Explorations in Cyber International Relations
Massachusetts Institute of Technology Harvard University

Rational Robustness for Mechanism Design

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

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

November 10, 2009

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.

MINERVA INITIATIVE
DEPARTMENT OF DEFENSE

Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2009-057 November 10, 2009

Rational Robustness for Mechanism Design (First Draft)

Jing Chen and Silvio Micali

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

CSAIL

Rational Robustness for Mechanism Design[*]

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

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

November 10, 2009

Abstract

The currently prevailing equilibrium-based approach to mechanism design suffers from a plurality of fundamental problems, and new conceptual frameworks are needed to solve or sufficiently alleviate them.

In this paper, we put forward *rational robustness*, a new solution concept/implementation notion that is not equilibrium-based; prove its fundamental structural theorems; and compare it with prior notions.

Our notion of implementation is specifically built so as to be robust against the problem of equilibrium selection. We prove it robust against other fundamental problems as well in different papers.

[*]Work partially supported by ONR Contract Number N00014-09-1-0597.

1 Introduction

1.1 A Conceptual Contribution

The primary contribution of this paper is conceptual: we introduce new notions for mechanism design and establish their fundamental structural properties. Conceptual contributions are harder to evaluate than technical ones. After all, a technical advancement is by definition immediately measurable, whether or not it will be superseded in a few months. The value of a new notion is instead best measured over several years. So it is understandable that, despite all claims to the contrary, conferences (including FOCS and STOC) might treat new notions with some skepticism. Yet, the risk of encouraging new notions for mechanism design is worth taking.

Traditionally based on equilibria, mechanism design is a beautiful and successful field, see in particular [9, 13, 14, 15, 11, 16], but needs a new conceptual platform in order to be robust against collusion, computation/communication complexity, and privacy. *Collusion* is a problem for traditional mechanisms because an equilibrium only guarantees that no individual player has any incentive to deviate from his envisaged strategy. But two or more players may have plenty of incentive to jointly deviate from their envisaged strategies, which is exactly what collusive players do. *Complexity* is a problem because traditional mechanisms often are too hard to play (because the players must act over exponentially many rounds or communicate exponentially many bits, or because —after the players are done playing— the outcome functions are too hard to evaluate). In practice this problem would imply that no one could live long enough to see the outcomes of such mechanisms. Moreover, it may not be solved by adopting some computationally tractable variants, because these may fail to be "incentive compatible" or sufficiently close approximations. *Privacy* is a problem in traditional mechanisms because the players are routinely asked to reveal their own utility functions. But, by so doing, privacy-valuing players will by definition receive a negative utility, thus putting into serious question the overall incentive compatibility of such mechanisms. Game theory, of which mechanism design is a major part, aims at building a science of human strategic behavior. And in the world as it is, rather than as it should be, humans *collude, have bounded computation/communication resources, and care a lot about privacy.* By ignoring these uniquely human truths, mechanism design cannot be sufficiently robust, and can actually fail to reach its ultimate research goals, or severely limit the extent to which it achieves them.

To be sure, there have been efforts to improve the robustness of various mechanisms against one or more of collusion, complexity, or privacy. But there is no effort underway to rebuild mechanism design from its foundations so as to demand and achieve reasonable robustness, in all these three fronts, from any mechanism.

In our opinion, the quest of *robust* mechanisms will be ultimately frustrated by the currently prevailing insistence on equilibria as the main, underlying solution concept. This solution concept has served us well so far, but new ones are needed to underwrite the design of robust mechanisms. Providing and studying such a candidate new solution concept is the very goal of this paper.

Coherently with what just said we put forward a solution concept that is *not equilibrium-based,* because it typically demands mechanisms to guarantee their desired properties also at profiles of strategies that are not equilibria. Only occasionally, our concept coincides with more traditional ones: such as the single equilibrium surviving the iterative elimination of strictly dominated strategies, or, in extensive-form games, the unique subgame-perfect equilibrium. Although applicable to normal-form games, where all players act once and simultaneously, our solution concept achieves its full power for games of *extensive form,* where the players act over several turns. We find this phenomenon quite natural for two reasons. First, because we want to define a rational play in a more stringent way than in a Nash equilibrium, and thus sequential mechanisms give us a "finer control on the players' rationality." (For instance, "empty threats" can be easily dismissed in extensive-form games, but not in their normal-form counterparts.) Second, because we want to preserve the players' privacy as much as possible, and thus interaction becomes crucial. (For instance, proofs can be *zero-knowledge* if prover and verifier talk back and forth [7], but no zero-knowledge proofs exist if the prover must communicate his proofs in a single message —unless additional infrastructure such as public and random string is available [2].)

Technically, our solution concept is based on a more elementary notion, **iterative elimination of distinguishably dominated strategies**, which may itself be of independent interest. In fact, it bridges a currently vast chasm: that between elimination of *strictly dominated* strategies (a procedure very meaningful, but unlikely to provide the foundation of many extensive-form mechanisms) and elimination of *weakly dominated* strategies (a procedure very general, but unlike to be sufficiently meaningful for providing a solid foundation for any mechanism).

We have been able to design a few mechanisms provably robust against collusion, complexity, and privacy [4, 3, 5]. But to enable the design of many more robust mechanisms, mechanism design needs a new, meaningful, and general solution concept. This paper is thus fully dedicated to this goal.

Our present notion actually is a strict generalization of those used in our previous robustness results.

1.2 The Problems of Equilibrium Selection and Players' Beliefs

Short of taking our word, it might not be immediately clear that our new notion of implementation will usher in mechanisms robust against collusion, complexity and privacy. (Although it might be reasonable to hope that non equilibrium-based mechanism design will succeed where equilibrium-based mechanism design failed.) But it should be obvious that our solution concept addresses directly a more basic problem in traditional mechanism design: *equilibrium selection.*

The Lethality of Equilibrium Selection Mechanism design aims at achieving a desired property P by leveraging the players' rationality. Traditional mechanism design interprets this goal as achieving P "at equilibrium." That is, it aims at engineering a game G with an equilibrium σ whose play yields an outcome for which P holds. Recall that an equilibrium σ is a profile of strategies, one for each player, such that each σ_i is player i's best strategy to play *if he believes that each other player j will choose strategy σ_j.* Since equilibria are regarded as predictions of how rational players will play, traditional mechanism design provides a rational hope that P will hold in the actual play of G. But this hope is *not* a guarantee. Typical games, including engineered ones, have multiple equilibria, and if P holds at some of them, it may not hold at others.

Accordingly, a more demanding notion of mechanism design has been considered, *full implementation (at equilibria)*, whose engineered games must be such that their desired properties hold at each of their possible equilibria. Full implementation too, however, can be astronomically far from guaranteeing a desired property P. Let us explain. Assume that the game G is of *normal form.* This means that all players act only once, choosing their strategies without any interaction with the other players. (In essence, each player can be thought as choosing his strategy in a separate room.) All games can be put in this form, and the overwhelming majority of the games arising in mechanism design are of this form. Assume now that G has just two equilibria, σ and τ, such that $\sigma_i \neq \tau_i$ for each player i. Then, two different predictions are equally rational about what each player i will do in his own room: (1) he will select strategy σ_i, and (2) he will select strategy τ_i. But since all players act independently, this also implies that it is rational to envisage that 2^n different plays of G can actually arise! Thus, if P holds at just two of them, namely the equilibria σ and τ, hoping that P will hold in an actual play of G has much more to do with miracles than with rationality.

The Fatal Attraction of Beliefs Equilibrium selection arises because the very general notion of an equilibrium is based on the player's beliefs (about what the others will do), but *players' beliefs may not "properly match."* Reliance on players' beliefs is as convenient as it is artificial and dangerous. On one hand, it guarantees that every finite game has an equilibrium, the famous Nash theorem. On the other, it causes equilibrium selection to be not an exogenous and occasional nuisance, but an *intrinsic* and *ever present* problem for equilibrium-based mechanism design.

As we shall soon see, reliance on players' beliefs will further tempt us —in new ways— to trade meaningfulness for convenience, and we will have to work harder to reject the offer.

Mechanism Design Free of Equilibrium Selection The problem of equilibrium selection totally vanishes when a mechanism yields a game with a (single) equilibrium in which each strategy (1) is *strictly dominant* or (2) survives *iterated elimination of strictly dominated strategies.* Unfortunately, the applicability of both types of mechanisms is limited.

Strictly dominant-strategy mechanisms are very rare (and provably cannot achieve many desired properties —e.g., [6, 18]). Accordingly, forcing designers to consider only such mechanisms would confine mechanism design to the catching of white elephants.

Mechanisms of the second type have been shown by Abreu and Matsushima [1] to be capable of essentially achieving all desired properties, but under very strong assumptions. In essence, ignoring other technical conditions, their general mechanism assumes that *each player perfectly knows the utility functions of all players.* This of course is a significant restriction and cannot be easily removed. In fact, iterated elimination of strictly dominated strategies is not even defined unless all players' utilities are common knowledge (at least in a non-Bayesian setting). Yet, even under such strong assumptions, their mechanism, as well as the extensive-form variant of Glazer and Perry [8], can be proved to be vulnerable to collusion, complexity, and privacy in intrinsic ways.

(Note that equilibrium selection continues to affect, although to a lesser extent, the notions of implementation in weakly dominant strategies, and implementation in undominated Nash equilibrium [12].)

In sum, to enable mechanism design to cast a wide and robust net, new and "belief-independent" notions of implementations should be sought.

1.3 Our Notion of Implementation

Mechanism design must be based on an underlying *solution concept.* Indeed, one must first define *rational plays* and then ensure that a given property P holds at the rational plays of his engineered game. The weakness of Nash equilibria as a solution concept is not being "closed under Cartesian product", which we have argued to be at the root of the problem of equilibrium selection. To eradicate this problem at its root, (1) we consider only solution concepts consisting of a *profile S of strategy subsets* —rather than a set of strategy profiles— and (2) we demand that a mechanism guarantees its desired property for any play $\sigma \in S_1 \times \cdots S_n$. But:

How should such strategy subsets S_i be chosen?

We break our answer in several stages.

Implementation in Iteratively Undominated Strategies Assume initially that (1) each player perfectly knows everyone's utilities, and (2) the iterative elimination of strictly dominated strategies (see [17]) leaves each player i with a single strategy σ_i. In this wonderful case, we chose $S_i = \{\sigma_i\}$ for each player i. This choice provides a perfectly rational answer to the above question, and indeed is the very implementation notion of the cited mechanism of Abreu and Matsushima [1].

Implementation in Surviving Strategies Let us now keep the first assumption above, but remove the second one. That is, let us assume that, after the iterative elimination of strictly dominated strategies, each player i is left with a subset L_i of surviving strategies. Then we choose $S_i = L_i$ for each i. By this choice, we demand that our desired property P holds for any strategy profile of surviving strategies, and thus refer to the corresponding notion of implementation as *implementation in surviving strategies.* Note that such an implementation already is *not equilibrium-based,* because —as for implementation in undominated strategies [10]— a profile of surviving strategies needs not to be an equilibrium at all.

Rationally Robust Implementation Still keeping for now the assumption of perfectly knowledgeable players, we want to eliminate strategies a bit more aggressively. One obvious temptation would be to adopt iterative elimination of *weakly dominated strategies*, but such a solution concept would be ill-defined. This is so because the final subsets of surviving strategies would be crucially dependent on the order in which the players choose to eliminate their strategies. Thus, if different players used different orders of elimination, a mechanism's desired property P might not be guaranteed at all, even if it held whenever all players eliminate strategies in the *same* order. To overcome this obstacle, it is tempting to rely upon a different type of beliefs: *beliefs about elimination orderings.* But we reject this easy temptation. No matter what beliefs the players might have (about playing, elimination ordering, or anything else), *even if* these beliefs happened to be miraculously and properly matched, assuming that a mechanism designer might be aware of them strikes us as totally artificial.

Accordingly, we instead consider a new elimination process, *iterative elimination of distinguishably dominated strategies.* Distinguishable domination is "in between" strict and weak domination (but coincides with strict domination for normal-form games). While the precise description of our elimination process $\mathscr{P}$ is better left to our technical sections, we wish to highlight here its intuitive properties. Namely,

0. *Order Independence.* Even if every player uses *his own* order of strategy elimination, the final strategy subset of each player is unique (up to "isomorphism").

1. *Perspective Independence.* If i eliminates (in his mind) a strategy σ_j of an opponent j, then a rational j will never play σ_j.

   (Thus, it is safe for i to further eliminate one of his own strategies σ_i based on his own elimination of σ_j, in the sense that he will never be "surprised" by j using σ_j against him in a real play.)

2. *Belief Independence.* After each player ends his own execution of $\mathscr{P}$, our desired property must hold for any play σ (not necessarily an equilibrium!) in the Cartesian product of the sets of surviving strategies.

Note that we do not demand that such a process $\mathscr{P}$ produces a profile of *minimal* strategy subsets for which the above three properties hold.[1] Indeed, as we are *also* eager to deal with mechanisms that do not rely on fully rational players, we find it important to consider also processes $\mathscr{P}$ that yield strategy subsets "barely" satisfying the above three properties.

Summary of Rationally Robust Implementation for Perfect-Knowledge, Independent Settings

- *For normal-form games,* rationally robust implementation coincides with implementation in surviving strategies, and thus with implementation in iteratively undominated strategies in the best case; and
- *For extensive-form games,* rationally robust implementation is different from implementation in surviving strategies and is stronger than backward induction, but coincides with unique subgame-perfect equilibrium in the best case.

By *strictly generalizing* these classical and very *meaningful* notions, rationally robust implementation aims at providing *meaningful flexibility,* that is, at enabling designers to construct mechanisms achieving new desired properties —such as robustness against collusion, complexity, and privacy— or achieving older properties in more practical ways.

Dealing with Imperfect Knowledge It is now time for us to remove the unrealistic assumption that the players have perfect knowledge of each other's utility functions, without introducing other debatable assumptions such as the availability of Bayesian information. But then, even the iterative elimination of distinguishably dominated strategies becomes ill-defined. Thus, we must extend our notion of implementation to the case of *imperfectly knowledgeable* players. To this end, we define what strategies the players can iteratively and safely eliminate, not only based on the common knowledge of rationality, but also on *whatever knowledge each player may have about the utility functions of his opponents.*

[1] For instance, if the iterated elimination of all strictly dominated strategies whose binary representation starts with 0 satisfied mechanism usefulness, we would be satisfied.

Dealing with Collusion Finally, we enlarge our implementation notion (and thus its underlying notion of an iterative elimination) to settings where different players belong to different collusive sets, and a player may not have any knowledge about these collusive sets (except his own, if he is a colluder).

Road Map We first present our notion of implementation in the simpler setting in which each player is (1) independent (i.e., does not collude with anyone else) and (2) perfectly informed about the utilities of his opponents. Secondly, we compare our notions with previous ones in the same setting. Thirdly, still in the same setting, we discuss stronger version of our notion applicable to non-fully rational players. Finally, we enlarge our notions to independent players with imperfect knowledge.

Our structural theorems about our notions are proven in our appendix.

Our notions for the collusive setting will appear in the final version of this paper.

2 Our Notions for Independent Players with Perfect Knowledge

A game G has two components: a context C and a mechanism M, $G = (C, M)$. The context describes the players, all possible outcomes, the players' utilities for each outcome, and the players' knowledge (and possibly their beliefs). The mechanism describes which strategies are available to the players (including their *opt-out* strategies) and how each profile of strategies yields an outcome. Through out this paper, all our games are finite, and thus so are the set of players, the set of all possible outcomes, and the set of strategies for each player.

We start by clarifying the contexts of the present setting.

Definition 1. *A context C with perfectly knowledgeable and independent players, a PKI context for short, consists of the following components:*

- *N, the finite set of* players: $N = \{1, \dots, n\}$
- *Ω, the set of possible outcomes. A member ω of Ω is referred to as an* outcome.
- *u_i, for each player i, is i's* utility function, *mapping outcomes to real numbers.*

All these components are common knowledge to all players.

As for the purest form of mechanism design, we insist that all knowledge lies with the players, and only with the players. Accordingly, the designer of a mechanism for the above context knows the sets N and Ω as usual, and knows as well that all players know the profile u of utility functions, but he himself has no information about u. (Notice that, aiming at robustness, we have no need for players' beliefs.)

Above, as for all contexts considered in this paper, we automatically assume that all players are independent whenever no coalitional structure is specified. Accordingly, in any game whose context coincides with the above C, each players i acts so as to maximizes his own u_i.

Essential Notation If σ is a profile of strategies in a game $G = (C, M)$, then $H(\sigma)$ denotes the *history* of σ, and $M(\sigma)$ denotes the *outcome* of σ. If M is of normal form then $H(\sigma)$ coincides with σ. If M is of extensive form, then $H(\sigma)$ consists of the sequence of decision nodes plus the terminal node of the game tree reached when executing σ. If M or some σ_i's are probabilistic, then $H(\sigma)$ and $M(\sigma)$ are both distributions.

Whenever we say that S is a profile of "strategy subsets", we mean that each S_i is a subset of i's strategies. For such an S, we define the Cartesian closure of S as $\overline{S} = S_1 \times \cdots \times S_n$, and we define $\overline{S_{-i}} = \prod_{j \neq i} S_j$.

Through out this paper, for every mechanism M, we denote by OUT_i the opt-out strategy of player i. In fact, every M must satisfy the following

opt-out condition: Each player i has an *opt-out* strategy OUT_i such that, for all subprofile σ_{-i} of strategies of the other players, $u_i(M(\text{OUT}_i \sqcup \sigma_{-i})) = 0$.[2]

[2] If M or the strategies of the other players are probabilistic, then the above utility for player i equals 0 with probability 1.

Let us now proceed to define a notion fundamental to rational robustness.

Definition 2. (Distinguishable Strategies.) *In a game G, let S be a profile of deterministic-strategy subsets and let σ_i and σ'_i be two different strategies for some 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*

$$H(\sigma_i \sqcup \tau_{-i}) \neq H(\sigma'_i \sqcup \tau_{-i}).^{3}$$

If this is the case, we say that τ_{-i} distinguishes *σ_i and σ'_i over S; else, that σ_i and σ'_i are* equivalent *over S.*

Note that, in all definitions of this section, we might as well assume that σ_i and σ'_i belong to S_i. In the proofs of our theorems, however, we shall need more generality. (Namely, we need to consider strategies σ_i and σ'_i that are distinguishable over S, while only one of them belongs to S_i.)

Note too that, if G is of normal-form, then, as long as each S_j, for $j \neq i$, is non-empty, any pair of different strategies of player i are distinguishable over S. (In fact, by definition, in a normal-form game the history of a strategy profile σ coincides with σ itself, so that any two different strategy profiles have different histories.) Therefore, the notion of distinguishable strategies can meaningfully come into play only for extensive-form mechanisms. This is indeed in accordance with our prior claim that mechanisms of extensive form are ideally suited for protecting mechanism design from the problems of collusion, complexity, and privacy.

We leverage our notion of distinguishable strategies in order to bridge the currently vast gap between strict and weak domination. (The following notion of "distinguishable domination" is again definable for all games, but really meaningful only for games of extensive form.)

Definition 3. (Distinguishably Dominated Strategies.) *Let $G = (C, M)$ be a game, i a player, σ_i and σ'_i two strategies of i, and S a profile of deterministic strategy subsets. We say that σ_i is* distinguishably dominated *(by σ'_i) over S —equivalently that σ'_i* distinguishably dominates *σ_i over S— if*

1. *σ_i and σ'_i are distinguishable over S; and*
2. *$\mathbb{E}[u_i(M(\sigma_i \sqcup \tau_{-i}))] < \mathbb{E}[u_i(M(\sigma'_i \sqcup \tau_{-i}))]$ for all sub-profiles τ_{-i} distinguishing σ_i and σ'_i over S.*

Notation For short,

- We refer to a distinguishably dominated strategy as a DD strategy.
- We write "$\sigma_i \simeq \sigma'_i$ over S" or "$\sigma_i \simeq_S \sigma'_i$" to denote that σ_i and σ'_i are equivalent over S.
- We write "$\sigma_i \prec \sigma'_i$ over S" or "$\sigma_i \prec_S \sigma'_i$" to denote that σ_i is DD by σ'_i over S.
- We write "$\sigma_i \preceq \sigma'_i$ over S" or "$\sigma_i \preceq_S \sigma'_i$" to denote that either $\sigma_i \prec \sigma'_i$ over S or $\sigma_i \simeq \sigma'_i$ over S.

Definition 4. (Iterative Elimination of DD Strategies.) *In a game G, let $\Sigma = \Sigma^0, \ldots, \Sigma^K$ be a sequence of profiles of deterministic-strategy subsets such that, for each $k < K$, there exists at least one player i such that (a) Σ_i^{k+1} is a proper subset of Σ_i^k, and (b) Σ_j^{k+1} is a subset of Σ_j^k for all $j \neq i$.*

We say that Σ is an iterative elimination of DD strategies *(IEDD for short) if for each player i and any strategy $\sigma_i \in \Sigma_i^k \setminus \Sigma_i^{k+1}$, there exists a strategy $\sigma'_i \in \Sigma_i^{k+1}$ such that $\sigma_i \preceq \sigma'_i$ over Σ^k.*

We say that an IEDD Σ is full *if, for each player i, (1) Σ_i^0 is the set of all strategies of i in G and (2) $\sigma_i \not\preceq \sigma'_i$ over Σ^K for all pairs of strateigs $\sigma_i, \sigma'_i \in \Sigma_i^K$. If $\Sigma = \Sigma^1, \ldots, \Sigma^K$ is a full IEDD, we refer to Σ^K as the* terminal set *of Σ.*

The above definition allows Σ^{k+1} to be generated from Σ^k by "simultaneously" eliminating multiple strategies of multiple players. Notice that this does not generate "any problem" in the sense that , as part of the proof of our first theorem, we show that Σ^{k+1} can always be generated from Σ^k by a *sequence* of elementary steps, in each of which only a single strategy is eliminated.[4]

[3] If $H(\sigma_i \sqcup \tau_{-i})$ and $H(\sigma'_i \sqcup \tau_{-i})$ are distributions over the histories of G, then the inequality means that the two distributions are different.

[4] In principle problems may arise in multiple ways. For instance, when a player i eliminates σ_i because $\sigma_i \prec \sigma'_i$ over Σ^k and there exists $\tau_{-i} \in \overline{\Sigma_{-i}^k}$ distinguishing the two, while another player j simultaneously eliminates τ_j.

Definition 5. (Rationally Robust Solutions, Strategies and Plays) *In a game G, we say that a profile $\mathcal{S}$ of strategy subsets is a* rationally robust solution *for G if there exists a full IEDD $\Sigma = \Sigma^0, \ldots, \Sigma^K$ such that $\mathcal{S} = \Sigma^K$.*

If this is the case, we refer to each $\sigma_i \in \Sigma_i^K$ as a rationally robust strategy *(over Σ), and to each profile $\sigma \in \overline{\Sigma^K}$ as a* rationally robust play *(over Σ).*

Definition 6. (Rationally Robust Implementation) *Let $\mathscr{C}$ be a class of PKI contexts, P a property over (distributions of) outcomes of contexts in $\mathscr{C}$, and M an extensive-form mechanism with simultaneous and public actions. We say that M is a* rationally robust implementation of P over $\mathscr{C}$ *if, for any $C \in \mathscr{C}$, there exists a rationally robust solution $\mathcal{S}$ for the game (C, M) such that:*

1. *for each player i, $\text{OUT}_i \notin \mathcal{S}_i$; and*
2. *for all plays $\sigma \in \overline{\mathcal{S}}$, P holds for $M(\sigma)$.*

Recall that each mechanism must satisfy the opt-out condition, and thus each player i will not be afraid to participate in the above game (C, M), because he always has the opportunity of playing his own opt-out strategy OUT_i and receive 0 utility. However, if OUT_j were rationally robust for some player j, then M should guarantee property P when j chooses to opt out. In an auction, depending on the property, this might be still possible, because auction outcomes are "separable," that is, it is possible to comply with player j choice to opt out by assigning him no goods and ask him to pay nothing, while "running the auctions for the remaining players. However, in general the only outcome in which a player j has zero utility when he chooses to opt out consists of forcing the "empty outcome" (in essence to cancel the game), making it impossible to guarantee any non-trivial property.

Requiring that no opt-out strategy is rationally robust implies that each player has a rationally robust strategy which is "safe" for him, that is ensuring him a non-negative utility, as long as all other players choose their rationally robust strategies. More precisely,

Proposition 1. *Let $\mathcal{S}$ be a rationally robust solution of a game (C, M). For any player i, if $\text{OUT}_i \notin \mathcal{S}_i$, then there exists $\tau_i \in \mathcal{S}_i$ such that $\mathbb{E}[u_i(M(\tau_i \sqcup \sigma_{-i}))] \geq 0$ for each subprofile $\sigma_{-i} \in \overline{\mathcal{S}_{-i}}$.*

Proof Sketch. Since $\mathcal{S}$ is the terminal set of a full IEDD $(\mathcal{S}^0, \ldots, \mathcal{S}^K)$ and $\text{OUT}_i \notin \mathcal{S}_i$, there must exists $k < K$ such that $\text{OUT}_i \in \mathcal{S}_i^k$ but $\text{OUT}_i \notin \mathcal{S}_i^{k+1}$. Accordingly, there exists $\tau_i^{k+1} \in \mathcal{S}_i^{k+1}$ such that $\text{OUT}_i \preceq \tau_i^{k+1}$ over $\mathcal{S}_i^k$, which implies that for any $\sigma_{-i}^{k+1} \in \overline{\mathcal{S}_{-i}^{k+1}}$, $\mathbb{E}[u_i(M(\tau_i^{k+1} \sqcup \sigma_{-i}^{k+1}))] \geq 0$. Thus if $\tau_i^{k+1} \in \mathcal{S}_i$, then setting $\tau_i = \tau_i^{k+1}$ satisfies the proposition. Otherwise, our game being finite, we can trace the "killer" of strategy τ_i^{k+1}, the "killer of the killer", and so on, until we reach the desired τ_i. ■

The above solution concept and implementation notion are certainly robust "in name". Let us now argue that they are also robust "in fact." To this end, our first theorem guarantees that full iterated elimination of DD strategies is *order independent.* That is, the terminal set of every possible full IEDD is essentially the same: not only will each player i end up with the same number of rationally robust strategies, but all sets of such strategies will be equivalent to one another, in the sense that they yield the same histories, and thus the same outcomes. More formally,

Theorem 1. *Let $G = (C, M)$ be a game, and $\mathcal{S} = (\mathcal{S}^0, \ldots, \mathcal{S}^K)$ and $\mathcal{T} = (\mathcal{T}^0, \ldots, \mathcal{T}^L)$ be two full IEDDs. Then there exists a profile ϕ such that*

- *for each player i, ϕ_i is a bijection from $\mathcal{S}_i^K$ to $\mathcal{T}_i^L$; and*
- *for any strategy profile $\sigma \in \overline{\mathcal{S}^K}$, defining $\phi_i(\sigma)$ to be $\phi_i(\sigma_i)$, we have $H(\phi(\sigma)) = H(\sigma)$.*

Proof. See Appendix A.

Theorem 1 implies that if a mechanism robustly implements a property P, then, for all possible IEDDs Σ, as long as all players choose to eliminate DD strategies in the order dictated by Σ, P will hold no matter which profile of surviving strategies is played out. But: *why should all players choose the same elimination order?*

Common knowledge of rationality is one thing, and common knowledge of elimination order a different thing altogether. To us, a notion of implementation is robust only if it does not rely on anything beyond common knowledge of rationality. Thus to demonstrate the robustness "in fact" of our implementation notion, we prove that, as long as a mechanism rationally robustly implements a property P, P will hold even when each player chooses his own full IEDD independently of all other players. More formally,

Theorem 2. *Let M be a rationally robust implementation of a property P over a class of PKI contexts $\mathscr{C}$, C a context in $\mathscr{C}$, and $\mathcal{S}^1, \ldots \mathcal{S}^n$ rationally robust solutions of the game (C, M). Then $\forall$ plays $\sigma \in \mathcal{S}^1_1 \times \cdots \times \mathcal{S}^n_n$,*

P holds for $M(\sigma)$.

Proof. See Appendix B.

3 Comparisons with Related Prior Notions

Let us compare rationally robust implementation with prior solution concepts in the current setting, that is, independent players with perfect knowledge. To this end, we consider separately the case of normal-form and extensive-form mechanisms, and make use of the following notation.

Notation In any game G,

- If S is a set of strategy profiles, then $H(S)$ denotes the set of histories $\{H(\sigma) : \sigma \in S\}$, and $\overline{S}$ denotes the Cartesian closure of S, that is $\overline{S} = S_1 \times \cdots \times S_n$, where $S_i = \{\sigma_i : \sigma \in S\}$.
(That is, we use the same symbol for the Cartesian closure of a profile of strategy subsets. Thus if X is such a profile, then $\overline{\overline{X}} = \overline{X}$.)
- $\mathscr{R}$ will consistently denote the closure of a rationally robust solution,
$\mathscr{S}$ the set of strategy profiles surviving iterative elimination of strictly dominated pure strategies,
$\mathscr{E}$ the set of all pure Nash equilibria, and
$\mathscr{PE}$ the set of all subgame-perfect equilibria (if G is of extensive form).

As noted in the second paragraph after Definition 2, in a game of normal form, any two different strategies σ_i and σ'_i of player i are distinguishable over any profile S of strategy subsets. Therefore $\sigma_i \preceq \sigma'_i$ over S if and only if $\sigma_i \prec \sigma'_i$ over S, which itself occurs if and only if σ_i is strictly dominated by σ'_i over S. Accordingly, any full IEDD is a (complete) iterative elimination of strictly dominant strategies. Thus, because the later notion is well known to be order independent and preserve all Nash equilibira, the following statement trivially holds.

Fact 1 *For normal-form games, $\mathscr{R} = \mathscr{S}$ and thus $\mathscr{R} \supset \mathscr{E}$ and $\mathscr{R} \supset \overline{\mathscr{E}}$.*

Since our notions can be defined relative to mixed strategies, analogous statements hold for such strategies. Let us now turn our attention to extensive-form games.

Fact 2 *For all extensive-form games, $\mathscr{PE} \subset \overline{\mathscr{PE}} \subset \mathscr{R} \subset \mathscr{S}$, and all inclusions are strict for some games.*

The latter statement, of course, holds "up to equivalent strategies." But in extensive-form games, different strategy profiles may yield the same *history.* Thus a stronger and more meaningful version of the last statement is as follows.

Fact 2′ *For all extensive-form games, $H(\mathscr{PE}) \subset H(\overline{\mathscr{PE}}) \subset H(\mathscr{R}) \subset H(\mathscr{S})$, and all inclusions are strict for some games.*

(An example where the second inclusion is strict is given in Appendix C. All other inclusions are trivial.)

Note that Fact 2′ is *absolute*, that is, it no longer requires the qualification "up to equivalent strategies." Since in games of extensive form different histories have different outcomes, Fact 2′ also shows our notion of *implementation* to be distinct from both subgame-perfect implementation and implementation in iteratively undominated strategies.

4 Independent Players with Perfect Knowledge but Less Rationality

Mechanism design must leverage the players' rationality. But if we want to maximize the relevance of implementation *theory,* we must be aware that in *practice* not all players are perfectly rational. Of course there is little that a mechanism can achieve if the players are dumb, but we must consider a mechanism M to be preferable to a mechanism M' if "it achieves the same property while requiring less rationality from the players."

Accordingly, in Appendix D we present a "hierarchy" of implementation notions, that starts from rationally robust implementation and requires less and less rationality.

5 Independent Players with Imperfect Knowledge

We now extend rationally robust implementation to a more realistic player model. The precise details of this model are not crucial, but we do need a model in order to present our notion formally.

Definition 7. *A context with imperfectly knowledgable and independent players (an IKI context for short) is a quadruple $C = (N, \Omega, u, K)$ where*

- *N is as usual the set of players;*
- *Ω is as usual the set of possible outcomes;*
- *u is as usual the profile of utility functions; and*
- *K is the* external knowledge *profile, where each K_i is a set of utility-functions subprofiles (for the players other than i) such that $u_{-i} \in K_i$.*

As usual, N and Ω are assumed to be common knowledge. In addition, each player i knows u_i and K_i (and no other information). We may refer to u_i also as i's internal knowledge*.*

Remarks

- We again prefer to assume that all knowledge lies with the players, and thus a mechanism designer knows nothing about the profiles u and K.
- For simplicity, the contexts defined above do not include "knowledge about other players' knowledge." (Such knowledge may however be properly utilized within our notions.)
- K_i, the external knowledge of player i, consists of all information that i has about the utilities of the other players. For simplicity, we define it above as a set guaranteed to include the true utility functions of the other players. More generally, K_i could be a distribution whose support includes u_{-i}. More generally yet, K_i may be a "partial" probability distribution.[5] Such generality will unnecessarily complicate the simple way in which we intend to derive our imperfect-knowledge notions from our perfect-knowledge ones.

Intuition When the players are perfectly informed, the statements and proofs of our theorems and lemmas in particular imply that a full IEDD can be computed by the following sequence of "big steps." First, the players simultaneously eliminate all their DD strategies, over the set of all possible strategies. Then, they simultaneously eliminate all their DD strategies, over all strategies surviving the previous step. And so on, until no strategy can be eliminated any more.

Let us now see what happens when one modifies the knowledge of the players, until we reach our desired knowledge setting. For simplicity, we let there be just two players: Alice and Bob.

1. *No External Knowledge.* Assume that neither Alice nor Bob has any knowledge about the other's utility function. Then, the best Alice can do is to eliminate all her DD strategies, assuming the full

[5] That is, starting with a distribution assigning a probability $p(u'_{-i})$ to each possible utility-function subprofile u'_{-i}, K_i may replace each $p(u'_{-i})$ with a subinterval of $[0, 1]$ including $p(u'_{-i})$.

strategy set for Bob. In fact, in order to eliminate any strategy for Bob, she must have at least some information about his utilities. Thus she cannot "iterate" and thus eliminate any further strategy of hers. Symmetrically, Bob can eliminate all his DD strategies, assuming the full strategy set for Alice, and nothing more. Although each one has acted separately, from a "global" point of view, they have performed the first "big step" of the above IEDD.

2. *Full, But Not Common, Knowledge.* Now assume that each of Alice and Bob knows exactly the other's utilities, but that this is not common knowledge. In fact, assume that Alice knows nothing about Bob's knowledge, and viceversa. Then, from a global point of view, they can perform exactly the first two big steps, but cannot eliminate any other strategy.
3. *Our External Knowledge.* Finally assume, as we do, that Alice has some knowledge about Bob's utility function, but no knowledge about Bob's knowledge. And viceversa. Then, they will end up "between Big Step 1 and Big Step 2. Namely, the best Alice can do is the following. First, pretending to be in Bob's shoes, she eliminates as many DD strategies for Bob as possible, based on her knowledge about Bob's utilities, *over the full set of strategies for herself.* (The latter restriction is necessary because, having no knowledge about Bob's knowledge, to be safe she must assume that Bob knows nothing about her!) Second, she eliminates all possible strategies for her, over what is left in Bob's strategy set.
   Bob will eliminates his own strategies in a symmetric way.
   That is all they can safely eliminate.

All is left is to explain what is meant by

> "pretending to be in Bob's shoes, she eliminates as many DD strategies for Bob as possible, based on her knowledge about Bob's utilities, over the full set of strategies for herself."

Essentially, this means that Alice considers, one by one, all possible utility functions for Bob in her external knowledge set K_A. For each $u'_B \in K_A$, pretending for a moment that it is Bob's true utility function, Alice eliminates all of Bob's DD strategies over her full strategy set, and thus computes the corresponding set of surviving strategies for Bob: $S_{u'_B}$. Then, she computes S_B, her best prediction for Bob's "truly surviving set", by taking the union of all $S_{u'_B}$. (Notice that S_B is a conservative choice, in the sense that it is guarantee to contain all strategies for Bob after *he himself* eliminates all his own DD strategies!) Finally, Alice eliminates all her own DD strategies over S_B.

Let us now proceed more formally.

Definition 8. (Compatible Contexts.) *We say that a context $C' = (N', \Omega', u', K')$ is* compatible with player *i in a game $G = (C, M)$ if, letting $C = (N, \Omega, u, K)$, we have $N' = N$, $\Omega' = \Omega$, $u_i = u'_i$ and $K_i = K'_i$.*

Thus if i is a player in a context C, then he knows nothing about C besides the set of players and the set of outcomes, N and Ω, and his own internal and external knowledge, u_i and K_i. Thus from his own perspective, every context C' compatible with him is equally likely. Such a C' is any context whose players, outcomes and his own utility function is as i knows them to be, and whose subprofile of utility functions for the other players belongs to K_i.

Recall that a mechanism consists of the strategies available to the players, and of the outcome function mapping strategy profiles to outcomes. Now assume that a designer comes up with a mechanism M for an IKI context C. Then, because he has no information about the players internal and external knowledge, the strategies envisaged by M can at most depend on the set of players N and the set of outcomes Ω. Following tradition, and without ambiguity, we denote by Σ^0_i the set of all strategies available to i, and by Σ^0 the corresponding profile. In particular, therefore, Σ^0 will be the same for the game (C, M) as for any other game (C', M) whose context C' is compatible with some player $i \in N$.

But if the profile Σ^0 is the same for all possible compatible contexts, the same cannot be said for the set of DD strategies over Σ^0. Indeed, distinguishable domination depends on the players utilities, and the player utilities of two contexts compatible with player i can be different for any player other than i. We must therefore always specify the context of interest when speaking of DD strategies.

Definition 9. (Level-1 Rationally Robust Plays) *Let M be a mechanism for a set of players N and a set of outcomes Ω. Then for any IKI context C for M,*

- *We denote by Σ^1_C the profile of strategy subsets such that each $\Sigma^1_{C,i}$ consists of all strategies in Σ^0_i that are not distinguishably dominated over Σ^0 in the game (C, M).*
- *We say that a strategy $\sigma_i \in \Sigma^1_{C,i}$ is* level-1 distinguishably dominated *if there is a strategy $\sigma'_i \in \Sigma^1_{C,i}$, such that σ'_i distinguishably dominates σ_i over $\Sigma^1_{C'}$ for all contexts C' compatible with i.*
- *We denote by $\Sigma^2_{C,i}$ the set of all strategies in $\Sigma^1_{C,i}$ that are* not level-1 distinguishably dominated.
- *We say that a strategy vector σ is an* level-1 rationally robust play *of the game (C, M) if $\sigma_i \in \Sigma^2_{C,i}$ for all player i.*

Definition 10. (Level-1 Rationally Robust Implementation.) *Let $\mathscr{C}$ be a class of IKI contexts, P be a property over (distributions of) outcomes of contexts in $\mathscr{C}$, and M an extensive-form mechanism with simultaneous and public actions. We say that M* level-1 rationally robustly implements *P if, for all contexts $C \in \mathscr{C}$*

1. *for each player i, $\text{OUT}_i \notin \Sigma^2_{C,i}$; and*
2. *for all level-1 rationally robust plays σ of the game (C, M), P holds for $M(\sigma)$.*

Above, we have extensively used the label "level-1" to emphasize how our notion depend on the fact that the external knowledge of our players is defined only at the first level. This labeling may make it easier for someone to extend our notions to contexts whose players have "knowledge about knowledge." Such generalization, however, will make our notions sufficiently heavier without —in our opinion— enhancing enough their practical applicability.

The rationale for requiring the opt-out strategy of player i not to belong to $\Sigma^2_{C,i}$ is similar to that for rationally robust implementation. Again, this requirement implies that there exists a strategy in $\Sigma^2_{C,i}$ which ensures player i a non-negative utility, as long as each other player j chooses his strategy from $\Sigma^1_{C,j}$.

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] M. Blum, A. De Santis, S. Micali, G. Persiano. Non-Interactive Zero Knowledge. *SIAM Journal on Computing*, Vol.20, No.6, pages 1084-1118, 1991.

[3] J. Chen, A. Hassidim, and S. Micali. Generating Perfect Revenue from Perfectly Informed Players. To appear in ICS2010.

[4] J. Chen and S. Micali. A New Approach to Auctions and Resilient Mechanism Design. *Symposium on Theory of Computing*, pages 503-512, 2009. Full version available at http://people.csail.mit.edu/silvio/Selected Scientific Papers/Mechanism Design/.

[5] J. Chen, S. Micali, and P. Valiant. Levaraging Collusion in Combinatorial Auctions. To appear in ICS2010.

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

[7] S. Goldwasser, S. Micali, and C. Rackoff. The Knowledge Complexity of Interactive Proof Systems. *SIAM Journal of Computing*, Vol.18, No.1, pages 186-208, Feb. 1989.

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

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

[10] M. Jackson. Implementation in Undominated Strategies: a Look at Bounded Mechanisms. *Review of Economic Studies*, Vol. 59, No. 201, pages 757-775, Oct., 1992.

[11] M. Jackson. Mechanism Theory. In U. Derigs ed.: *The Encyclopedia of Life Support Systems*, EOLSS Publishers: Oxford UK, 2003.

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

[13] E. Maskin. Nash Equilibrium and Welfare Optimality. *Review of Economic Studies*, Vol.66, No.1, pages 23-38, 1999.

[14] R. Myerson. Incentive Compatibility and the Bargaining Problem. *Econometrica*, Vol.47, No.1, pages 61-74, 1979.

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

[16] N. Nisan. Introduction to Mechanism Design (for Computer Scientists). In N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani eds.: *Algorithmic Game Theory*. Cambridge Univ. Press, 2007.

[17] M.J. Osborne and A. Rubinstein. *A Course in Game Theory*, MIT Press, 1994.

[18] 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.

Appendix

A The Proof of Theorem 1

Definition 11. *We define three binary relations among profiles of strategy subsets, the* elimination relation $\xrightarrow{e}$*, the* replacement relation $\xrightarrow{r}$*, and the* idle relation $\xrightarrow{i}$*, as follows. Let* $S = \prod_i S_i$ *and* $T = \prod_i T_i$ *be two profiles of strategy subsets. We write*

- $S \xrightarrow{e} T$*, if there exists a player* i *such that*
  (1) $T_j = S_j$ *for all* $j \neq i$*; and*
  (2) $S_i = T_i \cup \{\sigma_i\}$*, where* $\sigma_i \in S_i \setminus T_i$ *and* $\sigma_i \preceq_S \tau_i$ *for some* $\tau_i \in S_i$*.*
- $S \xrightarrow{r} T$*, if there exists a player* i *such that*
  (1) $T_j = S_j$ *for all* $j \neq i$*; and*
  (2) $S_i \setminus T_i = \{\sigma_i\}$ *and* $T_i \setminus S_i = \{\tau_i\}$*, where* σ_i *and* τ_i *are equivalent over* S*.*
- $S \xrightarrow{i} T$*, if* $S = T$*.*

Remark Notice that

- $S \xrightarrow{e} T$ if T is obtained from S by eliminating a single strategy σ_i such that $\sigma_i \preceq_S \tau_i$ for some $\tau_i \in S_i$.
- $S \xrightarrow{r} T$ if T is obtained from S by replacing a single strategy σ_i with an equivalent strategy τ_i.
- $S \xrightarrow{i} T$ if T is obtained from S by "doing nothing".

Definition 12. *If, in some game* G*,* $S \xrightarrow{r} T$*, then the* natural bijection $\phi : S \to T$ *is the profile of bijections defined as follows:*

- *For any player* j *such that* $S_j = T_j$*,* $\phi_j : S_j \to T_j$ *is the identity map.*
- *If* i *is the single player such that* $S_i \setminus T_i = \{\sigma_i\}$ *and* $T_i \setminus S_i = \{\tau_i\}$ *then* $\phi_i : S_i \to T_i$ *is the identity map every where, except that* $\phi_i(\sigma_i) = \tau_i$*.*
- *For any strategy profile* $\sigma \in S$*,* $\phi(\sigma) = (\phi_1(\sigma_1), \ldots, \phi_n(\sigma_n))$*.*

Lemma 1. *For any game* G*, if* $S \xrightarrow{r} T$ *and* ϕ *is the natural bijection from* S *to* T*, then*

1. $T \xrightarrow{r} S$*;*
2. ψ*, the natural bijection from* T *to* S*, equals* ϕ^{-1}*, where* ϕ^{-1} *is defined to be* $(\phi_1^{-1}, \ldots, \phi_n^{-1})$*; and*
3. $H(\phi(\sigma)) = H(\sigma)$ *for any strategy profile* $\sigma \in S$*.*

Proof. The proof is quite trivial. Indeed, let i be the single player such that $S_i \setminus T_i = \{\sigma_i\}$ and $T_i \setminus S_i = \{\tau_i\}$. We have that τ_i and σ_i are equivalent over T also, and S is obtained from T by replacing τ_i with σ_i. Moreover, for any strategy profile $\mu \in S$, if $\mu_i \neq \sigma_i$, then $\mu = \phi(\mu)$, and thus $H(\mu) = H(\phi(\mu))$; if $\mu_i = \sigma_i$, then $\phi(\mu) = \tau_i \sqcup \mu_{-i}$, and again $H(\mu) = H(\phi(\mu))$, since σ_i and τ_i are equivalent over S. *Q.E.D.*

Lemma 2. (Interchangeability of Replacements) *In any game* G*, if* $R \xrightarrow{r} S$ *by replacing* σ_i *with* σ'_i*, and* $S \xrightarrow{r} T$ *by replacing* τ_j *with* τ'_j*, then either*

(1) $i = j$*,* $\sigma'_i = \tau_j$*, and* $\sigma_i = \tau'_j$*, in which case* $R \xrightarrow{i} T$ *(i.e., "nothing done"); or*

(2) $i = j$*,* $\sigma'_i = \tau_j$*, and* $\sigma_i \neq \tau'_j$*, in which case* $R \xrightarrow{r} T$ *by replacing* σ_i *with* τ'_j *(i.e., "shortcut"); or*

(3) $i = j$*,* $\sigma'_i \neq \tau_j$*, and* $\sigma_i = \tau'_j$*, in which case* $R \xrightarrow{r} T$ *by replacing* τ_j *with* σ'_i *(i.e., "shortcut"); or*

(4) $i = j$*,* $\sigma'_i \neq \tau_j$*, and* $\sigma_i \neq \tau'_j$*, in which case* $R \xrightarrow{r} W$ *by replacing* τ_j *with* τ'_j*, and* $W \xrightarrow{r} T$ *by replacing* σ_i *with* σ'_i *(i.e., "switch").*

(5) $i \neq j$*, in which case* $R \xrightarrow{r} W$ *by replacing* τ_j *with* τ'_j*, and* $W \xrightarrow{r} T$ *by replacing* σ_i *with* σ'_i *(i.e., "switch").*

Proof. The proofs of (1) — (4) are absolutely trivial. The proof of (5) includes two steps.

Step 1. We prove that $R \xrightarrow{r} W$ by replacing τ_j with τ'_j.

It is easy to see that $\tau_j \in R_j$ (as $R_j = S_j$), and thus it suffices to show that $\tau_j \simeq_R \tau'_j$, that is, $H(\tau_j \sqcup \mu_{-j}) = H(\tau'_j \sqcup \mu_{-j})\ \forall \mu_{-j} \in R_{-j}$.

If $\mu_i \neq \sigma_i$, then $\mu_{-j} \in S_{-j}$ also, and we immediately have $H(\tau_j \sqcup \mu_{-j}) = H(\tau'_j \sqcup \mu_{-j})$, since $\tau_j \simeq_S \tau'_j$.

If $\mu_i = \sigma_i$, then we have

$$H(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}) = H(\tau_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}) = H(\tau'_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}),$$

where the first equality is because $\sigma_i \simeq_R \sigma'_i$, and the second is because $\tau_j \simeq_S \tau'_j$. Let $D^h, \ldots, D^0$ be the decision nodes in history $H(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$, from height h to height 0 respectively. We have that: (a) D^h is the root of the decision tree; (b) for any ℓ such that $i \in N^{D^\ell}$, σ_i and σ'_i coincide at node D^ℓ, that is, player i chooses the same action at D^ℓ according to these two strategies; and (c) for any k such that $j \in N^{D^k}$, τ_j and τ'_j coincide at node D^k. (b) and (c) are because the above two equalities respectively.

Consider the execution of $\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}$. We show that $D^h, \ldots, D^0$ are the decision nodes in $H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$, by induction. First, the execution starts at node D^h, since this is the root of the decision tree. Assume $D^h, \ldots, D^\ell$ are the decision nodes in $H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$. For any player $k \in N^{D^\ell}$, if $k = j$, then player k chooses his action according to τ'_j, but this is equivalent to say that he chooses his action according to τ_j, by (c); if $k = i$, then k chooses his action according to σ_i; if $k \neq i, j$, then k chooses his action according to μ_k. That is, all players in N^{D^ℓ} choose the same actions as in execution $\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}$, and thus the decision node in height $\ell - 1$ reached is execution $\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}$ is $D^{\ell-1}$. In sum, $D^h, \ldots, D^0$ are the decision nodes in $H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$, which implies that $H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}) = H(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$.

Step 2. We prove that $W \xrightarrow{r} T$ by replacing σ_i with σ'_i.

Again it is easy to see that $\sigma_i \in W_i$, and thus it suffices to show that $\sigma_i \simeq_W \sigma'_i$, that is, $H(\sigma_i \sqcup \mu_{-i}) = H(\sigma'_i \sqcup \mu_{-i})\ \forall \mu_{-i} \in W_{-i}$.

If $\mu_j \neq \tau'_j$, then the equality follows trivially.

If $\mu_j = \tau'_j$, then we need to show that $H(\sigma_i \sqcup \tau'_j \sqcup \mu_{-\{i,j\}}) = H(\sigma'_i \sqcup \tau'_j \sqcup \mu_{-\{i,j\}})$. From Step 1, we have that both two histories equal $H(\sigma_i \sqcup \tau_j \sqcup \mu_{-\{i,j\}})$, and we are done.

Q.E.D.

Lemma 3. (Preponement of Elimination) *In any game G, if $R \xrightarrow{r} S$ and $S \xrightarrow{e} T$, then either*
(1) $R \xrightarrow{e} T$ (i.e., "shortcut"); or
(2) there exists W such that $R \xrightarrow{e} W$ and $W \xrightarrow{r} T$ (i.e., "switch").

Proof. Let $R \xrightarrow{r} S$ by replacing σ_i with σ'_i, and $S \xrightarrow{e} T$ by eliminating τ_j because $\tau_j \preceq_S \tau'_j$.

If $i = j$ and $\sigma'_i = \tau_j$, then $R \xrightarrow{e} T$ by eliminating σ_i because $\sigma_i \preceq_R \tau'_j$, trivial.

If $i = j$ and $\sigma'_i = \tau'_j$, then $R \xrightarrow{e} W$ by eliminating τ_j because $\tau_j \preceq_R \sigma_i$, and $W \xrightarrow{r} T$ by replacing σ_i with σ'_i, trivial.

If $i = j$, $\sigma'_i \neq \tau_j$, and $\sigma'_i \neq \tau'_j$, then $R \xrightarrow{e} W$ by eliminating τ_j because $\tau_j \preceq_R \tau'_j$, and $W \xrightarrow{r} T$ by replacing σ_i with σ'_i, trivial.

Else, $i \neq j$. We first show that $R \xrightarrow{e} W$ by eliminating τ_j because $\tau_j \preceq_R \tau'_j$, and then $W \xrightarrow{r} T$ by replacing σ_i with σ'_i. First, it is easy to see that $\tau_j, \tau'_j \in R_j$, since $R_j = S_j$. Let μ_{-j} be an arbitrary strategy subprofile in R_{-j}. If $\mu_i \neq \sigma_i$, then $\mu_{-j} \in S_{-j}$ also, and thus $\mathbb{E}[u_j(M(\tau_j \sqcup \mu_{-j}))] < \mathbb{E}[u_j(M(\tau'_j \sqcup \mu_{-j}))]$ whenever $H(\tau_j \sqcup \mu_{-j}) \neq H(\tau'_j \sqcup \mu_{-j})$, since $\tau_j \preceq_S \tau'_j$. If $\mu_i = \sigma_i$, then

$$H(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}) = H(\tau_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}) \quad \text{and} \quad H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}) = H(\tau'_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}),$$

since $\sigma_i \simeq_R \sigma'_i$. Because $\sigma'_i \sqcup \mu_{-\{i,j\}} \in S_{-j}$, we have that

$$\mathbb{E}[u_j(M(\tau_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}))] < \mathbb{E}[u_j(M(\tau'_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}))]$$

whenever $H(\tau_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}}) \neq H(\tau'_j \sqcup \sigma'_i \sqcup \mu_{-\{i,j\}})$. Accordingly, we have that

$$\mathbb{E}[u_j(M(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}))] < \mathbb{E}[u_j(M(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}))]$$

whenever $H(\tau_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}}) \neq H(\tau'_j \sqcup \sigma_i \sqcup \mu_{-\{i,j\}})$. In sum, $\tau_j \preceq_R \tau'_j$, and $R \xrightarrow{e} W$ by eliminating τ_j. The second part (i.e., $W \xrightarrow{r} T$ by replacing σ_i with σ'_i) is trivial.
Q.E.D.

Theorem 1. *Let $G = (C, M)$ be a game, and $\mathcal{S} = (\mathcal{S}^0, \ldots, \mathcal{S}^K)$ and $\mathcal{T} = (\mathcal{T}^0, \ldots, \mathcal{T}^L)$ be two full IEDDs. Then there exists a profile ϕ such that*

- *for each player i, ϕ_i is a bijection from $\mathcal{S}_i^K$ to $\mathcal{T}_i^L$; and*
- *for any strategy profile $\sigma \in \mathcal{S}^K$, defining $\phi_i(\sigma)$ to be $\phi_i(\sigma_i)$, we have $H(\phi(\sigma)) = H(\sigma)$.*

Proof. The proof consists of proofs of several claims, which follow easily from the above lemmas.

Claim 1. *Let the sequence S, T be an IEDD, then there exists another IEDD $X^0, \ldots, X^L$ such that (1) $X^0 = S$; (2) $X^L = T$; and (3) $X^\ell \xrightarrow{e} X^{\ell+1}$ for all $\ell < L$.*

Proof. The proof is quite trivial. Notice that if $\sigma_i \preceq_S \sigma'_i$, then $\sigma_i \preceq_{S'} \sigma'_i$ for any S' such that $S'_j \subseteq S_j \forall j$. Therefore the strategies (even belonging to different players) eliminated simultaneously when generating T from S can be eliminated one by one and in any order. ■

According to Claim 1, without loss of generality, we assume that $\mathcal{S}^k \xrightarrow{e} \mathcal{S}^{k+1}$ for all $k < K$, and that $\mathcal{T}^\ell \xrightarrow{e} \mathcal{T}^{\ell+1}$ for all $\ell < L$, because the sequences $\mathcal{S}^k, \mathcal{S}^{k+1}$ and $\mathcal{T}^\ell, \mathcal{T}^{\ell+1}$ are both IEDDs. Because for each player i, $\mathcal{S}_i^0$ and $\mathcal{T}_i^0$ are both the set of all strategies of i in G, we have that $\mathcal{S}^0 = \mathcal{T}^0 \triangleq R$, and thus the following diagram:

$$\begin{tikzcd}
 & & R(= \mathcal{S}^0 = \mathcal{T}^0) \arrow[dl, "e"'] \arrow[dr, "e"] & & \\
 & \mathcal{S}^1 \arrow[dl, dashed, "e"'] & & \mathcal{T}^1 \arrow[dr, dashed, "e"] & \\
\mathcal{S}^K & & & & \mathcal{T}^L
\end{tikzcd}$$

Claim 2. *Let R, S, and T be three profiles of strategy subsets. If $R \xrightarrow{b_1} S$ and $R \xrightarrow{b_2} T$ with $b_1, b_2 \in \{e, r, i\}$, then there exists W such that $S \xrightarrow{c_1} W$ and $T \xrightarrow{c_2} W$ with $c_1, c_2 \in \{e, r, i\}$. In other words, we have the following diagram:*

$$\begin{tikzcd}
 & R \arrow[dl, "b_1"'] \arrow[dr, "b_2"] & \\
S \arrow[dr, "c_1"'] & & T \arrow[dl, "c_2"] \\
 & W &
\end{tikzcd}$$

Proof. The proof consists of complicated case analysis for b_1 and b_2, however, each case can be verified easily based on our lemmas. We only provide an example to show how such a case analysis looks like.

Assume that $R \xrightarrow{e} S$ and $R \xrightarrow{r} T$. Then by Lemma 1, we have that $T \xrightarrow{r} R$. By Lemma 3, we have that either (a) $T \xrightarrow{e} S$ or (b) $T \xrightarrow{e} W$ and $W \xrightarrow{r} S$. For case (a), letting $W = S$, we have that $S \xrightarrow{i} W$ and $T \xrightarrow{e} W$; for case (b), by Lemma 1 again, we have that $S \xrightarrow{r} W$, which together with $T \xrightarrow{e} W$ gives our claim. ■

Claim 3. *There exist a profile of strategy subsets W and two sequence of profiles of strategy subsets $X = X^0, \ldots, X^U$ and $Y = Y^0, \ldots, Y^V$ such that (1) $X^0 = \mathcal{S}^K$ and $X^U = W$; (2) $Y^0 = \mathcal{T}^L$ and $Y^V = W$; (3) $X^u \xrightarrow{b_u} X^{u+1}$ with $b_u \in \{e, r, i\}$ for all $u < U$; and (4) $Y^v \xrightarrow{c_v} Y^{v+1}$ with $c_v \in \{e, r, i\}$ for all $v < V$. In other words, we have the following diagram:*

$$\begin{tikzcd}
\mathcal{S}^K(= X^0) \arrow[dr, dashed, "b_0"' near start, "b_u"' near end] & & \mathcal{T}^L(= Y^0) \arrow[dl, dashed, "c_0" near start, "c_v" near end] \\
& W(= X^U = Y^V) &
\end{tikzcd}$$

Proof. This claim follows by using Claim 2 iteratively in the diagram we get in Claim 1. ■

Claim 4. *There exist $X = X^0, \ldots, X^U$ and $Y = Y^0, \ldots, Y^V$ satisfying Claim 3, such that $b_u = r$ for all $u < U$, and $c_v = r$ for all $v < V$. In other words, we have the following diagram:*

$$\begin{tikzcd}
\mathcal{S}^K(= X^0) \arrow[dr, dashed, "r"'] & & \mathcal{T}^L(= Y^0) \arrow[dl, dashed, "r"] \\
& W(= X^U = Y^V) &
\end{tikzcd}$$

Proof. Let $X = X^0, \ldots, X^U$ and $Y = Y^0, \ldots, Y^V$ be the two sequences in Claim 3. We first prove that $b_u \neq e \forall u < U$. By contradiction, assume that there exists $0 \leq u < U$ such that $b_u = e$. Consider the smallest such u. If $u = 0$, then there exists two strategies $\sigma_i, \sigma_i' \in \mathcal{S}_i^K$ such that $\sigma_i \preceq \sigma_i'$ over $\mathcal{S}_i^K$, contradicting the fact that $\mathcal{S}^K$ is the terminal set of $\mathcal{S}$. If $u > 0$, then we find a sequence X^{u-1}, X^u, X^{u+1} such that $X^{u-1} \xrightarrow{r} X^u$ and $X^u \xrightarrow{e} X^{u+1}$. By Lemma 3, we have that either $X^{u-1} \xrightarrow{e} X^{u+1}$, or $X^{u-1} \xrightarrow{e} Z$ and $Z \xrightarrow{r} X^{u+1}$. No matter which is the case, we have moved the elimination relation one step forward. Keep on doing so, we have that $\mathcal{S}^K \xrightarrow{e} Z$ for some Z, again contradicting the fact that $\mathcal{S}^K$ is the terminal set of $\mathcal{S}$.

By symmetry, we have that $c_v \neq e \forall v < V$. By omitting all idle relations (if exist) in the two sequences, we find two new sequences which still satisfy Claim 3 but contain only the replacement relation. ■

Let $X = X^0, \ldots, X^U$ and $Y = Y^0, \ldots, Y^V$ be the two sequences in Claim 4. According to the first property in Lemma 1, we have that $Y^{v+1} \xrightarrow{r} Y^v$ for any $v < V$. Therefore there exists a sequence $Z = Z^0, \ldots, Z^{U+V}$ such that $Z^0 = \mathcal{S}^K$, $Z^{U+V} = \mathcal{T}^L$, $Z^s \xrightarrow{r} Z^{s+1}$ for all $s < U + V$. Indeed, $Z^s = X^s$ for all $s \leq U$, and $Z^s = Y^{U+V-s}$ for all $s \geq U$. Accordingly, we have the following diagram:

$$\begin{tikzcd}
\mathcal{S}^K(= Z^0) \arrow[r, dashed, "r"] & \mathcal{T}^L(= Z^{U+V})
\end{tikzcd}$$

For each $s < U + V$, let ϕ^s be the natural bijection from Z^s to Z^{s+1}. Let $\phi = \phi^{U+V-1} \circ \cdots \circ \phi^0 \triangleq (\phi_1, \ldots, \phi_n)$, where the composition of two profiles ψ and ξ, denoted by $\xi \circ \psi$, is defined to be $(\xi_1 \circ \psi_1, \ldots, \xi_n \circ \psi_n)$. It is easy to see that each ϕ_i is a bijection from $\mathcal{S}_i^K$ to $\mathcal{T}_i^L$. It is also easy to see that by Lemma 1, for any $\sigma \in \mathcal{S}^K$, we have that

$$H(\phi(\sigma)) = H((\phi^{U+V-1} \circ \phi^{U+V-2} \circ \cdots \circ \phi^0)(\sigma)) = H((\phi^{U+V-2} \circ \cdots \circ \phi^0)(\sigma)) = \cdots = H(\phi^0(\sigma)) = H(\sigma).$$

Q.E.D.

B The Proof of Theorem 2

Theorem 2. *Let M be a rationally robust implementation of a property P over a class of PKI contexts $\mathscr{C}$, C a context in $\mathscr{C}$, and $\mathcal{S}^1, \dots \mathcal{S}^n$ rationally robust solutions of the game (C, M). Then $\forall$ plays $\sigma \in \mathcal{S}_1^1 \times \cdots \times \mathcal{S}_n^n$, P holds for $M(\sigma)$.*

Proof. By definition of rationally robust implementation, there exists a rationally robust solution $\mathcal{S}^0$ of the game (C, M), such that for all strategy profiles $\tau \in \prod_i \mathcal{S}_i^0$, P holds for $M(\tau)$.

As shown by the proof of Theorem 1, for each $0 \leq k < n$, there exists a sequence $\mathcal{Z}^k = \mathcal{Z}^{k,0}, \dots, \mathcal{Z}^{k,\ell_k}$ such that (1) $\mathcal{Z}^{k,0} = \mathcal{S}^k$, (2) $\mathcal{Z}^{k,\ell_k} = \mathcal{S}^{k+1}$, and (3) $\mathcal{Z}^{k,c} \xrightarrow{r} \mathcal{Z}^{k,c+1}$ for each $0 \leq c < \ell_k$. In other words, we have the following diagram:

$$\begin{tikzcd}
\mathcal{S}^0(=\mathcal{Z}^{0,0}) \arrow[r, "r"] & \mathcal{Z}^{0,1} \arrow[r, dashed, "r"] & \mathcal{S}^1(=\mathcal{Z}^{0,\ell_0}=\mathcal{Z}^{1,0}) \arrow[r, dashed, "r"] & \mathcal{S}^n(=\mathcal{Z}^{n-1,\ell_{n-1}})
\end{tikzcd}$$

Moreover, for each $0 \leq k < n$, there exists a profile of bijections $\phi^k : \mathcal{S}^k \to \mathcal{S}^{k+1}$. Indeed, $\phi^k = \phi^{k,\ell_k - 1} \circ \cdots \circ \phi^{k,0}$, where $\phi^{k,c}$ is the natural bijection from $\mathcal{Z}^{k,c}$ to $\mathcal{Z}^{k,c+1}$, for each $0 \leq c < \ell_k$.

Accordingly, for each $\sigma \in \mathcal{S}_1^1 \times \cdots \times \mathcal{S}_n^n$, there exists a strategy profile $\tau \in \mathcal{S}^0$ such that for each player i, $\sigma_i = \phi_i^{i-1} \circ \cdots \circ \phi_i^0(\tau_i)$. Indeed, each τ_i can be found by taking the inverse of the bijections $\phi_i^{i-1}, \dots, \phi_i^0$ consecutively. Notice that P holds for $M(\tau)$. Therefore it suffices to show that $H(\sigma) = H(\tau)$.

According to Lemma 1 (and as in the proof of Theorem 1), it suffices to show that there exists a sequence $\mathcal{X}^0, \dots, \mathcal{X}^K$ such that: (1) $\mathcal{X}^0 = \mathcal{S}^0$, and (2) $\mathcal{X}^k \xrightarrow{r} \mathcal{X}^{k+1}$ for each $k < K$, and $\psi(\tau) = \sigma \in \mathcal{X}^K$, where $\psi = \psi^{K-1} \circ \cdots \circ \psi^0$, and ψ^k is the natural bijection from $\mathcal{X}^k$ to $\mathcal{X}^{k+1}$.

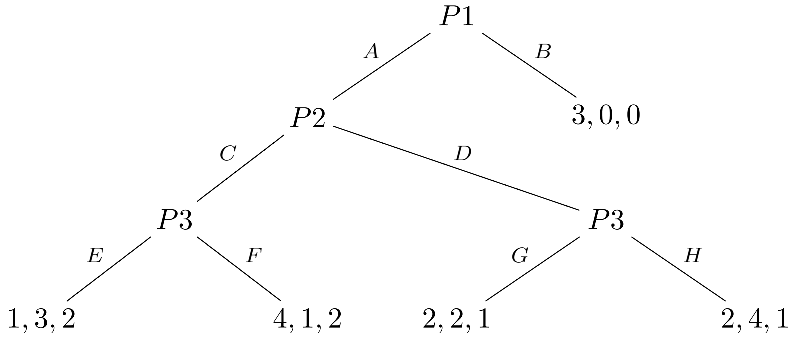
Without loss of generality, assume that $\sigma_i \neq \tau_i$ for each player i. It suffices to show that there exists a sequence $\mathcal{X} = \mathcal{X}^0, \dots, \mathcal{X}^n$ such that: (1) $\mathcal{X}^0 = \mathcal{S}^0$, and (2) for each $1 \leq k \leq n$, $\mathcal{X}^{k-1} \xrightarrow{r} \mathcal{X}^k$ by replacing τ_k with σ_k.

We do so by construct the sequence $\mathcal{X}$ explicitly. First of all, let $\mathcal{X}^0 = \mathcal{S}^0$. Since $\tau_1 \in \mathcal{S}_1^0 = \mathcal{Z}^{0,0}$, $\sigma_1 \in \mathcal{S}_1^1 = \mathcal{Z}^{0,\ell_0}$, and $\sigma_1 = \phi_1^0(\tau_1)$, there must exist a step $0 < c \leq \ell_0$ such that: (a) $\sigma_1 \in \mathcal{Z}_1^{0,c}$, (b) $\sigma_1 = \phi_1^{0,c-1}(\mu_1)$ for some $\mu_1 \in \mathcal{Z}^{0,c-1}$, and (c) μ_1 is the image of τ_1 under the sequence of bijections $\phi_1^{0,0}, \dots, \phi_1^{0,c-2}$ (or $\mu_1 = \tau_1$ if $c = 1$). Consider the smallest such c. If $c = 1$, then let $\mathcal{X}^1 = \mathcal{Z}^{0,1}$ and we are done. Otherwise, we have found a sub-sequence $\mathcal{Z}^{0,c-2}, \mathcal{Z}^{0,c-1}, \mathcal{Z}^{0,c}$ such that $\mathcal{Z}^{0,c-2} \xrightarrow{r} \mathcal{Z}^{0,c-1}$, and that $\mathcal{Z}^{0,c-1} \xrightarrow{r} \mathcal{Z}^{0,c}$ by replacing μ_1 with σ_1. According to Lemma 2 (2), (4), and (5), there exists $W^{0,c-1}$ such that $\mathcal{Z}^{0,c-2} \xrightarrow{r} W^{0,c-1}$ by replacing some strategy ν_1 with σ_1, and either $W^{0,c-1} \xrightarrow{r} \mathcal{Z}^{0,c}$ or $W^{0,c-1} \xrightarrow{i} \mathcal{Z}^{0,c}$. In other words, we have moved the place where σ_1 appears one step forward in the sequence from $\mathcal{S}^0$ to $\mathcal{S}^1$, while keeping the head and the tail unchanged. Keep on doing so, we find a new sequence (perhaps shorter than $\mathcal{Z}^0$) $W^{0,0} = \mathcal{S}^0, \dots, W^{0,c-1}, \mathcal{Z}^{0,c}, \dots, \mathcal{Z}^{0,\ell_0} = \mathcal{S}^1$ such that $W^{0,0} \xrightarrow{r} W^{0,1}$ by replacing τ_1 with σ_1. Let $\mathcal{X}^1 = W^{0,1}$, we are done. By a similar procedure, we move the place where σ_2 appears forward, so that it appears one step after $\mathcal{X}^1$, and we have found $\mathcal{X}^2$. In sum, the elements in sequence $\mathcal{X}$ can be constructed one by one. Given such a sequence, letting ψ be the corresponding bijection from $\mathcal{X}^0$ to $\mathcal{X}^n$, we have that $\sigma_i = \psi_i(\tau_i)$ for each i, and thus $H(\sigma) = H(\psi(\tau)) = H(\tau)$, which implies P holds for $M(\sigma)$.
Q.E.D.

C "Proof" of Fact 2′

Recall from Section 3:
Fact 2′ *For all extensive-form games, $H(\mathscr{PE}) \subset H(\overline{\mathscr{PE}}) \subset H(\mathscr{R}) \subset H(\mathscr{S})$, and all inclusions are strict for some games.*

It is easy to see that the first and third inclusions can be strict. Here we construct a game $\mathcal{G}$ where the second inclusion is strict, that is, $H(\overline{\mathscr{PE}}) \subsetneq H(\mathscr{R})$. Game $\mathcal{G}$ includes 3 players: $P1$, $P2$, and $P3$. They act sequentially and each acts only once. The players strategies are listed on the edges of the game tree, and their utilities are listed at every terminal node, with $P1$ comes first, as shown below.



It is easy to see that the closure of all sub-game perfect equilibria is $\overline{\mathscr{PE}} = \{B\}\times\{C, D\}\times\{EG, EH, FG, FH\}$. In particular, the history $\{A, C, F\}$ is not in $H(\overline{\mathscr{PE}})$. However, no strategy (of anybody) is distinguishably dominated, and no two strategies are equivalent, over the profile of strategy subsets of $\mathcal{G}$. Therefore $\mathscr{R} = \{A, B\} \times \{C, D\} \times \{EG, EH, FG, FH\}$, and clearly the history $\{A, C, F\}$ is in $H(\mathscr{R})$.

D A Stronger Notion of Implementation for Less Rational Players

Computing a full IEDD in an extensive-form mechanism (as said, the most suitable form of mechanisms for tackling collusion, complexity, and privacy problems) may require lots of iterations and force the players to always keep a global view of the whole game tree. Indeed, rationally robust implementation (as well as any other notions of implementation based on iterative elimination) may require quite a deal of rationality! However, once the game is almost over and a player i is about to take the last action in the game, we are in safer ground to assume that he will be able to analyze the rest of the game and choose his action in a rational manner.

Accordingly, we would be very happy if a mechanism M guarantees its desired property P as long as the players act rationally in just the last step of the game. Indeed, this requires less rationality than assuming that the players are able to analyze the whole game. The "next best" arises when M guarantees P as soon as the players act *sufficiently* rationally in the last two steps of the game. And so on.

Because we are interested in capturing not fully rational players, when we move towards the root of the game tree, from height k to height $k + 1$, we do not demand that the larger subgame be totally analyzed. Demanding so would imply that the players must "revisit" games of height k and thus demanding a lot of rationality. Indeed, we are interested in capturing an elementary, single-pass, bottom-up analysis, so that after processing the whole game, some DD strategies will still be left.

It is actually important to realize that (1) full iterative elimination of DD strategies, (2) the above sketched elimination process, and (3) backward induction are all *distinct* processes.[6]

Before presenting our stronger notions, let us develop a suitable notation for the class of mechanisms we focus on.

D.1 Extensive-Form Mechanisms With Simultaneous and Public Actions

Definition 13. *We say that mechanism M is of* extensive-form with simultaneous and public actions *if it satisfies the following criteria. As for any extensive-form mechanism, M specifies a* a game tree T, *whose internal nodes are* decision nodes *and whose leafs are* terminal nodes. *As usual too, each terminal node is associated with an outcome ω in Ω, or a distribution over Ω. For each decision node D our mechanism*

[6] Indeed, the third notion is as prone to equilibrium selection as plain Nash-equilibrium implementation. Only when the Cartesian closure of set S of subgame-perfect equilibria equals the set S itself is backward induction robust against equilibrium selection.

specifies a set of players acting at D, as well as a set of available actions for each player acting at D. Each action becomes common knowledge as soon as it is played by one of the acting players.[7]

In dealing with such a mechanism M we make use of the following notation.

- *Tree Height.* The height of a node in T is taken to be the number of edges in the longest path from the node to a leaf. (Thus a leaf has eight 0). The height of a tree is the height of its root.

- *Actors and Actions.* We denote by N^D the set of players acting at a decision node D, and by A_i^D the set of available actions for each player $i \in N^D$. Accordingly, the entire subprofile of action sets at node D is denoted by A^D.

- *Subgames and Substrategies.* Together with a context C, M not only specifies a game G, but also a subgame for each subtree of T. For each decision node D, we denote by G^D the subgame rooted at D. and by σ_i^D the restriction of σ_i to subgame G^D. (Thus, σ_i^D specifies which action i chooses among all those available to him for each node of the subtree in which he acts.) By Given a restricted strategy profile σ^D for G^D, the outcome of M obtained by executing σ^D is denoted by $M(\sigma^D)$.

D.2 Backward Robustness

Rather than eliminating all his DD strategies, our next notion only requires that a player is able to eliminate some of his strategies, by processing the game tree in a natural, bottom-up fashion.

Recall that a refinement of a game G is game coinciding with G, except that each player i has available only a subset of his original strategies.

Definition 14. (Backward-Robust Solutions) *Let $G = (C, M)$ be a game, where C is a perfect-knowledge context and M an extensive-form mechanism, with simultaneous and public actions. Then the* Backward-Robust solution *of G is defined to consist of all strategy profiles of $\overline{G}$, the refinement of G computed as follows:*

- *At each decision node D of height 1, the players in N^D refine their strategy sets in G^D by iteratively eliminating all their strictly dominated strategies in G^D.*

  $\widetilde{G}^D$ denotes the so computed refinement of G^D, and $\overline{G|_1}$ the refinement of G obtained by substituting, for each decision node D of height 1, subgame G^D with game $\widetilde{G}^D$.

- *For $h = 2$ to t, the height of G's game tree, compute the refinement $\overline{G|_h}$ of $\overline{G|_{h-1}}$ as follows:*
  For each decision node D of height h, the players in N^D compute the refinement $\widetilde{G}^D$ of $\overline{G|_{h-1}}^D$ by iteratively eliminating all their strictly dominated strategies in $\overline{G|_{h-1}}^D$.

  $\overline{G|_h}$ is then the game obtained by substituting, for each decision node D of height h, subgame $\overline{G|_{h-1}}^D$ with game $\widetilde{G}^D$.

Game $\overline{G}$ is then defined to be game $\overline{G|_t}$.

Remarks

- Notice that, when a decision node D has height 1, then iterative elimination of strictly dominated strategies coincides with "iterated elimination of strictly dominated actions." We emphasize, however that when D's height is $h > 1$, for computing $\widetilde{G}^D$, the players in N^D do not iteratively eliminate their strictly dominated *actions* in the action set A^D, but their strictly dominated *strategies* in the whole game $\overline{G|_{h-1}}^D$.

[7] We refrain from using the more standard term "perfect-information" to avoid confusion, as we use the term "perfect-knowledge" to refer to a context where the players' utilities are common knowledge to all players.

- Also notice that $\overline{G}$ is a rationally robust refinement of G in the sense that strategy profiles in $\overline{G}$ do *not* depend on the players' beliefs, but only on the assumption of common knowledge of rationality.
- Finally notice that $\overline{G}$ could be further refined by additional "bottom-up passes." Assume that, when refining $\overline{G|_{h+1}}^D$, a player i's strategy σ_i is eliminated. Then σ_i is of the form (a, σ'_i) where a is an action of i at node D and σ'_i is a strategy of i in some subgame $\widetilde{G}^{D'}$ of height h. Accordingly, eliminating σ_i implies eliminating σ'_i from game $\widetilde{G}^{D'}$. However, eliminating σ'_i may cause a strategy τ'_j of player j in $\widetilde{G}^{D'}$ to become strictly dominated. Yet, since the subgames of height h have already been processed, such τ'_j will continue to exist in $\overline{G}^{D'}$. This would not be a "problem" if D' were not reachable in $\overline{G}$. But if it were, then $\overline{G}$ could actually be further refined by another bottom-up process. Indeed, $\overline{G}$ is not a rationally robust solution of G.

Definition 15. (Backward-Robust Implementation) *Let $\mathscr{C}$ be a class of perfect-knowledge contexts, let P be a property over (distributions of) outcomes of contexts in $\mathscr{C}$, and let M be an extensive-form mechanism with simultaneous and public actions. We say that M is a* backward-robust implementation of P (over $\mathscr{C}$) *if, for any $C \in \mathscr{C}$,*

1. *for each player i, $\text{OUT}_i \notin \overline{\Sigma}_i$ where $\overline{\Sigma}_i$ is the set of strategies for player i in the backward-robust solution $\overline{G}$ of $G = (C, M)$; and*
2. *for any profile of strategies σ in $\overline{G}$, P holds for $M(\sigma)$.*

In essence, in a backward robust implementation, the players can work less but the mechanism must work more, in the sense that it must guarantee its desired property at a set of strategy profiles which includes a rationally robust solution. (Indeed, "the weaker the solution concept, the stronger the implementation!")

Let us now prove some crucial properties of our stronger notion of implementation. The first is that it implies rationally robust implementation.

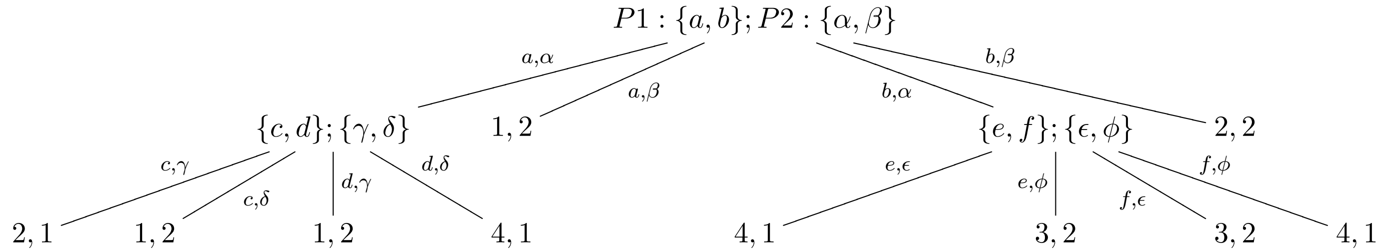
Theorem 3. (Backward-Robust Implementations are Rationally Robust) *Let $\mathscr{C}$ be a class of perfect-knowledge contexts, let P be a property over (distributions of) outcomes of contexts in $\mathscr{C}$, and let M be an extensive-form mechanism with simultaneous and public actions. If M is a backward-robust implementation of P over $\mathscr{C}$, then M is also a backward-robust implementation of P over $\mathscr{C}$.*

Proof. Let C be a context in $\mathscr{C}$, and $G = (C, M)$ the corresponding game. Let $\overline{G|_0} = G$, and let Σ^k be the profile of strategy subsets of $\overline{G|_k}$, for $0 \leq k \leq t$, with t the height of G's game tree. It is easy to see that for each $k < t$, each node D of height $k+1$, and each player i, if a strategy σ_i^D in subgame $\overline{G|_k}^D$ is (iteratively) strictly dominated and thus eliminated, then all strategies τ_i of i in $\overline{G|_k}$ with $\tau_i^D = \sigma_i^D$ are (iteratively) distinguishably dominated and can be eliminated simultaneously. Accordingly, there exists an IEDD $\mathcal{X}^0, \ldots, \mathcal{X}^{L_k}$ such that $\mathcal{X}^0 = \Sigma^k$ and $\mathcal{X}^{L_k} = \Sigma^{k+1}$. Connecting all these IEDD's, we have an IEDD $\mathcal{Y}^0, \ldots, \mathcal{Y}^L$ such that $\mathcal{Y}^0 = \Sigma^0$ and $\mathcal{Y}^L = \Sigma^t$. Because M backward robustly implements P, P holds for $M(\sigma)$ for all strategy profiles $\sigma \in \mathcal{Y}^L$. Keeping on eliminating DD strategies from $\mathcal{Y}^L$ until nothing can be eliminated, we get a full IEDD $\mathcal{Y}^0, \ldots, \mathcal{Y}^K$, with $K \geq L$. Since $\mathcal{Y}_i^K \subseteq \mathcal{Y}_i^L$ for every player i, we have that P holds for $M(\tau)$ for every $\tau \in \mathcal{Y}^K$. Therefore M is a rationally robust implementation of P over $\mathscr{C}$.
Q.E.D.

Another property is that our second notion of implementation is indeed strictly stronger than our first one. Namely,

Theorem 4. *There exist games G such that $H(\overline{G}) \supsetneq H(\mathscr{R})$.*

Proof. We construct such a game G explicitly. It includes 2 players, $P1$ and $P2$. The game tree is of height 2. At each decision node, each player has two actions available (e.g., $\{a, b\}$ for $P1$ and $\{\alpha, \beta\}$ for $P2$), and they act simultaneously. At each terminal node, the players utilities are listed, with $P1$ comes first, as shown below.



According to backward robust implementation: At height 1, no strategy can be eliminated from the two sub-games. At height 2, there are two strictly dominated strategies for $P1$, namely, ace (by bce) and acf (by bcf). After eliminating these two strategies, nothing can be eliminated any more. In particular, there are two backward robust strategies (of $P1$ and $P2$ respectively), ade and $\alpha\delta\epsilon$, yielding history $\{(a, \alpha), (d, \delta)\}$. Namely, $\{(a, \alpha), (d, \delta)\} \in H(\overline{G})$.

According to rationally robust implementation: In the first step, the same two strategies ace and acf of $P1$ are eliminated (they are strictly dominated, and thus distinguishable dominated) as above. In the second step, based on remaining strategies, there are two DD strategies for $P2$, namely $\alpha\delta\epsilon$ (by $\alpha\gamma\epsilon$) and $\alpha\delta\phi$ (by $\alpha\gamma\phi$), and they are eliminated. Notice that once these two strategies of $P2$ are eliminated, we can conclude that the history $\{(a, \alpha), (d, \delta)\}$ is not in $H(\mathscr{R})$, without even finishing the full IEDD. Accordingly, $H(\overline{G}) \supsetneq H(\mathscr{R})$.
Q.E.D.

We stress that backward robust implementation is indeed possible. In fact, some of our robust mechanisms guarantee their desired properties under this stronger notion.

Let us point out that one could consider even stronger notions of implementation, requiring even less rationality from the players. In particular, we consider mechanisms that, for each height h, guarantee a desired property P_h for any profile of strategies in $\overline{G|_h}$. Of course, the smaller h, the stronger the implementation notion, because

$$H(\overline{G|_1}) \supset H(\overline{G|_2}) \supset \cdots \supset H(\overline{G|_t}) = H(\overline{G}) \supset H(\mathscr{R})$$

where it is easy for the inclusions to be strict. Of course too, however, a rationality-quality tradeoff should be expected. That is, we interpret the fact that a property P could be guaranteed with lesser rationality as an indication that a stronger property P' could be guaranteed by a rationally robust implementation.