Non-Consequential Moral Preferences, Detail-Free Implementation, and Representative Systems[+]

Hitoshi Matsushima[*]

Faculty of Economics, University of Tokyo

First Version: March 4, 2002
This version: October 11, 2004

Abstract

We investigate implementation of social choice functions where the central planner has no knowledge about the detail of model specifications, and only a few individuals participate in the mechanism. In contrast with the standard model of implementation, each agent has non-consequential moral preference in that she prefers truth-telling to lying whenever the resulting consequence is unchanged. We show that with complete information, there exists a single, detail-free mechanism that can implement any social choice function whenever agents regard its value as being socially desirable. This result holds even if psychological cost for lying is close to zero. Non-consequential moral preferences play a very powerful role in eliminating unwanted equilibria in detail-free mechanism design with representative systems. We extend this result to incomplete information.

Keywords: Unique Implementation, Non-Consequential Moral Preferences, Detail-Free Mechanism Design, Representative Systems.

JEL Classification Numbers: C72, D71, D78, H41

[+] This paper is a revised version of the manuscript entitled "Universal Mechanisms and Moral Preferences in Implementation" (Discussion Paper CIRJE-F-254, Faculty of Economics, University of Tokyo, 2003). I would like to thank the co-editor and anonymous referees for their comments. The research for this paper was supported by Grant-In-Aid for Scientific Research (KAKENHI 15330036) from JSPS and MEXT of the Japanese Government.

[*] Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan. Fax: +81-3-3818-7082. E-mail: hitoshi@e.u-tokyo.ac.jp

1. Introduction

This paper demonstrates a new approach to the implementation problem, where a social choice function, which is defined as a mapping from states to consequences, is said to be *implementable* in any equilibrium concept such as Nash equilibrium and iterative dominance, if the central planner (the constitution designer, or the auctioneer) can design a mechanism, in which at every state, there exists the *unique* equilibrium outcome, and this outcome equals the associated value of the social choice function. This paper will assume that the central planner has *no* knowledge about the detail of model specifications such as the set of states, the utility functions, and the social choice function. Hence, the central planner can only design mechanisms that are *detail-free*, i.e., are independent of this detail. This paper will also assume that the central planner is faced with restrictions of *representative systems*, where she can invite only a few individuals in the society to participate in the mechanism as *agents*.

This paper will allow agents to have, not only preferences for consequences, but also *non-consequential moral preferences*. This implies that each agent's welfare depends, not only on the consequence of the central planner's choice, but also on whether she tells the truth or lies in the process of public decision. This point is in sharp contrast with the standard model of implementation where each agent has preference only for consequences, i.e., each agent's welfare at any state depends only on the central planner's choice of alternative and monetary transfer to her. For surveys on the implementation literature, see Moore (1992), Palfrey (1992), Osborne and Rubinstein (1994, Chapter 10), and Maskin and Sjöström (2002), for instance.[1] This paper will then show a very permissive result that *the central planner can design a single mechanism that is detail-free but is almighty in the sense that, irrespective of how the set of states to be specified, it can implement any social choice function as long as agents regard its value as being socially desirable at any state.* This result holds even if agents' moral preferences are very *weak* relatively to their preferences for consequences.

The mechanisms designed and used in the implementation literature have depended on the very detail of model specifications. In real situations, however, the central planner may not be informed of this detail in advance. Even if she could know this detail, it may be impossible to describe it on a document because of its complexity. This is the reason why the previous studies of implementation were not much successful as a practical theory, even if there are many important contributions on its own right. In order to put implementation theory into practice, it is inevitable to restrict our attentions only to *detail-free mechanism design*. As the experts in this literature have known, however, in the standard model, it is impossible for the central planner to design a mechanism that is detail-free but can implement multiple social choice functions in Nash equilibrium.

[1] There are a few exceptions. For example, Eliaz (2002) took into account factors other than individuals' preferences for consequences such as bounded rationality. Glazer and Rubinstein (1998) allowed agents to have non-consequential preferences.

In real situations, it may be costly for the central planner to invite all individuals in the society to participate in the mechanism. Hence, it is quite important to investigate implementation with restrictions of representative systems where only a few individuals are allowed to participate in the mechanism as agents. The previous works on implementation, however, have never investigated such representative systems, and instead have implicitly assumed that any individual participates in the mechanism whenever her preference influences the value of the social choice function. In the standard model, this assumption is even necessary for showing any possibility result. In fact, we can easily check that it is mostly impossible to implement any social choice function whenever there is an individual who does not participate in the mechanism but whose preference influences the value of this social choice function.

From these observations, we must conclude that as long as we stick to the standard model of implementation, no devices of mechanism design work in practical situations with limited knowledge of the central planner and with representative systems. Based on this, the main departure of this paper from the previous works is to give up the standard model and to investigate an alternative model where each agent has, not only preference for consequences, but also non-consequential preference. Here, each agent treats the same consequence of the central planner's choice *differently*, depending on the process that leads up to it. In particular, she strictly prefers truth-telling to lying whenever the resulting outcome is unchanged between truth-telling and lying. Whether she tells the truth or lies has its own *intrinsic* value for her welfare.

Surely, this non-consequentialist view is inconsistent with the neo-classical framework of "homo economicus", who cares only about the consequence in a purely selfish way.[2] This view, on the other hand, is very *consistent* with empirical and experimental evidences. For instance, the recent work by Gneezy (2002) did very important experimental researches, showing that most subjects in the laboratory prefer truth-telling to lying when whether lying or not does not much influences the consequence. For any real human being, decision to lie is more or less a matter of weighting costs and benefits.

This paper will show a new idea of mechanism design that describes the following decision procedure. The central planner will invite just *three* individuals as agents 1, 2, and 3, respectively. She will ask each agent to make *multiple* announcements about which alternative to be socially desirable. The number of their multiple announcements is set to be sufficiently large. The central planner will then pick up one announcement profile at random among all profiles except the first profile. If two or three agents recommend the same alternative, she will choose it. Otherwise, she will choose the status quo. She will fine any agent a small monetary amount whenever this agent is either agent 2 or agent 3 and is the first deviant from agent 1's first announcement. Clearly, this mechanism is detail-free. This mechanism is also consistent with representative systems, because we need just three individuals to participate in it.

---

[2] Preference for consequence does not necessarily imply pure self-interest. In fact, most models of *social preferences* falls under the category of preference for consequence. See Fehr and Schmidt (1999), for instance.

We consider the *complete* information environments, where which alternative to be socially desirable is common knowledge among agents. Each agent's preference is defined as a combination of preference for consequences and non-consequential moral preference. Here, each agent has positive psychological cost for lying, i.e., for not announcing the socially desirable alternative honestly. We assume that the more the number of dishonest announcements she makes, the more psychological cost she has. The total psychological cost can be close to zero even if all announcements by her are dishonest. In other word, the present paper will simply exclude the case that agents have preferences only for consequences as the standard model has assumed.

The introduction of non-consequential moral preferences in this way plays a very powerful role as follows. Irrespective of how the set of states to be specified, any social choice function can be implemented in iterative dominance (and therefore, in Nash equilibrium) by the single, detail-free mechanism with just three agents, whenever at any state these agents regard the associated value of this social choice function as being socially desirable. Agents' non-consequential moral preferences can be very week relatively to their preferences for consequences. Hence, we can say that implementation never works in the consequentialist model, whereas the best possibility result for implementation holds in our non-consequentialist model.

Even if we permit the central planner to have enough knowledge about the detail of model specifications, and we also permit all relevant individuals to participate in the mechanism, i.e., even if we do not take restrictions of detail-free mechanism design and representative systems into account, our model has the following great advantage over the standard model. It is clear from the very definition of implementability that distinct states must correspond to the same value of the social choice function if these correspond to the same preference profile of agents. For related arguments, see Maskin and Tirole (1999) and Tirole (1999). This implies that in the standard model, any implementable social choice function must depend *only* on agents' preferences for consequences. This, however, will very severely restrict the range of implementable social choice functions. As many important attempts to establish foundations of social choice and welfare such as Rawls (1971), Dworkin (1981), and Sen (1982, 1985, 1999) has emphasized, several ethical factors of the state are crucial to determine which social choice function to be right, and these factors should be sharply distinct from agents' preferences for consequences.[3] Even if the central planner could know these factors in advance, the relevance of them to social welfare may be too complicated to describe on a document. Hence, we must conclude that in the standard model, no ethically right social choice function is implementable. In contrast, in our model, whenever distinct states correspond to distinct values of the social choice function that agents regard as being socially desirable, then these states always correspond to distinct non-consequential moral preferences for each agent. This will be the driving force of implementing any ethically right social choice function in our model.[4]

---

[3] Rawls introduced primary goods. Dworkin introduced compensation and responsibility. Sen introduced liberty, functioning, and capabilities. These are concepts categorized into non-consequentialism. See Basu, Pattanaik, and Suzumura (1995), Sen (1999), and Suzumura (2002).

[4] The disadvantages possessed by the standard model matter irrespective of implementation being

In the latter part of this paper, we will extend the above arguments to the *incomplete* information environments. Each agent receives her private signal in advance, including only *partial* information about which alternative to be socially desirable. A state is defined as a profile of agents' private signals, and an alternative is defined as a bundle of characteristics. Each characteristic of the socially optimal alternative is known to an agent in advance through her private signal.

The central planner will ask each agent to make multiple announcements about what she knows about the socially desirable alternative. Each agent has non-consequential moral preference in that she has positive psychological cost for not honestly announcing the characteristic of the socially desirable alternative that she knows. Their costs can be as close to zero as possible.

Even with incomplete information, we can show a very permissive result that *there exists a single mechanism that can implements any social choice function in iterative dominance if this social choice function satisfies (a variant of) incentive compatibility. This mechanism is independent of the set of states, the probability function on the set of states, agents' state-contingent preferences for consequences, and even the social choice function.*

We would like to emphasize that the main role of non-consequential moral preferences in this paper is not to weaken the restriction of incentive compatibility but to eliminate unwanted equilibria. Even if agents' psychological costs are close to zero, their non-consequential moral preferences play this role perfectly.[5]

The organization of this paper is as follows. Section 2 shows the basic model. Section 3 considers the complete information environments. Subsection 3.1 defines implementation in iterative dominance. Subsection 3.2 specifies agents' utility forms. Subsection 3.3 designs a detail-free mechanism. Subsection 3.4 shows and proves the first main theorem of this paper. Subsection 3.5 discusses the implications of this theorem, where we will argue that the use of small fines is particularly important in representative systems, and without fines, even the canonical mechanisms do not work. We will also argue that the central planner does not even need to know the set of possible alternatives.

Section 4 considers the incomplete information environments. Subsection 4.1 defines implementation in iterative dominance in the Bayesian framework. Subsection 4.2 specifies utility forms. Subsection 4.3 designs a detail-free mechanism. Subsection 4.4 shows and proves the second main theorem. Subsection 4.4.5 generalizes this theorem by allowing the central planner to know the set of states and the social choice function in advance.

'virtual' or 'exact'. See Subsection 3.5.4.

[5] Several works such as Erard and Feinstein (1994), Alger and Ma (2003), and Deneckere and Severinov (2001) examined the case including agents who have non-consequential preference for honesty. These works assumed that their costs for lying are very large.

2. Basic Model

Let $N = \{1,...,n\}$ denote the set of *agents* who participate in the public decision procedure as being representative of all individuals in the society, where $n \geq 2$. We must note that N may be a proper subset of the set of individuals in the society who are influenced by the central planner's decision.

Let A denote the set of *alternatives*. Let Δ denote the set of simple lotteries over alternatives. Let M_i denote the set of *messages* for each agent $i \in N$. Let $M = \prod_{i \in N} M_i$ denote the set of message profiles. Let $m_i \in M_i$ and $m = (m_i)_{i \in N} \in M$.

Fix a positive real number $\varepsilon > 0$ arbitrarily, which can be close to zero. We regard ε as the upper bound of monetary fines. Given the set of message profiles M, a *mechanism* is defined by

$$G = (x, t),$$

where

$$x : M \to \Delta,$$
$$t = (t_i)_{i \in N},$$
$$t_i : M \to [-\varepsilon, \infty),$$

and t satisfies the *budgetary constraint* in the sense that

$$\sum_{i \in N} t_i(m) \leq 0 \text{ for all } m \in M.$$

The central planner will ask each agent $i \in N$ to announce a message $m_i \in M_i$. When agents announce $m \in M$, the central planner will choose an alternative according to the simple lottery $x(m) \in \Delta$, and she will make a monetary transfer $t_i(m) \in [-\varepsilon, \infty)$ to each agent $i \in N$.

3. Complete Information

This section investigates the following complete information environments. We assume $n = 3$, i.e.,

$$N = \{1,2,3\}.$$

We must note that there may be other individuals in the society who do not participate in the decision procedure but are influenced by the central planner's decision. This section will show that only *three* participants are enough for the central planner to implement any alternative as long as these participants regard it as being socially desirable. This implies that implementation works even under severe restrictions of representative systems.

Let $a^* \in A$ denote the *socially desirable alternative*. The fact that a^* is socially desirable is common knowledge among agents, whereas the central planner does not know which alternative to be socially desirable.

A *utility function* for each agent $i \in N$ is defined by

$$u_i : A \times R \times M_i \to R,$$

where $u_i(a, t_i, m_i)$ denote agent $i's$ utility when she announces message $m_i \in M_i$, the central planner chooses alternative $a \in A$, and the central planner makes monetary transfer $t_i \in R$ to agent i. We will allow each agent's announcement to have *intrinsic* value for her welfare, which implies that each agent may have *non-consequential* preference as well preference for consequences.

We assume the expected utility hypothesis. Let $u = (u_i)_{i \in N}$ denote a utility function profile. The utility function profile u is common knowledge among the agents, whereas the central planner does not know it.

3.1. Implementation in Iterative Dominance

A combination (G,u) defines a *game*. The solution concept used in this section is *iterative dominance*, where the agents announce any message profile that survives after iterative eliminations of undominated messages.

Let $M_i^{(0)} = M_i$ and $M^{(0)} = \prod_{i \in N} M_i^{(0)}$. Recursively, for every $r = 1,2,...$, let $M_i^{(r)}$ denote the set of messages $m_i \in M_i^{(r-1)}$ for each agent i that are *undominated with respect to* $M_{-i}^{(r-1)} = \prod_{j \in N/\{i\}} M_j^{(r-1)}$ in the sense that there exists no $m_i' \in M_i$ such that for every $m_{-i} \in M_{-i}^{(r-1)}$,

$$u_i(x(m), t_i(m), m_i) < u_i(x(m_i', m_{-i}), t_i(m_i', m_{-i}), m_i').$$

Let $M^{(r)} = \prod_{i \in N} M_i^{(r)}$ and $M^{(\infty)} = \bigcap_{r=0}^{\infty} M^{(r)}$. A message profile $m \in M$ is said to be *iteratively undominated in the game* (G,u) if

$$m \in M^{(\infty)}.$$

The socially desirable alternative $a^* \in A$ is said to *be implemented in iterative dominance in the game* (G,u) if there exists the unique iteratively undominated message profile m, and this profile satisfies that

$$x(m) = a^*,$$

and

$$x_i(m) = 0 \text{ for all } i \in N.$$

We must note that whenever a^* is implemented in iterative dominance in (G,u), then it is implemented also in *mixed strategy Nash equilibrium* in (G,u).

3.2. Assumptions

This section assumes that there exists a positive integer $K > 0$ such that

$$M_i = A^K \text{ for all } i \in N,$$

where K is sufficiently large. The central planner will ask each agent to make K announcements at once about which alternative to be socially desirable. Let

$$M_i = M_{i,1} \times \cdots \times M_{i,K}$$

where

$$M_{i,k} = A.$$

Let $m_i^* \in M_i$ denote the *honest* message for agent i such that

$$m_{i,k}^* = a^* \text{ for all } k \in \{1,...,K\},$$

where agent i K times honestly announces the socially desirable alternative. Let $m^* = (m_i^*)_{i \in N} \in M$ denote the honest message profile.

Fix a positive real number $d > 0$ arbitrarily, which is sufficiently large. We regard d as the upper bound of each agent's utility differences for consequences, as we will explain below. We assume that each agent $i's$ utility function u_i satisfies that there exist a function $v_i : A \to R$ and a positive real number $c_i > 0$ such that

(1) $$u_i(a,t_i,m_i) = v_i(a) + t_i - \frac{\#\{k \in \{1,...,K\} : m_{i,k} \neq a^*\}}{K} c_i,$$

and

(2) $$\max_{(a,a',i) \in A^2 \times N} |v_i(a) - v_i(a')| \leq d.$$

We regard c_i as the upper bound of agent $i's$ *psychological* cost for not announcing the socially desirable alternative honestly. Each agent's non-consequential moral preference is described by the psychological cost for not announcing the socially desirable alternative honestly, which is given by

$$\frac{\#\{k \in \{1,...,K\} : m_{i,k} \neq a^*\}}{K} c_i ,$$

and is increasing with respect to the number of agent $i's$ dishonest announcements $\#\{k \in \{1,...,K\} : m_{i,k} \neq a^*\}$.

The function v_i represents agent $i's$ preference for consequences. Inequalities (2) imply that d is an upper bound of agents' utility differences for consequences. From equalities (1), it follows that each agent's preference is defined as a combination of preference for consequences and non-consequential moral preferences.

The standard model of implementation assumes that c_i equals zero for all $i \in N$, i.e., each agent has only preference for consequences. In contrast, our model in this section assumes that for every $i \in N$, c_i must be positive but can be as close to *zero* as possible.

We assume that the number of announcements K is large enough to satisfy

(3) $$(K-1)\varepsilon > d .$$

This implies that the mechanism may depend on the upper bound d of utility differences for consequences. This section, however, designs mechanisms that never depend on any more details of model specifications.

3.3. Mechanism Design

Fix an alternative $\bar{a} \in A$ arbitrarily, which is regarded as the *status quo*. We design a mechanism $G^* = (x^*, t^*)$ as follows.

For every $m \in M$ and every $a \in A/\{\bar{a}\}$, let

$$x^*(m)(a) = \frac{\#\{k \in \{2,...,K\} : m_{i,k} = a \ \textit{for two or three agents}\}}{K-1},$$

and

$$x^*(m)(\bar{a}) = 1 - \sum_{a \in \Gamma} x^*(m)(a),$$

where

$$\Gamma = \{a \in A/\{\bar{a}\} : x^*(m)(a) > 0\},$$

i.e., $\Gamma \bigcup \{\bar{a}\}$ is the support of x^*. For every $k \in \{2,...,K\}$, with probability $\frac{1}{K-1}$, the central planner will pick up the $k-th$ announcement profile $(m_{1,k}, m_{2,k}, m_{3,k})$, and she will

then choose any alternative $a \in A$ if at least two agents announce it as their $k-th$ announcements, i.e.,

$$m_{i,k} = a \text{ for at least two agents.}$$

If there exists no such alternative, the central planner will choose the status quo $\bar{a}$. Note that $x^*(m)$ does not depend on agents' first announcements $(m_{1,1}, m_{2,1}, m_{3,1})$.

For every $m \in M$, let

$$t_1^*(m) = 0.$$

Hence, agent 1 is never fined. For every $i \in \{2,3\}$ and every $m \in M$, let

$$t_i^*(m) = -\varepsilon \text{ if there exist } k \in \{2,...,K\} \text{ such that } m_{i,k} \neq m_{1,1}, \text{ and}$$
$$m_{2,h} = m_{3,h} = m_{1,1} \text{ for all } h \in \{1,...,k-1\}.$$

and

$$t_i^*(m) = 0 \text{ if there exists no such } k.$$

Hence, each agent $i \in \{2,3\}$ is fined if and only if she is the first agent between agents 2 and 3 whose announcement is different from agent 1's first announcement.[6]

3.4. Possibility Theorem

We will show that the mechanism G^* implements *any* alternative in iterative dominance, as long as agents regard it as being socially desirable.

Theorem 1: *The socially desirable alternative $a^* \in A$ is implemented in iterative dominance in (G^*, u).*

Proof: Whenever agents announce the honest message profile m^*, then the central planner will chooses the socially desirable alternative a^*, and no agents will be fined, i.e.,

$$x^*(m^*) = a^*,$$

and

$$t_i^*(m^*) = 0 \text{ for all } i \in N.$$

Hence, all we have to do in this proof is to show that the honest message profile m^* is the unique iteratively undominated message profile.

Note that each agent $i \in \{1,2,3\}$ has incentive to announce

$$m_{i,1} = a^*,$$

[6] We must note that agents 2 and 3's first announcements $(m_{2,1}, m_{3,1})$ are redundant, and therefore, we can simply delete them with a minor change of our mechanism design.

because both $x^*(m)$ and $t_i^*(m)$ are independent of $m_{i,1}$, and because of moral preference.

Fix $h \in \{2,...,K\}$ and $m \in M$ arbitrarily, where we assume that all agents honestly announce from their first announcements to their $(h-1)-th$ announcements, i.e.,

$$m_{i,h'} = a^* \text{ for all } i \in N \text{ and all } h' \in \{1,...,h-1\}.$$

First, consider any agent $i \in \{2,3\}$. Suppose

$$m_{i,h} \neq a^*.$$

Let $m_i' \in M_i$ be the message for agent i defined by

$$m_{i,h}' = a^*,$$

and

$$m_{i,h'}' = m_{i,h'} \text{ for all } h' \in \{1,...,K\}/\{h\}.$$

Hence, m_i' is the same as m_i except $h-th$ announcement, and the $h-th$ announcement of m_i' is honest. If

$$m_{j,h} = a^* \text{ for all } j \in N/\{i\},$$

then $x^*(m)$ is independent of $m_{i,h}$, and $t_i^*(m_i', m_{-i}) - t_i^*(m) \geq 0$, which implies that agent i has incentive to announce m_i' instead of m_i because of her moral preferecne. If

$$m_{j,h} \neq a^* \text{ for some } j \neq i,$$

then it follows

$$t_i^*(m_i', m_{-i}) - t_i^*(m) = \varepsilon,$$

which, together with inequalities (2) and (3), implies that agent i has incentive to announce m_i' instead of m_i, because

$$u_i(x^*(m), t_i^*(m), m_i) - u_i(x^*(m_i', m_{-i}), t_i^*(m_i', m_{-i}), m_i')$$
$$\leq -\varepsilon + \frac{d}{K} < 0.$$

Next, Consider agent 1. Suppose

$$m_{i,h} = a^* \text{ for each } i \in \{2,3\},$$

and

$$m_{1,h} \neq a^*.$$

Let $m_1' \in M_1$ be the message for agent 1 defined by

$$m_{1,h}' = a^*,$$

and

$$m_{1,h'}' = m_{1,h'} \text{ for all } h' \in \{1,...,K\}/\{h\}.$$

Note that $x^*(m)$ is independent of $m_{1,h}$ and $t_1^*(m_1', m_{-1}) = t_1^*(m) = 0$, which implies that agent 1 has incentive to announce m_1' instead of m_1, because of moral preference.

The above arguments imply that the honest message profile m^* is the unique iteratively undominated message profile in (G^*,u). Hence, we have proved that the socially desirable alternative $a^* \in A$ is implemented in iterative dominance in (G^*,u).

Q.E.D.

3.5. Discussions

3.5.1. Implementation of Social Choice Functions

The introduction of non-consequential moral preferences to the model plays the following striking role in implementing social choice functions. Fix a set of states Ω arbitrarily, and we define a social choice function as a mapping from states to alternatives, i.e.,

$$f : \Omega \to A .$$

We assume that each agent's preference for consequences is *contingent on the state*, and therefore, we denote

$$v_i = v_i(\omega),\ u_i = u_i(\omega), \text{ and } u = u(\omega) .$$

Given an arbitrary set of message profile M, a social choice function f is said to be *implementable in Nash equilibrium (iterative dominance)* if there exists a mechanism G such that at any state $\omega \in \Omega$ the value of the social choice function $f(\omega) \in A$ is implemented in Nash equilibrium (iterative dominance) in the game $(G,u(\omega))$, i.e., there exists a Nash equilibrium (an iteratively undominated message profile) in $(G,u(\omega))$, and any Nash equilibrium (any iteratively undominated message profile) in $(G,u(\omega))$ induces $f(\omega)$ with no fines realized. From Theorem 1, it follows that *irrespective of how the set of state to be specified, any social choice function f is implementable in iterative dominance whenever at any state $\omega \in \Omega$ agents regard $f(\omega)$ as being socially desirable.*

It is clear from the above definition of implementability that if distinct states ω and $\omega' \neq \omega$ correspond to distinct values $f(\omega)$ and $f(\omega') \neq f(\omega)$, then these states must correspond to distinct preference profiles, i.e., there exist no real numbers ρ and λ such that

$$\rho > 0 \text{ and } u(\omega) = \rho u(\omega') + \lambda .$$

Note that if there exist such ρ and λ, then the set of Nash equilibria in $(G,u(\omega))$ equals the set of Nash equilibria in $(G,u(\omega'))$. Hence, implementability in Nash equilibrium implies $f(\omega') = f(\omega)$, which is a contradiction.

Since $c_i > 0$ for all $i \in N$, i.e., agents have non-consequential moral preferences, it follows that whenever distinct states correspond to distinct values, i.e., distinct socially desirable alternatives, then these states always correspond to distinct non-consequential moral preference for each agent. This point is in contrast with the standard model of implementation. The standard model assumes that $c_i = 0$ for all $i \in N$, i.e., agents have preferences only for consequences, which implies that if distinct states correspond to distinct values, then these states must correspond to distinct preference profiles for consequences. This will prevent a very wide variety of important social choice functions from being implementable in Nash equilibrium. For instance, as we have mentioned in the introduction, several authors such as Rawls, Dworkin, and Sen have emphasized that non-consequential factors of the state other than agents' preferences for consequences are very substantial in foundations of social choice and welfare. Hence, ethically important social choice functions are not implementable in the standard model, whereas these are all implementable in our model.

There is an important class of social choice functions that are non-ethical but not implementable in the standard model. For instance, Serrano and Vohra (2001) investigated the economic environments where agents' preferences are the same across states but their initial endowments depend on the state. They showed that no individually rational social choice function is implementable. Theorem 1 implies that any social choice function in this class is implementable in our model.

3.5.2. Detail-Free Mechanism Design

The mechanism G^* can be regarded as being *detail-free* in the sense that it does not depend on the detail of model specifications. In fact, it is independent of the set of states Ω, state-contingent preferences for consequence $(v_i(\cdot))_{i \in N}$, and the size of agents' psychological costs $(c_i)_{i \in N}$. Most of the mechanisms that have been designed in the implementation literature were *not* detail-free at all. On the other hand, some trading mechanisms in real situations such as auction formats may be detail- free.

The mechanism G^* does not depend on even *the social choice function* f. Theorem 1 implies that such a single mechanism as G^* can implement *any* social choice function f, as long as at any state $\omega \in \Omega$ agents regard its associated value $f(\omega)$ as being socially desirable, i.e., $a^* = f(\omega)$. This point is in contrast with the standard model of implementation where agents have preferences only for consequences. It is clear from Subsection 3.5.1 that given any set of states Ω and any state-contingent preference profile for consequences $(v_i(\cdot))_{i \in N}$, there exists *no* single mechanism that can implement two distinct social choice functions in Nash equilibrium. This is the reason why in the standard model, a mechanism must be well tailored to the detail of a particular specification of social

choice function. On the other hand, when agents have non-consequential moral preferences, their state-contingent preferences inevitably depend on how the social choice function f to be specified, because at any state $\omega \in \Omega$, each agent sustains positive psychological cost if and only if she announces other alternatives than $a^* = f(\omega)$. This dependence will be the driving force of keeping a mechanism independent of the specification of social choice function.

3.5.3. Representative Systems

Theorem 1 implies that whenever agents have non-consequential moral preferences, then implementation works very well even in *representative systems* where only a few individuals in the society are allowed to participate in the mechanism. In fact, this section needs just *three* individuals to participate in the mechanism.

This point is in contrast with the standard model where all individuals in the society have preferences only for consequences. In the standard model, in order to implement any social choice function, the central planner must invite all individuals whose preferences influence its value to participate in the decision procedure. As long as there is no proper subset of individuals whose preferences for consequences are sufficient statistics for necessary information about the other individuals' preferences for consequences, it is impossible for the central planner to implement any social choice function under restrictions of representative systems. On the other hand, agents' non-consequential moral preferences can be a sufficient statistic for the socially desirable alternative, which will be the driving force for making implementation compatible with restrictions of representative systems in our model.

3.5.4. Pareto Property and Small Fines

Since the set of agents is a proper subset of all individuals in the society, it might be rather absurd to require a social choice function to satisfy *Pareto Property* or *No Veto Power* among these agents. Even if all agents prefer the same alternative the best, the other individuals who do not participate in the mechanism may dislike it.

This point may make implementation difficult to solve by using canonical mechanisms in this literature such as modulo or integer mechanisms originated by Maskin (1999) and others. Most relevant authors have focused on social choice functions that satisfy No Veto Power. In fact, their arguments have crucially relied on No Veto Power.

Based on this observation, Theorem 1 implies that the introduction of *small fines* will play an important role under restrictions of representative systems. In the proof of Theorem 1, instead of the canonical mechanisms, we used a variant of mechanism design device

originated by Abreu and Matsushima (1992a), where each agent makes multiple announcements, and only the first deviants are fined a small monetary amount.

Abreu and Matsushima showed that in the standard model with small fines, any non-Paretian social choice function is *virtually* implementable in iterative dominance on the assumption that distinct states corresponding to distinct values of this social choice function correspond to distinct preference profiles. In contrast with Abreu and Matsushima, the present paper does not need this assumption, and our possibility theorem is not 'virtual' but '*exact*'. In order to incentivize agents to announce honestly, the original Abreu and Matsushima mechanism needed to have a positive probability with which the central planner will choose 'wrong' alternatives on purpose. On the other hand, our mechanism, by using only agents' small non-consequential moral preferences, succeeded to incentivize them not to lie at all without reducing the probability of right alternative choice.

We should be careful about the advantages that virtual implication has over exact implication. It is well known in the implementation literature that in the standard model, virtual implementation works much better than exact implementation, because exact implementation needs Monotonicity condition as being necessary, whereas virtual implementation does not.[7] We must note, however, that even virtual implementation needs to be based on the assumptions that the social choice function depends only on agents' preferences for consequences, that the mechanism can depend on the detail of model specifications, and that all relevant individuals participate in the mechanism as agents. Hence, our model has the great advantages over the standard model, irrespective of whether implementation being virtual or exact in the latter model.

3.5.5. Non-Consequential Moral Preferences

In Subsection 3.2, we have specified each agent $i's$ non-consequential moral preferences by the psychological cost $\frac{\#\{k \in \{1,...,K\} : m_{i,k} \neq a^*\}}{K} c_i$. The content of Theorem 1, however, does not much depend on this specification. For instance, replace equalities (1) with equalities for all $i \in N$ such that

$$u_i(a, t_i, m_i) = v_i(a) + t_i - q_i(m_i)c_i,$$

where $q_i : M_i \to R$ satisfies that for every $m_i \in M_i$ and every $m_i' \in M_i$,

$$q_i(m_i) > q_i(m_i') \text{ if there exists } k \in \{1,...,K\} \text{ such that } m_{i,k}' = a^* \neq m_{i,k}, \text{ and}$$

$$m_{i,h}' = m_{i,h} \text{ for all } h \in \{1,...,K\}/\{k\}.$$

This implies that the more the number of agent $i's$ honest announcements is, the less her psychological cost is. We can easily check that Theorem 1 holds even if we consider more general forms of non-consequential moral preferences like this.

---

[7] See Matsushima (1988) and Abreu and Sen (1991).

3.5.6. Set of Alternatives

This section showed that even if the central planner has no knowledge about the model specifications such as the set of states, agents' preferences for consequences, and the social choice function, she can design well-behaved mechanisms such as G^*. We do not even require the central planner to know *the set of alternatives* A in advance. In fact, the mechanism G^* can be simply described by the following document written by the central planner.

> "Tell me K times on what I should do for your society. I will pick up one announcement profile from the last $K-1$ profiles. If at least two of you recommend me to do the same thing, then I will do it. Otherwise, I will do nothing. I will fine you a small monetary amount ε if and only if you are either agent 2 or agent 3 and are the first deviant from agent 1's first announcement."

Clearly, this document does not say anything about the set of alternatives. This, however, can perform the same work as the mechanism G^* defined in Subsection 3.3.

4. Incomplete Information

This section investigates the following *incomplete* information environments. Before announcing a message, each agent receives her *private signal* denoted by ω_i. Let Ω_i denote the set of private signals for agent $i \in N$. Let $\omega = (\omega_i)_{i \in N}$ denote a *state*. Let $\Omega = \prod_{i \in N} \Omega_i$ denote the set of states. Let $p : \Psi \to [0,1]$ denote a probability measure on (Ω, Ψ), according to which the state is randomly determined, where Ψ is a $\sigma-$field. Let P denote the set of probability measures.

A social choice function $f : \Omega \to A$ is defined as a mapping from states to alternatives. At any state $\omega \in \Omega$, the agents regard $f(\omega)$ as the socially desirable alternative. The central planner never knows not only which alternative to be socially desirable at the current state but also what the social choice function is.

We redefine a *utility function* for each agent $i \in N$ by replacing $u_i : A \times R \times M_i \to R$ in Section 3 with

$$u_i : A \times R \times M_i \times \Omega \to R .$$

We allow *interdependent* values in that each agent's utility can depend on the other agents' private signals. We assume the expected utility hypothesis. Let $u = (u_i)_{i \in N}$ denote a utility function profile. The utility function profile u is common knowledge among the agents, whereas the central planner does not know it.

4.1. Implementation in Iterative Dominance

A combination (G, Ω, Ψ, p, u) defines a *Bayesian game*. A *strategy for each agent* $i \in N$ is defined as a function

$$s_i : \Omega_i \to M_i .$$

Let S_i denote the set of all strategies for agent i. We denote by $s = (s_i)_{i \in N}$ a strategy profile. Let $S \equiv \prod_{i \in N} S_i$, $s(\omega) = (s_i(\omega_i))_{i \in N}$, and $s_{-i}(\omega_{-i}) = (s_j(\omega_j))_{j \in N/\{i\}}$.

The solution concept used in this section is the *Bayesian* version of iterative dominance, which defined as follows. Let $S_i^{(0)} = S_i$ and $S^{(0)} = \prod_{i \in N} S_i^{(0)}$. Recursively, for every $r = 1,2,...$, let $S_i^{(r)}$ denote the set of strategies $s_i \in S_i^{(r-1)}$ for each agent i that are *undominated with respect to* $S_{-i}^{(r-1)} = \prod_{j \in N/\{i\}} S_j^{(r-1)}$ in that there exist no $m_i \in M_i$ and no $\omega_i \in \Omega_i$ such that for every $s_{-i} \in S_{-i}^{r-1}$,

$$E[u_i(x(s(\omega)),t_i(s(\omega)),s_i(\omega),\omega)\,|\,\omega_i]$$
$$< E[u_i(x(m_i,s_{-i}(\omega_{-i})),t_i(m_i,s_{-i}(\omega_{-i})),m_i,\omega)\,|\,\omega_i],$$

where $E[\cdot\,|\,\omega_i]$ implies the expected value conditional on agent $i's$ private signal ω_i. Let $S^{(r)}=\prod_{i\in N}S_i^{(r)}$ and $S^{(\infty)}=\bigcap_{r=0}^{\infty}S^{(r)}$. A strategy profile $s\in S$ is said to be *iteratively undominated in the Bayesian game* (G,Ω,Ψ,p,u) if

$$s\in S^{(\infty)}.$$

A social choice function $f\in F$ is said to *be implemented in iterative dominance in the Bayesian game* (G,Ω,Ψ,p,u) if there exists the unique iteratively undominated strategy profile s, and this profile satisfies that for every $\omega\in\Omega$,

$$x(s(\omega))=f(\omega),$$

and

$$x_i(s(\omega))=0 \text{ for all } i\in N.$$

We must note that whenever f is implemented in iterative dominance in (G,Ω,Ψ,p,u), then it is implemented also in *mixed strategy Bayesian Nash equilibrium* in (G,Ω,Ψ,p,u).

4.2. Assumptions

This section assumes

$$A=\prod_{i\in N}A_i.$$

We denote $a=(a_i)_{i\in N}\in A$, where a_i implies the $i-th$ *characteristic* of the alternative a. We assume that each agent i, by observing her private signal ω_i, can know the $i-th$ characteristic of the socially desirable alternative $f(\omega)$. These assumptions imply that the social choice function f is *decomposable* in the sense that there exists $(f_i)_{i\in N}$ such that

$$f_i:\Omega_i\to A_i \text{ for all } i\in N,$$

and

$$f(\omega)=(f_i(\omega_i))_{i\in N} \text{ for all } \omega\in\Omega,$$

where $f_i(\omega_i)$ is regarded as the $i-th$ characteristic of the socially desirable alternative $f(\omega)$ at state $\omega\in\Omega$, which depends only on agent $i's$ private signal $\omega_i\in\Omega_i$.

We assume that there exists a positive integer $K>0$ such that

$$M_i=A_i^K \text{ for all } i\in N,$$

where K is sufficiently large. Each agent $i\in N$ makes K announcements at once about what the $i-th$ characteristic of the socially desirable alternative is. Let $M_i=M_{i,1}\times\cdots\times M_{i,K}$ where $M_{i,k}=A_i$. We denote $s_i=(s_{i,k})_{k=1}^K$ where

$$s_{i,k} : \Omega_i \to M_{i,k} \text{ for all } k \in \{1,...,K\}.$$

Let $\hat{s}_i \in S_i$ denote the *honest* strategy for agent i such that

$$\hat{s}_{i,k}(\omega_i) = f_i(\omega_i) \text{ for all } k \in \{1,...,K\} \text{ and all } \omega_i \in \Omega_i.$$

Let $\hat{s} = (\hat{s}_i)_{i \in N} \in S$ denote the honest strategy profile.

Fix a positive real number $d > 0$ arbitrarily, which is sufficiently large. We assume that each agent $i's$ utility function u_i satisfies that there exist a function $v_i : A \times \Omega \to R$ and a positive real number $c_i > 0$ such that

$$u_i(a, t_i, m_i, \omega) = v_i(a, \omega) + t_i - \frac{\#\{k \in \{1,...,K\} : m_{i,k} \neq f_i(\omega_i)\}}{K} c_i,$$

and

$$\max_{(a,a',\omega,i) \in A^2 \times \Omega \times N} \left| v_i(a, \omega) - v_i(a', \omega) \right| \leq d.$$

We assume that the upper bound of monetary fines $\varepsilon > 0$ is smaller than c_i, i.e.,

(4) $\varepsilon < c_i$ for all $i \in N$.

Based on inequalities (4), we can assume that the number of announcements K is so large that there exists a positive integer $\hat{K} \in \{1,..., K-1\}$ such that

(5) $\frac{\hat{K}}{K} c_i > \varepsilon$ for all $i \in N$,

and

(6) $(K - \hat{K})\varepsilon > d$.

These inequalities imply that the mechanism depends on, not only the upper bound d of utility differences for consequences, but also agents' psychological costs $(c_i)_{i \in N}$. We, however, will design a mechanism that does not depend on any more details of model specifications.

4.3. Mechanism Design

We design a mechanism $\hat{G} = (\hat{x}, \hat{t})$ as follows. For every $m \in M$ and every $a \in A$,

$$\hat{x}(m)(a) = \frac{\#\{k \in \{\hat{K}+1,...,K\} : (m_{i,k})_{i \in N} = a\}}{K - \hat{K}}.$$

Hence, for every $k \in \{\hat{K}+1,...,K\}$, with probability $\frac{1}{K - \hat{K}}$, the central planner will pick up the profile of agents' $k-th$ announcements $(m_{i,k})_{i \in N} \in A$, and will choose it. Note that $\hat{x}(m)$ does not depend on agents' first $\hat{K}$ announcements $(m^1,..., m^{\hat{K}})$.

For every $i \in N$ and every $m \in M$,

$\hat{t}_i(m) = -\varepsilon$ if there exist $k \in \{2,...,K\}$ such that $m_{i,k} \neq m_{i,1}$, and
$(m_{j,k})_{j \in N} = (m_{j,1})_{j \in N}$ for all $h \in \{1,...,k-1\}$.

and

$\hat{t}_i(m) = 0$ if there exists no such k.

Hence, each agent $i \in N$ is fined if and only if she is the first agent whose announcement is inconsistent with her first announcement.

We must note that the mechanism $\hat{G}$ is *different* from the mechanism G^* in a substantial way. In G^* any agent other than agent 1 is fined if she firstly deviates from agent 1's first announcement, whereas in $\hat{G}$ any agent is fined if she firstly deviates from *her own* first announcement. This difference is inevitable, because agents have different partial information about the true socially desirable alternative. This will cause another difference between G^* and $\hat{G}$ such that in G^* only the first announcement profile is irrelevant to the alternative choice, whereas in $\hat{G}$ we need the first $\hat{K}$ multiple announcement profiles to be irrelevant.

4.4. Possibility Theorem

We will show that the mechanism $\hat{G}$ implements in iterative dominance any social choice function that satisfies a variant of *incentive compatibility* given by inequalities (7) below, as long as the agents regard its value at any state as being socially desirable.

Theorem 2: *A social choice function f is implemented in iterative dominance in $(\hat{G}, \Omega, \Psi, p, u)$ if for every $i \in N$, every $\omega_i \in \Omega_i$, and every $a_i \in A_i / \{f_i(\omega_i)\}$,*

$$E[v_i(f(\omega),\omega) \mid p,\omega_i] > E[v_i((a_i, f_{-i}(\omega_{-i})),\omega) \mid p,\omega_i] - \frac{(K-\hat{K})}{K} c_i . \tag{7}$$

Proof: Whenever agents announce according to the honest strategy profile $\hat{s}$, then at any state $\omega \in \Omega$ the central planner will choose the social desirable alternative $f(\omega)$, and no agents will be fined, i.e., for every $\omega \in \Omega$,

$\hat{x}(\hat{s}(\omega))(f(\omega)) = 1$,

and

$\hat{t}_i(\hat{s}(\omega)) = 0$ for all $i \in N$.

Hence, all we have to do in this proof is to show that $\hat{s}$ is the unique iteratively undominated strategy profile.

Fix $s \in S$ and $i \in N$ arbitrarily. Fix $\omega \in \Omega$ arbitrarily. Suppose

$s_{j,k}(\omega_j) \neq s_{j,k}(\omega_j)$ for some $j \in N / \{i\}$ and some $k \in \{2,...,\hat{K}\}$.

Then, agent i is never fined whenever she announces

$$m_{i,k} = f_i(\omega_i) \text{ for all } k \in \{1,...,\hat{K}\}.$$

Next, suppose

$$s_{j,k}(\omega_j) = s_{j,k-1}(\omega_j) \text{ for all } k \in \{2,...,\hat{K}\} \text{ and all } j \in N.$$

If

$$s_{i,k}(\omega_i) \neq f_i(\omega_i) \text{ for all } k \in \{1,...,\hat{K}\},$$

then by announcing $m_{i,k} = f_i(\omega_i)$ for all $k \in \{1,...,\hat{K}\}$ instead, agent i can save the amount $\frac{\hat{K}}{K} c_i$ of psychological cost. This amount is greater than the monetary fine ε because of inequality (5). If

$$s_{i,k}(\omega_i) \neq s_{i,k-1}(\omega_i) \text{ for some } k \in \{2,...,\hat{K}\},$$

then the central planner will fine agent i. Since she has non-consequential moral preference and her first $\hat{K}$ announcements do not influence the central planner's alternative choice, it follows from the above arguments that agent i is willing to replace the first $\hat{K}$ announcements $(s_{i,k}(\omega_i))_{k=1}^{\hat{K}}$ with $(\hat{s}_{i,k}(\omega_i))_{k=1}^{\hat{K}}$. Hence, we have proved that for every $i \in N$, if s_i is iteratively undominated, then it must hold that

$$s_{i,k} = \hat{s}_{i,k} \text{ for all } k \in \{1,...,\hat{K}\}.$$

Fix $\bar{k} \in \{\hat{K}+1,...,K\}$ arbitrarily. Suppose that

$$s_{j,k} = \hat{s}_{j,k} \text{ for all } j \in N \text{ and all } k \in \{1,...,\bar{k}-1\}.$$

Fix $\omega_i \in \Omega_i$ arbitrarily, and suppose

$$s_{i,\bar{k}}(\omega_i) \neq f_i(\omega_i).$$

Let $m_i \in M_i$ denote the message for agent i defined by

$$m_i^k = \hat{s}_{i,k}(\omega_i) \text{ for all } k \in \{1,...,\bar{k}\},$$

and

$$m_i^k = s_{i,k}(\omega_i) \text{ for all } k \in \{\bar{k}+1,...,K\}.$$

Suppose

$$s_{j,\bar{k}}(\omega_j) \neq f_i(\omega_j) \text{ for some } j \in N/\{i\}.$$

Then,

$$\hat{t}_i(s(\omega)) = -\varepsilon,$$

and

$$\hat{t}_i(m_i, s_{-i}(\omega_{-i})) = 0.$$

Inequality (6) implies that the expected value of agent $i's$ utility differences for consequences between the messages $s_i(\omega_i)$ and m_i is less than ε. Hence, agent i prefers announcing m_i instead of $s_i(\omega_i)$.

Next, suppose

$$s_{j,\bar{k}}(\omega_j) \neq f_i(\omega_j) \text{ for all } j \in N/\{i\}.$$

Then,

$$\hat{t}_i(s(\omega)) = -\varepsilon,$$

and

$$\hat{t}_i(m_i, s_{-i}(\omega_{-i})) \geq -\varepsilon.$$

Inequality (7), together with non-consequential moral preference, implies that agent i has strict incentive to make the honest announcement whenever the other agents make the honest announcements. Hence, agent i prefers to announce m_i instead of $s_i(\omega_i)$.

From the above arguments, we have proved that if s is an iteratively undominated strategy profile, then $s = \hat{s}$ must hold. Since the set of iteratively undominated strategy profiles S^{∞} is nonempty, we have completed the proof of Theorem 2.

Q.E.D.

We can say that the mechanism $\hat{G}$ is *detail-free*, because it is independent of the set of states Ω, the probability function p, agents' preferences for consequences $(v_i)_{i\in N}$, and even the social choice function f. In contrast with the mechanism G^*, however, the construction of the mechanism $\hat{G}$ is not independent of agents' psychological costs $(c_i)_{i\in N}$, because of inequalities (5).

Moreover, in contrast with the mechanism G^*, the construction of the mechanism $\hat{G}$ is not independent of the set of alternatives. Note that the sufficient condition (7) depend on how the set of possible $i-th$ characteristics A_i to be specified for each $i \in N$. Suppose that the central planner knows the range of the social choice function f, i.e.,

$$f(\Omega) = \{a \in A : a = f(\omega) \text{ for some } \omega \in \Omega\}.$$

Then, she can modify the construction of the mechanism $\hat{G}$ by replacing A with $f(\Omega)$. In this case, the sufficient condition (7) will be replaced with the following *weaker* condition. That is, for every $i \in N$, every $\omega_i \in \Omega_i$, and every $\omega_i' \in \Omega_i/\{\omega_i\}$,

$$E[v_i(f(\omega),\omega) \mid p, \omega_i] > E[v_i(f(\omega_i', \omega_{-i}),\omega) \mid p, \omega_i] - \frac{(K-\hat{K})}{K} c_i. \tag{8}$$

Note that if agents' psychological costs are very small, then inequalities (8) will be approximated by the *standard* condition of incentive compatibility such that for every $i \in N$, every $\omega_i \in \Omega_i$, and every $\omega_i' \in \Omega_i/\{\omega_i\}$,

$$E[v_i(f(\omega),\omega) \mid p, \omega_i] \geq E[v_i(f(\omega_i', \omega_{-i}),\omega) \mid p, \omega_i]. \tag{9}$$

4.5. Generalization

This section has focused on social choice functions that are decomposable in the sense of Subsection 4.2. This subsection will drop all assumptions in Subsection 4.2. We, instead, will assume that the central planner knows the set of states Ω and the social choice function f in advance, and consider the very general class of social choice functions that are not necessarily decomposable.

We assume that there exists a positive integer $K > 0$ such that

$$M_i = \Omega_i^K \text{ for all } i \in N,$$

where K is sufficiently large. Each agent $i \in N$ makes K announcements at once about which private signal she receives among Ω_i. We redefine the honest strategy $\tilde{s}_i \in S_i$ for agent i by

$$\tilde{s}_{i,k}(\omega_i) = \omega_i \text{ for all } k \in \{1,...,K\} \text{ and all } \omega_i \in \Omega_i.$$

Fix a positive real number $d > 0$ arbitrarily, which is sufficiently large. We assume that each agent $i's$ utility function u_i satisfies that there exist a function $v_i : A \times \Omega \to R$ and a positive real number $c_i > 0$ such that

$$u_i(a, t_i, m_i, \omega) = v_i(a, \omega) + t_i - \frac{\#\{k \in \{1,...,K\} : m_{i,k} \neq \omega_i\}}{K} c_i,$$

and

$$\max_{(a,a',\omega,i) \in A^2 \times \Omega \times N} \left| v_i(a, \omega) - v_i(a', \omega) \right| \leq d.$$

Hence, each agent has positive psychological cost for not honestly announcing her true private signal. We assume that inequalities (4), (5), and (6) hold for some $\hat{K} \in \{1,..., K-1\}$.

We design a mechanism $\tilde{G} = (\tilde{x}, \tilde{t})$ in the same way as $\hat{G}$. For every $m \in M$ and every $a \in A$,

$$\hat{x}(m)(a) = \frac{\#\{k \in \{\hat{K}+1,...,K\} : (m_{i,k})_{i \in N} = a\}}{K - \hat{K}}.$$

For every $i \in N$ and every $m \in M$,

$$\hat{t}_i(m) = -\varepsilon \text{ if there exist } k \in \{2,...,K\} \text{ such that } m_{i,k} \neq m_{i,1}, \text{ and}$$
$$(m_{j,k})_{j \in N} = (m_{j,1})_{j \in N} \text{ for all } h \in \{1,...,k-1\}.$$

and

$$\hat{t}_i(m) = 0 \text{ if there exists no such } k.$$

Note that the mechanism $\tilde{G}$ depends on the set of states Ω and the social choice function f. In spite of this, we would like to regard $\tilde{G}$ as being *detail-free*, because it is independent of the probability function p and agents' preferences for consequences $(v_i)_{i \in N}$.

In the same way as in Theorem 2, we can prove that a *social choice function f, irrespective of whether it is decomposable or not, is implemented in iterative dominance in*

$(\widetilde{G},\Omega,\Psi,p,u)$ *if inequalities (8) hold.* Note that if agents' psychological costs are very small, then inequalities (8) will be approximated by the standard condition of incentive compatibility, i.e., inequalities (9). In the standard model of implementation, incentive compatibility given by inequalities (9) is not sufficient for implementation in Bayesian Nash equilibrium, even if we use mechanisms that are not detail-free.[8] In fact, in the standard model, we need additional conditions that require the dependence of agents' preferences for consequences on the state and correlation among their private signals. See Jackson (1991), Abreu and Matsushima (1992b), Matsushima (1993), Duggan (1997), Serrano and Vohra (2000), and others. In contrast to the standard model, incentive compatibility is *sufficient* in the model of this subsection even if we use only detail-free mechanisms.

We can not see it appropriate in general to assume that any factor of the state relevant to the social choice function invariably includes information about either agents' preferences for consequences or the degree to which their private signals are correlated. Hence, we must admit that these additional conditions in the standard model are very restrictive, even if we can see them as being generic in the spaces of utility functions and probability distributions. For related arguments, see Neeman (2004), for instance.

[8] Bergemann and Morris (2003) and Matsushima and Ohashi (2004) investigated implementation in ex post equilibrium. These works used the mechanisms that do not depend on the probability function, but required ex post incentive compatibility, which is much stronger than inequalities (9).

References

Abreu, D. and H. Matsushima (1992a): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica* 60, 993-1008.

Abreu, D. and H. Matsushima (1992b): "Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information," mimeo.

Abreu, D. and A. Sen (1991): "Virtual Implementation in Nash Equilibrium," *Econometrica* 59, 997-1021.

Alger, I. and C. A. Ma (2003): "Moral Hazard, Insurance, and Some Collusion," *Journal of Economic Behavior and Organization* 50, 225-247.

Bergemann, D. and S. Morris (2004): "Robust Implemengation: The role of Large Type Spaces," mimeo, Yale University.

Basu, K., P. Pattanaik, and K. Suzumura (1995): *Choice, Welfare, and Development*, Oxford: Clarendon Press.

Deneckere, R. and S. Severinov (2001): "Mechanism Design and Communication Costs," mimeo.

Duggan, J. (1997): "Virtual Bayesian Implementation," *Econometrica* 65, 1175-1199.

Dworkin, R. (1981): "What is equality ? Part 1: Equality of Welfare, Part 2: Equality of Resources," *Philosophy and Public Affairs* 10, 185-246, 283-345.

Eliaz, K. (2002): "Fault Tolerant Implementation," *Review of Economic Studies* 69, 589-610.

Erard, B. and J. Feinstein (1994): "Honesty and Evasion in the Tax Compliance Game," *RAND Journal of Economics* 25, 1-19.

Fehr, E. and K. Schmidt (2003): "Theories of Fairness and Reciprocity: Evidence and Economic Applications," in *Advances in Economics and Econometrics: Eighth World Congress*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky.

Glazer, J. and A. Rubinstein (1998): "Motives and Implementation: On the Design of Mechanisms to Elicit Opinions," *Journal of Economic Theory* 79, 157-173.

Gneezy, U. (2004): "Deception: The Role of Consequences," mimeo.

Jackson, M. (1991): "Bayesian Implementation," *Econometrica* 59, 461-477.

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

Maskin, E. and T. Sjöström (2002): "Implementation Theory," in *Handbook of Social Choice and Welfare Volume 1*, ed. by K. Arrow, A. Sen, and K. Suzumura. Elsevier.

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.

Matsushima, H. (1993): "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory* 59, 107-121.

Matsushima, H. and Y. Ohashi (2004): "Belief-Free Implementation," under revision.

Moore, J. (1992): "Implementation in Environments with Complete Information," in *Advances in Economic Theory: Sixth World Congress*, ed. by J.-J. Laffont. Cambridge University Press.

Neeman, Z. (2004): "The Relevance of Private Information in Mechanism Design," *Journal of Economic Theory*, forthcoming.

Osborne, M. and A. Rubinstein (1994): *A Course in Game Theory*, MIT Press.

Palfrey, T. (1992): "Implementation in Bayesian Equilibrium: the Multiple Equilibrium Problem in Mechanism Design," in *Advances in Economic Theory: Sixth World Congress*, ed. by J.-J. Laffont, Cambridge University Press.

Rawls, J. (1971): *A Theory of Justice*, Cambridge: Harvard: Harvard University Press.

Sen, A. (1982): *Choice, Welfare and Measurement*, Oxford: Blackwell.

Sen, A. (1985): *Commodities and Capabilities*, Amsterdam: North-Holland.

Sen, A. (1999): "The Possibility of Social Choice," *American Economic Review* 89, 349-378.

Serrano, R. and R. Vohra (2000): "Type Diversity and Virtual Bayesian Implementation," Working Paper No. 00-16, Department of Economics, Brown University.

Serrano, R. and R. Vohra (2001): "Some Limitations of Virtual Bayesian Implementation," *Econometrica* 69, 785-792.

Suzumura, K. (2002): "Introduction," in *Handbook of Social choice and Welfare, Volume 1*, ed. by K. Arrow, A. Sen, and K. Suzumura, Elsevier Science.

Tirole, J. (1999): "Incomplete Contracts: Where do we stand?," *Econometrica* 67, 741-781.