Simple Mechanisms and Preferences for Honesty[*]

Navin Kartik, Olivier Tercieux, and Richard Holden[†]

November 22, 2013

Abstract

We consider full implementation in complete-information environments when agents have an arbitrarily small preference for honesty. We offer a condition called *separable punishment* and show that when it holds and there are at least two agents, any social choice function can be implemented by a simple mechanism in two rounds of iterated deletion of strictly dominated strategies.

1 Introduction

What social objectives can be achieved in decentralized economies? In his celebrated work, Maskin (1999, circulated in 1977) formally introduced the notion of implementing social choice functions (hereafter, SCFs) through a suitably-constructed mechanism or game form when agents know some "state of the world" that a social planner does not.[1] Maskin (1999) focussed on outcomes that obtain in Nash equilibria and found that only SCFs that satisfy a fairly demanding property, *(Maskin-)monotonicity*, are implementable.

[*] We thank Amanda Friedenberg, the referees, and an Associate Editor for helpful comments.

[†] Kartik: Columbia University, Department of Economics; email: nkartik@columbia.edu. Tercieux: Paris School of Economics; email: tercieux@pse.ens.fr. Holden: University of New South Wales, School of Economics, Australian School of Business and NBER; email: richard.holden@unsw.edu.au.

[1] "Implementation" without qualification in this paper always refers to full-implementation, which means that *every* outcome should coincide with the social objective; this contrasts with partial-implementation where only *some* outcome need be desirable.

Moreover, sufficiency of this property—along with some other mild conditions—has only been established in general environments using so-called "integer games" or "tail-chasing mechanisms", which are unappealing for well-known reasons (see, for example, Jackson, 1992).

For both the above reasons, a sizable literature has examined implementation in other solution concepts, often refinements of Nash equilibrium. The scope for implementation expands substantially under either subgame-perfect Nash equilibria (Abreu and Sen, 1990; Moore and Repullo, 1988) or undominated Nash equilibria (Palfrey and Srivastava, 1991; Jackson et al., 1994); furthermore, some of these implementing mechanisms are "well behaved". However, these implementations of non-monotonic SCFs are problematic because they are not robust to even slight departures from underlying common knowledge assumptions (Chung and Ely, 2003; Aghion et al., 2012).[2]

More recently, a burgeoning literature studies the scope for implementation when players have (small) preferences for honesty. Loosely speaking, a player is said to have a preference for honesty if he prefers a "truthful" message/report when his message does not affect the outcome of the mechanism. What now determines implementability is preferences over the joint space of messages and outcomes. Kartik and Tercieux (2012) observe that preferences for honesty render any SCF monotonic on this extended space. This suggests that even Nash implementation may be quite permissive when players have preferences for honesty. However, existing results suffer from some of the weaknesses mentioned above. For example, Dutta and Sen (2011), Lombardi and Yoshihara (2011), Kartik and Tercieux (2012), and Korpela (2012) use integer games; Ortner (2012) invokes particular refinements of Nash equilibrium and requires five or more players; Matsushima (2008a) uses mechanisms that randomize over outcomes and only studies settings with three or more players; while Matsushima (2008c) also uses randomization and further assumes stronger conditions on the nature of preferences for honesty.[3]

[2] An alternative approach that also yields permissive results is that of virtual or approximate implementation (Matsushima, 1988; Abreu and Sen, 1991; Abreu and Matsushima, 1992). As is well-recognized, however, a weakness of this approach is that mechanisms must randomize over outcomes and the resulting outcome may be very inefficient, unfair, or "far" from the desired outcome, even if this only occurs with small ex-ante probability. Furthermore, the mechanisms used here either rely on integer games (Matsushima, 1988; Abreu and Sen, 1991) or quite critically on the linearity assumption of expected utility (Abreu and Matsushima, 1992). These results also become much less permissive when the implementation problem concerns only two players, a case that is important for bargaining and bilateral contracting.

[3] Both Matsushima (2008a) and Matsushima (2008c) achieve implementation in iterated deletion of

In this paper, we also study implementation when players have a preference for honesty. Our contribution is to derive a strong positive result for a wide class of environments of economic interest. Specifically, we show that so long as there are *two or more players* and the environment satisfies a condition we call *separable punishment*, any SCF is implementable in *two rounds of iterative deletion of (strictly) dominated strategies* by a simple and well-behaved—indeed, essentially a direct—mechanism. Roughly speaking, the separable-punishment condition requires that one can find a player with a preference for honesty, say j, and another player, say i, such that for any outcome in the range of the SCF, there is an alternative outcome under which, in any state, j is indifferent between the two outcomes while i strictly prefers the socially-desired outcome to the alternative. In this sense, it is possible to suitably "punish" one player without punishing the other. We provide examples of economic problems in which this condition is naturally satisfied; in particular, it holds both in a standard exchange economy and also whenever a mechanism can augment large-enough monetary punishments on players.

The idea that some form of separability in the environment can simplify the implementation problem is not new; Abreu and Matsushima (1994), Jackson et al. (1994) and Sjöström (1994) are important antecedent contributions in this vein.[4] We defer a more detailed comparison to Section 4.

The remainder of the paper is organized as follows. Section 2 presents a simple example to highlight the main idea and discusses the merits of the mechanism we propose. The general setting and result are provided in Section 3. Section 4 relates our work to other approaches.

dominated strategies following the approach of Abreu and Matsushima (1992, 1994). A virtue of these papers is that the planner only needs to be able to impose small fines on players. Note that Matsushima (2008a) studies settings with complete information, as does most of the literature cited earlier, whereas Matsushima (2008c) tackles settings with incomplete information relying on the expected utility hypothesis. While we focus on complete-information environments here, an earlier version of this paper (Holden et al., 2013) showed that the logic can be extended to settings of incomplete information under weak assumptions on how players evaluate lotteries (in particular, without assuming expected utility) so long as there is *non-exclusive information*, a notion that is familiar in the Bayesian implementation literature (e.g. Postlewaite and Schmeidler, 1986).

[4] In work concurrent to ours, Lombardi and Yoshihara (2013) focus exclusively on "price-quantity implementation" in pure exchange economies when some players have a preference for honesty.

2 An Example

Consider a setting with two players, 1 and 2, and two states, θ' and θ''. The socially desired outcomes at state θ' and θ'' are respectively denoted by $f(\theta')$ and $f(\theta'')$. The mechanism designer can augment outcomes with transfers, but he does not want to use such transfers on the (predicted) path of play. Agents' preferences are quasi-linear and state independent: i's utility at state $\theta \in \{\theta', \theta''\}$ given outcome a and transfer t_i is $v_i(a) - t_i$. Because preferences are state-independent and the goal is full implementation, f is not implementable in virtually any solution concept unless $f(\theta') = f(\theta'')$.

We will study a direct mechanism where each player announces a value of the state. The outcome selected by the mechanism only depends on player 1's announcement: if he claims the state is θ then the outcome selected is $f(\theta)$. In this sense, player 1 is a dictator over outcomes. Regarding transfers: if players disagree on their announcements, then only player 1 is fined an amount t_1; if players agree, there are no transfers.

Now we introduce the notion that player 2 has a preference for honesty by supposing that his payoff increases by $\varepsilon > 0$ when making an "honest" announcement. Formally, if the true state is $\theta \in \{\theta', \theta''\}$, the payoff matrix in the game induced by the mechanism is as follows, where we denote the possible announcements for each player by θ and $\neg\theta$:

(1,2)	θ	$\neg\theta$
θ	$v_1(f(\theta)), v_2(f(\theta)) + \varepsilon$	$v_1(f(\theta)) - t_1, v_2(f(\theta))$
$\neg\theta$	$v_1(f(\neg\theta)) - t_1, v_2(f(\neg\theta)) + \varepsilon$	$v_1(f(\neg\theta)), v_2(f(\neg\theta))$

Table 1 – Payoffs when the true state is θ

It is clear that in this game player 2 has a strictly dominant strategy to announce the truth, θ. Furthermore, provided the fine t_1 is large enough, it is then iteratively strictly dominant for player 1 to also announce θ. Consequently, in either state, both players telling the truth is the unique profile of strategies surviving two rounds of iterative deletion of strictly dominated strategies.

Implementation in two rounds of iterative deletion of strictly dominated strategies is an appealing solution concept for multiple reasons. It is a robust solution concept: when a "small amount of incomplete information" is introduced, no undesirable outcomes appear (Dekel et al., 2006; Oury and Tercieux, 2012), unlike with refinements of Nash equilibrium, as noted earlier. In addition, researchers often focus only on *pure-strategy* Nash implementation, and one may be justifiably concerned that even if a simple mechanism works for this solution concept, it would need complicated augmentation to deal with mixed strategies. Our mechanism obviates this concern. Furthermore, the fact that only two rounds of deletion are required means that players only need mutual knowledge—rather than common knowledge—that strictly dominated strategies will not be played (cf. Matsushima, 2008b).

While we only needed to assume above that one player (viz., player 2) has a preference for honesty, the mechanism does exploit the identity of this player. If, however, *both* players have a preference for honesty—which we view as reasonable—then the mechanism is in fact "detail free" in the sense that the planner can choose either player to act as "dictator" and does not need to know much about players' preferences: all that he needs to do is impose a sufficiently large fine on the dictator when announcements do not coincide. Of course, there is a minimal requirement that the planner must know what amount of fine will be large enough.

In the sequel, we show how these ideas can be extended to more general settings of complete information; readers interested in the case of incomplete information can consult our earlier working paper (Holden et al., 2013).

3 The Main Result

There is a set of states Θ, a set of outcomes or allocations A, and a finite set of players $I = \{1, \ldots, n\}$ with $n \geq 2$. A social choice function (SCF) is a mapping $f : \Theta \to A$. Given any function α whose domain is Θ, let $\alpha(\Theta) := \bigcup_{\theta \in \Theta} \alpha(\theta)$. The primitives specify (ordinal) preferences over A in each state θ for each player j, captured by a linear order $\succeq^A_{j,\theta}$. Given a space of message profiles, $M = M_1 \times \cdots \times M_n$, players have preferences defined over the joint space of allocations and message profiles: in each state θ, a player j has preferences over $A \times M$ denoted by $\succeq_{j,\theta}$. In the standard framework, $(a, m) \succeq_{j,\theta} (a', m')$ if and only if $a \succeq^A_{j,\theta} a'$. A mechanism is a pair (M, g) where $g : M \to A$. To simplify the exposition we will focus below on pure strategies only; all the concepts and results can be extended to cover mixed strategies with very weak assumptions on how players evaluate lotteries, as should be clear from the arguments we make.

We now formalize our general notion of a preference for honesty.

Definition 1. Given a space of message profiles, M, j has a *preference for honesty* on M if there is an injective function $m_j^* : \Theta \to M_j$ such that for any $g : M \to A$ and $\theta \in \Theta$:

If

$$\forall m_{-j}, m_j, \tilde{m}_j : \; g(m_{-j}, m_j) \sim_{j,\theta}^{A} g(m_{-j}, \tilde{m}_j) \tag{1}$$

then

$$\forall m_{-j} \text{ and } \forall m_j \neq m_j^*(\theta) : \; (g(m_{-j}, m_j^*(\theta)), m_{-j}, m_j^*(\theta)) \succ_{j,\theta} (g(m_{-j}, m_j), m_{-j}, m_j). \tag{2}$$

The key idea here is that because of the antecedent (1), the condition only has bite on preferences over the subset of $A \times M$ among which j is "materially indifferent" over the allocations; in this sense, it captures small or even lexicographic considerations. The message $m_j^*(\theta)$ is what j considers "truthful" in state θ. A leading example is when $M_j = \Theta$ and j's preferences are as follows: (i) if $a \succeq_{j,\theta}^{A} a'$, $m_j = \theta$, and $m_j' \neq \theta$, then $(a, m) \succ_{j,\theta} (a', m')$; (ii) otherwise, $(a, m) \succeq_{j,\theta} (a', m')$ if and only if $a \succeq_{j,\theta}^{A} a'$. In this case, $m_j^*(\theta) = \theta$ and our definition reduces to that of Dutta and Sen (2011) and is very similar to Kartik and Tercieux (2012, Example 2).

It is worth highlighting that a player's preference for honesty is defined with respect to a particular space of message profiles; in particular, Definition 1 does not require player j's message space to coincide with the set of states (although its cardinality must be at least as large).[5] To see why this may be substantively relevant, suppose each state is a profile of preferences over allocations. Then, if player j were asked to report the state (i.e. $M_j = \Theta$) he may not have a strict preference for truth-telling because he is reporting other players' allocation-preferences too; but if he is asked to only report his own allocation-preferences (i.e. that component of the state), then his preference for truth-telling may have bite. So long as player j's allocation-preferences are distinct in

[5] Furthermore, the definition also allows j's preferences to depend on the messages sent by other players beyond how these affect allocations.

every state, this setting would satisfy Definition 1.

We next introduce a domain restriction that will play a key role.

Definition 2. There is *separable punishment* if there is a function $x : \Theta \to A$ and players i and $j \neq i$ such that for all $\theta \in \Theta$ and $\theta' \in \Theta$: $x(\theta') \sim^A_{\theta,j} f(\theta')$ and $x(\theta') \prec^A_{\theta,i} f(\theta)$.

In words, separable punishment requires for each state θ', there be an alternative to the socially desired outcome, $x(\theta') \neq f(\theta')$, such that in any state θ, player j is indifferent between $x(\theta')$ and $f(\theta')$ while player i finds $x(\theta')$ strictly worse than $f(\theta')$. Separable punishment may be reminiscent of various "bad/worst outcome" conditions in the literature (e.g. Moore and Repullo, 1990; Jackson et al., 1994), but it differs in three ways: first, it allows for state-dependent alternative allocations; second, it requires that each state's alternative allocation keep player j indifferent rather than making him worse off, and furthermore satisfy this indifference no matter the true state; and third, it does not require that a state's alternative allocation must be "bad" for player i relative to all allocations in the range of the SCF, but rather only with respect to the state's socially desired alternative.

Generally, separable punishment is more likely to hold when there are transferable private goods, and indeed, there are natural and well-studied economic environments that satisfy separable punishment. We provide two examples. Consider first an economy with transfers and quasi-linear preferences. Here the outcome space $A = B \times \mathbb{R}^n$ consists of pairs (b, t) where b is some fundamental allocation and $t = (t_i)_{i \in I}$ is a vector of transfers. For each player i and state θ, preferences $\succeq^A_{i,\theta}$ over outcomes (b, t) are represented by $v_i(b, \theta) - t_i$. Assume that for some player i, the function $v_i(\cdot, \cdot)$ is bounded uniformly over b and θ, i.e. there is a constant $C \in \mathbb{R}_+$ satisfying $|v_i(b, \theta)| \leq C$ for all b and θ. Given the SCF $f : \Theta \to B \times \mathbb{R}^n$, for any θ let $f_b(\theta)$ be first component of $f(\theta)$ and $f_{t_i}(\theta)$ be the transfer specified for player i. One can now easily check that the requirement of separable punishment is satisfied with the function $x(\cdot)$ defined by $x(\theta) = (f_b(\theta), t')$ where t'_i is chosen sufficiently large while $t'_j = f_{t_j}(\theta)$ for all $j \neq i$. This setting subsumes prominent settings in the literature such as Moore and Repullo (1988, Section 5).

Second, consider an exchange economy with $\ell \geq 2$ commodities. There is an aggregate endowment vector $\omega_\ell \in \mathbb{R}^\ell_{++}$. An outcome a is an allocation $(a_1, ..., a_n) \in \mathbb{R}^{\ell n}$ such that $a_i \geq 0$ and $\sum_{i \in I} a_i \leq \omega_\ell$.[6] For each player i and state θ, preferences $\succeq^A_{i,\theta}$ over outcomes are assumed to be strictly increasing in i's component, i.e., $a_i > a'_i \implies a \succ^A_{i,\theta} a'$. Assume that at each state the social choice function f allocates each player a strictly positive amount of some commodity. It is now straightforward to verify that separable punishment is satisfied with the function $x(\cdot)$ defined by $x(\theta) = a'$ such that $a'_i = 0$ while $a'_j = f_j(\theta)$ for all $j \neq i$, where $f_j(\theta)$ denotes player j's component of the allocation $f(\theta)$.

We are now in a position to state the main result.

Theorem 1. *Assume separable punishment and fix i and j from that definition. Suppose further there is a message space (M_i, M_j) such that (i) there is an injective function $h_i : \Theta \to M_i$, and (ii) player j has a preference for honesty on (M_i, M_j). Then the SCF f can be implemented in two rounds of iterated deletion of strictly dominated strategies.*

Proof. Fix i, j and M_i, M_j from the theorem's hypotheses. Pick an arbitrary $\theta^* \in \Theta$ and define the mechanism $((M_i, M_j), g)$ where

$$g(m_i, m_j) = \begin{cases} f(h_i^{-1}(m_i)) & \text{if } m_i \in h_i(\Theta) \text{ and } m_j = m_j^*(h_i^{-1}(m_i)) \\ x(h_i^{-1}(m_i)) & \text{if } m_i \in h_i(\Theta) \text{ and } m_j \neq m_j^*(h_i^{-1}(m_i)) \\ x(\theta^*) & \text{otherwise.} \end{cases}$$

Consider any state θ. For any given m_i, j is indifferent over all the outcomes he can induce, so condition (1) is satisfied. Hence, j's preference for honesty implies (2) and it is strictly dominant for j to send message $m_j^*(\theta)$.

Now fix $m_j = m_j^*(\theta)$. It follows from the definition of $g(\cdot)$ that if player i reports $m_i = h_i(\theta)$, then because $m_j^*(h_i^{-1}(m_i)) = m_j^*(\theta)$ (using the injective property of $h_i(\cdot)$), he induces $f(h_i^{-1}(m_i)) = f(\theta)$. The definition of $g(\cdot)$ combined with the injective property of both $m_j^*(\cdot)$ and $h_i(\cdot)$ further implies that if player i sends any $m_i \neq h_i(\theta)$, he will induce $x(\theta')$ for some θ'. The separable punishment condition implies that $x(\theta')$ for any θ' is strictly worse for i than $f(\theta)$ in state θ. It follows that the unique best response for player i is to send message $h_i(\theta)$.

Therefore, the unique strategy profile surviving two rounds of iterative deletion of

[6] Each a_i is a vector with l components; $\geq$ and $>$ on $\mathbb{R}^l$ are the standard component-wise partial orders.

strictly dominated strategies is $m_j = m_j^*(\theta)$ and $m_i = h_i(\theta)$, which yields the outcome $f(\theta)$, as desired. ∎

The logic behind Theorem 1 is similar to that presented in the example of Section 2. Indeed, if we were to assume that each M_i is the set of states of the world, we could just let $h_i(\cdot)$ in the proof of Theorem 1 be the identity mapping. In addition, if we (naturally) assume that the function $m_j^*(\cdot)$ is the identity mapping, the mechanism in the proof simplifies to the following: denoting player i's announcement of the state by θ_i, choose outcome $f(\theta_i)$ if player j announces the same state; otherwise choose $x(\theta_i)$. By the separable punishment condition, player j is indifferent between all his messages. So it is uniquely optimal for player j to tell the truth; in turn, separable punishment further implies that the unique best response for player i is to also tell the truth. As discussed at the end of Section 2, this mechanism is robust, direct, and simple: there are only two rounds of elimination of strictly dominated strategies and obviously no use of integer games or related ideas.

4 Discussion and Further Connections to the Literature

4.1 Separable punishment

Dutta and Sen (2011) provide a separability condition under which they establish that any social choice function can be implemented by a mechanism that does not use integer games so long as there are three or more players. Their condition is logically incomparable with our separable punishment condition. However, separability conditions in the literature generally incorporate settings such as public-good environments with transfers and quasi-linear preferences,[7] but Dutta and Sen's notion excludes this standard environment (as they note) while ours does not. This is one reason our result applies to a class of economic problems that Theorem 4 in Dutta and Sen (2011) does not. Furthermore, our result applies when there are only two players, which is important for some applications such as bilateral contracting. In fact, Dutta and Sen (2011, page 166) discuss

[7] See, for example, Jackson et al. (1994). Our condition is also logically incomparable with that of Jackson et al. (1994), but both conditions hold in the pure exchange economy and transferable-utility settings discussed following Definition 2.

whether a strengthening of their separability condition would be sufficient for implementation with a small preference for honesty when there are only two players. They conjecture that the answer to this question must be negative. Our Theorem 1 disproves their conjecture because their suggested strengthening is stronger than our separable punishment condition.[8]

The message of Theorem 1 is related to a result in Ben-Porath and Lipman (2012). They show that in a complete-information setting where any pair of states can be distinguished via some player's hard evidence, a planner who can use large off-path fines can implement any SCF in subgame-perfect equilibria of a perfect-information mechanism. They note that this conclusion also holds if players have a small cost of forging evidence. Preferences for honesty are a special case of costly evidence fabrication (cf. Kartik and Tercieux, 2012). Due to the additional structure, Theorem 1 derives a much simpler mechanism and stronger conclusion than Theorem 1 of Ben-Porath and Lipman.[9]

4.2 Strict versus weak dominance

Prior literature, notably Abreu and Matsushima (1994), Jackson et al. (1994), and Sjöström (1994), on implementation using the solution concept of iterative deletion of weakly dominated strategies (IDWDS) has established that separable environments allow for permissive results using "bounded" mechanisms, even without preferences for honesty. Specifically, the mechanisms in these papers achieve unique implementation through truth-telling, i.e., in any state of the world, truth-telling is the unique strategy profile surviving IDWDS and yields an outcome that coincides with the social objective.

Implementation using the solution concept of IDWDS suffers from multiple drawbacks that implementation using iterative deletion of strictly dominated strategies (IDSDS)

[8] Dutta and Sen (2011) call an environment separable if there exists an alternative $w \in A$ with the following property: for all $a \in A$ and $J \subseteq N$, there exists $a^J \in A$ such that for any θ, $a^J \sim^A_{\theta,j} w$ for all $j \in J$ and $a^J \sim^A_{\theta,i} a$ for all $i \notin J$. The strengthening they propose for the two-player case consists in assuming that w is the "worst" outcome relative to outcomes in the range of the SCF, i.e., $w \prec^A_{\theta,i} f(\theta)$ for each player i and state θ. It is straightforward that this would imply our separability condition with the function $x(\cdot)$ defined as $x(\theta) = f(\theta)^{\{i\}}$ for all θ.

[9] Ben-Porath and Lipman (2012) use a perfect-information mechanism. It is straightforward to see that our Theorem 1 can also be proved with a perfect-information mechanism: simply view the game form used in the Theorem's proof as the normal-form representation of a perfect-information game form where player j moves first followed by player i.

usually does not. First, the order in which weakly dominated strategies are eliminated can affect the final outcome. This is a problem for the mechanisms of both Jackson et al. (1994) and Sjöström (1994): the reason that certain "lies" for a player are weakly dominated by truth-telling in these mechanisms is only because opponents may themselves be playing weakly dominated strategies. If one were to first eliminate these weakly dominated strategies for opponents, then some lies for a player would no longer be weakly dominated, and consequently truth-telling would not obtain as the unique outcome of the deletion process.[10]

Second, there are conceptual tensions between the use of IDWDS by players and their knowledge of the implications of this concept. Specifically, there are games in which it is unclear whether players who know the set of strategies surviving IDWDS (in some order) would in fact find it compelling to use only these strategies on the basis of weak dominance.[11] In the context of implementation theory, we claim that any "permissive result" that achieves unique implementation through IDWDS will be subject to such a critique.

To make our claim precise, fix an arbitrary mechanism, an arbitrary state of the world, and suppose that $(s_1^*, \ldots, s_n^*)$ is the unique strategy profile surviving IDWDS (in some order). We will say that $(s_1^*, \ldots, s_n^*)$ is *consistent* if for every player i, when each player $j \neq i$ is assumed to play s_j^*, the only weakly undominated strategy for i is s_i^*.[12] Consistency in this sense ensures that players' use of IDWDS is compatible with their knowledge of the consequences of IDWDS.[13] While an arbitrary game need not possess

[10] As we were reminded by an Associate Editor, IDSDS can also be sensitive to the order of deletion in an arbitrary infinite game. However, even when the environment is not finite, this is not a concern with the mechanism used to prove Theorem 1 because of its property that player j has a unique strictly dominant strategy and player i has a unique best response to this strategy of j. Indeed, this pair of strategies is the unique maximal reduction in the sense of Dufwenberg and Stegeman (2002).

[11] This is true even when there is a unique strategy profile surviving IDWDS, and it is the same profile no matter the order of deletion. Example 8 in Samuelson (1992) illustrates the point. More generally, Samuelson (1992) discusses various issues that must be resolved regarding the use of IDWDS as a solution concept. See Brandenburger et al. (2008) and Keisler and Lee (2011) for more recent developments.

[12] In other words, if one considers a modified game in which player i's set of available (pure) strategies is M_i while each player $j \neq i$ only has available s_j^*, i's set of weakly undominated strategies (pure or mixed) in this modified game should be $\{s_i^*\}$.

[13] An implication of Theorem 4 of Samuelson (1992) is that, in his framework of knowledge, if there is common knowledge of admissibility (i.e. the deletion of all weakly dominated strategies) and common knowledge of admissibility yields a unique strategy profile, this outcome can be known to players if and only if the strategy profile is consistent in the aforementioned sense.

a consistent strategy profile surviving IDWDS, it is desirable from a mechanism design perspective to design game forms under which IDWDS does yield a consistent strategy profile in each state, as this ensures that the planner's recommendation of what strategy profile to play in each state on the basis of IDWDS is "self-enforcing".

Accordingly, let us say that a SCF is implementable in IDWDS by truth-telling *in a consistent way* if there is a mechanism such that in every state, truth-telling is the unique profile surviving IDWDS (in some order), and it is consistent while also leading to the socially desired objective. It is straightforward that in any state where truth-telling is the unique strategy profile surviving IDWDS and is consistent, truth-telling must be a strict Nash equilibrium.[14] Thus, if a SCF is implementable in IDWDS by truth-telling in a consistent way, it is also implementable when using strict Nash equilibrium as the solution concept. However, Cabrales and Serrano (2011, Theorem 3) have shown that SCFs that are implementable in strict Nash equilibrium satisfy a small variation of Maskin-monotonicity that they call "quasi-monotonicity". It follows that *any* mechanism that implements a non-(quasi-)monotonic SCF in IDWDS by truth-telling fails to do so in a consistent way. This concern applies, in particular, to the mechanisms of Abreu and Matsushima (1994), Jackson et al. (1994), and Sjöström (1994) whenever these mechanisms are used to implement non-(quasi-)monotonic SCFs through IDWDS.

It is natural to wonder whether one couldn't just take a mechanism that achieves implementation through IDWDS without a preference for honesty and use it to achieve implementation through IDSDS when there is a preference for honesty. The answer is generally negative with respect to the minimal notion of preference for honesty considered in this paper. In particular, one can check that the mechanisms of Abreu and Matsushima (1994), Jackson et al. (1994), and Sjöström (1994) would not achieve implementation in IDSDS even if all players had a preference for honesty in the sense of Definition 1 holding with their message spaces. The point is that while preferences for honesty do transform certain "primitive" indifferences into strict preferences, an implementing mechanism must create the right kinds of primitive indifferences to exploit them. Using the aforementioned authors' mechanisms to achieve IDSDS with a prefer-

[14] If some player i has multiple best responses when every $j \neq i$ tells the truth, weak dominance does not pin down a unique strategy for i in the modified game in which all $j \neq i$ can only tell the truth while i can choose any of his original strategies.

ence for honesty would require strengthening the notion of a preference for honesty in an appropriate way (that depends on the mechanism); this is illustrated by the approach of Matsushima (2008a) and Matsushima (2008c).

We end by reiterating other merits of the mechanism used in Theorem 1 besides the solution concept. It is essentially a direct mechanism—unlike those of Jackson et al. (1994) and Abreu and Matsushima (1994)—and requires only two players. As already mentioned, the two-player setting is important because of its relevance to bargaining and bilateral contracting, yet it is not accommodated by any of the aforementioned papers that achieve implementation through IDWDS without a preference for honesty. In addition, in some salient cases the mechanism is "detail free" as explained at the end of Section 2. Finally, Theorem 1 applies even when players' preferences over allocations do not change across states while the socially preferred alternative does. As also highlighted by Ben-Porath and Lipman (2012), there are interesting problems that have this property;[15] for obvious reasons, no result in the literature without a preference for honesty (or related ideas, such as costly signaling or evidence) can accomodate nontrivial SCFs in these cases, and furthermore, simply using those mechanisms without modification would not generally work even when there is a preference for honesty.

References

ABREU, D. AND H. MATSUSHIMA (1992): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993–1008.

——— (1994): "Exact Implementation," *Journal of Economic Theory*, 64, 1–19.

ABREU, D. AND A. SEN (1990): "Subgame perfect implementation: A necessary and almost sufficient condition," *Journal of Economic Theory*, 50, 285–299.

——— (1991): "Virtual Implementation in Nash Equilibrium," *Econometrica*, 59, 997–1021.

AGHION, P., D. FUDENBERG, R. HOLDEN, T. KUNIMOTO, AND O. TERCIEUX (2012): "Subgame-Perfect Implementation under Information Perturbations," *Quarterly Journal of Economics*, 127, 1843–1881.

[15] To take just one example, consider assigning a settlement that a defendant must pay a plaintiff: a social planner may wish to condition the transfer (the allocation) on what damage was actually caused (the state of the world) even though neither player's preferences varies with the damage level.

BEN-PORATH, E. AND B. L. LIPMAN (2012): "Implementation with Partial Provability," *Journal of Economic Theory*, 147, 1689–1724.

BRANDENBURGER, A., A. FRIEDENBERG, AND H. J. KEISLER (2008): "Admissibility in Games," *Econometrica*, 76, 307–352.

CABRALES, A. AND R. SERRANO (2011): "Implementation in Adaptive Better-response Dynamics: Towards a General Theory of Bounded Rationality in Mechanisms," *Games and Economic Behavior*, 73, 360–374.

CHUNG, K.-S. AND J. ELY (2003): "Implementation with Near-Complete Information," *Econometrica*, 71, 857–871.

DEKEL, E., D. FUDENBERG, AND S. MORRIS (2006): "Topologies on Types," *Theoretical Economics*, 1, 275–309.

DUFWENBERG, M. AND M. STEGEMAN (2002): "Existence and Uniqueness of Maximal Reductions under Iterated Strict Dominance," *Econometrica*, 70, 2007–2023.

DUTTA, B. AND A. SEN (2011): "Nash Implementation with Partially Honest Individuals," *Games and Economic Behavior*, 74, 154–169.

HOLDEN, R., N. KARTIK, AND O. TERCIEUX (2013): "Simple Mechanisms and Preferences for Honesty," Columbia University Economics Discussion Paper No. 1213-18.

JACKSON, M. O. (1992): "Implementation in Undominated.Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, 59, 757–75.

JACKSON, M. O., T. R. PALFREY, AND S. SRIVASTAVA (1994): "Undominated Nash Implementation in Bounded Mechanisms," *Games and Economic Behavior*, 6, 474–501.

KARTIK, N. AND O. TERCIEUX (2012): "Implementation with Evidence," *Theoretical Economics*, 7, 323–355.

KEISLER, H. J. AND B. S. LEE (2011): "Common Assumption of Rationality," Unpublished.

KORPELA, V. (2012): "Bayesian Implementation with Partially Honest Individuals," Unpublished.

LOMBARDI, M. AND N. YOSHIHARA (2011): "Partially-honest Nash implementation: Characterization results," Unpublished.

——— (2013): "Natural Implementation with Partially Honest Agents," Unpublished.

MASKIN, E. (1999): "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, 66, 23–38.

MATSUSHIMA, H. (1988): "A New Approach to the Implementation Problem," *Journal of Economic Theory*, 45, 128–144.

——— (2008a): "Behavioral Aspects of Implementation Theory," *Economics Letters*, 100, 161–164.

——— (2008b): "Detail-free mechanism design in twice iterative dominance: Large economies," *Journal of Economic Theory*, 141, 134–151.

——— (2008c): "Role of Honesty in Full Implementation," *Journal of Economic Theory*, 139, 353–359.

MOORE, J. AND R. REPULLO (1988): "Subgame Perfect Implementation," *Econometrica*, 56, 1191–1220.

——— (1990): "Nash Implementation: A Full Characterization," *Econometrica*, 58, 1083–1099.

ORTNER, J. (2012): "Direct Implementation with Minimally Honest Individuals," Unpublished.

OURY, M. AND O. TERCIEUX (2012): "Continuous Implementation," *Econometrica*, 80, 1605–1637.

PALFREY, T. R. AND S. SRIVASTAVA (1991): "Nash Implementation Using Undominated Strategies," *Econometrica*, 59, 479–501.

POSTLEWAITE, A. AND D. SCHMEIDLER (1986): "Implementation in differential information economies," *Journal of Economic Theory*, 39, 14–33.

SAMUELSON, L. (1992): "Dominated strategies and common knowledge," *Games and Economic Behavior*, 4, 284–313.

SJÖSTRÖM, T. (1994): "Implementation in Undominated Nash Equilibria without Integer Games," *Games and Economic Behavior*, 6, 502–511.