Columbia University

Department of Economics
Discussion Paper Series

Simple Mechanisms and Preferences for Honesty

Richard Holden, Navin Kartik, Olivier Tercieux

Discussion Paper No.: 1213-18

Department of Economics
Columbia University
New York, NY 10027

March 2013

Simple Mechanisms and Preferences for Honesty

Richard Holden, Navin Kartik, and Olivier Tercieux*

January 9, 2013

Abstract

We consider full implementation in abstract 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. We also extend our result to settings of incomplete information so long as there is non-exclusive information.

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.

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

[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 equilibria. Perhaps most prominently, the scope for implementation expands substantially when one considers 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, the mechanisms used in some of these papers are insulated from the critiques levied against the mechanisms used for Nash implementation.

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). An alternative approach that also yields permissive results is that of virtual or approximate implementation (Matsushima, 1988; Abreu and Sen, 1991; Abreu and Matsushima, 1992b). As is well-known, 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.[2] Furthermore, many of these results become much less permissive when the implementation problem concerns only two players, a case that is important due to its relevance for bilateral contracting.

More recently, a burgeoning literature studies the scope for implementation when players have preferences for honesty. Loosely, a (small) preference for honesty means that a player has an intrinsic preference for "truthful" messages/reports when his message does not change the outcome of the mechanism.[3] Intuitively, what now determines

[2]Moreover, Matsushima (1988) and Abreu and Sen (1991) also use integer games. Abreu and Matsushima (1992b) achieve implementation in iterative deletion of (strictly) dominated strategies with finite mechanisms (in a very broad class of finite environments), but rely quite critically on the linearity assumption of expected utility. Furthermore, they use potentially-long chains of iterative dominance reasoning, which has been criticized by, for example, Glazer and Rosenthal (1992); see Abreu and Matsushima (1992a) for a response and Sefton and Yavas (1996) for experimental evidence bearing on the debate. Abreu and Matsushima (1994) use similar ideas as Abreu and Matsushima (1992b) to obtain exact implementation in iterative deletion of weakly-dominated strategies under certain conditions.

[3]In this paper, we focus on settings where preferences for honesty are *small* in the sense that they only play a role when players are indifferent (or nearly indifferent) over outcomes. Of course, in some

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 at least one of the weaknesses mentioned above. For example, Dutta and Sen (2011), Lombardi and Yoshihara (2011), and Kartik and Tercieux (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.[4]

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 general 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 called *separable punishment*, any SCF is implementable in *two rounds of iterative deletion of (strictly) dominated strategies* by a well-behaved 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.

While we defer until later a detailed discussion of the result, a few brief comments are worth noting up front. First, what makes the separable-punishments condition restrictive yet powerful is that the alternative allocation referred to above must be state-

applications, it may also be reasonable that players have large preferences for honesty.

[4]Both Matsushima (2008a) and Matsushima (2008c) achieve implementation in iterated deletion of dominated strategies following the approach of Abreu and Matsushima (1992b, 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 concentrate on complete-information environments in this paper, Section 4 shows how our ideas can be extended to a class of incomplete information environments under very weak assumptions on how players evaluate lotteries (in particular, without assuming expected utility).

independent. We give examples of economic problems in which this condition is naturally satisfied; in particular, it holds both in a standard exchange economy and also in any problem when the mechanism has the ability to augment large-enough monetary punishments on players. The latter is a setting that has received quite some interest, for example in the incomplete contracts literature (see Section 5). Second, implementation in two rounds of iterative deletion of dominated strategies is an appealing solution concept. Since it is weaker than rationalizability, it is a robust solution concept (unlike refinements of Nash, as noted earlier); moreover, the fact that only two rounds of deletion are required also means that players only need only mutual knowledge — rather than common knowledge — that dominated strategies will not be played (cf. Matsushima, 2008b). Third, the mechanism is extremely simple, and is essentially a direct mechanism. Moreover, if, for example, *all* players have a preference for honesty and large monetary punishments are available, implementation is "detail free" in the sense of not depending on fine details of the environment. Finally, we emphasize that the result applies to settings with only two players; not only is this inherently important, but inter alia also disproves a conjecture by Dutta and Sen (2011) that we discuss in Section 3.

The remainder of the paper is organized as follows. Section 2 presents a simple example to highlight the main idea. The general setting and result are provided in Section 3. Section 4 develops an extension to settings of incomplete information using the concept of non-exclusive information (Postlewaite and Schmeidler, 1986), while making minimal assumptions on how players evaluate lotteries. Finally, Section 5 concludes with some comments on robustness and implications for a debate in the incomplete contracts literature.

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 (that he is only willing to use off the equilibrium path). 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 represented by $v_i(a) - t_i$.

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 player 1 claims the state is θ then the outcome selected is $f(\theta)$. In this sense, player 1 is a dictator over outcomes. In terms of transfers, if players disagree on their announcements, then only player 1 is fined an amount t_1; if players agree, there is no transfer to either player.

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

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.

Clearly, the assumption of only two states is not important for the argument. Notice also that we only needed to assume that one player (viz., player 2) has the preference for honesty, but the mechanism exploits 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.

It is noteworthy that not only is the mechanism used here a direct mechanism, but also that it works through (iterative) strict dominance. The latter is important because 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, and moreover, only requires two players.

To summarize the message of this section: by constructing a mechanism wherein one player is indifferent between all strategies in terms of her material payoff, a small preference for honesty can be exploited to ensure that it is a dominant strategy for her to tell the truth. The mechanism then ensures that the unique best response for the other player is to also tell the truth. Below, we will identify a property called *separable punishment* that permits such a construction more generally under complete information, and then extend the result to some settings of incomplete information.

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

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

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 $a \succeq^A_{j,\theta} a'$, $m_j = \theta$, 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 $a \succeq^A_{j,\theta} a'$. 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 every state, this setting would satisfy Definition 1.

We next introduce the domain restriction that will play a central role in our main result.

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 says requires for each state θ', there be an alterna-

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

tive 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 differs from various "bad/worst outcome" conditions in the literature (e.g. Moore and Repullo, 1990; Jackson et al., 1994) 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 $v_i(b, \theta) - t_i$. Further 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 $f(\theta)$.

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

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

The essence of 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. Plainly, this mechanism is an extremely simple direct mechanism: there is only one round of messages and obviously no use of integer games or related ideas.

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. In fact, Dutta and Sen (2011, page 166) discuss 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 (2011). They show that in a complete-information setting where any pair of states can be distinguished via hard evidence by some player, 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 their conclusion also holds if players have a small cost of forging

[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.

[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.

evidence. Settings with preferences for honesty are a special case of settings with 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 Incomplete Information

In this section, we depart from the complete-information assumption. The goal is to identify relatively simple conditions under which the logic underlying Theorem 1 can be extended to settings of incomplete information.

4.1 Setting and Definitions

In a setting of incomplete information, each player i has a set of types T_i, with a generic element denoted by t_i. For technical convenience, each T_i is assumed to be a finite set. The set of profiles of types is denoted $T := T_1 \times \cdots T_n$ and P is the prior probability over T. We let $T^* \subseteq T$ be the set of types with positive prior probability. A social choice function is now $f : T^* \to A$, where A is the space of allocations. Without loss, we assume that $P(t_i) := \sum_{t_{-i} \in T_{-i}} P(t_i, t_{-i}) > 0$ for each i and t_i. Let the conditional probability distribution given any t_i be $P(t_{-i} \mid t_i) := \frac{P(t_i, t_{-i})}{P(t_i)}$. To avoid some inessential complications involving conditioning on zero-probability events, we will further assume that for each j and each t_{-j}, $P(t_{-j}) := \sum_{t_j \in T_j} P(t_j, t_{-j}) > 0$. An *incomplete information environment* is a tuple of players, spaces of allocations and type profiles, and a prior, i.e. $\langle I, A, T, P \rangle$. Throughout what follows, we fix such an environment.

A key ingredient that helps extend our earlier argument to the current setting is that no player has exclusive information, a notion that is familiar in the Bayesian implementation literature (e.g. Postlewaite and Schmeidler, 1986). Informally, there is *non-exclusive information* (NEI) if, whenever the designer learns the profile of types of players other

[9] Ben-Porath and Lipman (2011) 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.

than j, he can infer what the true type of player j is. Formally:

Definition 3. The incomplete-information environment satisfies NEI if for each j and t_{-j} there is some t_j such that $P(t_j \mid t_{-j}) = 1$.

It is important to note that the complete-information environment considered in the previous section can be viewed as an incomplete-information environment satisfying NEI: set $T_i = \Theta$ for all i and $P(t_1, \ldots, t_n) > 0$ if and only if $(t_1, \ldots, t_n) = (\theta, \ldots, \theta)$ for some $\theta \in \Theta$. Note also that when there are only two players, NEI implies complete information.

Given a profile of types t, each agent j has an *ex-post* preference order $\succeq^A_{t,j}$ over allocations, with $\sim^A_{t,j}$ denoting the corresponding ranking of indifference. Given a space of message profiles, $M = M_1 \times \cdots \times M_n$, and a profile of types, t, a player j also has ex-post preferences over $A \times M_{-j}$ denoted by $\succeq_{t,j}$. To simplify the exposition, we will focus hereafter on direct mechanisms, i.e. assume the message space for each player i is $M_i = T_i$.[10] We can now define preferences for honesty under incomplete information:

Definition 4. Agent j has a *preference for honesty* on T if for any $g : T \to A$ and $t_j \in T_j$: If for all t_{-j} for which $P(t_{-j} \mid t_j) > 0$, it holds that

$$\forall t'_{-j}, t'_j : \; g(t'_{-j}, t_j) \sim^A_{(t_j, t_{-j}),j} g(t'_{-j}, t'_j), \tag{3}$$

then for all t_{-j} for which $P(t_{-j} \mid t_j) > 0$, it also holds that

$$\forall t'_{-j} \text{ and } \forall t'_j \neq t_j : \; (g(t'_{-j}, t_j), t'_{-j}, t_j) \succ_{(t_j, t_{-j}),j} (g(t'_{-j}, t'_j), t'_{-j}, t'_j). \tag{4}$$

Intuitively, the above definition says that if player j with a given type t_j is indifferent over all allocations he can induce when taking as given an arbitrary report of his opponents, then he strictly prefers to tell the truth. Note that the antecedent is demanding, and hence the requirement overall is weak. It is not hard to verify that, for direct message spaces, the above definition reduces to Definition 1 when there is complete information.

[10] One can extend the subsequent development to indirect mechanisms in a manner similar to Section 3.

Next, we state the domain restriction that is the analog of Definition 2 for incomplete-information environments:

Definition 5. Fix an incomplete information environment satisfying NEI and some SCF $f : T^* \to A$. There is *Bayesian separable punishment* if there is a function $x : T \backslash T^* \to A$ and a player i such that for all $j \neq i$ and for all $t \in T^*$:

1. $x(t'_j, t'_{-j}) \sim^A_{t,j} f(t^*_j(t'_{-j}), t'_{-j})$ for all t'_{-j} and t'_j satisfying $(t'_j, t'_{-j}) \notin T^*$, where $t^*_j(t'_{-j})$ is the unique type such that $(t^*_j(t'_{-j}), t'_{-j}) \in T^*$;

2. $x(t'_i, t_{-i}) \prec^A_{t,i} f(t_i, t_{-i})$ for all t'_i satisfying $(t'_i, t_{-i}) \notin T^*$.

The intuition behind this condition is related to that underlying separable punishment in the complete-information environment: one should be able to find allocations for certain profiles of types that keep some players indifferent relative to certain socially-desired allocations, while being worse for other players than some socially-desired allocations. To see that Definition 5 strictly generalizes Definition 2, suppose $T_i = \Theta$ for each i and the prior $P(t) > 0$ if and only if $t = (\theta, ..., \theta)$. Assume the complete-information separable punishment per Definition 2, yielding a function $x(\theta)$. Now define the function $\hat{x} : T \backslash T^* \to A$ as follows: for any $t \in T \backslash T^*$, $\hat{x}(t) := x(t_i)$. It is readily verified that the function $\hat{x}(\cdot)$ satisfies the requirement of Definition 5.

It is worth highlighting, however, that in a setting with genuine incomplete information, Bayesian separable punishment is more demanding in spirit that its complete-information counterpart because Definition 5 uses a universal quantifier over players $j \neq i$ while Definition 2 only has an existential quantifier. The reason is straightforward: under incomplete information, we will need to elicit the type of each of $n-1$ players (with the last player taken care of by NEI), whereas in the complete information setting, it suffices to elicit the type of just one player.

Nevertheless, there are interesting economic environments with incomplete information that satisfy Bayesian separable punishment. For instance, assume player 1 is a seller and the other $n-1$ other players are buyers. Assume further that the seller has no exclusive information. The outcome space is $A_1 \times \prod_{j \neq 1} A_j$ where for $j \neq 1$, $A_j = \mathbb{R}^2_+$ specifies the price paid to the seller and the quantity of goods obtained by each buyer

while $A_1 = \mathbb{R}_+$ specifies a fine paid by the seller. Each buyer j cares only about the price he pays and the quantity of goods he gets, i.e., he only cares about the j^{th} component of the outcome space; the seller cares, of course, about the price paid by each buyer and the fine he must pay (he may also care about the quantities of goods he need to provide, if, for example, there is a cost of production). Let f be any social choice function such that for any $t \in T^*$, $f(t)$ involves no fine paid by the seller. Assume that a sufficiently severe fine on the seller is worse for him than any outcome in the range of the social choice function f. Then, construct the function x as follows.[11] For all $j \neq 1$ and $(t'_j, t'_{-j}) \notin T^* : x_j(t'_j, t'_{-j}) = f_j(t''_j, t'_{-j})$ where t''_j is the unique type of player j satisfying $(t''_j, t'_{-j}) \in T^*$. In addition, for all $(t'_i, t'_{-i}) \notin T^*$, $x_1(t'_i, t'_{-i})$ is a large enough fine such that it is worse for player 1 than $f_1(t''_i, t'_{-i})$ where t''_i is the unique type of player i satisfying $(t''_i, t'_{-i}) \in T^*$. It is straightforward that $x(\cdot)$ so-defined satisfies the requirement of Definition 5.

4.2 The Result

To state our result for incomplete information, we must take some stand on how players evaluate lotteries (even though, as before, for simplicity alone we continue to focus on pure strategies only). We will use a very weak notion of iterative deletion of strictly dominated strategies. More precisely, a message t'_j for player j of type t_j is said to be *ex-post strictly dominated* by another message t''_j if, for any profile t_{-j} that has positive probability under $P(\cdot \mid t_j)$, and for any announcement t'_{-j} of j's opponents, it holds that $(g(t''_j, t'_{-j}), t''_j, t'_{-j}) \succ_{(t_j, t_{-j}), j} (g(t'_j, t'_{-j}), t'_j, t'_{-j})$. For each profile of types, we can compute the set of ex-post strictly dominated messages for each player. Based on this notion, iterative deletion of ex-post strictly dominated strategies can be defined in the usual way.[12] This is a coarse solution concept: for an arbitrary direct mechanism, many strategies would survive this process of elimination. Hence, full implementation in iterative deletion of ex-post dominated strategies is a demanding notion. Nevertheless, we have:

[11] In what follows, we use standard notation and write $x_i(\cdot)$ as well as $f_i(\cdot)$ for the projections of $x(\cdot)$ and $f(\cdot)$ on A_i.

[12] This is reminiscent of the notion of implementation used by Bergemann and Morris (2008). However, their notion does not require that the profile of types t_{-j} have positive probability for player j of type t_j, because agents do not have priors in their setting.

Theorem 2. *Fix an incomplete information environment satisfying NEI and any SCF $f : T^* \to A$. Assume Bayesian separable punishment and fix i from that definition. Assume further that each player $j \neq i$ has a preference for honesty on T. Then the SCF f can be implemented in two rounds of iterated deletion of ex-post strictly dominated strategies.*

The proof is in the Appendix. To see how this result generalizes Theorem 1,[13] recall that in the complete-information case, the set of types for each player i is Θ and NEI is satisfied. Note also that under the assumptions of Theorem 1, we could restrict the problem to only two players i and j as defined in that theorem. In this smaller environment, NEI is still satisfied. Finally, as explained earlier, separable punishment implies Bayesian separable punishment.

5 Conclusion

While the logic of our proofs rely on some players being indifferent over the outcomes they can induce through their messages, our conclusions are not knife-edged. The reason is that our results deliver implementation in rationalizable strategies (requiring only two rounds of iterative deletion of dominated strategies). Given an incomplete information game, Dekel et al. (2006) show that all rationalizable actions in any nearby game derived by small perturbations to lower-order beliefs and arbitrary perturbations to higher-order beliefs must in fact be rationalizable actions in the original complete-information game.[14] Although our setting differs from theirs because we consider lexicographic preferences over the joint space of messages and outcomes, our results would also continue to hold for a large class of perturbations of economically-natural environments satisfying separable punishment.[15] To see this, consider in particular any setting where the designer can augment (large) individual transfers to an underlying outcome

[13] Theorem 1 also applies to indirect mechanisms. We ignore this aspect in the comparison since it is not essential.

[14] More precisely: the correspondence of interim-correlated rationalizable strategies is upper hemi-continuous in the product topology in the universal type space.

[15] In fact, given that the k^{th} round of elimination of the rationalizability process only depends on the k^{th} first order beliefs (Dekel et al., 2007), it can be shown that our result would continue to hold so long as the perturbation are small for only first and second order beliefs; they can be arbitrary for third and higher order beliefs.

space and players have quasi-linear preferences over underlying outcomes and their own transfer.[16] In the mechanism used in the proof of Theorem 1, player j is outcome-indifferent between all his messages because they only affect the transfer to the other player, i, and hence has a strictly dominant strategy to report the truth owing to his preference for honesty. Hence, virtually no matter how j's preferences over underlying outcomes are perturbed — so long as the transfer to i does not matter to j — truth-telling would continue to be strictly dominant. The mechanism's construction then implies that i's unique iteratively strictly dominant strategy would be to tell the truth even after any perturbation to his preferences.

Consequently, our main results would continue to hold if, for example, agents have slight uncertainty about either their own or others' payoffs from outcomes, or about whether others have a preference for honesty. Such robustness contrasts with many implementation results that rely rather heavily on higher-order beliefs (see Oury and Tercieux, 2012).

We conclude by noting that our results has some relevance for a debate on foundations of incomplete contracting. Theories of incomplete contracts (e.g. the Property Rights Theory of Grossman and Hart, 1986) that rely on information being observable to the contracting parties but not verifiable to a third party (such as a court of law) suffer from a "message game critique". Maskin and Tirole (1999), for example, show how one can effectively make information verifiable in the unique subgame-perfect equilibrium of a suitable mechanism following Moore and Repullo (1988), at least if renegotiation can be eliminated and some other conditions are satisfied.[17] This critique has been called into question by Aghion et al. (2012) who show that the conclusion is not robust to an arbitrarily small weakening of common knowledge between the contracting parties.[18] By contrast, Theorem 1 does not suffer this weakness because of the robustness

[16]The linearity is not important but is rather a convenient shorthand.

[17]The issue of renegotiation is, of course, important but is itself debated in the literature from various perspectives, such as whether it can be eliminated, how best to model it, and whether accounting for it precludes ex-ante efficiency or not. For our purposes, it suffices to note that the message game critique, if valid, forces one to find alternative foundations for incomplete contracts than mere non-verifiability, whether the alternative is based on renegotiation or other reasons.

[18]Specifically, they show that an arbitrarily small perturbation that departs from common knowledge eliminates the "truth-telling" equilibrium, and furthermore, any extensive-form mechanism admits an undesirable sequential equilibrium in the perturbed game given a non-monotonic social choice function.

properties discussed above. In this sense, in contracting environments when separable punishment is available, small preferences for honesty resuccitates the "message game critique" of just using non-verifiable but observable information as a foundation for incomplete contracts. We emphasize two caveats: first, while separable punishment holds in some contracting environments of interest (e.g. Maskin and Tirole, 1999, Section 4), the condition is not implied by joint punishments such as a no-trade outcome, for reasons discussed immediately after Definition 2. Second, Aghion et al. (2012, Section 5) point out in an example that introducing significant amount of asymmetric information at the ex-post stage can provide a justification for the Property Rights Theory when players have standard preferences; it is easy to verify that their example would go through even under small preferences for honesty. Our Theorem 2 reveals that a failure of non-exclusive information would be key to any such justification; indeed, their example has just two players.

Appendix: Proof of Theorem 2

Fix player i from the Theorem's hypotheses. Consider the direct mechanism whose outcome function is given by

$$g(t) = \begin{cases} f(t) & \text{if } t \in T^* \\ x(t) & \text{otherwise.} \end{cases} \tag{5}$$

Fix any player $j \neq i$ and any type t_j of this player. Consider any profile of other types for players other than j, t_{-j}, such that $(t_j, t_{-j}) \in T^*$, and fix an arbitrary message profile for these players, t'_{-j}.

Claim: For any t'_j, $g(t_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} g(t'_j, t'_{-j})$.

Proof: The claim is trivially true if $t'_j = t_j$, so pick an arbitrary $t'_j \neq t_j$. Assume first that $(t_j, t'_{-j}) \in T^*$. Then, (5) implies $g(t_j, t'_{-j}) = f(t_j, t'_{-j})$. Further, $(t'_j, t'_{-j}) \notin T^*$ and $g(t'_j, t'_{-j}) = x(t'_j, t'_{-j})$, where the first equality is by NEI and the second by (5). Since $(t_j, t'_{-j}) \in T^*$, Bayesian separable punishment (Definition 5) now implies that $g(t'_j, t'_{-j}) = x(t'_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} f(t_j, t'_{-j}) = g(t_j, t'_{-j})$, as claimed.

Next, assume $(t_j, t'_{-j}) \notin T^*$ and $(t'_j, t'_{-j}) \notin T^*$. Then, (5) yields $g(t_j, t'_{-j}) = x(t_j, t'_{-j})$ and $g(t'_j, t'_{-j}) = x(t'_j, t'_{-j})$. By NEI, there is a unique type $\hat{t}_j$ satisfying $P(\hat{t}_j \mid t'_{-j}) = 1$, and moreover $(\hat{t}_j, t'_{-j}) \in T^*$. Bayesian separable punishment then implies $g(t'_j, t'_{-j}) = x(t'_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} f(\hat{t}_j, t'_{-j})$ as well as $g(t_j, t'_{-j}) = x(t_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} f(\hat{t}_j, t'_{-j})$. Hence, by transitivity, $g(t_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} g(t'_j, t'_{-j})$, as desired.

Finally, it remains to consider $(t_j, t'_{-j}) \notin T^*$ and $(t'_j, t'_{-j}) \in T^*$. Then, (5) yields $g(t_j, t'_{-j}) = x(t_j, t'_{-j})$ and $g(t'_j, t'_{-j}) = f(t'_j, t'_{-j})$. By NEI and Bayesian separable punishment, it follows that $x(t_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} f(t'_j, t'_{-j})$, and hence $g(t_j, t'_{-j}) \sim^A_{(t_j,t_{-j}),j} g(t'_j, t'_{-j})$, as desired. ‖

Since j has a preference for honesty on T, the Claim above implies the following for all t_{-j} for which $(t_j, t_{-j}) \in T^*$:

$$\forall t'_{-j} \text{ and } \forall t'_j \neq t_j : \ (g(t'_{-j}, t_j), t'_{-j}, t_j) \succ_{(t_j,t_{-j}),j} (g(t'_{-j}, t'_j), t'_{-j}, t'_j).$$

In other words, it is ex-post strictly dominant for player j (who, recall was an arbitrary player different from i) to report the truth.

Now, consider player i of type t_i and consider any profile t_{-i} such that $(t_i, t_{-i}) \in T^*$. NEI and (5) imply that when all his opponents tell the truth, a truthful report from i yields the outcome $f(t_i, t_{-i})$ while lying by reporting $t'_i \neq t_i$ yields $x(t'_i, t_{-i}) \prec^A_{(t_i,t_{-i}),i} f(t_i, t_{-i})$, where the strict preference inequality is by Bayesian separable punishment. Hence, given that it is ex-post strictly dominant for all his opponents to report their types truthfully, it is iteratively ex-post strictly dominant for player i to also report truthfully. *Q.E.D.*

References

ABREU, D. AND H. MATSUSHIMA (1992a): "A Response to Glazer and Rosenthal," *Econometrica*, 60, 1439–1442.

——— (1992b): "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 Value Perturbations," Unpublished.

BEN-PORATH, E. AND B. L. LIPMAN (2011): "Implementation with Partial Provability," Unpublished.

BERGEMANN, D. AND S. MORRIS (2008): "Ex post implementation," *Games and Economic Behavior*, 63, 527–566.

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.

——— (2007): "Interim Correlated Rationalizability," *Theoretical Economics*, 2, 15–40.

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

GLAZER, J. AND R. W. ROSENTHAL (1992): "A Note on Abreu-Matsushima Mechanisms," *Econometrica*, 60, 1435–38.

GROSSMAN, S. J. AND O. D. HART (1986): "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, 94, 691–719.

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*, forthcoming.

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

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

MASKIN, E. AND J. TIROLE (1999): "Unforeseen Contingencies and Incomplete Contracts," *Review of Economic Studies*, 66, 83–114.

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*, forthcoming.

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.

SEFTON, M. AND A. YAVAS (1996): "Abreu-Matsushima Mechanisms: Experimental Evidence," *Games and Economic Behavior*, 16, 280–302.