_i$":

(i) for *each* $t \in \mathscr{B}_i$, s_i is strictly dominated with respect to $S(t)$ by *some* σ_i, and

(ii) for *each* $t \in \mathscr{B}_i$, s_i is strictly dominated with respect to $S(t)$ by *the same* σ_i.

Although both ways are based on the players' set-theoretic beliefs $\mathscr{B}$, we have adopted the latter one. The reason is that, when he eliminates a strategy s_i dominated according to **(ii)**, player i is sure to have a better strategy to play, namely σ_i, no matter which type profile in $\mathscr{B}_i$ might be the right one. But the same is not true when he eliminates a strategy dominated according to **(i)**.

Example[11] Consider a mechanism M played by two players, where the true type profile is $\theta = (\theta_1, \theta_2)$, and the belief of player 1 is $\mathscr{B}_1 = \{(\theta_1, \theta_2), (\theta_1, \theta_2')\}$. (Since we are going to analyze only player 1's behavior, we do not need to specify $\mathscr{B}_2$ nor the other possible type profiles.) The mechanism gives player 1 the pure strategies a, b, and c, and player 2 the pure strategies d and e. For each type profile in $\mathscr{B}_1$, the players' utilities under M are as follows.

(θ_1, θ_2)

2 1	d	e
a	2,0	2,1
b	-100,0	3,1
c	3,0	-100,1

(θ_1, θ_2')

2 1	d	e
a	2,1	2,0
b	-100,1	3,0
c	3,1	-100,0

Notice that, in Step 1 of our notion, player 1 cannot eliminate any strategy. Player 2 instead would eliminate d (strictly dominated by e) if his true type were θ_2, and e (strictly dominated by d) if his true type were θ_2'. Let us now consider Step 2. If we adopted definition **(i)** in Step 2.2, then player 1 should eliminate strategy a, because it is strictly dominated by b with respect to his candidate type profile (θ_1, θ_2), and by c with respect to his other candidate type profile (θ_1, θ_2'). However, whether player 1 should play b or c in place of a really depends on whether (θ_1, θ_2) or (θ_1, θ_2') is the true type profile. If he makes the wrong choice, then his loss is huge compared with his possible gain: namely, -100 versus 3. Without any "likelihood" associated with each candidate type profile in his belief $\mathscr{B}_1$, it might be reasonable and safer for player 1 to use a to always get utility 2. (Thus, if M banked on player 1 not choosing a in order to implement its desired social choice correspondence, it may not implement it in a robust sense.)

Mutual Belief of Rationality Implementation in dominant or undominated strategies only requires that every player is rational. Conservative strict implementation instead additionally requires that every player believes that his opponents are rational. However, it does not require "higher-level" beliefs of rationality, let alone common belief. That is,

> *Conservative strict implementation solely relies on rationality and* ***mutual*** *belief of rationality.*

In essence, our notion is only "slightly" weaker than implementation in strictly dominant strategies, yet is defined carefully to explicitly leverage the players' beliefs about others in a robust way.

[11] We thank Paul Valiant for this example.

1.6. The Second-Belief Benchmark is Conservatively Strictly Implementable

Finally, we prove that conservative strict implementation succeeds where classical notions fail. Namely, under our solution concept, we exhibit a mechanism $\mathcal{M}$, the *second–belief mechanism*, that guarantees second-belief revenue, within an arbitrarily small additive value ϵ, in all single-good auction contexts. Our mechanism is uniformly specified for all values ϵ, numbers of players n, and valuation bounds V: $\mathcal{M} = \mathcal{M}_{\epsilon,n,V}$. Formally,

Theorem 3. *For any $\epsilon \in (0, 1]$, n, and V, $\mathcal{M}_{\epsilon,n,V}$ conservatively strictly implements $2^{nd} - \epsilon$ for $\mathscr{C}_n^V$.*

The second-belief mechanism is defined in Section 4 and analyzed in Section 5. In Section 6 we address three concerns raised about our mechanism.

2. Related Work

Works centered on true-type prior distributions, known or not to the designers/players, are unrelated to our set-theoretic framework. There are, however, relevant works whose probabilistic assumptions are less central.

Other Models of Incomplete Information Postlewaite and Schmeidler [14] 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 [11] if the state space is finite.

Chung and Ely [7] model a player's belief about the state of the world via a *distribution*, but assume that he prefers one outcome ω to another ω' if he locally prefers ω to ω' in every state that is possible according to his belief. In this sense, what matters is the support of the distribution, which is set-theoretic. 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 set-theoretic 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.

Impossibility Results Several impossibility results have been proved for implementation in dominant strategies: for instance, for many forms of elections (see Gibbard [8] and Sattherwaite [16]), for maximizing social welfare in a budget-balanced way (see Green and Laffont [10] and Hurwicz [12]), and for maximizing revenue in general settings of quasi-linear utilities (see Chen, Hassidim and Micali [3]). As for mechanisms working in undominated strategies, Jackson [13] 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).

Prior-Free Mechanisms Prior-free mechanisms for auctions have also been investigated —in particular, by Baliga and Vohra [1], Segal [17], and Goldberg, Hartline, Karlin, Saks, and Wright [9], although the first two of them do not consider auctions of a single good. The term "prior-free" seems to suggest that this approach be relevant to our set-theoretic 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 would automatically fail to guarantee it.

Our Own Prior Work In [4] we studied mechanisms leveraging only (what we now call) *external correct beliefs*, and, as already mentioned, 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 [6], we were able to extend our combinatorial-auction mechanism so as to leverage also, to a moderate extent, the internal knowledge of the players.[12] 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. Perhaps interestingly, our prior mechanisms were of extensive form, and we still do not know whether equivalent, normal-form ones exist.

3. Conservative Strict Implementation

The following two auxiliary definitions envisage a game with context $C = (n, \Omega, \Theta, u, \theta, \mathscr{B})$ and mechanism M (whose strategy-profile set is denoted by S as usual).

[12] The emphasis of [6] actually was the possibility of leveraging the internal knowledge of coalitions rather than individual ones.

Definition 5. *Let i be a player, t_i a type of i, and $T = T_1 \times \cdots \times T_n$ a set of pure strategy profiles. Then,*

- *We say that a strategy $s_i \in T_i$ is **strictly t_i-T-dominated** by another strategy*[13] *$\sigma_i \in \Delta(T_i)$, in symbols $s_i <_T^{t_i} \sigma_i$, if for all strategy subprofiles $s_{-i} \in T_{-i}$, $u_i(t_i, M(s_i, s_{-i})) < u_i(t_i, M(\sigma_i, s_{-i}))$.*
- *We denote by $S(t_i)$ the set of pure strategies of i that are not strictly t_i-S-dominated, and, for any type profile t, we set $S(t) = S(t_1) \times \cdots \times S(t_n)$ and $S(t_{-i}) = S(t_1) \times \cdots \times S(t_{i-1}) \times S(t_{i+1}) \times \cdots \times S(t_n)$.*

Accordingly, s_i is strictly dominated by σ_i in the traditional sense if $s_i <_S^{\theta_i} \sigma_i$, and $S(t_i)$ represents the strategies of i that would survive elimination of strictly dominated strategies (in the traditional sense) if his true type were t_i. Also note that, for any $t \in \mathscr{B}_i$, $S(t_i) = S(\theta_i)$, because $t_i = \theta_i$, while $S(t_j)$ and $S(\theta_j)$ may be very different for $j \neq i$. Thus, in general $S(t) \neq S(\theta)$ for $t \neq \theta$.

Definition 6. *A strategy $s_i \in S(\theta_i)$ is **conservatively strictly dominated** if there exists another strategy $\sigma_i \in \Delta(S(\theta_i))$ that strictly θ_i-$S(t)$-dominates s_i for all $t \in \mathscr{B}_i$. Else, s_i is **conservatively strictly rational**.*

We are now ready to define our notion of implementation.

Definition 7. *We say that a mechanism M **conservatively strictly implements** a social choice correspondence F for a class of contexts $\mathscr{C}$ if, for all contexts $C \in \mathscr{C}$ and for all profiles s of conservatively strictly rational strategies in (C, M), denoting by $\mathscr{B}$ the belief profile of C, we have that $M(s) \in F(\mathscr{B})$.*

4. The Second-Belief Mechanism

For any $\epsilon \in (0, 1]$, n, and V, the mechanism $\mathcal{M}_{\epsilon,n,V}$ is described below. Note that the mechanism applies to any context in $\mathscr{C}_n^V$, and is of normal form because the players act simultaneously and only once: in Step **1**. Steps **a** through **e** are just "conceptual steps taken by the mechanism". The expression "$X := x$" denotes the operation that sets or resets variable X to value x.

Mechanism $\mathcal{M}_{\epsilon,n,V}$

a: *Set $a := 0$, and $P_i := 0$ for all players i.*
COMMENT. Upon termination, after all proper resettings, (a, P) will be the final outcome.

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

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.*
COMMENT. Player a gets the good for sure, and thus the mechanism never leaves the good unassigned.

d: *If $e_a = 0$ then $P_a := \max_{j \neq a} v_j$; otherwise $P_a := v_a$.*
COMMENT. If a's announcement is about himself, then he pays the second-highest v_i, else the highest.

e: *For each player i, $P_i := P_i - \delta_i$, where $\delta_i = \frac{\epsilon}{2n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right]$.*
COMMENT. Each player i receives a reward δ_i.

Remark As promised, it is clear that $\mathcal{M}_{\epsilon,n,V}$ is uniformly and efficiently constructible on inputs ϵ, n, and V. In addition, it is very simple. In light of our impossibility results about implementing $\epsilon 2^{nd}$ under classical solution concepts, this simplicity suggests that conservative strict implementation can be quite powerful.

5. Analysis of The Second-Belief Mechanism

Theorem 3. *For any $\epsilon \in (0, 1]$, n, and V, $\mathcal{M}_{\epsilon,n,V}$ conservatively strictly implements $2^{nd} - \epsilon$ for $\mathscr{C}_n^V$.*

Proof. Arbitrarily fix $\epsilon \in (0, 1]$, n, V, $C = (n, V, \theta, \mathscr{B}) \in \mathscr{C}_n^V$, and a strategy profile s. Denoting $\mathcal{M}_{\epsilon,n,V}$ by $\mathcal{M}$ for short, it suffices for us to prove that, if s is conservatively strictly rational in the game $(C, \mathcal{M})$, then

$$REV(\mathcal{M}(s)) \geq 2^{nd}(\mathscr{B}) - \epsilon. \qquad (1)$$

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 $t_i \in \{0, \ldots, V\}$ of i, if $s_i \in S(t_i)$ then $v_i \geq t_i$.*

PROOF OF CLAIM 1. Assume for sake of contradiction that $s_i \in S(t_i)$ and $v_i < t_i$. We show that s_i is strictly t_i-S-dominated by $s_i' = (0, t_i)$. By definition, this implies that $s_i \notin S(t_i)$, 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 δ_i and δ_i' be the rewards that player i receives in Step **e**, in the plays of (s_i, s_{-i}') and (s_i', s_{-i}') respectively. By the construction of $\mathcal{M}$ we have that

$\delta_i = \frac{\epsilon}{2n}\left[\frac{v_i}{1+v_i} + \frac{1-e_i}{(1+V)^2}\right]$, and $\delta_i' = \frac{\epsilon}{2n}\left[\frac{t_i}{1+t_i} + \frac{1}{(1+V)^2}\right]$.

Accordingly,

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

[13] As usual, for any set A, $\Delta(A)$ is the set of probabilistic distributions over A.

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

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

Let (a, P) and (a', P') be the outcomes of (s_i, s'_{-i}) and (s'_i, s'_{-i}) respectively, and let $s'_j = (e'_j, v'_j)$ for each player $j \neq i$. Below we distinguish three cases.

Case 1. $a' \neq i$.

In this 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.

Case 2. $a' = i$ and $a = i$.

In this case, we have that: (i) $P'_i = \max_{j \neq i} v'_j - \delta'_i$; (ii) $P_i = \max_{j \neq i} v'_j - \delta_i$ if $e_i = 0$ and $P_i = v_i - \delta_i$ otherwise; and (iii) $v_i \geq \max_{j \neq i} v'_j$.

According to (ii) and (iii), $P_i \geq \max_{j \neq i} v'_j - \delta_i$. This fact, combined with (i) and the fact that $\delta'_i > \delta_i$, implies that $P_i > \max_{j \neq i} v'_j - \delta'_i = P'_i$, which in turn implies that $u_i(t_i, (s_i, s'_{-i})) = t_i - P_i < t_i - P'_i = u_i(t_i, (s'_i, s'_{-i}))$, as desired.

Case 3. $a' = i$ and $a \neq i$.

In this case, we have that: (i) $P'_i = \max_{j \neq i} v'_j - \delta'_i$; (ii) $P_i = -\delta_i$; and (iii) $t_i \geq \max_{j \neq i} v'_j$. Accordingly, $u_i(t_i, (s_i, s'_{-i})) = -P_i = \delta_i < \delta'_i \leq (t_i - \max_{j \neq i} v'_j) + \delta'_i = t_i - P'_i = u_i(t_i, (s'_i, s'_{-i}))$, 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 t_i-S-dominated by s'_i, contradicting the fact that $s_i \in S(t_i)$. Therefore Claim 1 holds. □

CLAIM 2. $\forall$ *player* i *and* $\forall$ *type* $t_i \in \{0, \ldots, V\}$ *of* i, *if* $s_i = (1, t_i)$ *then* $s_i \notin S(t_i)$.

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

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

The analysis below is very similar to that of Claim 1. Indeed, in the plays of (s_i, s'_{-i}) and (s'_i, s'_{-i}) respectively, we denote by δ_i and δ'_i the rewards that player i receives in Step **e**, 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 that

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

and we distinguish three cases as before:

- If $a' \neq i$, then $a \neq i$ as well, and we have that

$$\begin{aligned} u_i(t_i, (s_i, s'_{-i})) &= -P_i = \delta_i \\ &< \delta'_i = -P'_i = u_i(t_i, (s'_i, s'_{-i})). \end{aligned}$$

- If $a' = i$ and $a = i$, then $P_i = t_i - \delta_i \geq \max_{j \neq i} v'_j - \delta_i > \max_{j \neq i} v'_j - \delta'_i = P'_i$, and we have that

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

- Otherwise, we have that $a' = i$ and $a \neq i$, which implies that

$$\begin{aligned} u_i(t_i, (s_i, s'_{-i})) &= -P_i = \delta_i < \delta'_i \\ &\leq (t_i - \max_{j \neq i} v'_j) + \delta'_i = t_i - P'_i \\ &= u_i(t_i, (s'_i, s'_{-i})). \end{aligned}$$

In sum, s_i is strictly t_i-S-dominated by s'_i, and Claim 2 holds. □

CLAIM 3. $\forall$ *player* i, *if* s_i *is conservatively strictly rational in game* $(C, \mathcal{M})$, *then* $v_i \geq smp_i$.

PROOF OF CLAIM 3. Assume for sake of contradiction that s_i is conservatively strictly rational and $v_i < smp_i$. By definition we have that $s_i \in S(\theta_i)$, and thus by Claim 1 we have that $v_i \geq \theta_i$. Accordingly,

$$\theta_i < smp_i.$$

Let $s'_i = (1, smp_i)$. Without loss of generality, we assume that $s'_i \in S(\theta_i)$, and prove that s_i is conservatively strictly dominated by s'_i, which contradicts the fact that s_i is conservatively strictly rational.[14] To prove this, it suffices to show that $\forall t \in \mathscr{B}_i$, s_i is strictly θ_i-$S(t)$-dominated by s'_i. Arbitrarily fixing a type profile $t \in \mathscr{B}_i$ and a strategy subprofile $s'_{-i} \in S(t_{-i})$, it suffices to show that

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

To do so, letting δ_i and δ'_i be the rewards that player i receives in Step **e**, in the plays of (s_i, s'_{-i}) and (s'_i, s'_{-i}) respectively, we have that

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

[14] If $s'_i \notin S(\theta_i)$, then by well known properties of strict dominance we have that there exists a strategy $\sigma'_i \in \Delta(S(\theta_i))$ such that s'_i is strictly dominated by σ'_i in game $(C, \mathcal{M})$. By similar analysis, it can be proved that s_i is conservatively strictly dominated by σ'_i, which again contradicts the fact that s_i is conservatively strictly rational.

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

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

Let (a, P) and (a', P') be the final outcomes of (s_i, s'_{-i}) and (s'_i, s'_{-i}) respectively, let $s'_j = (e'_j, v'_j)$ for each player $j \neq i$, and let $\star(t) = \operatorname{argmax}_j t_j$ with ties broken lexicographically. Because $t \in \mathscr{B}_i$, by the definition of smp_i we have that $smp_i \leq \max_j t_j = t_{\star(t)}$, which together with the fact that $t_i = \theta_i < smp_i$ implies that $\star(t) \neq i$. Further because $s'_{\star(t)} \in S(t_{\star(t)})$, we have that

$$v'_{\star(t)} \geq t_{\star(t)} \geq smp_i > v_i,$$

where the first inequality holds by Claim 1 and the last one by our hypothesis about v_i.

If $v'_{\star(t)} > smp_i$, then by the construction of $\mathcal{M}$ we have that the pair announced by player $\star(t)$ is ordered before that by player i in both plays. If $v'_{\star(t)} = smp_i$, then we have that $v'_{\star(t)} = t_{\star(t)}$, and thus $e'_{\star(t)} = 0$ by Claim 2, which again implies that the pair announced by player $\star(t)$ is ordered before that by player i in both plays. Accordingly, no matter which is the case, we always have that

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

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 conservatively strictly rational then Inequality 1 holds, which implies Theorem 3. To do so, recall that $2^{nd}(\mathscr{B})$ is the second highest value in $\{smp_1, \ldots, smp_n\}$. Let $\star = \operatorname{argmax}_i smp_i$ and $\star' = \operatorname{argmax}_{i \neq \star} smp_i$, with ties broken lexicographically. By definition,

$$smp_{\star'} = 2^{nd}(\mathscr{B}).$$

Because s is conservatively strictly rational, by Claim 3 we have that for each i,

$$v_i \geq smp_i.$$

By the construction of $\mathcal{M}$ we have that for each reward δ_i in Step **e**,

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

Letting (a, P) be the outcome of s, we have that for each $i \neq a$,

$$P_i = -\delta_i.$$

If $a = \star$, then by the construction of $\mathcal{M}$ we have that

$$\begin{aligned} P_a &\geq \max_{j \neq a} v_j - \delta_a = \max_{j \neq \star} v_j - \delta_a \\ &\geq \max_{j \neq \star} smp_j - \delta_a = smp_{\star'} - \delta_a = 2^{nd}(\mathscr{B}) - \delta_a, \end{aligned}$$

where the first inequality holds because P_a equals either $\max_{j \neq a} v_j - \delta_a$ or $v_a - \delta_a$, and $v_a \geq \max_{j \neq a} v_j$, and the second inequality because $v_j \geq smp_j$ for each j.

If $a \neq \star$, then we have that

$$\begin{aligned} P_a &\geq \max_{j \neq a} v_j - \delta_a \geq v_\star - \delta_a \\ &\geq smp_\star - \delta_a \geq smp_{\star'} - \delta_a = 2^{nd}(\mathscr{B}) - \delta_a. \end{aligned}$$

Accordingly, whether or not a equals $\star$, we always have $P_a \geq 2^{nd}(\mathscr{B}) - \delta_a$, and thus

$$\begin{aligned} REV(\mathcal{M}(s)) &= P_a + \sum_{i \neq a} P_i \geq 2^{nd}(\mathscr{B}) - \delta_a - \sum_{i \neq a} \delta_i \\ &> 2^{nd}(\mathscr{B}) - n \cdot \epsilon/n = 2^{nd}(\mathscr{B}) - \epsilon. \end{aligned}$$

Therefore Theorem 3 holds. *Q.E.D.*

6. Three Concerns About the Second-Belief Mechanism “in Practice”

A concern raised about the second-belief mechanism is that “ϵ rewards” may not be enough motivation for the players to participate. When the relevant players opt to “stay at 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 the second-price mechanism. (Saying otherwise requires an alternative notion of rationality.[15]) Second, as we have already observed, conservative beliefs are implicit in any context, whether or not a designer 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 $\mathscr{B}$. Consider a player i who does not believe that his valuation is the highest. Then i concludes that he will receive “ϵ 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

[15] To be sure, such alternative notions exist: in particular, ϵ-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 ϵ-alternative. This is so because, when the revenue benchmark equals the highest valuation minus ϵ, by definition the sum of the players' utilities must be at most ϵ. Therefore any ϵ-alternative notion of rationality will make the players indifferent between participating and opting out. And when players opt out, the mechanism cannot guaranteed its desired benchmark.

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 the second-price mechanism, he should also participate the second-belief one.

Another (related) concern was raised for the case in which the players only have very unprecise external beliefs. In this case, while the revenue generated by the second-price mechanism is equal to the second-highest valuation, $2^{nd}(\theta)$, the one generated by the second-belief mechanism is "$2^{nd}(\theta) - \epsilon$." 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). By contrast, the second-price mechanism only guarantees that the players' utilities are ≥ 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 ϵ 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 "ϵ rewards" as well, and thus will miss its target revenue in its purest form.

A 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 $\theta_i \geq 2^{nd}(\mathscr{B})$, and (4) i further believes that $2^{nd}(\mathscr{B}) > 2^{nd}(\theta)$, 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.

Acknowledgment

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, and to Elchanan Ben-Porath, Sergiu Hart, and Philip Reny for helping us clarify our connections to ex-post equilibria. (Any remaining lack of clarity is solely due to us!)

References

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

[2] 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.

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

[4] 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.

[5] J. Chen and S. Micali. Mechanism Design with Set-Theoretic Beliefs. Full version, available at
http://people.csail.mit.edu/silvio/Selected Scientific Papers/ Mechanism Design.

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

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

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

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

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

[11] 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.

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

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

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

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

[16] 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.

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