Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2010-060 December 20, 2010

Conservative Rationalizability and The Second-Knowledge Mechanism

Jing Chen and Silvio Micali

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Conservative Rationalizability and The Second-Knowledge Mechanism*

Jing Chen
CSAIL, MIT
Cambridge, MA 02139, USA
jingchen@csail.mit.edu

Silvio Micali
CSAIL, MIT
Cambridge, MA 02139, USA
silvio@csail.mit.edu

December 20, 2010

Abstract

In mechanism design, the traditional way of modeling the players' incomplete information about their opponents is "assuming a Bayesian." This assumption, however, is very strong and does not hold in many real applications. Accordingly, we put forward

(1) a set-theoretic way to model the knowledge that a player might have about his opponents, and

(2) a new class of mechanisms capable of leveraging such more conservative knowledge in a robust way.

In auctions of a single good, we show that such a new mechanism can perfectly guarantee a revenue benchmark (always lying in between the second highest and the highest valuation) that no classical mechanism can even approximate in any robust way.

*This work is funded by the Office of Naval Research under award number N00014091059. (Any opinions, findings, and conclusions or recommendations expressed in this email are those of the author (s) and do not necessarily reflect the views of the Office of Naval Research.)

1 Introduction

As introduced by Hurwicz [17], mechanism design aims at leveraging the players' information and rationality so as to produce outcomes satisfying a desired property that depends on the players' *types.* Traditional mechanism design implements this goal "at equilibrium", that is, by engineering a game whose equilibrium outcomes satisfy the desired property. This approach has proved very successful when the players have *complete information* about each other's true types [19, 20, 1, 14]. But things are more complex when the players have *incomplete information* about the true types of their opponents, which is indeed the case in most applications. In such settings, without knowing exactly everyone's utility for every possible outcome, a player may not even be able to verify whether a given strategy profile σ is an equilibrium. Let us review prior ways to bypass this specific difficulty.

Prior Approaches to Mechanism Design in Settings of Incomplete Information

1. *Mechanism Design in Dominant Strategies.* This approach does not model the "external knowledge" of the players, that is, the knowledge that each player may have about his opponents. Instead, it considers only mechanisms where each player has a "best strategy" for him to choose no matter what his opponents might do. The applicability of this approach is limited by the fact that the existence of such mechanisms is far from being guaranteed. (In particular, they do not exist for many forms of elections [13, 23] or for maximizing revenue in general settings of quasi-linear utilities [8].)
2. *Implementation in Undominated Strategies.* This approach does not model the players' external knowledge either. Instead, it considers only mechanisms guaranteeing their desired properties as long as no player chooses for himself a weakly dominated strategy. The applicability of this approach is limited too because, as shown by Jackson [18], such mechanisms must be *bounded* in order to be meaningful, and the set of properties implementable by bounded mechanisms is quite constrained.
3. *Bayesian Mechanism Design.* At the core of Bayesian mechanism design is the assumption that the players types have been generated according to a joint distribution $\mathcal{D}$, the *prior.* Different assumptions are then made to specify who knows what about $\mathcal{D}$ and/or to restrict $\mathcal{D}$ —for example, see [2, 11, 16, 21, 24]. These assumptions, however, may not hold in many practical applications. Particularly for games played only once, the very ones considered in this paper. (For this reasons, prior-free mechanisms have been developed —see in particular [2, 24, 15]— but they use dominant-strategy equilibria as their solution concept.)

As these approaches are not always applicable in settings of incomplete information, there is room to explore alternative ones. Putting forward and exemplifying such an alternative approach is the goal of this paper.

Our Approach Our approach has the following three components, of possibly independent interest.

1. *A Set-Theoretic Knowledge Model.* Differently from implementation in undominated strategies, we aim at leveraging the knowledge that the players may have about their opponents. But differently from Bayesian mechanism design, we do not rely on distributions to model the uncertainty that each player has about his opponents' types. Our knowledge model is totally *set theoretic*, and relies on *weaker* and thus "safer" assumptions than the Bayesian one.
2. *Conservative Rationalizability.* We put forward the notion of *conservative rationalizability,* so as to capture which strategies are safe for rational players to use in settings of *incomplete information,* conservatively modeling their knowledge about their opponents in our set-theoretic way. We note that Pearce [22] and Bernheim [7] have independently proposed the original notion of rationalizability for normal-form games, and Pearce explicitly also for extensive-form games. These notions, however, apply only to settings of *complete information* where the players' types are common knowledge among the players. We thus develop our notion by properly generalizing theirs.

3. *Safe Mechanisms.* In settings of incomplete information, we consider and design only mechanisms that are *safe*, that is, guaranteeing their desired properties not at equilibrium, but at each profile of conservatively rationalizable strategies. Thus, differently from mechanisms working at equilibrium, safe mechanisms are invulnerable to any problem of belief mismatch (see Footnote 4). And differently from mechanisms working with undominated strategies, safe mechanisms do leverage not only the players' rationality, but also their possible knowledge about their competitors.

For completeness, we model also the *"knowledge about knowledge"* that the players might have. As we shall see, however, the mechanisms constructed in this paper renounce to leverage the players' knowledge about knowledge (or their beliefs). On the contrary, to be really on the "safe side", our mechanisms work no matter what (beliefs and) knowledge about knowledge the players may have!

Our Results We demonstrate the power of our new approach by showing that

> **Theorem 1 (Informal Statement):** There exists a safe mechanism that, in any auction of a single good, guarantees a revenue benchmark, the *second-knowledge benchmark*, always lying between the highest and second highest true valuation of the item for sale.

Thus our mechanism returns at least as much revenue as the second-price one, and possibly much higher. That is, the revenue performance of our mechanism grows with the conservative knowledge the players have about their opponents, but gracefully degrades to the performance of the second-price mechanism when the players' knowledge about their opponents is zero. In other words: *in an auction of a single good, there is nothing to lose, but possibly something to gain, by running our mechanism instead of the second-price one.*

Beyond economics, mechanism design has enormous potential applications in networks, computer science, and engineering in general. It is a subtle and difficult field, and thus new approaches to it should be explored. The usefulness of ours is actually emphasized by the *simplicity* of our mechanism and yet by the *impossibility* for prior approaches to *guarantee* the same performance. Namely, we prove the following results.

> **Theorem 2 (Informal Statement):** Not even a minuscule fraction of the second-knowledge revenue benchmark can be guaranteed "at equilibrium" —except than in an extremely fragile sense; and

> **Theorem 3 (Informal Statement):** Not even a minuscule fraction of the second-knowledge revenue benchmark can be guaranteed in undominated strategies.

2 Preliminaries

We find it useful to decompose a game G into two components: (1) a *context* C, describing the possible outcomes and the players, including their utilities for each outcome, their information about themselves and the others, and their beliefs; and (2) a *mechanism* M, describing the strategies available to the players and how each profile of strategies yields an outcome. That is, $G = (C, M)$.

Our notions apply to games of incomplete information whose contexts are pretty standard —except for the information available to the players, which is described in the next section—, and whose mechanisms are very general — namely, finite extensive-form mechanisms of perfect recall. Our desired revenue benchmark is however achieved by much simpler mechanisms: namely, (finite) mechanisms of observable actions [12], OA mechanisms for short. In such mechanisms players can act simultaneously at each decision node, and all actions become immediately public after being played. We thus describe our notions and results for games with OA mechanisms, so as to avoid the unnecessary complexity of dealing with information sets.

Context Notation To describe the classical portion of a context C, we use the following notation.

- $\mathcal{N} = \{1, \ldots, n\}$ is the set of players.
- Ω is the set of outcomes.

- $\mathcal{T}_i$ is the set of all possible (payoff) types for each player i, and $\mathcal{T} = (\mathcal{T}_1, \ldots, \mathcal{T}_n)$.
- u_i is the utility function for each player i, mapping from $\mathcal{T}_i \times \Omega$ to $\mathbb{R}$, and $u = (u_1, \ldots, u_n)$.
- θ is the the profile of true types.

Mechanism Notation To describe a mechanism M, we use the following notation.

- T is the underlying game tree. The height of a node is taken to be the number of edges in the longest path from the node to a leaf (thus a leaf has height 0), and the height of T is the height of its root.
- D is a generic internal node (decision node) of T, and P^D is the subset of players acting at D.
- A_i^D is the set of actions available at D to each player $i \in P^D$.
- Σ_i is the set of all pure strategies of each player i, and $\Sigma = (\Sigma_1, \ldots, \Sigma_n)$. A strategy (or strategy profile) is always pure if represented by a lowercase Latin letter, while can be either pure or mixed if represented by a lowercase Greek letter.
- For each player $i \in P^D$, $\sigma_i(D)$ is the distribution over A_i^D induced by a strategy σ_i.
- Given a strategy profile s, $M(s)$ is the outcome associated with the terminal node obtained by playing M where each player i uses s_i.
- Given a strategy profile σ, $M(\sigma)$ is the distribution of outcomes $\{M(s) : s \leftarrow \sigma\}$, and $\mathcal{U}_i(t_i, M(\sigma))$ denotes $\mathbb{E}_{s \leftarrow \sigma}[u_i(t_i, M(s))]$ for each player i and each $t_i \in \mathcal{T}_i$.
- $u_i(M(s))$ denotes $u_i(\theta_i, M(s))$, and $\mathcal{U}_i(M(\sigma))$ denotes $\mathcal{U}_i(\theta_i, M(\sigma))$, for simplicity only.

Histories In a game with an observable-action mechanism M, the history of a strategy profile σ, denoted by $H(\sigma)$, consists of the distribution over the sequences of nodes in M's game tree reached in a play of M under σ. For any decision node X, we say X belongs to the history of σ, written as $X \in H(\sigma)$, if the play of M under σ reaches X with positive probability. If S is a set of pure strategy profiles, then $H(S) = \{H(s) : s \in S\}$.

Profiles of Strategy Sets By saying that S is a *profile of strategy sets,* we mean that, for each player i, $S_i \subseteq \Sigma_i$. The *Cartesian closure* of such a profile S, denoted by $\overline{S}$, is defined to be $S_1 \times \cdots \times S_n$. Similarly, for each player i we define $\overline{S_{-i}}$ to be the Cartesian product $\prod_{j \neq i} S_j$.

3 Our Model of Incomplete Information and Conservative Knowledge

Our model of incomplete information is set-theoretic and very simple. Striving for a uniform notation, it consists of a *knowledge sequence* K: a sequence of profiles $K^0, K^1, K^2, \ldots$, where for each player i

- K_i^0 represents i's knowledge about his own true (payoff) type, that is, $K_i^0 = \theta_i$.
- K_i^1 is a set of profiles, representing all possible candidates for K^0 known to i, satisfying the following two conditions:
  1. *Consistency:* $T_i = K_i^0$ for all $T \in K_i^1$; and
  2. *Genuineness:* $K^0 \in K_i^1$.
- K_i^2 is a set of profiles, representing all possible candidates for K^1 known to i, satisfying the following two properties:
  1. *Consistency:* $T_i = K_i^1$ for all $T \in K_i^2$; and
  2. *Genuineness:* $K^1 \in K_i^2$.

And so on.

We refer to each K^ℓ as the players' *knowledge of level* ℓ. We sometimes refer to K^0 as the players' *internal knowledge*, to K^1 as the players' *conservative knowledge*[1], and to the sequence $K^2, K^3, \ldots$ as the players' *higher-level* knowledge. We define $K_i = K_i^0, K_i^1, \ldots$, and refer to it as player i's *total knowledge.*

Discussion In this paper we design mechanisms leveraging only the players' conservative knowledge.[2] However, to ensure being on the safe side, we demand that our mechanisms work no matter what the players' higher-level knowledge might be. Let us thus make a few remarks about conservative knowledge.

- *Conservative Knowledge is Guaranteed Knowledge.* Each K_i^1 models completely and genuinely the *guaranteed* (i.e., non-Bayesian and even non-probabilistic) knowledge of i about the actual profile θ of true types. Completely, because it coincides with all possible candidates for θ in i's mind. Genuinely, because it always uniquely determines θ_i and always "includes" θ_{-i}.
- *Conservative Knowledge is Weaker Than Bayesian Knowledge.* If a player i really knows the distribution D_i from which θ_{-i} has been drawn, then K_i^1 is the set of all profiles (θ_i, t_{-i}), where t_{-i} belongs to D_i's support.
- *Conservative Knowledge is Totally Separate from the Designer's Knowledge.* In traditional Bayesian mechanism design, the designer is assumed to know the distribution D from which the players' types have been generated. In our set-theoretic knowledge model, we instead allow for a total asymmetry between the knowledge of the players and that of the designer. In particular, the mechanism proposed in this paper assumes that all knowledge resides with the players: the designer is totally ignorant. (Of course, sometimes better results can be proven, assuming that the designer too has some knowledge.)
- *Conservative Knowledge is by Itself Purely Individual.* No information about θ is assumed to be common knowledge, even among the players only. Saying that the conservative knowledge of a player i is K_i^1 solely implies that i knows that $\theta \in K_i^1$. If another player j has knowledge about K_i^1, this is separately specified in j's higher-level knowledge, and cannot be inferred either from i's or j's conservative knowledge, or from the entire conservative knowledge profile K^1.
- *Conservative Knowledge Always Exists.* The conservative knowledge model does not assume any lower- or upper-bound on the amount of information that a player i may have about θ_{-i}. When K_i^1 consists of all possible type profiles, that is, when $K_i^1 = \{\theta_i\} \times \overline{\mathcal{T}_{-i}}$, i has no knowledge about θ_{-i}. When $K_i^1 = \{\theta\}$, player i is perfectly informed about his opponents. (Strictly speaking therefore, conservative knowledge is not an assumption at all.)
- *Conservative Knowledge Can Be Arbitrarily Augmented By The Players.* Differently from the Bayesian setting, each player i is free to acquire additional knowledge about his opponents without endangering our mechanisms. We do not rely in any way —nor attempt to define— on the "original" knowledge of player i. Player i is welcome to refine his candidate set for θ up to the moment when the mechanism is played: K_i^1 is by definition taken to consist of i's candidate set when our mechanisms start.
- *(Conservative) Knowledge as a Richer Type.* At its most general form, each player i's type is a comprehensive description of i in the strategic situation at hand. Thus, in our treatment, we are essentially separating i's *payoff* type, θ_i, from his *conservative knowledge* type, K_i^1, and more generally his *knowledge* type, K_i.
- *Compatibility of Knowledge and Beliefs about Payoff Types.* Our set-theoretic knowledge model is compatible with the players additionally having (probabilistic knowledge or) beliefs, beliefs about beliefs, etc., about each other's payoff types. Such beliefs can be arbitrary —in particular, of a probabilistic

[1] In fact, higher-level knowledge consists of "knowledge about knowledge", and we believe that in most practical setting such knowledge can be safely equated to zero.

[2] We in fact believe that leveraging higher-level knowledge, although possible and theoretical interesting, might not be of practical relevance. And while our aims are primarily theoretical, potential applicability is a crucial component of our motivation in this research.

nature, or even "false"— but cannot "contradict K_i"[3]. Accordingly, such beliefs do not alter the set of rationalizable strategies, and are omitted for simplicity from our contexts.

Our Contexts Clarified In accordance with our model, a context C of incomplete information consists of a tuple $(\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$, where $\mathcal{N}$, Ω, $\mathcal{T}$, u, and θ are as usual, and K is a knowledge sequence such that $\theta = K^0$. In any such context: (1) the first 4 components and the independence and rationality of the players are common knowledge to everyone, including the mechanism designer; (2) each player i has no additional information beyond θ_i and K_i; and (3) the designer, unless otherwise specified, has no information about θ or K.

4 Conservative Rationalizability and Safe Mechanism Design

As already said, in settings of incomplete information, we put forward mechanisms that leverage the knowledge the players may have about their opponents' types. Further, our mechanisms achieve their desiderata not *at equilibrium* but at all profiles of *rationalizable* strategies. Accordingly, our mechanisms solely rely on the players' rationality and thus are immune to any *belief mismatch* problem.[4]

To be general, our approach requires a general notion of rationalizability in settings of incomplete information. Rationalizability, however, has been studied only in settings of *complete information.* (See Pearce [22], Bernheim [7], Battigalli [3], Battigalli and Siniscalchi [4, 5, 6], Shimoji and Watson [25], and the authors [9].[5]) Accordingly, we first recall this classical notion —but in a new and more convenient form— and then extend it to settings of *incomplete information.*

4.1 Distinguishable Dominance and Complete-Information Rationalizability

Rationalizable strategies are defined constructively, via a greedy iterative procedure, as follows:

(a) Define a suitable notion of dominance among the strategies of any given player; and

(b) At each round, for each player i, eliminate all dominated strategies of i.

The rationalizable strategies are then those still standing when no strategy can be further eliminated.

All definitions of rationalizability proposed in the literature follow (implicitly or explicitly) the above construction. All of those for extensive-form games are provably equivalent (at least for games of perfect recall), and they differ only in the conceptual Step (a), that is, in their choice of the underlying notion of dominance. The original one of Pearce, referred to as Pearce-rationalizability below, was quite complex, and the more recent one of Shimoji and Watson significantly simpler. Here we shall use our own notion of dominance, the most recent one and —in our opinion— the simplest one. Essentially, a pure strategy a of player i is distinguishably dominated (DD) by another, possibly mixed, strategy b of i if

1. for some strategy subprofile σ_{-i} the (distributions of) terminal nodes reached by (a, σ_{-i}) and (b, σ_{-i}) do not coincide and,
2. for all such σ_{-i}, i's (expected) utility is smaller with a than with b.

We restate this more formally below, collecting some useful language and notation along the way.

[3]For instance, if i believes that the type subprofile of his opponents has been drawn from some distribution D, K_i^1 should coincide with θ_i together with D's support.

[4]When there are multiple equilibria, even if a mechanism ensures that its desired property $\mathcal{P}$ holds at each one of them, it is quite possible that no equilibrium will be reached, and thus that $\mathcal{P}$ is far from being guaranteed in a rational play. For instance, consider a normal-form mechanism with just two players and two equilibria, σ and τ. Then, the first player, believing that σ will arise, may play his strategy σ_1, while the second one, believing that τ will prevail, will play τ_2. Accordingly, independent of the players' rationality, the final play will be (σ_1, τ_2), which may not be an equilibrium at all. This is a serious problem for mechanism design at equilibrium, and grows dramatically with the number of players and the number of equilibria.

[5]Bernheim focuses on normal-form games, Pearce also on extensive-form games, and all others on extensive-form games.

Definition 1. (Distinguishability and Indistinguishability [9]) *In a game G, let S be a profile of strategy sets and let σ_i and σ_i' be two different strategies of the same player i. Then we say that σ_i and σ_i' are* distinguishable *over S if $\exists \tau_{-i} \in \overline{S_{-i}}$ such that*

$$H(\sigma_i \sqcup \tau_{-i}) \neq H(\sigma_i' \sqcup \tau_{-i}).$$

If this is the case, we say that τ_{-i} distinguishes *σ_i and σ_i' over S. Else, we say that σ_i and σ_i' are* indistinguishable *over S, and write "$\sigma_i \simeq \sigma_i'$ over S" or "$\sigma_i \simeq_S \sigma_i'$" to express this fact more concisely.*

Definition 2. (Distinguishable Dominance [9]) *Let $G = (C, M)$ be a game, i a player, σ_i a pure strategy of i, σ_i' a strategy of i, and S a profile of strategy sets. We say that σ_i is* distinguishably dominated *(DD for short) by σ_i' over S if*

1. *σ_i and σ_i' are distinguishable over S; and*
2. *$\mathcal{U}_i(M(\sigma_i \sqcup \tau_{-i})) < \mathcal{U}_i(M(\sigma_i' \sqcup \tau_{-i}))$ for all sub-profiles τ_{-i} distinguishing σ_i and σ_i' over S.*

We write "$\sigma_i \prec \sigma_i'$ over S" or "$\sigma_i \prec_S \sigma_i'$" to express this fact more concisely. We further write "$\sigma_i \preceq \sigma_i'$ over S" or "$\sigma_i \preceq_S \sigma_i'$" to express the fact that either $\sigma_i \simeq_S \sigma_i'$ or $\sigma_i \prec_S \sigma_i'$.

Remark. Distinguishable dominance is weaker than strict dominance which requires that $\mathcal{U}_i(M(\sigma_i \sqcup \tau_{-i})) < \mathcal{U}_i(M(\sigma_i' \sqcup \tau_{-i}))$ for all $\tau_{-i} \in \overline{S_{-i}}$, and is stronger than weak dominance which is recalled in Appendix C.

Definition 3. (Greedy Iterative Elimination of DD Strategies [9]) *We say that a profile of strategy sets S survives the* greedy *iterative elimination of DD strategies if there exists a sequence of profiles of strategy sets $S^0, \ldots, S^K = S$ such that*

1. *$S^0 = \Sigma$;*
2. *$\forall k < K$ and $\forall i$, $S_i^k \setminus S_i^{k+1}$ consists of all strategies of i distinguishably dominated within S^k; and*
3. *Each S_i^K contains no strategy distinguishably dominated within S^K.*

Theorem 0 ([9]) *Let S be the profile of strategy sets surviving the greedy iterative elimination of DD strategies. Then each S_i coincides with i's set of Pearce-rationalizable strategies.*

4.2 Conservative Rationalizability

Let us now define how the players of an extensive-form game (C, M) of incomplete information can refine their strategy sets based only on their conservative knowledge.

Let $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$, and consider a player i. Since the OA mechanism M is common knowledge, i knows Σ and thus the initial strategies of every player. Notice, however, that i does not exactly know the context C. He knows $\mathcal{N}$, Ω, $\mathcal{T}$, and u, because they too are common knowledge, but only knows his own component of the remaining quantities: that is, θ_i and K_i.

As we wish to focus on leveraging just the conservative knowledge, K^1, letting $K_i^1 = \{\theta', \theta'', \theta''', \ldots\}$, we explain how i refines his original strategy set Σ_i to get his conservatively rationalizable strategy set, Σ_i^2.

Player i starts by considering each candidate type profile in K_i^1 together with the original profile of strategy sets:

$$\theta', (\Sigma_1, \ldots, \Sigma_i, \ldots, \Sigma_n) \qquad \theta'', (\Sigma_1, \ldots, \Sigma_i, \ldots, \Sigma_n) \qquad \theta''', (\Sigma_1, \ldots, \Sigma_i, \ldots, \Sigma_n) \quad \ldots$$

Then for each pair, he eliminates all DD strategies for each player with respect to the corresponding candidate type profile, to obtain the new pairs

$$\theta', (\Sigma_1', \ldots, \Sigma_i^1, \ldots, \Sigma_n') \qquad \theta'', (\Sigma_1'', \ldots, \Sigma_i^1, \ldots, \Sigma_n'') \qquad \theta''', (\Sigma_1''', \ldots, \Sigma_i^1, \ldots, \Sigma_n''') \quad \ldots$$

(Focusing on the first new pair, notice that the surviving strategy sets of i's opponents appear "primed" to emphasize that they are based on a conjecture: namely the candidate type profile θ'. By contrast, the strategy set of i does not appear "primed," because it has been calculated not based on a conjecture, but on the truth. This is so because the ith component of each candidate type profile coincides with i's true type θ_i. Moreover, in each new pair, i's surviving strategy set is the same: Σ_i^1. This is so because eliminating all DD strategies of i yields the same surviving strategies for i whenever i's type and the initial strategy sets of all players are the same —here, θ_i and Σ.)

At this point, i considers all pure-mixed pairs (s_i, τ_i) of his Σ_i^1 strategies and
IF τ_i distinguishably dominates s_i in all cases, that is, if

$$s_i \prec_{(\Sigma_1',\ldots,\Sigma_i^1,\ldots,\Sigma_n')} \tau_i \qquad s_i \prec_{(\Sigma_1'',\ldots,\Sigma_i^1,\ldots,\Sigma_n'')} \tau_i \qquad s_i \prec_{(\Sigma_1''',\ldots,\Sigma_i^1,\ldots,\Sigma_n''')} \tau_i \qquad \ldots$$

THEN i eliminates s_i from Σ_i^1. The surviving strategies are *i's conservative rationalizable strategies,* and are denoted by Σ_i^2.

Discussion

- Notice that, for example, $s_i \prec_{(\Sigma_1',\ldots,\Sigma_i^1,\ldots,\Sigma_n')} \tau_i$ does not necessarily imply that $s_i \prec_{(\Sigma_1'',\ldots,\Sigma_i^1,\ldots,\Sigma_n'')} \tau_i$. This is so because distinguishable dominance requires comparing s_i and τ_i against all possible strategy subprofiles of i's opponents, which in the first case come from $\overline{\Sigma'_{-i}}$, and in the second case from $\overline{\Sigma''_{-i}}$.
- For s_i to be eliminated from Σ_i^1, it does not suffice for it to be distinguishably dominated in each case, but by different strategies in different cases. That is, it does not suffice that
  $s_i \prec_{(\Sigma_1',\ldots,\Sigma_i^1,\ldots,\Sigma_n')} \tau_i' \quad s_i \prec_{(\Sigma_1'',\ldots,\Sigma_i^1,\ldots,\Sigma_n'')} \tau_i'' \quad s_i \prec_{(\Sigma_1''',\ldots,\Sigma_i^1,\ldots,\Sigma_n''')} \tau_i''' \quad \ldots$
  The reason is the following. When $x \prec y$ we can safely eliminate x because "we can always use y instead of x whenever we feel like using x." (And if we then eliminate y because $y \prec z$, we can always use z instead of x.) But since i does not know whether the true type profile out there is —say— θ' or θ'', he would not know whether he should use τ_i' or τ_i'' instead of s_i. Only when s_i is dominated by the *same* τ_i is i sure that, no matter what the true type profile is, it is always safe for him to use τ_i instead of s_i.
- To compute i's conservatively rationalizable strategies, we do not iterate the above refinement process at all. In particular, we do not require that player i compute Σ_j^2 for any other player j and then "Σ_i^3" based them. This is so because, to iterate our process in any way, we would have to rely also on i's *higher knowledge.* And in this paper we renounce to do so.

4.3 Safe Implementation

Having defined rationalizable strategies, it is trivial to define safe mechanisms relative to them.

Definition 4. (Safe Implementation.) *Let $\mathcal{P}$ be a property over (distributions of) outcomes of incomplete-information contexts belonging to a class $\mathscr{C}$. We say that a mechanism M* safely implements $\mathcal{P}$ over $\mathscr{C}$ *if, for all contexts $C \in \mathscr{C}$ and all profiles σ of conservatively rationalizable strategies of the game (C, M), $\mathcal{P}$ holds for $M(\sigma)$.*

Since conservative knowledge is level-1 knowledge, above we actually define "level-1 safe" mechanisms. By using the players' higher-level knowledge, we could extend the above definition to "higher-level safe" mechanisms. Although *we* choose not to leverage higher-level knowledge, *the players* may actually choose to use their higher-level knowledge to further refine their strategies. But then a safe mechanism will continue to work: any profile of so refined strategies continues to be a profile of conservatively rationalizable strategies!

5 The Second-Knowledge Benchmark

Single-Good Auctions The context $(\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$ of an auction of a single good g is so specified:

- Each outcome $\omega \in \Omega$ is a pair (a, P), where a is an *allocation*, in our case an integer in $\{0, 1, \ldots, n\}$, and P is a *profile of prices*, that is, a profile of real numbers. If $a = 0$ then g remains unallocated (i.e., unsold), else a is the player getting the good. If P_i is positive, then it is the price paid by player i, else $|P_i|$ is the payment received by i.
- $\mathcal{T}_i = \{0, 1, \ldots, T\}$ for each player i. Player i having type t_i means that he values t_i the good for sale.
- For each $t_i \in \mathcal{T}_i$ and each outcome $\omega = (a, P)$, $u_i(t_i, \omega) = t_i - P_i$ if $i = a$, and $-P_i$ otherwise.

The components θ and K are defined as usual from the above three.

For outcome $\omega = (a, P)$, the *revenue* of ω, $REV(\omega)$, is $\sum_i P_i$, and the *social welfare* of ω, $SW(\omega)$, is θ_a (or 0 if $a = 0$).

The Classical Second-Valuation Benchmark In single-good auctions, the famous second-price mechanism is (weakly) dominant-strategy truthful, DST for short, and guarantees revenue equal to the second-highest valuation of g. This is a very significant revenue benchmark, in light of the well known fact that no DST mechanism can guarantee revenue greater than any constant fraction of the highest valuation. In many auctions, however, the players may have significant knowledge about the true valuations of their competitors, and thus it would be nice to transform this additional knowledge into additional revenue.

The New Second-Knowledge Benchmark In a single-good auction context $(\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$, for any $i, j \in \mathcal{N}$, we define $K_i^1(j) = \min_{k \in K_i^1} k_j$ and call it the *conservative knowledge of player i about j*. That is, $K_i^1(j)$ is the largest lowerbound for θ_j known to i. Notice that $K_i^1(i) = \theta_i$.

Let us now define the *maximum known welfare* function, MKW. For any subset S of the players, we define $MKW(K_S^1)$ to be the maximum social welfare that one could guarantee, by properly allocating the good, if he had the whole conservative knowledge of the players in S, that is:

$$MKW(K_S^1) = \max_{i \in S} \max_{j \in \mathcal{N}} K_i^1(j).$$

Our benchmark then is

$$MKW(K_{-\star}^1)$$

where $\star$ denotes the player with the highest true valuation for the good, and is referred to as the *star player*. That is, our benchmark coincides with the maximum social welfare that one could guarantee given the conservative knowledge of all players, except the star player.

We refer to $MKW(K_{-\star}^1)$ as the *second-knowledge* benchmark because it coincides with the "second highest piece of conservative knowledge", that is, the second-highest value in $\{K_i^1(j) : i, j \in \mathcal{N}\}$. In fact, the highest value is $\theta_\star$, because $K_\star^1(\star) = \theta_\star$, and, by the genuineness condition, each piece of conservative knowledge is less than or equal to $\theta_\star$.

Note that our benchmark coincides with the larger one between the maximum knowledge that the other players have about the star player, and the second-highest valuation. That is,

$$MKW(K_{-\star}^1) = \max \left\{ \max_{i \neq \star} K_i^1(\star), \max_{i \neq \star} \theta_i \right\}.$$

6 Our Results

6.1 Second-Knowledge Revenue Can Be Guaranteed by Safe Mechanisms

We prove that it is always possible to safely implement the second-knowledge revenue benchmark. More formally, defining "always" as "for all single-good auction contexts", and "implementing second-knowledge revenue within ϵ" as "implementing the property $\mathcal{P}_\epsilon = \{\omega : \mathbb{E}[REV(\omega)] \geq MKW(K_{-\star}^1) - \epsilon\}$" we have

Theorem 1. *$\forall \epsilon \in (0,1)$ $\exists$ a mechanism $\mathcal{M}_\epsilon$ always safely implementing second-knowledge revenue within ϵ.*

The proof of Theorem 1 is provided in Appendix A. As an extra bonus, each $\mathcal{M}_\epsilon$ is (1) constructable in a uniform fashion on input ϵ; (2) very simple; and (3) observable-action, and thus implementable as usual, that is, without any trusted mediators.

6.2 Second-Knowledge Revenue Cannot Be Guaranteed at Equilibrium

We show that, due to a serious and *intrinsic* problem of "equilibrium selection", no mechanism can guarantee at equilibrium even an arbitrarily small fraction of the second-knowledge revenue benchmark. In essence we prove that, for all $\epsilon > \epsilon' > 0$, if —despite the mentioned difficulty of verifying equilibria in incomplete-information settings— each player i can find a strategy E_i such that $E = (E_1, \ldots, E_n)$ is an equilibrium of M, and if at E the revenue collected is at least a fraction ϵ of our benchmark, then at least one player j can find a new strategy E'_j such that, (a) $E' = E'_j \sqcup E_{-j}$ is also an equilibrium of M; and (b) the revenue collected in E' is less than a fraction ϵ' of our benchmark. Since the utility of player j is guaranteed to be the same under E and E', if he believes that his opponents will play the strategy subprofile E_{-j}, choosing E_j rather than E'_j is not a question of rationality, but of "kindness" to the mechanism designer.

Notice that our result does not rule out the existence of a mechanism M achieving the second-knowledge revenue benchmark in its unique equilibrium or all equilibria. However, our result implies that if such a mechanism M exists, then for at least one context and one player i, the equilibrium strategies of i in this context depend not only on i's own true type, but also on the other players' true types, and in particular on knowledge that player i does not have.[6] Therefore such strategies are *incomputable* by i, never mind the computational complexity of finding them. Accordingly, if the only "robustness" that the designer can provide is that the second-knowledge benchmark is achieved at all equilibria, then he must be aware that he is far from achieving the benchmark in reality, since it is hard for him to see an equilibrium played out. Indeed, if one cares about mechanism robustness, our result proves that

There are inherent limitations to what is achievable at equilibrium in settings of incomplete information. Let us now proceed a bit more formally.

Inability of imposing arbitrary prices Generating high revenue is trivial if a mechanism can force the players to accept negative utilities. Thus, to be meaningful, revenue impossibility results apply to mechanisms M satisfying the following

Opt-Out Condition: *Each player i has a strategy out_i such that for any strategy subprofile σ_{-i},*

$$\mathcal{U}_i(M(out_i \sqcup \sigma_{-i})) = 0.$$

Equilibrium-Based Mechanisms for Contexts of Incomplete Information In a contexts of complete information, knowing the true types —and thus the utilities— of everyone, each player can verify whether a profile of strategies σ is an equilibrium. Thus, by cycling through all possible strategy profiles, the players can in principle compute all possible equilibria. But in a general context of *incomplete* information, equilibria are unverifiable, and thus the players may be *unable* —even in principle!— to compute any equilibrium, let alone choosing the same one. Therefore, we let the designer of an equilibrium-based mechanism provide —if he can— the players with a way to compute the desired equilibrium strategies from their own types. More precisely, define a *strategy finder* to be a function $F(\cdot,\cdot,\cdot)$ mapping each player i, type t_i of i, and knowledge K_i of i, to a strategy $F(i, t_i, K_i)$ for i. Then,

[6] As an example, for any $\epsilon \in (0,1)$, consider the mechanism $\mathcal{M}_\epsilon$ whose construction is provided in Appendix A. Consider the context where there are two players, $\theta_1 = T-1$ and $\theta_2 = 1$, K_1^1 consists of all type profiles t such that $t_1 = T-1$ and $t_2 \geq 0$, K_2^1 consists of all type profiles t such that $t_2 = 1$ and $t_1 \geq 10$, all higher-level knowledge is empty. In such a context, for any Nash equilibrium σ and any pure strategy s_2 in the support of σ_2, s_2 consists of player 2 announcing his knowledge about player 1's true type to be $\geq T-1$. However, player 2 knowing that such a strategy σ_2 is an equilibrium strategy implies that he knows that player 1's true type is at least $T-1$, which is certainly beyond player 2's knowledge K_2^1.

Definition 5. *We say that a mechanism M* implements at equilibrium *a property $\mathcal{P}$ for a class of contexts $\mathscr{C}$ if there exists a strategy finder F such that, $\forall C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K) \in \mathscr{C}$*

1. *$\sigma = (F(1, \theta_1, K_1), \ldots, F(n, \theta_n, K_n))$ is a Nash equilibrium in (C, M); and*
2. *$\mathcal{P}$ holds for $M(\sigma)$.*

If this is the case we say that F helps *M.*

We say that M deterministically implements at equilibrium *$\mathcal{P}$ if the strategy finder F is deterministic.*

As the inputs θ_i and K_i to a strategy finder F can be considered as i's "extended true type," such an F is in line with the revelation principle [21].

Total vulnerability to equilibrium selection Consider a mechanism M trying to achieve a property $\mathcal{P}$ at equilibrium. Without specifying a strategy finder, it would be problematic for a play of M to end up at equilibrium. Specifying a strategy finder F certainly improves the chance of ending up at equilibrium. However, some mechanisms cannot be helped by specifying a strategy finder.

Definition 6. *Let M be a mechanism (deterministically) implementing at equilibrium a property $\mathcal{P}$ for a class of contexts $\mathscr{C}$. We say that M is* totally vulnerable to equilibrium selection *if there exist a context $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$ in $\mathscr{C}$ and a player $j \in \mathcal{N}$ such that, for any (deterministic) strategy finder F helping M, letting σ be the equilibrium of (C, M) where each $\sigma_i = F(i, \theta_i, K_i)$, then there exists a strategy σ'_j such that the following 2 properties hold:*

1. *$\sigma'_j \sqcup \sigma_{-j}$ is an additional equilibrium of (C, M); and*
2. *$\mathcal{P}$ does not hold for $M(\sigma'_j \sqcup \sigma_{-j})$.*

Notice that Property 1 implies that $\mathcal{U}_j(M(\sigma'_j \sqcup \sigma_{-j})) = \mathcal{U}_j(M(\sigma))$. We are finally ready to state and prove our theorems, and we start with the deterministic version. For any $\epsilon \in (0, 1]$, define the property $\epsilon\mathcal{MKW} = \{\omega : \mathbb{E}[REV(\omega)] \geq \epsilon MKW(K^1_{-\star})\}$. Then,

Theorem 2. *For any $\epsilon \in (0, 1]$, any $n > 1$, and any deterministic mechanism M that deterministically implements $\epsilon\mathcal{MKW}$ at equilibrium for all n-player single-good auction contexts, M is totally vulnerable to equilibrium selection.*

The proof of Theorem 2 is given in Appendix B. We shall actually prove a constructive version of Theorem 2. Namely, the crucial player j of Definition 6 not only realizes that he has the option of an alternative strategy σ'_j, but can also find it easily. In addition, we shall prove that the gap between the required revenue and the revenue achieved in a "competing" equilibrium can be arbitrarily large.

Now we state the probabilistic version of our theorem, that is,

Theorem 3. *For any $\epsilon \in (1/2, 1]$, any $n > 1$, and any mechanism M that implements $\epsilon\mathcal{MKW}$ at equilibrium for all n-player single-good auction contexts, M is totally vulnerable to equilibrium selection.*

The proof of Theorem 3 can be easily derived from that of Theorem 2, and is omitted in this version of our paper.

6.3 Second-Knowledge Revenue Cannot Be Guaranteed in Undominated Strategies

Assuming familiarity with the notions of bounded mechanisms and undominated strategies (recalled in Appendix C anyway) and recalling that revenue impossibility applies to mechanisms with opt-out strategies, we have

Theorem 4. *For any $\epsilon \in (0, 1]$ and $n > 1$, no deterministic bounded mechanism can implement $\epsilon\mathcal{MKW}$ in undominated strategies for all n-player single-good auction contexts.*

The proof of Theorem 4 is given in Appendix C. Notice that Theorem 4 rules out mechanisms achieving (any constant fraction of) the second-knowledge revenue benchmark in weakly dominant strategies.

7 Potential Critiques and Conclusions

The significance of our results (hopefully not their correctness) and that of our approach could be questioned in a variety of ways. We try to predict and answer some of this potential criticism in Appendix D.

Mechanism design is a fascinating field. To achieve its full potential, and to maintain its impetus and vitality, it must constantly develop new approaches. We put forward such an approach in this paper. The new approach is quite conservative, and yet provably capable of guaranteeing desiderata that were previously out of reach, or reachable only in a fragile sense. We believe and hope that it will guarantee plenty of other desiderata as well.

References

[1] D. Abreu and H. Matsushima. Virtual Implementation in Iteratively Undominated Strategies: Complete Information. *Econometrica*, Vol. 60, No. 5, pages 993-1008, Sep., 1992.

[2] S. Baliga and R. Vohra. Market research and market design. Advances in Theoretical Economics, Volume 3, Issue 1, Article 5, 2003.

[3] P. Battigalli. On rationalizability in extensive games. Journal of Economic Theory 74, pages 4–71, 1997.

[4] P. Battigalli and M. Siniscalchi. An epistemic characterization of extensive form ratinalizability. Social Science Working Paper 1009, California Institute of Technology, 1997.

[5] P. Battigali and M. Siniscalchi. Interactive beliefs and forward induction. Manuscript. 2000.

[6] P. Battigali and M. Siniscalchi. Rationalization and Incomplete Information. The B.E. Journal of Theoretical Economics, Volume 3, Issue 1, 2003.

[7] B. D. Bernheim. Rationalizable Strategic Behavior. Econometrica, 52(1984), 1007-1028.

[8] J. Chen, A. Hassidim, and S. Micali. Robust Perfect Revenue from Perfectly Informed Players. *Innovations in Computer Science*, pages 94-105, Beijing, 2010.

[9] J. Chen and S. Micali. Safe Rationalizabilty and Safe Mechanism Design. Draft available at `http://people.csail.mit.edu/silvio/Selected Scientific Papers/Mechanism Design.`

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

[11] P. Dhangwatnotai, T. Roughgarden, and Q. Yan. Revenue maximization with a single sample. *Proceedings of the 11th ACM conference on Electronic commerce*, pages 129–138. ACM, 2010.

[12] D. Fudenberg and J. Tirole. Game Theorey. The MIT Press, 1991.

[13] A. Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, Vol. 41, No. 4, pages 587-602, Jul. 1973.

[14] J. Glazer and M. Perry. Virtual Implementation in Backwards Induction. *Games and Economic Behavior*, Vo.15, pages 27-32, 1996.

[15] A. Goldberg, J. Hartline, A. Karlin, M. Saks, and A. Wright. Competitive auctions. Games and Economic Behavior, Volume 55, Issue 2, pages 242-269, 2006.

[16] J.D. Hartline and T. Roughgarden. Simple versus optimal mechanisms. *Proceedings of the tenth ACM conference on Electronic commerce*, pages 225–234. ACM, 2009.

[17] L. Hurwicz. On Informationally Decentralized Systems. *Decision and Organization*, edited by C.B. McGuire and R. Radner, North Holland, Amsterdam, 1972.

[18] M. Jackson. Implementation in Undominated Strategies: A Look at Bounded Mechanisms. The Review of Economic Studies, Vol. 59, No. 4, pp. 757-775, 1992.

[19] M. Jackson, T. Palfrey, S. Srivastava. Undominated Nash Implementation in Bounded Mechanisms. *Games and Economic Behavior*, Vol.6, pages 474-501, 1994.

[20] J. Moore and R. Repullo. Subgame Perfect Implementation. Econometrica, Vol. 56, No. 5, pages 1191-1220, 1988.

[21] R. Myerson. Optimal Auction Design. *Mathematics of Operation Research*, Vol.6, No.1, pages 58-73, 1981.

[22] D. Pearce. Rationalizable strategic behavior and the problem of perfection. Econometrica 52, pages 1029-1050, 1984.

[23] M. Satterthwaite. Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, Vol.10, No.2, pages 187-217, Apr., 1975.

[24] I. Segal. Optimal pricing mechanisms with unknown demand. American Economic Review, 93(3), pages 509-529, 2003.

[25] M. Shimoji and J. Watson. Conditional dominance, rationalizability, and game forms. Journal of Economic Theory 83, pages 161-195, 1998.

Appendix

A Proof of Theorem 1

Let us now present an observable-action mechanism $\mathcal{M}_\epsilon$ that safely achieves the second-knowledge revenue within ϵ. (Numbered steps are taken by the players, steps marked by letters are conceptual steps taken by $\mathcal{M}_\epsilon$.)

Mechanism $\mathcal{M}_\epsilon$

a: *Set $a = 0$, and $P_i = 0\ \forall i$.*
Comment. (a, P) will be the final outcome of $\mathcal{M}_\epsilon$.

1: *Each player i announces a profile V^i of non-negative integers, publicly and simultaneously with others.*
Comment. V^i_i is i's "self-declared valuation", and V^i_j is i's "knowledge about j" for each $j \neq i$.

b: *Let $w = \text{argmax}_i V^i_i$, $CP = \max_{j\neq w} V^j_j$, $bip_w = \text{argmax}_{j\neq w} V^j_w$, and $KP = V^{bip_w}_w$.*
Comment. Ties are broken lexicographically. Player w is the "winner" (the one selected to get the good), CP is w's "classical price", bip_w is the "best informed player" about w, and KP is w's "suggested price".

c: *If $KP \leq CP$, reset a to w, P_w to CP, and go to Step e; otherwise (i.e., $KP > CP$) continue to Step 2.*

2: *Player w publicly announces YES or NO.*
Comment. Player w indicates whether he wants to get the good with price $KP - \frac{\epsilon}{n+1}$.

d: *If player w announced YES, then reset a to w and P_w to $KP - \frac{\epsilon}{n+1}$; otherwise reset P_{bip_w} to KP.*
Comment. If player w announced NO, then the good remains unsold, and player bip_w is punished.

e: *Reset each P_i to be $P_i - \delta_i$, where $\delta_i = \frac{\epsilon \sum_j V^i_j}{(n+1)(1+\sum_j V^i_j)}$.*
Comment. Each player i receives a reward δ_i.

Remark As promised, the above mechanism is very simple. And our analysis about this mechanism is not very short only because, wishing to highlight every aspect of our notion for at least once, we have chosen not to take any "shortcuts." Undoubtedly, one can be more succinct in future proofs.

Lemma 1. *Letting $\mathscr{C}$ be the class of all single-good auction contexts, $\mathcal{M}_\epsilon$ safely implements $\mathcal{P}_\epsilon$ over $\mathscr{C}$.*

Proof. Given our mechanism $\mathcal{M}_\epsilon$, it suffices to prove that for any context $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K) \in \mathscr{C}$ and for any strategy profile $\sigma \in \overline{\Sigma^2}$,

$$REV(\mathcal{M}_\epsilon(\sigma)) \geq MKW(K^1_{-\star}) - \epsilon.$$

Before proceeding any further, let us first clarify the structure of the game tree T specified by $\mathcal{M}_\epsilon$. We have that: (1) T is of height 2; (2) the root of T is the only decision node of height 2, where all players act simultaneously as specified by Step 1; and (3) each decision node D of height 1 corresponds to a profile of actions taken by the players at Step 1, according to which Step 2 is reached —therefore P^D consists of a single player, whose available actions are to announce YES and to announce NO.

We prove Lemma 1 based on the following three claims, whose proof we postpone a bit.

Claim 1. *For each player i and each $\sigma_i \in \Sigma^1_i$, σ_i satisfies the following property:*
For each decision node D of height 1 such that $P^D = \{i\}$ and $D \in H(\sigma_i \sqcup \tau_{-i})$ for some $\tau_{-i} \in \overline{\Sigma_{-i}}$, $\sigma_i(D)$ consists of i announcing YES if and only if $\theta_i \geq KP$.

Claim 2. *For each player i and each $\sigma_i \in \Sigma^2_i$, at Step 1, σ_i instructs player i to announce $V^i_j \geq K^1_i(j)$ for each $j \neq i$.*

Claim 3. *For each player i and each $\sigma_i \in \Sigma_i^2$, at Step 1, σ_i instructs player i to announce $V_i^i \geq \theta_i$.*

Now we are ready to prove Lemma 1. Recall that $\star$ is the player with the highest valuation, that is $\star = \text{argmax}_i\, \theta_i$. Let σ be a strategy profile in $\overline{\Sigma^2}$.

If the winner w in the execution of σ is not $\star$, then the classical price CP is at least $V_\star^\star$, which is at least $\theta_\star$ according to Claim 2. By definition, we have that $\theta_\star \geq MKW(K_{-\star}^1)$, and thus

$$CP \geq MKW(K_{-\star}^1).$$

If $KP \leq CP$, then the winner gets the good and pays CP. Since the total reward given to the players in Step e is $\sum_i \delta_i < \frac{n\epsilon}{n+1} < \epsilon$, we have that

$$REV(\mathcal{M}_\epsilon(\sigma)) = CP - \sum_i \delta_i > MKW(K_{-\star}^1) - \epsilon.$$

If $KP > CP$, then no matter whether the winner announces YES or NO, the revenue is at least $KP - \frac{\epsilon}{n+1}$ (exactly $KP - \frac{\epsilon}{n+1}$ from the winner if he announces YES, or KP from the best informed player otherwise), minus the total reward. That is,

$$REV(\mathcal{M}_\epsilon(\sigma)) \geq KP - \frac{\epsilon}{n+1} - \sum_i \delta_i > CP - \frac{\epsilon}{n+1} - \frac{n\epsilon}{n+1} \geq MKW(K_{-\star}^1) - \epsilon.$$

If the winner is $\star$, then Claim 2 implies that $KP \geq \max_{i\neq\star} K_i^1(\star)$, and Claim 3 implies that $CP \geq \max_{i\neq\star} \theta_i$. Recall that

$$MKW(K_{-\star}^1) = \max\left\{\max_{i\neq\star} K_i^1(\star), \max_{i\neq\star} \theta_i\right\}.$$

Thus $\max\{CP, KP\} \geq MKW(K_{-\star}^1)$, and a detailed case analysis as above leads to the desired conclusion, finishing the proof of Lemma 1. ■

We now proceed to prove Claims 1, 2, and 3.

Proof of Claim 1. We focus on proving half of Claim 1, that is, if $\theta_i \geq KP$ then i announces YES (the other half, that is if $\theta_i < KP$ then i announces NO, is totally symmetric). By contradiction, assume that there exists a decision node D of height 1 such that: $P^D = \{i\}$, $D \in H(\sigma_i \sqcup \tau_{-i})$ for some $\tau_{-i} \in \overline{\Sigma_{-i}}$, $\theta_i \geq KP$ at D, and $\sigma_i(D)$ consists of i announcing NO. We prove that σ_i is distinguishably dominated over Σ, which contradicts the hypothesis that $\sigma_i \in \Sigma_i^1$.

Let σ_i' be a strategy of i such that σ_i' coincides with σ_i at every decision node of i, except at node D, where $\sigma_i'(D)$ consists of i announcing YES. We prove that $\sigma_i \prec_\Sigma \sigma_i'$. To do so, notice that each strategy subprofile $\tau_{-i} \in \overline{\Sigma_{-i}}$ belongs to one of the following two types.

Type 1. $D \notin H(\sigma_i \sqcup \tau_{-i})$.

For such a τ_{-i}, we have that $H(\sigma_i \sqcup \tau_{-i}) = H(\sigma_i' \sqcup \tau_{-i})$, and τ_{-i} does not distinguish σ_i and σ_i'.

Type 2. $D \in H(\sigma_i \sqcup \tau_{-i})$.

For such a τ_{-i}, we have that $D \in H(\sigma_i' \sqcup \tau_{-i})$ also, and τ_{-i} distinguishes σ_i and σ_i'.

By hypothesis, there exists τ_{-i} of Type 2, and thus $\sigma_i \not\succeq_\Sigma \sigma_i'$. Accordingly, it is left to show that for any τ_{-i} of Type 2,

$$u_i(\mathcal{M}_\epsilon(\sigma_i \sqcup \tau_{-i})) < u_i(\mathcal{M}_\epsilon(\sigma_i' \sqcup \tau_{-i})). \tag{1}$$

To see why this is true, recall that for any outcome (a, P), $u_i(a, P) = \theta_i - P_i$ if $a = i$, and $-P_i$ otherwise. In the execution of $\sigma_i \sqcup \tau_{-i}$, i announces NO at node D, therefore he doesn't get the good and doesn't pay the price $KP - \frac{\epsilon}{n+1}$; while in the execution of $\sigma_i' \sqcup \tau_{-i}$, i announces YES, therefore he gets the good and pays the price $KP - \frac{\epsilon}{n+1}$. Because the reward i receives in Step e depends only on his announcement in Step 1, where σ_i and σ_i' coincide, we have that i receives the same amount of reward in both executions. Therefore the difference between $u_i(\mathcal{M}_\epsilon(\sigma_i' \sqcup \tau_{-i}))$ and $u_i(\mathcal{M}_\epsilon(\sigma_i \sqcup \tau_{-i}))$ is precisely $\theta_i - (KP - \frac{\epsilon}{n+1})$, which is > 0, because $\theta_i \geq KP$ and $\epsilon > 0$. Accordingly, Equation 1 holds, concluding the proof of Claim 1. □

Proof of Claim 2. By contradiction, assume that σ_i instructs i to announce $V_j^i < K_i^1(j)$ for some $j \neq i$. Let σ_i' be such that σ_i' coincides with σ_i everywhere, except that at Step 1, σ_i' instructs i to announce his knowledge about j to be $K_i^1(j)$. Denote this knowledge by $\widehat{V}_j^i$ to differentiate from i's announcement according to σ_i. Arbitrarily fix a type profile $\theta' \in K_i^1$, and let Σ' be the profile of strategy sets $(\Sigma_1', \ldots, \Sigma_i^1, \ldots, \Sigma_n')$, where each Σ_k' $(k \neq i)$ consists of all strategies of player k that are not distinguishably dominated over Σ, when k's true type is θ_k'. We prove that $\sigma_i \prec_{\Sigma'} \sigma_i'$, contradicting the hypothesis that $\sigma_i \in \Sigma_i^2$.

To see why this is true, recall that $\theta_i' = \theta_i$, and $\theta_k' \geq K_i^1(k)$ for any $k \neq i$. Notice that σ_i and σ_i' differ at the root, and thus any strategy subprofile $\tau_{-i} \in \overline{\Sigma_{-i}'}$ distinguishes them. Therefore we have to prove that Equation 1 holds for any such τ_{-i}. Arbitrarily fix such a τ_{-i}, we have that $u_i(\mathcal{M}_\epsilon(\sigma_i' \sqcup \tau_{-i}))$ and $u_i(\mathcal{M}_\epsilon(\sigma_i \sqcup \tau_{-i}))$ may be affected by at most three terms as listed below:

1. *Purchase.* i may be the winner and get the good, paying the corresponding price — either the classical price, or the suggested price with an $\frac{\epsilon}{n+1}$ discount;
2. *Punishment.* i may not be the winner but instead be the best informed player of the winner, and be punished by the suggested price due to the winner announcing NO; and
3. *Reward.* i receives reward in Step e, depending on his announced knowledge in Step 1.

Let us now look at the effects of these terms. First of all, notice that in Step 1, the only difference between execution $\sigma_i \sqcup \tau_{-i}$ and execution $\sigma_i' \sqcup \tau_{-i}$ is i's knowledge about j —one is $V_j^i < K_i^1(j)$, the other is $\widehat{V}_j^i = K_i^1(j)$. Therefore the winner and the classical price are the same in both executions, and if player $k \neq j$ is the winner, then his best informed player and the suggested price are the same in both executions. Second, notice that by the arbitrary choice of C in the prove of Lemma 1, the conclusion of Claim 1 also applies to any player $k \neq i$ with true type θ_k', and to any strategy in Σ_k'. (In particular, Claim 1 applies to player j with true type θ_j', and to any strategy in Σ_j'.) Accordingly, in both executions, if the winner k has to announce YES or NO, then he announces YES if and only if θ_k' is greater than or equal to the suggested price, independent of anything else. Third, because $\theta' \in K_i^1$, we have that $\theta_j' \geq K_i^1(j)$. Therefore in both executions, if j is the winner and i is the best informed player, then the suggested price is always $\leq K_i^1(j)$ and j always announces YES if Step 2 is reached. With this said, we have that:

1. If i is the winner in execution $\sigma_i \sqcup \tau_{-i}$, then he is also the winner in execution $\sigma_i' \sqcup \tau_{-i}$, with the same classical price and the same suggested price. Thus if he gets the good in the former, then he also gets it in the latter, paying the same price. Therefore the effect of *Purchase* is the same on both utilities under consideration.
2. If the winner is $k \notin \{i, j\}$ and i is the best informed player, then i gets the same amount of punishment (maybe 0) in both executions; if the winner is j and i is the best informed player, then j never announces NO and i is never punished. Therefore the effect of *Punishment* is again the same on both utilities under consideration.
3. i receives reward $\delta_i = \frac{\epsilon(V_j^i + \sum_{\ell \neq j} V_\ell^i)}{(n+1)(1+V_j^i+\sum_{\ell \neq j} V_\ell^i)}$ in Step e of execution $\sigma_i \sqcup \tau_{-i}$, and $\widehat{\delta}_i = \frac{\epsilon(\widehat{V}_j^i + \sum_{\ell \neq j} V_\ell^i)}{(n+1)(1+\widehat{V}_j^i+\sum_{\ell \neq j} V_\ell^i)}$ in Step e of execution $\sigma_i' \sqcup \tau_{-i}$. Since $\widehat{V}_j^i = K_i^1(j) > V_j^i$, we have that $\delta_i < \widehat{\delta}_i$, that is the *Reward* term is smaller in execution $\sigma_i \sqcup \tau_{-i}$ than in execution $\sigma_i' \sqcup \tau_{-i}$.

Combining the effects of these three terms together, we have that $u_i(\mathcal{M}_\epsilon(\sigma_i \sqcup \tau_{-i})) < u_i(\mathcal{M}_\epsilon(\sigma_i' \sqcup \tau_{-i}))$ as desired, concluding the proof of Claim 2. $\square$

Proof of Claim 3. Again by contradiction, assume that σ_i instructs i to announce $V_i^i < \theta_i$. Let σ_i' be such that σ_i' coincides with σ_i everywhere, except that at Step 1, σ_i' instructs i to announce θ_i as his self-declared valuation. Denote this value by $\widehat{V}_i^i$ to differentiate from i's announcement in σ_i. Again arbitrarily fix a type profile $\theta' \in K_i^1$, and let Σ' be the same as in the proof of Claim 2. We prove that $\sigma_i \prec_{\Sigma'} \sigma_i'$, contradicting the hypothesis that $\sigma_i \in \Sigma_i^2$.

To do so, similar to the proof of Claim 2, we have to prove that for each $\tau_{-i} \in \overline{\Sigma_{-i}'}$, Equation 1 holds. Notice that in both executions $\sigma_i \sqcup \tau_{-i}$ and $\sigma_i' \sqcup \tau_{-i}$, i's utility is again affected by three terms: *Purchase*,

Punishment, and *Reward*. It is easy to see that i receives more *Reward* in execution $\sigma'_i \sqcup \tau_{-i}$. A detailed case analysis shows that i's utility caused by *Purchase* and *Punishment* in execution $\sigma'_i \sqcup \tau_{-i}$ is at least as large as that in execution $\sigma_i \sqcup \tau_{-i}$ —to see this, notice that by announcing θ_i as his self-declared valuation, i increases the chance where he is the winner and gets some positive utility by *Purchase*, and reduces the chance where he is the best informed player and gets negative utility by *Punishment*. In sum, here we also have $u_i(\mathcal{M}_\epsilon(\sigma_i \sqcup \tau_{-i})) < u_i(\mathcal{M}_\epsilon(\sigma'_i \sqcup \tau_{-i}))$, concluding the proof of Claim 3. $\square$

B Proof of Theorem 2

We focus on the case where $n = 2$, since the other cases are very similar[7]. By contradiction, assume that there exists ϵ and a mechanism M implements $\epsilon\mathcal{MKW}$ at equilibrium for all 2-player single-good auction contexts. In our proof we are going to consider different contexts, and thus different games. Accordingly, to avoid confusion we use $\mathcal{U}_i^G$ to denote player i's expected utility function in a specific game G.

Let F be a strategy finder that helps M. We shall consider three games: the "real" game G, for which we want to prove the equilibrium-selection problems of M, and two auxiliary "mental" games, G' and G''.

Game G. $G = (C, M)$, where $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$ is the 2-player single-good auction context such that

- $\theta_1 = H$ and $\theta_2 = h$, where $h > 1$ and $H = 4h/\epsilon^2$;
- K_1^1 consists of all type profiles t such that $t_2 \geq 0$, and K_2^1 of all t such that $t_1 \geq 3h/\epsilon^2$.
- K_i^ℓ ($\ell > 1$) is *empty* for each player i — in other words, K_i^ℓ ($\ell > 1$) consists of everything that do not contradict $K_i^0, \ldots, K_i^{\ell-1}$.

For G it is easy to verify that

- $\star = 1, \quad K_2^1(1) = 3h/\epsilon^2, \quad MKW(K_{-\star}^1) = 3h/\epsilon^2$;
- Setting $\sigma = (F(1, \theta_1, K_1), F(2, \theta_2, K_2))$ and $(a, P) = M(\sigma)$ we have
  (a) $REV(a, P) \geq \epsilon MKW(K_{-\star}^1) = 3h/\epsilon$ (by our hypothesis on M)
  (b) $\mathcal{U}_i^G(a, P) \geq 0$ for each i (because M satisfies the opt-out condition and σ is an equilibrium)
  (c) $a = 1, \quad 3h/\epsilon \leq P_1 \leq H, \quad P_2 \leq 0$ (else 1's or 2's utility would be negative despite point b).

The Target To prove Theorem 2, we construct a strategy σ'_2 for player 2 such that the following two properties hold:

1. (σ_1, σ'_2) is an equilibrium of G, referred to as the *competing equilibrium*; and
2. $REV(M(\sigma_1, \sigma'_2)) < 3h/\epsilon$.

To construct σ'_2 we consider the following auxiliary game.

Game G'. $G' = (C', M)$, where $C' = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K')$ is such that $K'_1 = K_1$, while $(K')_2^1$ consists of all type profiles t such that $t_1 \geq 2h/\epsilon$ and $(K')_2^\ell$ ($\ell > 1$) is empty. That is, C' coincides with C everywhere except at player 2's conservative knowledge.

For G' it is easy to verify that

- $\star = 1, \quad (K')_2^1(1) = 2h/\epsilon, \quad MKW((K')_{-\star}^1) = 2h/\epsilon$.
- Setting $\sigma' = (F(1, \theta_1, K'_1), F(2, \theta_2, K'_2))$ and $(a', P') = M(\sigma')$ we have that

$$REV(a', P') \geq \epsilon MKW((K')_{-\star}^1) = 2h.$$

[7] Although we are dealing with equilibria, where there is often "safety in (player) numbers," this is not the case here.

Proof of the first target property. By definition of equilibrium, we have

$$\mathcal{U}_2^{G'}(M(\sigma_1', \sigma_2')) \geq \mathcal{U}_2^{G'}(M(\sigma_1', \sigma_2)); \tag{2}$$

and

$$\mathcal{U}_2^{G}(M(\sigma_1, \sigma_2)) \geq \mathcal{U}_2^{G}(M(\sigma_1, \sigma_2')). \tag{3}$$

Because player 2's true type is the same in G and G', we have $\mathcal{U}_2^{G'}(\cdot) = \mathcal{U}_2^{G}(\cdot)$, and thus Equation 2 implies

$$\mathcal{U}_2^{G}(M(\sigma_1', \sigma_2')) \geq \mathcal{U}_2^{G}(M(\sigma_1', \sigma_2)). \tag{4}$$

Because $\sigma_1' = \sigma_1$, since they are both equal to $F(1, \theta_1, K_1)$, we have

$$\mathcal{U}_2^{G}(M(\sigma_1, \sigma_2)) = \mathcal{U}_2^{G}(M(\sigma_1', \sigma_2)) \qquad \text{and} \qquad \mathcal{U}_2^{G}(M(\sigma_1', \sigma_2')) = \mathcal{U}_2^{G}(M(\sigma_1, \sigma_2')).$$

These two equalities, together with Equations 3 and 4, imply

$$\mathcal{U}_2^{G}(M(\sigma_1, \sigma_2)) = \mathcal{U}_2^{G}(M(\sigma_1, \sigma_2')). \tag{5}$$

Equation 5 implies that σ_2' is a best response of player 2 to σ_1 in game G, because σ_2 is so. Moreover, because σ_1' is a best response of player 1 to σ_2' in game G', so is σ_1. Further, because player 1's true type is the same in G and G', we have that $\mathcal{U}_1^{G}(\cdot) = \mathcal{U}_1^{G'}(\cdot)$, and thus σ_1 is also a best response of player 1 to σ_2' in game G. That is,

$$(\sigma_1, \sigma_2') \textit{ is an equilibrium of } G.$$

Thus target property 1 holds.

Proof of target property 2. Let us now prove that $REV(M(\sigma_1, \sigma_2')) < 3h/\epsilon$.

Assume, for contradiction, that $REV(M(\sigma_1, \sigma_2')) \geq 3h/\epsilon$. Similar as before, because (σ_1, σ_2') is an equilibrium of G, we have that

$$a' = 1 \qquad 3h/\epsilon \leq P_1' \leq H \qquad P_2' \leq 0.$$

Therefore

$$\mathcal{U}_1^{G}(a', P') = \theta_1 - P_1' \leq H - 3h/\epsilon.$$

We prove that there exists a strategy σ_1'' for player 1 such that $\mathcal{U}_1^{G}(M(\sigma_1'', \sigma_2')) > H - 3h/\epsilon$, contradicting the fact that (σ_1, σ_2') is an equilibrium of G. To do so, we consider another auxiliary game.

Game G''. $G'' = (C'', M)$, where $C'' = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta'', K'')$ is such that: (1) $\theta_1'' = 2h/\epsilon$ and $\theta_2'' = h$; and (2) $(K'')_1^1$ consists of all type profiles t such that $t_2 \geq 0$, $(K'')_1^\ell$ ($\ell > 1$) is empty, and $K_2'' = K_2'$.

In other words, C'' coincides with C' everywhere except at player 1's true type (player 1's knowledge is adjusted only to ensure that his knowledge is consistent with his true type).[8]

For G'' it is easy to verify that

- $\star = 1, \quad MKW((K'')_{-\star}^1) = 2h/\epsilon = MKW((K')_{-\star}^1)$.
- Setting $\sigma'' = (F(1, \theta_1'', K_1''), F(2, \theta_2'', K_2''))$ and $(a'', P'') = M(\sigma'')$ we have
  (a) $REV(a'', P'') \geq \epsilon MKW((K'')_{-\star}^1) = 2h$, and thus
  (b) $a'' = 1, \ 2h \leq P_1'' \leq \theta_1'' = 2h/\epsilon$, and $P_2'' \leq 0$.

[8]Notice that C'' is a well defined context, because $\theta'' \in (K'')_1^1$ and $\theta'' \in (K'')_2^1$.

Now, because $\theta_2'' = \theta_2$ and $K_2'' = K_2'$, we have that $\sigma_2' = \sigma_2''$. In fact,

$$\sigma_2' = F(2, \theta_2, K_2') = F(2, \theta_2'', K_2'') = \sigma_2''.$$

Moreover, because a mechanism's output depends only on the input strategy profile, and not on the whole underlying game, we have that $(a'', P'') = M(\sigma_1'', \sigma_2')$ in game G. Therefore

$$\mathcal{U}_1^G(M(\sigma_1'', \sigma_2')) = \theta_1 - P_1'' \geq H - 2h/\epsilon > H - 3h/\epsilon.$$

Thus target property 2 holds, and the proof of Theorem 2 is complete. ■

Corollary 1. *For any $\epsilon \in (0, 1]$, any $n > 1$, and any $\epsilon' \in (0, \epsilon)$, there exist an n-player single-good auction context $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$ and a player $j \in \mathcal{N}$ such that, for any mechanism M implementing $\epsilon\mathcal{MKW}$ at equilibrium for all n-player single-good auction contexts and for any strategy finder F that helps M, letting σ be the equilibrium of (C, M) where each $\sigma_i = F(i, \theta_i, K_i)$, then there exists a strategy σ_j' such that the following 2 properties hold:*

- *$\sigma_j' \sqcup \sigma_{-j}$ is an additional equilibrium of (C, M); and*
- *$\mathbb{E}[REV(M(\sigma_j' \sqcup \sigma_{-j}))] \leq \epsilon' MKW(K_{-\star}^1)$.*

Proof. Because M implements at equilibrium $\epsilon\mathcal{MKW}$, it also implements at equilibrium $\epsilon'\mathcal{MKW}$, for all n-player single-good auction contexts. Moreover, any strategy finder F that helps M for the former also helps M for the latter. Therefore the corollary immediately follows by applying Theorem 2 to $\epsilon'\mathcal{MKW}$. □

C Proof of Theorem 4

First let us recall the definitions of undominated strategies and bounded mechanisms from [18], with the latter changed slightly to match our formalization of contexts and games.

Definition 7. *Let G be a game, i a player, and σ_i, σ_i' two strategies of player i. We say that σ_i is* weakly dominated *by σ_i' if $u_i(\sigma_i' \sqcup \tau_{-i}) \geq u_i(\sigma_i \sqcup \tau_{-i})$ for all strategy subprofile τ_{-i}, and $u_i(\sigma_i' \sqcup \hat{\tau}_{-i}) > u_i(\sigma_i \sqcup \hat{\tau}_{-i})$ for some $\hat{\tau}_{-i}$.*

We say that σ_i is undominated *if it is not weakly dominated by any strategy of i. We say that a strategy profile σ is undominated if each σ_i is undominated.*

Definition 8. *Let M be a mechanism for a class of contexts $\mathscr{C}$. We say that M is* bounded *if for any context $C \in \mathscr{C}$, any player i, and any strategy σ_i of i in the game (C, M), if σ_i is weakly dominated, then it is weakly dominated by some undominated strategy of i in (C, M).*

Now let us be formal about what it means for a mechanism to implement a property at undominated strategies.

Definition 9. *We say that a mechanism M* implements in undominated strategies *a property P for a class of contexts $\mathscr{C}$ if for all contexts $C \in \mathscr{C}$ and for all strategy profiles σ which is undominated in game (C, M), P holds for $M(\sigma)$.*

Notice that our notion of implementation in undominated strategies is weaker than that of [18], which requires that the set of outcomes where P holds and the set of outcomes produced by the set of undominated strategy profiles coincide with each other. Therefore our impossibility result automatically rules out implementation in undominated strategies in the sense of [18]. We are finally ready to prove Theorem 4.

Proof. We focus on the case where $n = 2$, and proceed by contradiction. Assume that there exists $\epsilon \in (0, 1]$ and a bounded mechanism M that satisfies the Opt-Out condition and implements $\epsilon\mathcal{MKW}$ in undominated strategies for all 2-player single-good auction contexts. Recall that Σ_1 and Σ_2 denote the sets of all strategies

for players 1 and 2 respectively. Similar to the proof of Theorem 2. we are going to consider different contexts, and thus different games. Accordingly, to avoid confusion we use $\mathcal{U}_i^G$ to denote player i's expected utility function in a specific game G.

We shall consider two games: the "real" game G, for which we want to prove that there exists undominated strategy profile σ such that $\epsilon\mathcal{MKW}$ does not hold for $M(\sigma)$, and an auxiliary "mental" games G'.

Game G. $G = (C, M)$, where $C = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta, K)$ is the 2-player single-good auction context such that

- $\theta_1 = H$ and $\theta_2 = h$, where $h > 1$ and $H > 2h/\epsilon^2$;
- K_1^1 consists of all type profiles t such that $t_2 \geq h$, and K_2^1 of all t such that $t_1 \geq H$.
- K_i^ℓ ($\ell > 1$) is empty for each player i.

For G, letting $UD^G = (UD_1^G, UD_2^G)$ be the profile of sets of undominated strategies, it is easy to verify that

- $\star = 1, \quad K_2^1(1) = H, \quad MKW(K_{-\star}^1) = H$;
- $\forall \sigma \in \overline{UD^G}$, we have that $REV(M(\sigma)) \geq \epsilon MKW(K_{-\star}^1) = \epsilon H > 2h/\epsilon$ (by our hypothesis about M).

Our goal is to demonstrate the existence of a strategy profile $\hat{\sigma} \in \overline{UD^G}$ such that $REV(M(\hat{\sigma})) \leq 2h/\epsilon$. To do so, we consider the following auxiliary game.

Game G'. $G' = (C', M)$, where $C' = (\mathcal{N}, \Omega, \mathcal{T}, u, \theta', K')$ is such that $\theta_1' = 2h/\epsilon$, $\theta_2' = h$, $K_1' = K_1$, while $(K')_2^1$ consists of all type profiles t such that $t_1 \geq 2h/\epsilon$ and $(K')_2^\ell$ ($\ell > 1$) is empty. That is, C' differs from C at player 1's valuation and player 2's conservative knowledge.

For G', letting $UD^{G'} = (UD_1^{G'}, UD_2^{G'})$ be the profile of sets of undominated strategies, it is easy to verify that

- $\star = 1, \quad (K')_2^1(1) = 2h/\epsilon, \quad MKW((K')_{-\star}^1) = 2h/\epsilon$.
- For any $\sigma' \in \overline{UD^{G'}}$, we have that

$$REV(M(\sigma')) \geq \epsilon MKW((K')_{-\star}^1) = 2h.$$

The existence of the desired $\hat{\sigma}$. Consider game G'. Because M satisfies the Opt-Out condition, we have that there exists a strategy $\sigma_1' \in UD_1^{G'}$ such that

$$\forall \tau_2 \in \Sigma_2, \quad \mathcal{U}_1^{G'}(M(\sigma_1', \tau_2)) \geq 0. \tag{6}$$

To see why this is true, notice that $\mathcal{U}_1^{G'}(M(out_1, \tau_2)) = 0$ for all $\tau_2 \in \Sigma_2$. If $out_1 \in UD_1^{G'}$ then take $\sigma_1' = out_1$ and we are done. Otherwise there exists $\sigma_1' \in UD_1^{G'}$ such that out_1 is weakly dominated by σ_1' (since M^2 is bounded), implying that $\mathcal{U}_1^{G'}(M(\sigma_1', \tau_2)) \geq \mathcal{U}_1^{G'}(M(out_1, \tau_2)) = 0$ for all $\tau_2 \in \Sigma_2$, as desired.

Similarly, there exists a strategy $\sigma_2' \in UD_2^{G'}$ such that

$$\forall \tau_1 \in \Sigma_1, \quad \mathcal{U}_2^{G'}(M(\tau_1, \sigma_2')) \geq 0. \tag{7}$$

Combining Equations 6 and 7 and letting $\omega' = (a', (P_1', P_2')) = M(\sigma_1', \sigma_2')$, we have that

$$\mathcal{U}_1^{G'}(\omega') \geq 0, \quad \text{and} \quad \mathcal{U}_2^{G'}(\omega') \geq 0. \tag{8}$$

Since $\sigma_1' \in UD_1^{G'}$ and $\sigma_2' \in UD_2^{G'}$, we have that

$$REV(\omega') \geq 2h. \tag{9}$$

Combining Equations 8 and 9 we have that

$$a' = 1, \quad 2h \leq P_1' \leq 2h/\epsilon, \quad \text{and} \quad P_2' \leq 0. \tag{10}$$

To see why this is true, notice that no matter which player gets the good, his price can not be higher than his valuation, since otherwise his utility is negative. For the same reason, the player who does not get the good must have non-positive price. Finally, if player 2 gets the good, then the revenue is at most $0 + h < 2h$, which contradicts Equation 9.

Now consider game G. By Equation 10, we have that $\mathcal{U}_1^G(\omega') = H - P_1' \geq H - 2h/\epsilon$. Thus there exists a strategy $\hat{\sigma}_1 \in UD_1^G$ such that

$$\mathcal{U}_1^G(M(\hat{\sigma}_1, \sigma_2')) \geq H - 2h/\epsilon. \tag{11}$$

To see why this is true, notice that if $\sigma_1' \in UD_1^G$ then take $\hat{\sigma}_1 = \sigma_1'$ and we are done. Otherwise σ_1' is weakly dominated by some $\hat{\sigma}_1 \in UD_1^G$, and we have that $\mathcal{U}_1^G(M(\hat{\sigma}_1, \sigma_2')) \geq \mathcal{U}_1^G(M(\sigma_1', \sigma_2')) = \mathcal{U}_1^G(\omega') \geq H - 2h/\epsilon$, as desired.

Because player 2's valuation is the same in G and G', his utilities are the same in the two games for any outcome, and his sets of undominated strategies are the same as well. That is,

$$\mathcal{U}_2^G(\cdot) = \mathcal{U}_2^{G'}(\cdot), \quad \text{and} \quad UD_2^G = UD_2^{G'}. \tag{12}$$

Combining Equations 7 and 12, we have that $\sigma_2' \in UD_2^G$ and $\forall \tau_1 \in \Sigma_1, \mathcal{U}_2^G(M(\tau_1, \sigma_2')) \geq 0$. In particular,

$$\mathcal{U}_2^G(M(\hat{\sigma}_1, \sigma_2')) \geq 0.$$

Letting $\hat{\sigma} = (\hat{\sigma}_1, \sigma_2')$, and $\omega = (a, (P_1, P_2)) = M(\hat{\sigma})$, we have that

$$\hat{\sigma} \in \overline{UD^G}, \quad \mathcal{U}_1^G(\omega) \geq H - 2h/\epsilon, \quad \text{and} \quad \mathcal{U}_2^G(\omega) \geq 0,$$

where the second inequality is precisely Equation 11.

Accordingly, if $a = 1$ then $P_1 \leq 2h/\epsilon$ and $P_2 \leq 0$, which implies $REV(\omega) \leq 2h/\epsilon$. If $a = 2$ then $P_1 \leq 0$ and $P_2 \leq h$, which also implies $REV(\omega) \leq 2h/\epsilon$. In sum, $REV(\omega) \leq 2h/\epsilon$, contradicting the fact that M implements $\epsilon\mathcal{MKW}$ in undominated strategies for all 2-player single-good auction contexts, and Theorem 4 follows. ■

D Answers to Potential Criticism

We believe our impossibility results to be uncontroversial, and can only hope that everyone else will agree.

Our mechanism, however, is somewhat unorthodox, and as such it may raise some concerns from a more standard perspective. Below we try to alleviate at least some of these concerns.

- *Players will never "tell on each other."*

  To leverage the players' conservative knowledge, our mechanism needs the player to reveal what they know about their opponents. And to ensure that this happens, it rewards them for the "external knowledge" they provide.

  Some (with many more friends than competitors) may object that any mechanism so constructed is doomed to failure. We disagree on three grounds.

  1. Technically speaking, many classical mechanism in settings of complete information rely on the players to report not only their own types, but also the types of their opponents. Examples include the celebrated results of Moore and Repullo [20]; Jackson, Palfrey, and Srivastava [19]; Abreu and Matsushima [1]; and Glazer and Perry [14].
  2. In elections, politicians are only too happy to tell on (the "types" of) their opponents. In fact, they are even happy to volunteer false information about (the "types" of) their opponents in order to be elected! Similarly, companies are often happy to provide negative information about their

opponents to advance their chances in highly contested games. This being the case, note that our mechanism only encourages the players to volunteer true information about their opponents, risking immediate punishment for declaring false information.

3. We find no moral problem when the players legitimately use all information at their disposal in order to advance their own case and enable the designer to reach his legitimate goals (which includes revenue maximization in a capitalistic society). In any case, mechanism design is about leveraging the players' rationality, not their morality or generosity.

In sum, our mechanism is totally in line with mainstream economic thinking. And if this thinking changes, it will force one to formally and openly *restate* the new goals of mechanism design.

- *Knowledge of Small Utility Will Discourage Participation.*

  Our mechanism may be critized because leaves very small utilities to the players. We note that this is a necessary byproduct of our goal of maximizing revenue, and in particular at achieving the second-knowledge benchmark. A seller should perhaps be weary of using a "revenue-maximizing" mechanism in which the players rush to participate knowing they will receive great utilities!

  Each player has knowledge about his opponents *independently* of the mechanism used. And he will use this knowledge to determine whether it is worth for him to participate in any possible mechanism. Our mechanism at least rewards each player for his external knowledge, and thus he will always have some incentive to participate in our mechanism. But if he knows that another player's valuation is higher than his own, the same player may not participate at all in —say— an auction conducted using the second-price mechanism.

  Finally, our mechanism enables the designer to arbitrarily choose ϵ, the maximum reward that a player can get. We find it important to prove that ϵ can be chosen to be arbitrarily small. But if there are few players, or if the seller feels generous, ϵ can easily be chosen so as to make everyone happy! In particular, our mechanism can be easily changed so as to set aside any *fraction* ϵ of the final revenue to reward the players for their contributed knowledge.

- *Players may fear to contribute their knowledge for fear of being punished.*

  This is not a worry as long as one assumes common knowledge of rationality, a traditional assumption in mechanism design. Indeed, our knowledge is *set-theoretic* and a player can guarantee himself *positive* utility by "bidding the highest knowledge he has."

- *You can't punish players.*

  All is fair in love, war, and mechanism design. This is the great appeal of mechanism design. This said, it is true that if —say— the second-knowledge mechanism is played over the Internet and a player from a remote country overbids his knowledge and must thus pay \$1M, it may be hard to oblige him to do so. But by the same token, if the same player bid \$2M in a second-price auction of a single good, where the second highest bid is \$1M, he must be obliged to pay \$1M. Will that be any easier?

  Our point is that mechanism design is a mathematical formalization whose practical meaningfulness crucially depends on a larger real infrastructure (with courts, bailiffs, etc.). Without it very few mechanisms would have practical value, and the second-knowledge one would be in merry company.

- *What has this to do with computation?*

  First of all, an axiomatic and defiant answer. Theory of computation will continue to be the driving force it currently is, only if it will continue to define its own borders. Should these borders become fixed, somehow, it will become an ordinary field —with its own custodians of the "sacred", but confined, flame. The mere fact that this paper is a contribution of computation theorists testifies that its subject matter is computation related.

Second, mechanism design has started influencing the development of Internet protocols, that have traditionally been part of computer science —and cryptography in particular. It has started fueling alternative approaches to secure computation. And so on. Without playing with a full deck of cards, that is, without adopting alternative and flexible models for mechanism design, our successes in these new applications will be limited at best.

Third, lots of beautiful mechanisms are currently been developed that will provide us better auctions, but not —narrowly speaking!— with new insights about the nature of computation. We believe that these mechanisms too will prove invaluable to the future of computation.

Again, if after all these explanations, one "senses" that the mechanism is fundamentally flawed, then it will serve the crucial role of forcing all of us to dramatically and explicitly *revise* the goals of mechanism design.