Available online at www.sciencedirect.com
ScienceDirect
Games and Economic Behavior 64 (2008) 487–513
GAMES and Economic Behavior
www.elsevier.com/locate/geb

# Mechanism design with partial state verifiability

Raymond Deneckere [a,*], Sergei Severinov [b]

[a] Department of Economics, University of Wisconsin, Madison, WI, USA
[b] Department of Economics, University of Essex, UK

Received 20 April 2005
Available online 12 February 2008

**Abstract**

We study implementation in environments where agents have limited ability to imitate others. Agents are randomly and privately endowed with type-dependent sets of messages. So sending a message becomes a partial proof regarding type. For environments where agents can send any combination of available messages, we develop an Extended Revelation Principle and characterize the incentive constraints which implementable allocations must satisfy. When not all message combinations are feasible, static mechanisms no longer suffice. If a 'punishment' allocation exists for each agent, then implementable allocations can be characterized as equilibria of a "Revelation Game," in which agents first select from the menus of allocation rules, then the mediator requests each agent to send some verifying messages. When a punishment allocation fails to exist for some agent, dynamic games in which agents gradually reveal their evidence implement a larger set of outcomes. The latter result provides a foundation for a theory of debate.

*JEL classification:* C7; D82

## 1. Introduction

The literature on implementation in incomplete information environments studies how a principal can effectively elicit private information regarding preferences from agents. This paper considers a different dimension on which agents may have private information, and along which the principal may be able to screen, namely the ability/inability of different agents to substantiate their own and other agent's claims regarding the state of the world.

[*] Corresponding author at: University of Wisconsin-Madison, Department of Economics, 1180 Observatory Drive, Madison, WI, USA.
*E-mail addresses:* rjdeneck@wisc.edu (R. Deneckere), sseverinov@gmail.com (S. Severinov).

0899-8256/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.geb.2007.12.006

The premise that agents can costlessly and effortlessly manipulate information is prevalent in economics. Yet in practice, the ability of agents to do so is often limited. There are several reasons for this. First, for psychological or ethical reasons it may be costly for some individuals to misrepresent the truth. Lying may cause stress or discomfort ("blushing," "feeling wrong"), producing a disutility. The physical symptoms associated with emotional discomfort that people experience when lying have been extensively studied by behavioral psychologists (see, e.g., Ekman, 1973). Experimental evidence confirms that a nonnegligible part of the population chooses not to lie regarding private information, even when this would increase their monetary payoffs.[1] As for economic consequences of such behavior, Erard and Feinstein (1994) argue that "some taxpayers appear to be inherently honest, willing to bear their full tax burden even when faced with financial incentives to underreport their income. ...(The presence of) such inherently honest taxpayers is supported by econometric evidence and survey findings..." Deneckere and De Palma (1995) and Alger and Ma (2003) emphasize that agents with stronger ethical views are less prone to colluding and misrepresenting their condition.[2]

Second, agents may be asked to support their claims with some form of evidence. Failure of an individual to produce evidence known to be available in a certain state of the world then provides proof that this state of the world has not occurred. Important applications include analysis of provability (Lipman and Seppi, 1995) and court proceedings (Bull and Watson, 2004a and Squintani, 2004).[3]

Finally, misrepresenting the truth may require costly physical actions. For example, a sharecropper who misrepresents the crop may have to hide part of it or borrow some from a third party.

Building on this motivation, this paper studies mechanism design in environments where participants have limited ability to misrepresent their information. In our model, each agent is randomly and privately endowed with a preference parameter affecting everyone's utility, and with a set of "verifying messages." Because particular verifying messages are available to an agent only in some states of the world, they possess direct informational content: they verify that the state of the world must be such that they are available to that agent.

We develop a general method for implementation and characterize the set of implementable social choice rules in such environments. In particular, we derive the incentive constraints that need to be satisfied by implementable allocations. This set of incentive constraints is smaller than in the standard environment because of the availability of verifying messages. Describing the incentive constraints precisely constitutes one of the central contributions of our paper, as it transforms the task of characterizing the set of implementable social choice rules from that of constructing appropriate mechanisms or disproving their existence to the much simpler task of deriving the set of solutions to a linear program.

[1] For example, Gneezy (2002) reports experiments with deception games in which responders were known to largely follow the sender's recommendation. Yet the proportion of informed senders who chose *not* to mislead opponents even though misleading was in the senders' best interests varied from 48% to 83% across experiments. Survey evidence paints a similar picture, with a core group of people having no qualms at all about inflating insurance claims, but an even greater fraction considering it unacceptable to do so (Tennyson, 1997).

[2] Other papers arguing that ethical considerations may lead some agent types to behave honestly are Alger and Renault (2006) and Kartik et al. (2007). Severinov and Deneckere (2006) argue that naive behavior—which has the same behavioral consequences as honesty—may be due to bounded rationality. Chen (2000) maintains that individuals have a tendency to keep promises, and shows that this may cause optimal contracts to be incomplete.

[3] Che and Gale (2000) study trading mechanisms with budget constrained buyers. Such buyers can credibly disclose information about their budget by posting a bond.

We first study environments in which there are no restrictions on the number or the combinations of messages that agents can send from their set of verifying messages. We derive an "Extended Revelation Principle" which shows that all implementable social choice functions can be implemented via a mechanism in which each agent sends all her verifying messages, and makes a choice from a menu of allocations that is contingent on the set of verifying messages she sent. Since choosing from a menu is equivalent to making a cheap-talk claim regarding her preference parameter, in equilibrium each agent truthfully reveals her preference parameter and provides all evidence she has available.

Using the Extended Revelation Principle we characterize the set of incentive constraints that have to be imposed upon implementable allocations (Corollary 1). This set of incentive constraints reflects the intuition that partial verifiability alleviates the incentive problem and elicits agents' information more easily. Nevertheless, we demonstrate that the set of implementable allocations is not always monotone in the degree of verifiability.

We then study environments in which agents are unable to send all possible subsets of their set of verifying messages. An interesting class of such environments occurs when the mechanism specifies exogenous limits on the number of messages an agent is allowed to send, as in Glazer and Rubinstein (2001, 2004). Implementation in such environments is complicated by the fact that generally agents are no longer able to fully demonstrate their communication abilities, necessitating the use of dynamic mechanisms and randomization in the set of messages sent.

We provide a necessary and sufficient condition—the existence of an agent-specific allocation which gives the lowest utility to all types of that agent—under which the set of implementable outcomes can be characterized in terms of truthful and obedient equilibria of a particular 'revelation game.' In this game, agents first send cheap talk messages (or choose from a menu). As a function of these reports, the principal randomly requests agents to send a combination of verifying messages. This result extends Theorem 6 of Bull and Watson (2004b) which considers similar restrictions on feasible sets of agents' verifying messages (evidence) but limits consideration to games and mechanisms in which each agent can send messages at most once along any path of the game. In contrast, we allow agents to disclose evidence more than once along the path of the game. This highlights the important role of player specific punishments (which are unnecessary in the setting of Bull and Watson, 2004b).

Using our characterization result for this environment, we then identify the set of incentive constraints that any implementable social choice function must satisfy.

Finally, we analyze the case in which agent-specific worst allocations fail to exist. In this case, dynamic mechanisms in which agents take turns to send their verifying messages, and some agents send messages more than once, implement a larger set of social choice functions than mechanisms in which each agent sends verifying messages only once. Thus, our paper provides a foundation for a theory of debates between asymmetrically informed parties.

The prior work most closely related to the present paper is Green and Laffont (1986) and Lipman and Seppi (1995). Green and Laffont (1986) study a model in which agents are limited in the direct claims on type they can make, and show that it may be optimal to induce agents to lie when they are allowed to submit only one such claim. Our paper demonstrates that, when the latter restriction is imposed, it is generally impossible to translate an arbitrary situation with partial verifiability into a direct revelation model with a single type announcement (see Proposition 1). Consequently, there is a need to develop a valid method of implementation in environments with partial verifiability, providing the impetus for the rest of our analysis.

Lipman and Seppi (1995) analyze a model in which a group of symmetrically informed agents with conflicting interests sequentially submit claims, and derive conditions under which the principal is able to elicit the truth. In contrast, our agents are asymmetrically informed, and we study conditions under which sequential revelation of evidence is (or is not) necessary.

Our paper is also closely related to recent work by Bull and Watson (2004b) and Forges and Koessler (2005). These papers differ from ours in both setting and motivation, and their analysis is complementary to ours, but there nevertheless is significant overlap in results. In the Conclusions we discuss the relation between these papers and ours in more detail.

In our model verifying evidence is either available to the agent or not, and she cannot generate it endogenously. In contrast, the literature on "costly state falsification," including Lacker and Weinberg (1989), Maggi and Rodriguez-Clare (1995), Crocker and Morgan (1998), and Sanchirico and Triantis (2004), studies the design of mechanisms in situations where the agent incurs a cost increasing in the magnitude of misrepresentation of the true state of the world. Kartik (2004) explores a signaling model in which untruthful signals are costly in a similar way. He shows that the outcome converges to the most informative cheap talk equilibrium as the cost of misrepresentation converges to zero. In a companion paper (Deneckere and Severinov, 2003), we analyze a screening model in which the agent can generate multiple pieces of evidence or has to pass several tests, incurring a cost which increases in the magnitude of falsification on each test (piece of evidence).

The rest of the paper is organized as follows. In Section 2 we present the model where agents can send all combinations of feasible messages. Section 3 develops our "Extended Revelation Principle," and studies the extent to which the number of messages sent in the mechanism can be reduced without limiting the scope of implementation. Section 4 analyzes situations in which agents are unable to send some combinations of verifying messages. Section 5 concludes. All proofs are relegated to Appendix A.

## 2. The model

We consider an asymmetric information environment with one principal and $L \geqslant 1$ agents. The set of public decisions (allocations) in this environment is $X$, with typical element $x$. Each agent $i \in \{1, \dots, L\}$ privately observes the outcome of a random variable $\theta_i \in \Theta_i$ that affects the utilities of all agents. Letting $\theta = (\theta_1, \dots, \theta_L)$ and $\Theta \equiv \prod_{i=1}^{L} \Theta_i$, the utility of agent $i$ when allocation $x$ is implemented is then given by $u_i(x, \theta)$.

Let $\mathcal{C}$ denote the space of "verifying messages," i.e. *the set of messages that can be submitted by some type of some agent in some states of the world, but that are not available to every type in every state of the world.* Every element of $\mathcal{C}$ is a minimal message unit, denoted by $m$. All restrictions on agents' communication abilities are embodied in their *feasible message sets* $\mathcal{M}_i \subset \mathcal{C}$. An agent with feasible message set $\mathcal{M}_i$ can send a collection of verifying messages $\{m_1, \dots, m_n\}$ if and only if $m_j \in \mathcal{M}_i$ for each $j = 1, \dots, n$.[4]

Each agent's set of verifying messages is her private information. So, a full description of agent $i$'s type includes both her preference parameter $\theta_i$ and her feasible message set $\mathcal{M}_i$. Accordingly, we define an agent's type as $t_i = (\theta_i, \mathcal{M}_i)$. Our approach that the message set $\mathcal{M}_i$ forms an integral part of agent $i$'s type description is a significant departure from the existing literature which maintains that an agent's preference parameter completely describes her type. We denote the set of all possible types of agent $i$ by $T_i \subset \Theta_i \times 2^{\#\mathcal{C}}$ (where $2^{\#\mathcal{C}}$ is the power set of $\mathcal{C}$). Let $t = (t_1, \dots, t_L)$ denote the profile (vector) of agents' types (we will also refer to $t$ as

[4] In Section 4 we consider restrictions on the *collections* of verifying messages that agents can send.

the state of the world), and let $T \subset T_1 \times \cdots \times T_L$ be the type space. Note that the inclusion could be strict because some type profiles may be infeasible.

The prior probability $F(\cdot)$ over $T$ is common knowledge among all parties, with $F(t) > 0$ for all $t \in T$. The marginal probability distribution over $i$'s types is denoted by $F_i(t_i)$. Our environment is thus completely described by an $(L+4)$-tuple $\{X, \mathcal{C}, T, F(\cdot), u_1(\cdot), \ldots, u_L(\cdot)\}$.

We do not assume that there is a deterministic relation between $\theta_i$ and $\mathcal{M}_i$. In particular, letting $I_i(\mathcal{M}_i) = \{\theta_i \mid (\theta_i, \mathcal{M}_i) \in T_i\}$ and $K_i(\theta_i) = \{\mathcal{M}_i \mid (\theta_i, \mathcal{M}_i) \in T_i\}$, it may or may not be that $I_i(\mathcal{M}_i)$ or $K_i(\theta_i)$ are singletons. We use the symbol $\mathcal{N}_i$ to denote the collection of all possible sets of verifying messages of agent $i$, i.e. $\mathcal{N}_i = \{\mathcal{M}_i \mid \exists \theta_i \in \Theta_i\colon\ (\theta_i, \mathcal{M}_i) \in T_i\}$. Note that $\theta_i \in I_i(\mathcal{M}_i)$ if and only if $F_i(\theta_i, \mathcal{M}_i) > 0$.

Several comments on this general model are in order. First, define agent $i$'s message to be "cheap talk" if it is available to agent $i$ in every state of the world $t \in T$. Such cheap talk messages do not contain credible information regarding the state of the world, and are explicitly excluded from the set $\mathcal{C}$. In contrast, any message $m_i \in \mathcal{M}_i$ sent by agent $i$ has direct informational content which can be exploited by the mechanism designer: it proves that agent $i$'s type cannot be such that $m_i$ is unavailable to her, and so partially verifies $i$'s type.

In practice, verifying messages exist for a host of reasons. Elements of the set $\mathcal{C}$ could be documents (contracts, receipts, or other legal records, etc.), physical items (such as evidence collected by investigators), or human acts (such as passing a test or demonstrating a skill). Alternatively, direct written or verbal communication can have verifiability if some agents are unable to lie, or can shade the truth only partially when describing the state of the world.

Our model makes an explicit distinction between any literal meaning a verifying message may have, on the one hand, and its real informational content, on the other hand. Although both agents and the principal could attach a literal meaning to some verifying messages (as is the case with documents or verbal testimony), sending those messages does not necessarily prove their literal meaning to be true. Rather, the informational content of a verifying message derives solely from the fact that it is available to the sender only in some states of the world.

We assume that the decision whether or not to send a verifying message from $\mathcal{M}_i$ is under the sole control of agent $i$. Consequently, the principal must provide agent $i$ with incentives to produce these messages. In designing these incentives, the principal must respect the constraint that type $t_i$ can only provide verifying messages belonging to $\mathcal{M}_i$. However, we do not impose further restrictions on the type of mechanisms that the principal can design. In particular, in line with the rest of the mechanism design literature, we assume that the principal can offer menus of different allocations from which agents can choose. The choice of an allocation from a menu is an action that does not interfere with an agent's ability to provide evidence in the form of verifying messages. As far as implementation is concerned, choosing from a menu plays the same role as cheap talk: both ensure that self-selection can occur. Thus, without loss of generality, cheap talk statements about type are available to all agents. Throughout the paper, we distinguish between such cheap talk statements regarding type and any evidence an agent can submit.[5]

Menus allow the mechanism designer to screen types who have different preferences parameters, $\theta_i$, but identical sets of verifying messages. Such types can provide the same proof and hence can imitate each other freely. So they can be screened further only by relying on self-selection

[5] In applying our model to cases where some agents have ethical concerns about lying, we distinguish between literal claims regarding the state of the world (which are ethically significant, and thus contained in $\mathcal{M}_i$), and implicit cheap talk claims resulting from their choices from a menu of allocations. Honest agents are free to choose any element from such a menu of allocations, yet are restricted in the literal claims they can make.

via menus. Also, in the model of Section 4, menus allow the mechanism designer to cross-check the agents' reports, and determine the collections of verifiable messages agents should send.

Green and Laffont (1986) analyze a principal-agent model in which the agent is restricted in her ability to make direct claims about her preference type. However, these authors do not explain how an underlying environment with verifying claims could give rise to such a model. Furthermore, their analysis is limited to static mechanisms in which the agent can send only one feasible claim. In contrast, we permit arbitrary dynamic mechanisms, and do not restrict agents in the number of verifying claims they can submit.

Our first proposition establishes that direct revelation mechanisms, which Green and Laffont (1986) focus on, are inadequate for implementation when agents can only send a single verifying claim. Specifically, there is then no general equivalence between models in which agent $i$'s feasible message set $\mathcal{M}_i$ is drawn from some arbitrary message space $\mathcal{C}$ and models of direct communication in which $i$'s feasible message set is drawn from $\Theta_i$.

**Proposition 1.** *Suppose that each agent is restricted to sending a single message in the mechanism. Then there exists an environment $\{X, \mathcal{C}, T, F(\cdot), u_1(\cdot), \ldots, u_L(\cdot)\}$ with feasible message sets $\mathcal{M}_i \subseteq \mathcal{C}$ for which there is no equivalent environment $\{X, \bigcup_{i=1}^{L} \Theta_i, T', F(\cdot), u_1(\cdot), \ldots, u_L(\cdot)\}$, with a bijection $b$ mapping the type space $T$ to $T'$ so that to each agent type $(\theta_i, \mathcal{M}_i) \in T_i$ there corresponds a type $t_i' = (\theta_i, M_i) \in T'$ with $M_i \subset \Theta_i$, and so that the sets of implementable social choice functions $f(\cdot): \Theta \mapsto X$ in the two environments coincide.*

In Section 4, we also show that when agents are restricted to making a single or a few verifying claims, static mechanisms generally do not suffice for implementation.

## 3. Analysis of the basic model

A social choice function (s.c.f.) is a mapping from the set of agent types into the outcome space $X$, i.e. $f(\cdot): T \mapsto X$. We assume that agents are expected utility maximizers.[6] Our goal in this section is to provide a method for characterizing the set of social choice functions implementable in a Bayesian equilibrium of some (possibly dynamic) mechanism. As a first step towards this goal, let us design a mechanism $G$ (or, more precisely, a class of mechanisms with the same game form)[7] which can be used to implement any implementable social choice function. For ease of exposition, we present $G$ as a two-stage mechanism, but argue later that $G$ can also be implemented as a static mechanism.

In the first stage of mechanism $G$ the agents are requested to simultaneously report all their feasible verifying messages to the principal. In the second stage, each agent is offered a menu of allocations to choose from, or, equivalently, is asked to send a cheap-talk message announcing her preference parameter. The purpose of the second stage is to use the self-selection method to assign different allocations to agent-types with the same set of verifying messages, but different preference parameters. Thus, the number of elements in the menu offered to agent $i$ who sent a set of verifying messages $\mathcal{M}_i$ in the first stage is equal to $\#I_i(\mathcal{M}_i)$. The outcome assigned by the

[6] Random allocations can be accommodated by letting $X$ be a space of probability distributions over a space of 'elementary' outcomes $Y$, and $u_i(x, \theta)$ the expected utility of lottery $x$.

[7] A mechanism is a collection of strategy spaces for the agents, $\{S_i\}_{i=1}^{L}$, and an outcome function $g(\cdot): \prod_{i=1}^{L} S_i \mapsto X$. Mechanisms from the class $G$ have the same strategy spaces, but differ in outcome function.

mechanism is determined by the verifying messages sent by the agents in stage 1 and the menu choices/cheap-talk reports in stage 2.

The submission of verifying messages in the first stage can be seen as a 'password' necessary to access a particular menu of allocations. For this reason we will refer to mechanisms of class $G$ as 'password' mechanisms. It is important to note that the types of agent $i$ who cannot send the set of verifying messages $\mathcal{M}_i$ will not be able to gain access to the menu contingent on $\mathcal{M}_i$ and thus will not be able to obtain the corresponding outcomes.

A strategy in which agent $i$ sends *all* her verifying messages and chooses the element from the menu labeled by her true preference parameter $\theta_i$ is called *truthfully revealing*. We then have:

**Theorem 1** *(Extended Revelation Principle). Any social choice function implementable in Bayesian equilibrium of some mechanism can be implemented in Bayesian equilibrium of mechanism $G$, with all agents using truthfully revealing strategies. Furthermore, there exist social choice functions that can be implemented only via a mechanism requiring agents to send all verifying messages.*

Theorem 1 establishes in what sense the Revelation Principle can be extended to the costly communication case. One can focus on mechanisms in which every agent truthfully reveals her type by sending all her verifying messages and then making the corresponding menu choice. Moreover, in some cases it may be *necessary* to require that agents send all their verifying messages, in order to fully exploit their limited abilities to manipulate information.

The intuition behind Theorem 1 is easy to understand. When two types of an agent have different sets of verifying messages, one of them (say, type $A$) cannot send all the verifying messages that are feasible for the other (say, type $B$). If the mechanism exploits this property, then an implementable social choice function need not satisfy the standard incentive constraint that type $A$ of this agent gets a higher payoff from the allocation designed for her than from the allocation designed for type $B$ of this agent. The larger is the set of incentive constraints that are eliminated in this way, the larger is the set of implementable social choice functions. Because the mechanisms of class $G$ require agents to send all their verifying messages, they eliminate the maximal possible number of incentive constraints. Therefore mechanism $G$ implements any social choice function implementable via any other mechanism.[8]

Mechanism $G$ can be implemented statically, by having agents simultaneously report both their verifying messages and the element of the menu they would select in stage two. Alternatively, agents could first make choices from the menus, and then present all their verifying claims establishing access to the respective menu. In practice, each of these timings seems to be used.

[8] Applying our approach to the motivating example studied by Green and Laffont (1986), we can implement a larger set of social choice functions than are implementable using the approach used by these authors. Indeed, consider an environment with a single agent, a decision set $X = \{x_1, x_2, x_3\}$, and three types $t_1 = (\theta_1, \{m_1, m_2\})$, $t_2 = (\theta_2, \{m_2, m_3\})$, $t_3 = (\theta_3, \{m_3\})$. The payoff structure is as follows:

$$u(x_1, \theta_i) < u(x_3, \theta_i) < u(x_2, \theta_i) \quad \forall i \in \{1, 2, 3\}.$$

Consider a social choice function $f(t_1) = x_1$, $f(t_2) = x_2$, $f(t_3) = x_3$. Then $f(\cdot)$ is not implementable via any mechanism in which the agent sends only one verifying message $m \in \{m_1, m_2, m_3\}$. To see this, suppose otherwise. Then all three agent types must send different verifying messages in equilibrium. Since type $t_3$ can only send message $m_3$, it follows that type $t_2$ must send message $m_2$. However, type $t_1$ would then imitate the message sent by type $m_2$. On the other hand, $f(\cdot)$ is implementable via the mechanism with the following outcome function $\tilde{G}(\cdot)$: $\tilde{G}(m_2, m_3) = x_2$, $\tilde{G}(m_3) = x_3$, and $\tilde{G}(\mathcal{S}) = x_1$, where $\mathcal{S}$ is any other report.

For example, sellers facing customers who may be "honest" or "naive" often engage in customer interviews eliciting information regarding their willingness to pay prior to presenting them with various options (Severinov and Deneckere, 2006). In contrast, the IRS requires taxpayers to include some hard evidence along with their income reports, and in court hearings the defendant first enters a plea (a cheap talk claim) and then presents evidence.

In mechanism $G$, a type $(\theta_i, \mathcal{M}_i)$ of agent $i$ can obtain the allocation designed for type $(\theta_i', \mathcal{M}_i')$ if and only if the former can fully mimic the latter, i.e. if $\mathcal{M}_i' \subset \mathcal{M}_i$. This precisely delineates which incentive constraints must be imposed on the social choice function:

**Corollary 1.** *A social choice function $f : T \to X$ is implementable in Bayesian equilibrium if and only if there exists a decision rule $\tilde{f} : \prod_{i=1}^{L} T_i \setminus T \to X$ such that the following incentive constraints hold for all i, all $t_i = (\theta_i, \mathcal{M}_i) \in T_i$, and all $t_i' = (\theta_i', \mathcal{M}_i') \in T_i$ s.t. $\mathcal{M}_i' \subseteq \mathcal{M}_i$:*

$$\sum_{t_{-i} \in T_{-i}} u_i\big(f(t_i, t_{-i}), \theta\big) F_i(t_{-i}|t_i)$$
$$\geqslant \sum_{t_{-i}:(t_i', t_{-i}) \in T} u_i\big(f\big(t_i', t_{-i}\big), \theta\big) F_i(t_{-i}|t_i) + \sum_{t_{-i}:(t_i', t_{-i}) \notin T} u_i\big(\tilde{f}\big(t_i', t_{-i}\big), \theta\big) F_i(t_{-i}|t_i). \tag{1}$$

Inequality (1) requires some explanation. In the standard case where agents have no verifying claims, we have $\mathcal{M}_i = \phi$ for all $t_i \in T_i$, and so inequality (1) must hold for all $t_i' \in T_i$. The presence of verifying messages therefore expands the set of implementable social choice functions. The role of the rule $\tilde{f}$ is to assign an allocation when the joint type report is infeasible, i.e. $t \notin T$. No such $\tilde{f}$ is needed if either $T = \prod_{i=1}^{L} T_i$ or $L = 1$. In the single-agent case, we can drop the agent subscript, and condition (1) takes on a particularly simple form:

$$u\big(f(\theta, \mathcal{M}), \theta\big) \geqslant u\big(f(\theta', \mathcal{M}'), \theta\big), \quad \text{for all } (\theta', \mathcal{M}') \in T \quad \text{s.t.} \quad \mathcal{M}' \subset \mathcal{M}.$$

Condition (1) also simplifies if there exists a worst outcome, i.e. an allocation $\underline{x}$ s.t. $u_i(\underline{x}, \theta) \leqslant u_i(x, \theta)$, for all $x \in X$, $\theta \in \Theta$ and $i = 1, \dots, L$.[9] Then we may set $\tilde{f}(t) \equiv \underline{x}$, as this makes it easier to satisfy the corresponding incentive constraints.

Requiring agents to submit verifying claims can have a dramatic impact on the set of implementable social choice functions, as the following condition demonstrates:

*Non-Nested Range Condition* (**NNRC**). For any agent $i$ and any $(\theta_i, \mathcal{M}_i, \theta_{-i}, \mathcal{M}_{-i}) \in T$, we have $(\theta_i', \mathcal{M}_i', \theta_{-i}, \mathcal{M}_{-i}) \notin T$ whenever $\theta_i' \neq \theta_i$ and $\mathcal{M}_i' \subseteq \mathcal{M}_i$.[10]

Under **NNRC**, an agent's untruthful cheap-talk announcement of *her preference parameter* produces an inconsistency with the verifying messages submitted by other agents and their reported preference parameters, when those agents report truthfully and submit all their verifying messages. Thus, when there exists a worst outcome,[11] we have:

[9] Such an allocation exists in environments with transferable utility, since the mechanism design can serve as a collector of fines. If we were also concerned with budget-balanced implementation, then the existence of a worst outcome would involve some loss of generality, even in environments with transferable utility.

[10] The analogue of NNRC in Lipman and Seppi (1995) is their *two-way disprovability* condition.

[11] In fact, we only need an allocation that is worse than the equilibrium allocation for all types of all agents.

**Lemma 1.** *Suppose that* **NNRC** *holds and there exists a worst outcome $\underline{x}$. Then any social choice rule $f(\theta_1,\dots,\theta_L):\Theta\mapsto X$ is implementable in Bayesian Equilibrium.*[12]

**NNRC** implies that any agent's preference parameter can be perfectly inferred from knowledge of the type profile of other agents. So, **NNRC** is similar to the concept of non-exclusive information introduced by Postlewaite and Schmeidler (1986).

**NNRC** can be weakened by requiring that an agent's deviation be detected with positive probability.

**WNNRC**: *Consider any $(\theta_i,\mathcal{M}_i)\in T_i$ and $\theta_i'\neq\theta_i$. Then for all $\mathcal{M}_i'\subseteq\mathcal{M}_i$ there exists $(\theta_{-i},\mathcal{M}_{-i})$ such that $(\theta_i,\mathcal{M}_i,\theta_{-i},\mathcal{M}_{-i})\in T$ but $(\theta_i',\mathcal{M}_i',\theta_{-i},\mathcal{M}_{-i})\notin T$.*

Essentially, **WNNRC** says that when $\theta_i$ shifts to $\theta_i'$, then some combination of other agents' types becomes infeasible. Lemma 1 still holds provided the principal can impose sufficiently large punishments (for example, when monetary transfers are available and agents do not have limited liability). Large punishments may be required because an agent's deviation is only detected with positive probability.

**NNRC** and **WNNRC** provide strong restrictions on agents' communication abilities under which all social choice functions are implementable. Meanwhile, in the traditional environment where all messages are feasible and the types are distributed independently, there are no verifying claims, and the set of implementable social choice functions is minimal. However, the set of implementable social choice functions is not necessarily monotonic in an agent's ability to prove claims, as the following example demonstrates:

**Example 1.** There is a single agent, with three possible types $(\theta_1,\mathcal{M}_1)$, $(\theta_2,\mathcal{M}_2)$ and $(\theta_3,\mathcal{M}_3)$, where $\mathcal{M}_1=\{m_1,m_2\}$, $\mathcal{M}_2=\mathcal{M}_3=\{m_1,m_2,m_3\}$. If we reduce $\mathcal{M}_2$ to $\{m_1,m_2\}$, then type $(\theta_2,\mathcal{M}_2)$ is no longer able to mimic type $(\theta_3,\mathcal{M}_3)$. However, type $(\theta_1,\mathcal{M}_1)$ can now mimic type $(\theta_2,\mathcal{M}_2)$, whereas before she was unable to do so. Hence the net effect on the set of implementable social choice functions is ambiguous.

Our next lemma provides a general condition under which reducing some type's ability to provide verifying claims makes implementation easier.

**Lemma 2.** *Let $t_i=(\theta_i,\mathcal{M}_i)\in T_i$ and $m_i\in\mathcal{M}_i$. Then reducing the set of verifying messages*[13] *of agent-type $t_i$ to $\mathcal{M}_i\setminus\{m_i\}$ does not reduce the set of implementable social choice functions if and only if there does not exist a type $t_i'=(\theta_i',\mathcal{M}_i')$ such that $\mathcal{M}_i\setminus\mathcal{M}_i'=m_i$.*

### *3.1. Minimizing communication*

Thus far, the focus of our model has been on communication costs implicitly associated with agents' inability to send certain verifying messages, not with the physical cost of sending and receiving those messages. In practice, however, producing proofs or evidence and examining them may be costly to agents and the mechanism designer. For this reason, in court proceedings

[12] Note that Lemma 1 focuses on social choice functions that depend *only on agents' preference parameters, not on their communication abilities.* If one wishes to implement social choice functions that also depend on communication abilities, then the following stronger version of NNRC is sufficient: Consider any agent $i$ and any $(\theta_i,\mathcal{M}_i,\theta_{-i},\mathcal{M}_{-i})\in T$. Then $(\theta_i',\mathcal{M}_i',\theta_{-i},\mathcal{M}_{-i})\notin T$ whenever $(\theta_i,\mathcal{M}_i)\neq(\theta_i',\mathcal{M}_i')$.

[13] Note that we make this change without modifying the probability distribution of type profiles $F(\cdot)$.

lawyers may wish to present the minimal amount of evidence necessary to establish their client's claims. Similarly, judges may wish to curtail the amount of evidence presented.

In this subsection, we determine the smallest collection of messages any agent type must send in order for the set of implementable social choice functions not to be curtailed. Intuitively, the solution involves agents sending only messages that have maximal informational content. In our context, if agent $i$ sends message $m \in \mathcal{M}_i$, she proves that her feasible message set does not belong to $Z_i(m) = \{\mathcal{M}_i' \in \mathcal{N}_i\colon\ m \notin \mathcal{M}_i'\}$. Thus, by sending a collection of messages $M \subset \mathcal{M}_i$ agent $i$ establishes that her feasible message set does not belong to $Z_i(M) = \bigcup_{m \in M} Z_i(m) = \{\mathcal{M}_i' \in \mathcal{N}_i \mid M \backslash \mathcal{M}_i' \neq \phi\}$.

Let us now eliminate redundancies in informational content amongst messages in $\mathcal{M}_i$ by selecting the smallest subset $M \subset \mathcal{M}_i$ such that $Z_i(M) = Z_i(\mathcal{M}_i)$ (if there are several such smallest subsets, fix one arbitrarily). Denote this set by $\mathcal{A}_i(\mathcal{M}_i)$.

The "unravelling" result of Grossman (1981) and Milgrom (1981) provides an illustration of our algorithm. Consider a single agent, who has $n$ possible preference types $\theta_1, \dots, \theta_n$. The feasible message set of type $\theta_i$ is $\mathcal{M}_i = \{m_1, \dots, m_i\}$. In this example, we have $Z(\mathcal{M}_i) = \{\mathcal{M}_1, \dots, \mathcal{M}_{i-1}\}$. However, since $Z(m_i) = Z(\mathcal{M}_i)$, type $\theta_i$ only needs to send the single message $m_i$. This message has maximal informational content in the sense defined above.

Next, construct a "minimal communication" mechanism $H$. In this mechanism, the agents are assigned the same allocations as in mechanism $G$ of Theorem 1. However, agent $i$ with a set of verifying messages $\mathcal{M}_i$ is requested to send only the collection of messages $\mathcal{A}_i(\mathcal{M}_i)$, instead of all $\mathcal{M}_i$. Theorem 2 states the implication for implementation in Bayesian equilibrium:

**Theorem 2.** *A social choice function is implementable if and only if it is implementable via mechanism $H$. Furthermore, some such social choice functions are not implementable via mechanisms in which some agent $i$ of type $(\theta_i, \mathcal{M}_i)$ sends less than $\#\mathcal{A}_i(\mathcal{M}_i)$ messages.*

In mechanism $H$, agent-type $t_i$ sends exactly one message that is not available to any type $t_i'$ that cannot send all messages available to $t_i$. So, by Corollary 1 mechanism $H$ minimizes the amount of communication without sacrificing the scope of implementability.[14] Theorem 2 has the following corollary.

**Corollary 2.** *A necessary and sufficient condition for agent type $(\theta_i, \mathcal{M}_i)$ to send a single 'identifier' message $m_i(\mathcal{M}_i)$ in mechanism $H$ is as follows: $m_i(\mathcal{M}_i) \notin \mathcal{M}_i'$, for all $\mathcal{M}_i' \in \mathcal{N}_i$ such that $\mathcal{M}_i \setminus \mathcal{M}_i' \neq \emptyset$.*[15]

### 3.2. Applications

This subsection exhibits the application of our approach to two well-known economic problems. First, consider the following version of Myerson–Satterthwaite bilateral bargaining:

[14] Bull and Watson (2004a) characterize minimal document collections in court trials between two contracting parties who have complete information, when courts can enforce budget-balanced transfers. They show that if agent 1 gets a higher transfer in state $\theta$ than in state $\theta'$, then either agent 1 is able to present a document in state $\theta$ that is not available to her in state $\theta'$, or agent 2 is able to present a document in state $\theta'$ that is not available to her in state $\theta$.

[15] We do not require that an 'identifier' message $m_i(\mathcal{M}_i)$ could not be imitated by an agent with different communication abilities $\mathcal{M}_i' \neq \mathcal{M}_i$. Rather, as in the unraveling example above, it must be the case that if $m_i(\mathcal{M}_i) \in \mathcal{M}_i'$, then $\mathcal{M}_i \subset \mathcal{M}_i'$.

**Example 2** *(Bilateral Bargaining with Limited Ability to Misrepresent).*

A buyer with valuation $b$ and a seller with cost $c$ bargain over the sale of a single good. The traders' types are independently distributed, and can take on one of $n$ possible values, respectively denoted by $b_1 < b_2 < \cdots < b_n$ and $c_1 < c_2 < \cdots < c_n$, where $b_{i-1} \leqslant c_{i-1} < b_i \leqslant c_i$ for all $i$. If trade takes place at price $p$, then the buyer's and seller's surplus are $b_i - p$ and $p - c_j$, respectively. Traders can falsify evidence regarding their types by at most one grid point. Thus $M^b(b_i) = \{b_{i-1}, b_i, b_{i+1}\}$ for $i = 2, \ldots, n-1$, $M^b(b_1) = \{b_1, b_2\}$, and $M^b(b_n) = \{b_{n-1}, b_n\}$. Similarly, $M^s(c_j) = \{c_{j-1}, c_j, c_{j+1}\}$ for $j = 2, \ldots, n-1$, $M^s(c_1) = \{c_1, c_2\}$ and $M^s(c_n) = \{c_{n-1}, c_n\}$.

These restrictions on the traders' falsification abilities imply that there exists an ex-post budget balanced, ex post individually rational incentive compatible mechanism with efficient trade. Indeed, these properties are satisfied by the following social choice function: the good is transferred from seller $c_j$ to buyer $b_i$ iff $i > j$, at a price $p(b_i, c_j) \in [c_j, b_i]$, which is non-decreasing in $b_i$ and $c_j$. According to Corollary 1, the only incentive constraints that have to be satisfied are that buyer type $b_2$ ($b_{n-1}$) be unwilling to mimic buyer type $b_1$ ($b_n$), and that seller type $c_2$ ($c_{n-1}$) be unwilling to imitate seller type $c_1$ ($c_n$). But buyer type $b_2$ has no incentive to mimic type $b_1$, as $b_1$ never gets to trade. Further, buyer type $b_{n-1}$ has no incentive to mimic $b_n$, as she cannot profit from trading with seller type $c_{n-1}$, and would trade at a lower price with all types $c_j$ with $j < n-1$. Similar statements holds for seller types $c_2$ and $c_{n-1}$.

**Example 3** *(Honest Agents).*

Next, we illustrate our approach in an environment where, due to honesty or bounded rationality, some agents are unable to misrepresent their types, while others have unlimited ability to do so. Alger and Ma (2003) and Erard and Feinstein (1994) have analyzed models of this kind, but their mechanisms do not fully exploit the restrictions on the agents' communication abilities.

To illustrate our approach, consider an adverse selection problem in which an agent's preference parameter is either $\theta_H$ (high income or productivity, good health, low cost) or $\theta_L$ (low income or productivity, poor health, high cost). Additionally, an agent can either be 'strategic' (can submit verifying messages, or evidence, consistent with every preference parameter) or 'honest' (can submit verifying messages, or evidence, consistent only with her true preference parameter). Corollary 1 implies that the only incentive constraints that must be imposed in the optimal mechanism are that a 'strategic' agent does not wish to imitate any other type. So, the mechanism designer does not need to leave any surplus to the 'honest' agents. Using mechanism $G$, the designer first distinguishes 'honest' agents from 'strategic' ones, since each of the former can submit only one verifying message corresponding to her true preference parameter, while each of the latter can submit any and all verifying messages. The mechanism designer further screens the 'strategic' agents by letting them choose from a menu.

Alger and Ma (2003) and Erard and Feinstein (1994) do not use menus or additional cheap talk messages to further screen the agents who submit the same 'evidence,' and so the cardinality of their allocation space does not exceed the cardinality of the preference parameter space. Because of this, their mechanisms do not attain the maximal level of profitability for the principal, and leave informational rents to the 'honest' agents under certain probability distributions over types.

An application of Example 3 to optimal taxation is available as an on-line supplement, at http://www.severinov.com/verifiability_supplement.pdf.

**Example 4** *(Unique Implementation).*

An interesting application of our results pertains to standard complete information environments where each agent is informed of the realization of $\theta = (\theta_1, \ldots, \theta_L)$ and $\mathcal{M}_i = \Theta$ for all $i$. Maskin (1999) has shown that a necessary and sufficient condition for a social choice function $f : \Theta \to X$ to be **uniquely** implementable in Nash equilibrium is that it be monotonic.

Monotonicity is a rather strong requirement. We will now show that if there is a small probability that one agent is 'honest,' then the monotonicity condition can be dispensed with. More specifically, let us modify the classical Nash environment to one that satisfies **WNNRC** as follows: there exists a single agent (without loss of generality let it be agent 1) who is 'honest' with probability $\epsilon$, in which case $\mathcal{M}_1 = \{\theta\}$, and 'strategic' with probability $1-\epsilon$, in which case $\mathcal{M}_1 = \Theta$. All other agents are 'strategic,' i.e. $\mathcal{M}_i = \Theta$ for $i \neq 2$. We then have:

**Lemma 3.** *Suppose that all agents have complete information, $u_i(\cdot)$ is continuous in $x$ with $u_i(\underline{x}, \theta_i) = -\infty$ and $X$ is path connected. Also, suppose that agent 1 is honest with probability $\varepsilon > 0$. Consider any social choice function $f : \Theta \to X$ s.t. $u_i(f(\theta), \theta_i) > -\infty$ for all $\theta \in \Theta$ and $i = 1, \ldots, L$. Then for any $\delta > 0$ there exist a social choice function $\hat{f} : \Theta \to X$ satisfying $|u_i(f(\theta), \theta_i) - u_i(\hat{f}(\theta), \theta_i)| \leqslant \delta$ for all $i$ and $\theta \in \Theta$, and a mechanism $g$ with a unique Bayes–Nash equilibrium, such that the outcome is $f(\theta)$ when agent 1 is 'strategic' and $\hat{f}(\theta)$ when agent 1 is 'honest.'*

**Proof of Lemma 3.** Let $\hat{f}(\theta)$ be such that $0 < u_i(f(\theta), \theta_i) - u_i(\hat{f}(\theta), \theta_i) \leqslant \delta$. Agent 1—who may be 'honest' or 'strategic'—is asked to submit two reports, agent 2 is asked to provide a single report, and all other agents do not submit any reports. Let the respective reports be denoted by $\{\theta^{11}, \theta^{12}\}$ and $\theta^2$. Now consider the following allocation rule:

$$g(\theta^{11}, \theta^{12}, \theta^2) = \begin{cases} f(\theta^2), & \text{if } \theta^{11} \neq \theta^{12}, \\ \hat{f}(\theta^2), & \text{if } \theta^{11} = \theta^{12} = \theta^2, \\ \underline{x}, & \text{if } \theta^{11} = \theta^{12} \neq \theta^2. \end{cases}$$

Under the allocation rule $g(\cdot)$, it is a strictly dominant strategy for the 'strategic' type of agent 1 to send a report with $\theta^{11} \neq \theta^{12}$. Consequently, agent 2's expected payoff from reporting $\theta$ truthfully strictly exceeds her expected payoff from reporting $\theta' \neq \theta$, i.e. $u_2(f(\theta), \theta_i) > (1-\epsilon)u_2(f(\theta'), \theta_i) + \epsilon u_2(\underline{x}, \theta_i) = -\infty$. Hence the outcome of the unique Bayes–Nash equilibrium of mechanism $g$ is $f(\theta)$ when agent 1 is strategic, and $\hat{f}(\theta)$ when agent 1 is honest. □

## 4. Further limits on communication

So far, we have assumed that any agent $i$ with feasible communication set $\mathcal{M}_i$ can send any subset of $\mathcal{M}_i$. This assumption is often satisfied in practice: if an agent possesses two different pieces of evidence, she typically has the option to present both.

However, there are important cases in which agents cannot send all possible combinations of feasible messages. Such restrictions may be due to cost, limited capacity of the communication channels, limited attention span of the participants, etc.[16]

To extend our model to such cases we need to modify our notion of an agent's type. For any agent $i$ let $\mathcal{E}_i \subset 2^{\#\mathcal{C}}$ denote the set of all collections of verifying messages she can send in

[16] One example is Glazer and Rubinstein's (2004) analysis of debates where participants can make a limited number of statements. Also, Sanchirico (2001) considers positive message costs in a court context.

combination. An element $E_i$ of $\mathcal{E}_i$ will be referred to as *feasible message combination*. In the model of the previous sections, $\mathcal{E}_i = \{\mathcal{M}_i^{'} \mid \mathcal{M}_i^{'} \subseteq \mathcal{M}_i\}$. At the other extreme agent $i$ may be allowed to send only one basic message from the set $\mathcal{M}_i$. Then $\mathcal{E}_i = \{\{m_i\} \mid m_i \in \mathcal{M}_i\}$. Message combinations that do not belong to $\mathcal{E}_i$ are infeasible or too costly to send.

Agent $i$'s type is described by a pair $t_i = (\theta_i, \mathcal{E}_i)$. As before, let $T_i$ denote the set of all possible types of agents $i$, $T \subset \prod_{i=1,\dots,L} T_i$ the set of all possible type profiles, and $F(t)$ the probability distribution over agents' types. The dependence of agent $i$'s set of feasible message combinations on her type $t_i$ is indicated by $\mathcal{E}_i(t_i)$. Let $\mathcal{E}(t) = \prod_{i=1}^{L} \mathcal{E}_i(t_i)$ be the profile of feasible message combinations, and $\mathcal{E} = \{\mathcal{E}(t)\colon\ t \in \prod_{i=1}^{L} T_i\}$ the set of all conceivable profiles.

A feasible mechanism is a dynamic game form with the property that at every information set where player $i$ can send verifying messages, type $t_i$ is restricted to sending message combinations that belong to $\mathcal{E}_i(t_i)$ . Along any path of this game the union of verifying messages that a player-type $t_i$ can send at different information sets must also belong to $\mathcal{E}_i(t_i)$.

Let $h$ be the outcome function mapping the set of final histories of this game into $X$. We say that a social choice function $f : T \to X$ is implementable in Bayesian (Perfect Bayesian) equilibrium if there exist a dynamic mechanism and an associated Bayesian (Perfect Bayesian) equilibrium $\sigma$ such that $f(t) = h(\sigma(t))$.

Implementation in this environment raises new interesting issues. Particularly, a mechanism needs to specify which feasible message combinations an agent should send. The answer will typically depend on the other agents' actions and the verifying messages they report in the mechanism. This suggests that dynamic mechanisms, in which agents have multiple opportunities to send verifying messages and can use reporting strategies contingent on messages sent by other agents in prior stages, can implement more outcomes than static mechanisms.

Contingent strategies are not necessary when all combinations of messages are feasible, for then the agents can just report all their verifying messages simultaneously and at once. More generally, Bull and Watson (2004b) have shown that dynamic game forms are not needed if their "evidentiary normality" condition holds. This condition is analogous to the Nested Range conditions of Green and Laffont (1986), and requires that for each agent-type there be a message combination s.t. any other type who can mimic it, can mimic all message combinations of the former.

In our analysis, we refrain from imposing any restrictions on the sets $\mathcal{E}_i$ and allow for an arbitrary structure of feasible message combinations for different types. In this general set-up, we show that a necessary and sufficient condition under which mechanisms with a single stage of simultaneous presentation of verifying messages are sufficient is:

**Assumption 1.** For every $i = 1, \dots, L$ there exists $\underline{x}_i \in X$ such that $u_i(\underline{x}_i, \theta) \leqslant u_i(x, \theta)$, for all $\theta \in \Theta$ and $x \in X$.

Assumption 1 holds in many applications such as standard screening environments.

### *4.1. The revelation mechanism*

Theorem 3 will establish that when Assumption 1 holds, it is enough to consider a simple 'Revelation Mechanism' $R$ without sequential presentation of verifying messages. Mechanism $R$ has the following structure:

**Stage 1.** Agents simultaneously and privately choose from individual menus offered by the mechanism. Selection of a specific menu element subjects the agent to a corresponding request for evidence (verifying messages) in stage 2. We label each element $c_i$ in the menu $C_i$ offered to agent $i$ by a different type of agent $i$, so $C_i = \{c_i(t_i)\colon\ t_i \in T_i\}$. Thus, an agent's choice from the menu serves as a cheap-talk announcement of her type.

**Stage 2.** The mechanism specifies (recommends) to each agent which combination of verifying messages she should send. These recommendations are given to agents confidentially, but the recommendation rule is common knowledge. Specifically, the recommendation rule uses agents' choices in stage 1 as an input, and is described by a random mapping $\mu = \prod_{i=1}^{L} T_i \mapsto \Delta(\mathcal{E})$, such that $\mu(.|t) \in \Delta(\mathcal{E}(t))$ for all $t \in T$. The projection of $\mu(.|t)$ on $\Delta(\mathcal{E}_i(t_i))$ denoted by $\mu_i(E_i \mid t_1, \dots, t_n)$ gives the probability with which the mechanism recommends that agent $i$ send a set of messages $E_i$ if in the first stage agents have chosen the profile of elements $(c(t_1), \dots, c(t_n))$ from their respective menus.

**Stage 3.** Agents simultaneously send collections of verifying messages to the mechanism, and the outcome is determined according to the outcome function of the mechanism.

The outcome function is a pair $g(\cdot \mid E, t)$ and $g^p(\cdot \mid E, t)$ of probability distributions over outcomes in $X$, where $E = (E_1, \dots, E_n)$, and $g(\cdot \mid E, t)$ and $g^p(\cdot \mid E, t)$ are implemented when (i) in stage 1, agents choose a profile of menu elements $(c_1(t_1), \dots, c_n(t_n))$, (ii) in stage 3, agents send a profile of message combinations $E$ which is, respectively, recommended and not recommended by the mechanism. Thus, given its game form, mechanism $R$ is completely described by a triple $(\mu, g, g^p)$.

Mechanism $R$ significantly reduces the complexity of the implementation task, because agents send their verifying messages only once, and do so simultaneously. In contrast, in a general dynamic mechanism agents may have to take decisions regarding which messages to send at many different nodes of the game tree.

Let us now provide some intuition why mechanism $R$ is optimal. In the first stage, agents announce their types by making choices from the menus. These announcements are cheap talk. The second stage in which agents send verifying messages serves both to check agents' own type announcements and to cross-check the type announcements made by other agents. Since each agent can report only one feasible message combination, and since the current framework permits an agent to have several non-nested feasible message combinations, it is advantageous to design the mechanism in such a way that the recommendations issued to agents are made contingent on other agents' choices from the menus (i.e. their reported types). Thus, asking agents to make choices from the menus before sending their verifying messages affords greater flexibility, as an agent who announced a particular type in stage 1 can be asked to send various feasible message combinations to cross-check different announced types of the other agents.

Abstracting from cross-checking other agents' reports for a moment, the fact that an agent may have several non-nested feasible message combinations also implies that the recommendations issued in stage 2 are likely to be random. That is, an agent may be required to send several of her feasible message combinations with a positive probability. Random recommendations make imitation by other types of the same agent more difficult. To illustrate this point, consider the following example:

**Example 5.** There is a single-agent who has three possible types: $t_1 = (\theta_1, E_1, E_2)$, $t_2 = (\theta_2, E_1)$, and $t_3 = (\theta_3, E_2)$, where $E_1 = \{m_1, m_2\}$ and $E_2 = \{m_1, m_3\}$. The set of outcomes is $X = \{a, b, d\}$, and the agent is an expected utility maximizer, with a utility function given by $u(a, \theta_i) = 10$, $u(b, \theta_i) = 6$, $u(d, \theta_i) = 0$, for $i \in \{1, 2, 3\}$.

Consider implementation of the social choice function $f(t_1) = a$, $f(t_2) = f(t_3) = b$. Observe that $f(\cdot)$ is not implementable via any mechanism in which $t_1$ does not have to randomize and sends either $E_1$ or $E_2$ with probability 1. In this case, either $t_2$ or $t_3$ will imitate $t_1$'s strategy by sending the same message set as $t_1$. In fact, it is easy to see that $t_2$ ($t_3$) will follow $t_1$'s strategy to the extent feasible if this strategy requires $t_1$ to send $E_1$ ($E_2$) with probability of at least 0.6. However, $f_1(\cdot)$ is implementable via mechanism $R$, in the first stage of which the agent is offered a menu consisting of three elements. The recommendation rule $\mu(\cdot|t)$ and the outcome functions of the mechanism $g(\cdot \mid E, t)$ and $g^p(\cdot \mid E, t)$ (probability distributions over outcomes) are described as follows:

Menu Element 1. $\mu(\cdot|t_1)$: $\mu(E_i|t_1) = 1/2$; $g(a \mid E_i, t_1) = 1$ for $i \in \{1,2\}$, and $g^p(d \mid E, t_1) = 1$.
Menu Element 2. $\mu(\cdot|t_2)$: $\mu(E_1|t_2) = 1$; $g(b \mid E_1, t_2) = 1$ and $g^p(d \mid E, t_2) = 1$.
Menu Element 3. $\mu(\cdot|t_3)$: $\mu(E_2|t_3) = 1$, $g(b \mid E_2, t_3) = 1$ and $g^p(d \mid E, t_3) = 1$.

The strategy space of player $i$ of type $t_i$ in mechanism $R$ is given by $\Sigma_i(t_i) = \{(c_i, \delta_i) \mid c_i \in C_i, \delta_i : \mathcal{E}_i \to \mathcal{E}_i(t_i)\}$. We will say that a strategy of agent-type $t_i$ is truthful and obedient if it prescribes that $t_i$ choose element $c_i(t_i)$ from the menu, and obey the recommendation from $\mu_i(\cdot \mid t_i, t_{-i})$.

**Theorem 3.** *Suppose Assumption* 1 *holds. Then a social choice function implementable in Bayesian equilibrium of some, possibly dynamic, mechanism is also implementable in a Perfect Bayesian equilibrium of mechanism R in which all agent-types follow truthful and obedient strategies.*

Theorem 3 shows that as long as Assumption 1 holds, from the viewpoint of implementation there is no loss of generality in considering only incentive compatible revelation mechanisms $R$. Thus, we extend Theorem 6 in Bull and Watson (2004b) which restricts consideration to dynamic games in which the agents may disclose evidence at most once along any path of the game tree. In contrast, Theorem 4 considers a larger class of games and mechanisms in which players can reveal verifying information at several nodes in the game. In this case Assumption 1 becomes necessary for mechanism $R$ to be sufficient for implementation.

Corollary 3 describes the incentive constraints that must be imposed on implementable social choice functions:

**Corollary 3.** *Suppose Assumption* 1 *holds. Then a social choice function $f : T \to X$ is implementable in Bayesian equilibrium if and only if the following condition holds*:

*For all $t \in \prod_i T_i$, there exist a probability distribution $\mu(\cdot|t) \in \Delta(\mathcal{E})$, and a probability distribution $g(\cdot \mid E, t) \in \Delta(X)$, such that $\mu(\cdot|t) \in \Delta(\mathcal{E}(t))$ and $g(f(t) \mid E, t) = 1$ for all $E \in \operatorname{supp} \mu(\cdot|t)$ and all $t \in T$,*[17] *and such that for every $i = 1, \ldots, L$, and every $t_i, t_i' \in T_i$ we have*

$$\begin{aligned}
&\sum_{t_{-i} \in T_{-i}} u_i\big(f(t_i, t_{-i}), \theta\big) F(t_{-i}|t_i) \\
&\quad \geqslant \sum_{t_{-i} \in T_{-i}} \Bigg\{ \sum_{E: E_i \in \mathcal{E}_i(t_i)} \sum_{x \in X} u_i(x, \theta) g\big(x \mid E, (t_i', t_{-i})\big) \mu\big(E \mid t_i', t_{-i}\big) \\
&\qquad + \sum_{E: E_i \notin \mathcal{E}_i(t_i)} u_i(\underline{x}_i, \theta) \mu\big(E \mid t_i', t_{-i}\big) \Bigg\} F(t_{-i}|t_i).
\end{aligned} \tag{2}$$

[17] Recall that our analysis allows for the possibility that $T \neq \prod_i T_i$.

To understand the corollary, suppose that mechanism $R$ with outcome triple $(\mu, g, g^p)$ implements $f(\cdot)$. Then, we must have $g(f(t) \mid E, t) = 1$ for all $E \in supp\, \mu(t)$ and all $t \in T$. Also, as shown in the proof of Theorem 3, the largest set of social choice functions is implemented when $g^p(\cdot)$ assigns $\underline{x}_i$, the worst outcome of agent $i$, when $i$ is the only agent who does not send the recommended set of messages. Therefore, if $i$ deviates in the choice from the menu, her optimal continuation strategy is to send the recommended set of messages whenever she is able to do so. Thus, type $t_i$'s payoff from choosing a menu element $c(t_i')$ is given by the right-hand side of (2). Since $f(\cdot)$ is implementable, the family of incentive constraints (2) must hold.

In checking the implementability of a social choice function, Corollary 3 transforms the complicated task of finding an appropriate mechanism or disproving its existence, to the simpler one of checking the existence of a solution to a linear program. In an important special case, it permits a very simple characterization of the set of implementable outcomes:

**Corollary 4.** *Suppose that $u_i(\underline{x}_i, \theta) = -\infty$ for all $i \in 1, \ldots, L$, and that $f(\cdot)$ is such that $u_i(f(t), \theta_i) > -\infty$. Then* (2) *becomes*

$$\sum_{t_{-i} \in T_{-i}} u_i\big(f(t_i, t_{-i}), \theta\big) F_i(t_{-i}|t_i)$$
$$\geqslant \sum_{t_{-i} \in T_{-i}} u_i\big(f\big(t_i', t_{-i}\big), \theta\big) F_i(t_{-i}|t_i), \quad \textit{for all } t_i' = \big(\theta_i', E_i'\big) \in T_i \quad \textit{s.t.} \quad E_i' \subset E_i.$$

Notice the analogy to Corollary 1: the only incentive constraints that need to be imposed upon a social choice function is that type $t_i$ not prefer the allocation of any type $t_i'$ that she can *fully* imitate, i.e. such that any feasible message combination of type $t_i'$ is available to type $t_i$. This is because the distribution $\mu_i(\cdot|t)$ can be chosen to assign equal weights to all elements of $\mathcal{E}_i(t_i)$. Then if type $t_i$ selects the menu element $c(t_i')$ and is unable to send some message combination feasible for $t_i'$, she is punished harshly.

### *4.2. The role of a debate in mechanisms*

Next, we show that when Assumption 1 does not hold, then dynamic mechanisms in which agents take turns sending verifying messages and some agents send such messages more than once can implement more outcomes than the revelation mechanism $R$:

**Theorem 4.** *Suppose Assumption* 1 *does not hold. Then there exist environments and social choice functions implementable in Perfect Bayesian equilibrium of a dynamic mechanism in which some agents send verifying messages at several information sets along some paths of the game, but not implementable in Bayesian equilibrium via the Revelation Mechanism $R$.*

Note that, to make out result stronger, we prove the impossibility of implementation via mechanism $R$ using a weaker concept of *Bayesian equilibrium*, but construct a dynamic mechanism implementing the desired social choice function in *Perfect Bayesian* equilibrium.

If a social choice function is not implementable in Bayesian equilibrium via mechanism $R$, then, as shown by Bull and Watson (2004b), it cannot be implemented in Bayesian equilibrium via any mechanism in which each agent sends verifying message/presents evidence only once along any path of the game. Thus, our Theorem 4 implies that, without Assumption 1, in some environments it is necessary to have agents present evidence more than once in the course of the game, taking turns in doing so. So the mechanism has to be designed in the form of a debate. This result provides a foundation for a theory of debates in environments with asymmetric information. In contrast, Theorem 3 shows that such dynamic mechanisms are not needed when Assumption 1 holds.

The intuition behind Theorem 4 is as follows. Suppose that some agent $i$ can present only two messages from her set of verifying messages, and the desired social choice function is implementable only if she does so. Specifically, suppose that $i$'s first message provides a partial verification of her type, while her second message can complete the screening if it is chosen properly and cross-checked against the claims of other agents. So, which second message agent $i$ is requested to send depends on the claims submitted by the other agent(s). In a dynamic mechanism, agent $i$ could initially send just her first verifying message. Then at some later point in the mechanism she receives a request from the mechanism for a specific second verifying message, and submits it on the equilibrium path. Importantly, which second message is requested gives agent $i$ additional information about the types of the other agent(s). Therefore, this sequence of events is important for implementability. Specifically, an alternative sequence where agent $i$ is told which second message she has to send before she has communicated her first message, could, in fact, undermine the incentive compatibility of sending the 'right' first verifying message. But this alternative sequence of events does, in fact, occur in the Revelation mechanism or any mechanism where agents submit verifying messages only once. In the proof of Theorem 4 we show that such early release of information to some agent could make implementation impossible in mechanism $R$. On the other hand, when some agent sends verifying messages more than once, the mechanism designer can make requests for specific verifying messages, and thereby release additional information to the agent, after the agent has already sent other verifying messages. This makes it easier to ensure that the agent still has incentives to send her first bunch of messages, which increases the scope of implementability.

## 5. Conclusion

We have studied implementation in environments where some agents have limited ability to imitate the behavior of other types and manipulate their private information. The principal can use such limitations to better screen the agents, and elicit information regarding their preference parameters more efficiently. The agents' abilities to prove claims play an important role in determining their payoffs, with those who possess larger sets of verifying messages receiving higher informational rents.

As noted in the introduction, our paper is closely related to the recent work of Bull and Watson (2004b) and Forges and Koessler (2005). Bull and Watson (2004b) consider a setting in which agents can present state-contingent *collections* of pieces of evidence, and focus on the role of "evidentiary normality." They establish a version of the revelation principle, by constructing a mechanism for which there is a one-to-one mapping between cheap talk report/evidence pairs produced in equilibrium and players' types. They also show that in settings with complete information, evidentiary normality permits a translation into the direct revelation model studied by Green and Laffont (1986). Finally, Bull and Watson (2004b) were first to raise the important question whether static mechanisms suffice for implementation when players cannot send all possible message combinations. Their pioneering work shows that a simple three-stage dynamic mechanism suffices, when the designer is limited to game forms in which each player can send evidence only once along any path of the tree.

There are several important differences between our work and Bull and Watson (2004b). First, under evidentiary normality Bull and Watson (2004b) establish equivalence results similar to our Theorem 1, but their equivalence results are established for static mechanisms and dynamic mechanisms in which every agent can send her evidence (verifying messages in our terminology) only once. Our Theorem 1 implies that the latter restriction on dynamic mechanisms is unnecessary, i.e. the scope of our Theorem 1 is broader. Our paper also goes further by identifying the conditions under which all verifying messages have to be sent.

Second, the analysis of dynamic mechanisms in Bull and Watson (2004b) when evidentiary normality does not hold (the counterpart of the environment studied in our Section 5) is also restricted to a class of mechanisms in which each agent is allowed to present evidence only once along any path of the game tree. Bull and Watson (2004b) show that any social choice rule implementable via a mechanism of this class can also be implemented via a *special three-stage mechanism*. As we demonstrate, their restriction on the class of dynamic mechanisms is substantive. We establish a necessary and sufficient condition under which one can restrict attention to the Revelation mechanism $R$—which is similar to the three-stage mechanism of Bull and Watson (2004b)—and, hence, to mechanisms in which every agent presents evidence only once along any path of the game tree. This condition requires all types of any particular agent to receive the lowest utility from the same allocation. Our Theorem 4 establishes that, when this condition fails, dynamic mechanisms in which agents take turns submitting verifying messages, with some agents submitting such messages more than once, implement a larger set of social choice functions than are implementable via our Revelation mechanism $R$ or the *three-stage mechanism* of Bull and Watson (2004b).

A third important difference is that we characterize the set of implementable allocations in terms of the incentive constraints that need to be satisfied (see Corollaries 1, 3, 4). This is important as the Revelation Principle has proven to be of value in standard environments largely because it permits such a characterization. Indeed, the task of determining which social choice functions are implementable then boils down to solving a linear program.

Finally, our model allows for a stochastic relationship between an agent's preference parameter and her set of verifying messages. In Bull and Watson (2004b), an agent's type determines both her preferences and her evidence (set of verifying messages in our terminology). This implies that either there is a deterministic relationship between an agent's preference parameter and her evidence, in which case our model allows for a broader range of uncertainty and incomplete information, or one has to broaden their notion of type so that it includes both a description of the agent's preference parameter and her set of verifying messages.[18] However, in this case there is a qualitative difference in the amounts of communication required by mechanism our $G$ and its counterpart in Bull and Watson (2004b). In the broader interpretation of their model, an agent would make a cheap talk announcement of her preference parameter *and* the evidence she has available, prior to presenting that evidence. In mechanism $G$ an agent makes a cheap talk statement *only about her preference parameter*.[19]

Forges and Koessler (2005) analyze the communication equilibria of a fixed game. In a communication stage preceding the game, in each of a fixed number of periods players simultaneously send a cheap-talk message and a verifying message to the mediator, and then receive

[18] We are thankful to a referee for suggesting that the model of Bull and Watson (2004b) admits such an interpretation.

[19] The structure of mechanism $G$ allows us to further reduce the amount of required communication in special cases. For example, when an agent's set of verifying messages uniquely determines her preference parameter no cheap-talk message is required.

a message back from him. At the end of the communication stage, players select their actions in the game. Theorem 3.1 of Forges and Koessler (2005) characterizes the Nash equilibria of this communication game in terms of truthful and obedient equilibria of a three-stage game in which players first simultaneously submit a cheap talk claim regarding their type and a "certificate" regarding their type (a report from a type-dependent message space) to the mediator, who then makes private recommendations to players on the actions to be taken in the game. Thus, the relationship between Theorem 3.1 of Forges and Koessler (2005) and our Theorem 1 is akin to the relationship between communication equilibria of a Bayesian game (where players' strategy choices determine the outcome, and the mediator facilitates communication) and the set of implementable outcomes of a Bayesian collective choice problem where the mechanism designer can construct any game and introduce any communication protocol. In particular, the mechanism designer can require players to play the communication game of Forges and Koessler (2005) and commit to the action rule chosen by players in the associated communication equilibrium. Forges and Koessler (2005) do not analyze general restrictions on the message combinations that player can send (as we do in Section 4), but they do investigate the interesting case in which there is a single period of communication (so players can only send one verifying message). Importantly, their Theorem 3.3 establishes that under a "minimal closure condition" lengthening the communication stage will not expand the set of equilibrium outcomes. The latter condition is analogous to evidentiary normality in Bull and Watson's (2004b) setting and to Corollary 2 in our setting.

We end this section with some suggestions for future work. The model and methodology presented in this paper open up an avenue for interesting applications to many areas, such as adjudication procedures, contract design, negotiation and price discrimination. In these applications, costs of information transmission or limited attention span of the participants impose restrictions on the number of messages that can feasibly or optimally be elicited. The results of Section 4 permit progress in these intriguing areas of research. In particular, our own research focuses on the optimal design of debating rules in settings of incomplete information.

## Acknowledgments

We thank Yeon Koo Che, Bart Lipman, Larry Samuelson, Robert Wilson and seminar participants at the University of Wisconsin, Caltech, UCLA, Florida, Stanford GSB, USC, Duke, the Summer 2001 European Meetings of the Econometric Society, the Winter 2001 North American Econometric Society Meetings, and the 2001 Canadian Economic Theory Conference for helpful comments and suggestions. The research on this paper was completed while the second author was at Fuqua School of Business, Duke University, and he thanks the School for the support.

## Appendix A

**Proof of Proposition 1.** Suppose that there is one agent, of three possible types: $(\theta_1, \{m_1, m_2, m_3, m_4\})$, $(\theta_2, \{m_2, m_3, m_4, m_5\})$, or $(\theta_3, \{m_1, m_3, m_5\})$.

The outcome space is $X = \{x_1, x_2, x_3\}$, and preferences are: $x_3 \prec_{\theta_1} x_2 \prec_{\theta_1} x_1$, $x_3 \prec_{\theta_2} x_1 \prec_{\theta_2} x_2$, and $x_3 \prec_{\theta_3} x_1 \prec_{\theta_3} x_2$. Define social choice functions $f^1(\cdot)$, $f^2(\cdot)$, $f^3(\cdot)$, $f^4(\cdot)$ as follows:

$$f^1(\theta_1) = x_1,\ f^1(\theta_2) = x_2,\ f^1(\theta_3) = x_3; \qquad f^2(\theta_1) = x_1,\ f^2(\theta_2) = x_3,\ f^2(\theta_3) = x_1;$$
$$f^3(\theta_1) = x_3,\ f^3(\theta_2) = x_1,\ f^3(\theta_3) = x_1; \qquad f^4(\theta_1) = x_3,\ f^4(\theta_2) = x_3,\ f^4(\theta_3) = x_2.$$

Note that $f^1(\cdot)$ is implementable via the mechanism with outcome function $g^1(m_2) = x_1$, $g^1(m_4) = x_2$, $g^1(m_1) = g^1(m_3) = g^1(m_5) = x_3$. Also $f^2(\cdot)$ is implementable via the mechanism with outcome function $g^2(m_1) = x_1$, $g^2(m_2) = g^2(m_3) = g^2(m_4) = g^2(m_5) = x_3$, and $f^3(\cdot)$ is implementable via the mechanism $g^3(m_5) = x_1$, $g^3(m_1) = g^3(m_2) = g^3(m_3) = g^3(m_4) = x_3$. On the other hand, $f^4(\cdot)$ is not implementable because any message available to $\theta_3$ is available either to $\theta_1$ or $\theta_2$, and each of these two types prefers to receive $x_2$ rather than $x_3$.

Now, let us show that there is no model in which type $\theta_i$ has message spaces $\mathcal{M}(\theta_i) \subset \Theta_i$ satisfying $\theta_i \in \mathcal{M}(\theta_i)$ for $i = 1, 2, 3$, that has the same set of implementable allocations. We will henceforth refer to such model as 'direct' because in such a model agents send direct messages about their preference parameter.

Observe that for the social choice function $f^1(\cdot)$ to be implementable in a direct model, it must be that $\mathcal{M}(\theta_3) = \{\theta_3\}$. Indeed, if $\theta_1 \in \mathcal{M}(\theta_3)$, then such implementation must rely on an outcome function $\hat{g}(\cdot)$ such that $\hat{g}(\theta_1) = \hat{g}(\theta_3) = x_3$. But then since only message $\theta_2$ can map into an outcome different from $x_3$ it would be impossible for types $\theta_1$ and $\theta_2$ to obtain their respective allocations. A symmetric argument rules out $\theta_2 \in \mathcal{M}(\theta_3)$.

Observe also that the social choice function $f^2(\cdot)$ is implementable in a direct model only if $\theta_3 \notin \mathcal{M}(\theta_2)$. Indeed, since $\mathcal{M}(\theta_3) = \{\theta_3\}$, then the outcome function $\hat{g}(\cdot)$ implementing $f^2(\cdot)$ in a direct model must satisfy $\hat{g}(\theta_3) = x_1$. So, if $\theta_3 \in \mathcal{M}(\theta_2)$, then type $\theta_2$ could do better by sending $\theta_3$ and obtaining the allocation $x_1$. A symmetric argument establishes that implementability of $f^3(\cdot)$ in a direct model requires that $\theta_3 \notin \mathcal{M}(\theta_1)$.

Now consider any direct model satisfying the restrictions $\mathcal{M}(\theta_3) = \{\theta_3\}$, $\theta_3 \notin \mathcal{M}(\theta_1)$ and $\theta_3 \notin \mathcal{M}(\theta_2)$. In any such model, the social choice function $f^4(\cdot)$ is implementable via the mechanism $g(\theta_3) = x_2$ and $g(\theta_1) = g(\theta_2) = x_3$, contradicting that the set of implementable allocations coincides with the one in the model with message spaces $\mathcal{M}(\theta_1), \mathcal{M}(\theta_2), \mathcal{M}(\theta_3)$. □

**Proof of Theorem 1.** Fix an environment $\{X, \mathcal{C}, T, F(\cdot), u_1(\cdot), \ldots, u_L(\cdot)\}$, and suppose that the social choice function $f(\cdot): T \to X$ is implementable in Bayesian equilibrium via some mechanism $\gamma$ with strategy space $S^\gamma = \prod_{i=1}^L S_i^\gamma$ and outcome function $g^\gamma$: $S^\gamma \mapsto X$. Let $s^*(t) = (s_1^*(t_1), \ldots, s_L^*(t_L))$ be the strategy profile in the equilibrium of the mechanism $\gamma$ which implements $f(\cdot)$. Thus, we have $f(t) = g^\gamma(s^*(t))$.

The proof relies on cheap-talk messages, but we indicate below how the mechanism can be implemented via menus rather than cheap-talk messages. Construct the mechanism $G^\gamma(\cdot)$ as follows. If the agents send a profile of collections of verifying messages $(\hat{\mathcal{M}}_1, \ldots, \hat{\mathcal{M}}_L)$ in the first stage and then announce a profile of preference parameters $(\hat{\theta}_1, \ldots, \hat{\theta}_L)$ such that $(\hat{\mathcal{M}}_i, \hat{\theta}_i) \in T_i$ for all $i = 1, \ldots, L$, then the mechanism $G^\gamma(\cdot)$ implements the outcome $g^\gamma(s_1^*(\hat{\theta}_1, \hat{\mathcal{M}}_1), \ldots, s_L^*(\hat{\theta}_L, \hat{\mathcal{M}}_L))$. Note that $s_i^*(\hat{\theta}_i, \hat{\mathcal{M}}_i)$ is well-defined for all $i \in \{1, \ldots, L\}$, since $(\hat{\theta}_i, \hat{\mathcal{M}}_i) \in T_i$.

If $(\hat{\mathcal{M}}_i, \hat{\theta}_i)$ is such that $\hat{\mathcal{M}}_i \notin \mathcal{N}_i$ (i.e. there it no type $t_i \in T_i$ whose full set of verifying messages is $\hat{\mathcal{M}}_i$) and $(\hat{\mathcal{M}}_j, \hat{\theta}_j) \in T_j$ for every agent $j \neq i$, then the mechanism $G^\gamma$ assigns the outcome which will be attained in mechanism $\gamma$ when agent $i$ follows some arbitrarily fixed strategy that involves sending no verifying messages at any information set in $\gamma$ and every other agent $j \neq i$ follows the strategy $s_j^*(\hat{\theta}_j, \hat{\mathcal{M}}_j)$. Finally, if there are at least two agents, $i$ and $i'$, such that $\hat{\mathcal{M}}_i \notin \mathcal{N}_i$ and $\hat{\mathcal{M}}_{i'} \notin \mathcal{N}_{i'}$, then $G^\gamma$ assigns some arbitrarily fixed outcome.

Since $f(t) = g^\gamma(s^*(t))$, the mechanism $G^\gamma(\cdot)$ implements $f(t)$ if all agents follow truthfully revealing strategies. To complete the proof let us show that these strategies constitute a Bayesian equilibrium of $G^\gamma(\cdot)$. So, suppose that all agents other than $i$ follow truthfully revealing strategies. Consider the best response by agent $i$ of type $t_i = (\mathcal{M}_i, \theta_i) \in T_i$. If this agent-type sends a collection of messages $\hat{\mathcal{M}}_i$ and announces preference parameter $\hat{\theta}_i \in \Theta_i$, then $\hat{\mathcal{M}}_i \subseteq \mathcal{M}_i$. Further, if $(\hat{\mathcal{M}}_i, \hat{\theta}_i) \in T_i$, then $G^\gamma(\cdot)$ assigns the outcome $g^\gamma(s_1^*(\theta_1, M_1), \dots, s_i^*(\hat{\theta}_i, \hat{M}_i), \dots, s_L^*(\theta_L, \mathcal{M}_L))$ whenever the type profile of all agents other than $i$ is given by $((\theta_1, M_1), \dots, (\theta_{i-1}, M_{i-1}), (\theta_{i+1}, M_{i+1}), \dots, (\theta_L, M_L))$. But since $s_i^*(\cdot)$ is agent $i$'s best response strategy in $\gamma$, type $t_i = (\mathcal{M}_i, \theta_i)$ gets a (weakly) higher payoff by following the truthfully revealing strategy in $G^\gamma$, i.e. when $(\hat{\mathcal{M}}_i, \hat{\theta}_i) = (\mathcal{M}_i, \theta_i)$.

Second, suppose that agent $i$ announces $\hat{\mathcal{M}}_i$ such that there is no $t_i \in T_i$ with the set of verifying messages $\hat{\mathcal{M}}_i$. Since the other agents follow truthfully revealing strategies, $G^\gamma(\cdot)$ assigns an arbitrarily fixed outcome of $\gamma$ that obtains when $i$ does not send any verifying messages. This outcome if worse for agent $i$ than the outcome that she gets from a truthfully revealing strategy, because the strategy profile $s^*(\cdot)$ constitutes an equilibrium of $\gamma$.

Note that the same equilibrium outcome is attained in $G^\gamma$ if, instead of cheap talk, in the second stage agent $i$ is asked to choose from a menu of allocation functions

$$\{f(\hat{\theta}_i, \hat{\mathcal{M}}_i, \cdot) : T_{-i} \to X\}_{\hat{\theta}_i \in I_i(\mathcal{M}_i)}$$

after she reports some $\mathcal{M}_i' \in \mathcal{N}_i$ in the first stage. This is a menu of allocation functions since the agent does not know the reports of the other agents and the final allocation $f(\hat{\theta}_1, \hat{\mathcal{M}}_1, \dots, \hat{\theta}_L, \hat{\mathcal{M}}_L)$ is contingent on every agent's report. Choosing a particular element from this menu is equivalent to announcing some preference parameter $\hat{\theta}_i$.

To show that mechanisms of class $G$ permit to implement a larger set of social choice rules than mechanisms in which agents do not have to send all their verifying messages, consider the following example. There is a single agent with three types $t_1 = (\theta_1, \{m_1, m_2\})$, $t_2 = (\theta_2, \{m_2, m_3\})$, $t_3 = (\theta_3, \{m_3, m_1\})$. The decision space is given by $X = \{x_1, x_2, x_3, \underline{x}\}$. The outcome $\underline{x}$ is the worst for every type, and the rest of the payoff structure is as follows:

$$u_1(x_1, \theta_1) < u_1(x_2, \theta_1) < u_1(x_3, \theta_1),$$
$$u_2(x_2, \theta_2) < u_2(x_3, \theta_2) < u_2(x_1, \theta_2),$$
$$u_3(x_3, \theta_3) < u_3(x_1, \theta_3) < u_3(x_2, \theta_3).$$

Consider the social choice function $f(t_1) = x_1$, $f(t_2) = x_2$, $f(t_3) = x_3$. Then a mechanism which assigns some allocation $x_i$, $i \in \{1, 2, 3\}$, after receiving only one of the messages $\hat{m} \in \{m_1, m_2, m_3\}$ cannot implement $f(\cdot)$. One of the types $t_j \neq t_i$ who can send message $\hat{m}$ will find is strictly profitable to do so, since this will yield allocation $x_i$ rather than $x_j$.

However, $f(\cdot)$ can be implemented via the following mechanism $G(\cdot)$: if the agent reports $\{m_i, m_{(i+1) \bmod 3}\}$ she receives the allocation $x_i$, otherwise she receives the allocation $\underline{x}$. □

**Proof of Lemma 1.** Consider any social choice rule $f(\cdot)$ and the corresponding mechanism $G$. By NNRC, if $(\theta_i, \mathcal{M}_i, \theta_{-i}, \mathcal{M}_{-i}) \in T$ and $(\theta_i', \mathcal{M}_i')$ is such that $\theta_i \neq \theta_i'$ and $\mathcal{M}_i' \subseteq \mathcal{M}_i$, then $(\theta_i', \mathcal{M}_i', \theta_{-i}, \mathcal{M}_{-i}) \notin T$, and so $G(\theta_i', \mathcal{M}_i', \theta_{-i}, \mathcal{M}_{-i}) = \underline{x}$. Consequently, all incentive constraints hold in $G$, and $f(\cdot)$ is implementable. □

**Proof of Lemma 2.** The condition of the lemma guarantees that any type $t_i'$ of agent $i$, who is unable to mimic $t_i$ when $t_i$'s set of verifying messages is $\mathcal{M}_i$ remains unable to mimic $t_i$ after $t_i$'s set of verifying messages shrinks to $\mathcal{M}_i \setminus \{m_i\}$. This together with the fact that the probability distribution $F(\cdot)$ does not change imply that the set of incentive constraints which an implementable social choice function has to satisfy (see Corollary 1) remains the same after $\mathcal{M}_i$ shrinks, and so the set of implementable choice functions remains unchanged. Conversely, when the condition of the lemma is violated, it is easy to construct examples where profitable deviations appear after agent $i$'s set of verifying messages shrinks. □

**Proof of Theorem 2.** Note that mechanism $H$ is well-defined, i.e. any collection of messages sent in $H$ by agent $i$ corresponds to a unique $\mathcal{M}_i \in \mathcal{N}_i$. Precisely, if $\mathcal{M}_i'$ is s.t. $\mathcal{M}_i' \neq \mathcal{M}_i$ then $\mathcal{A}_i(\mathcal{M}_i') \neq \mathcal{A}_i(\mathcal{M}_i)$. To see this, suppose otherwise. First, consider $\mathcal{M}_i'$ s.t. $\mathcal{M}_i \setminus \mathcal{M}_i' \neq \emptyset$. Then $\mathcal{M}_i' \in Z_i(\mathcal{M}_i) = Z_i(\mathcal{A}_i(\mathcal{M}_i))$. Thus if we had $\mathcal{A}_i(\mathcal{M}_i') = \mathcal{A}_i(\mathcal{M}_i)$, we would obtain the contradiction that $\mathcal{M}_i' \in Z_i(\mathcal{A}_i(\mathcal{M}_i'))$. Next, suppose that $\mathcal{M}_i \subset \mathcal{M}_i'$ and the inclusion is strict. Then $\mathcal{M}_i \in Z_i(\mathcal{M}_i') = Z_i(\mathcal{A}_i(\mathcal{M}_i'))$. But then $\mathcal{A}_i(\mathcal{M}_i') = \mathcal{A}_i(\mathcal{M}_i)$ would imply the contradiction that $\mathcal{M}_i \in Z_i(\mathcal{A}_i(\mathcal{M}_i))$.

Since mechanisms $G$ and $H$ differ only in the sets of verifying messages which agents send (their passwords), and since each agent $i$ can send the collection of messages $\mathcal{A}_i(\mathcal{M}_i)$ in mechanism $H$ iff she can send the collection of messages $\mathcal{M}_i$ in mechanism $G$, any s.c.f. implementable via $G$ is also implementable via $H$. The reverse follows by Theorem 1.

To establish that in general agent $i$ of type $(\theta_i, \mathcal{M}_i)$ has to send at least $\#\mathcal{A}_i(\mathcal{M}_i)$ messages, consider the single-agent example studied in the proof of Theorem 1. We have $\#\mathcal{A}(\mathcal{M}(\theta_i)) = 2$ for all $i = 1, 2, 3$. In the proof of Theorem 1, we have established the existence of a social choice function that can be implemented only if every type sends at least two messages. □

**Proof of Theorem 3.** Suppose that the social choice function $f : T \to X$ is implementable via some, possibly dynamic, mechanism $\Gamma$ with outcome function $h$, i.e. there exists a Bayesian equilibrium $\sigma^*(\cdot) = (\sigma_1^*(\cdot), \ldots, \sigma_L^*(\cdot))$ of $\Gamma$ such that $f(t) = h(\sigma^*(t))$. Let us construct a recommendation rule $\mu(\cdot)$ and outcome functions $g(\cdot)$ and $g^p(\cdot)$ so that the mechanism $R$ associated with $(\mu, g, g^p)$ has a Perfect Bayesian equilibrium in which all agent-types follow truthful and obedient strategies and the outcome of which is given by $f(\cdot)$.

First, let $\mu(\cdot|t)$ be the probability distribution over $\mathcal{E}(t)$, such that for any $E = (E_1, \ldots, E_L) \in \mathcal{E}(t)$, $\mu(E|t)$ equals the probability that on the path of the game associated with mechanism $\Gamma$ agent $i$ $(i = 1, \ldots, L)$ sends a collection of messages the union of which is equal to $E_i$ when the agents' type profile is given by $t = (t_1, \ldots, t_L)$ and they play the profile of strategies $\sigma^*(t)$.

Next, let $\tilde{f}(\cdot \mid E, t)$ denote the probability distribution over outcomes of $\Gamma$ conditional on the type profile being $t$, the agents following the strategy profile $\sigma^*(t)$, and the realized collection of messages along the path of the game being $E$. Set $g(\cdot \mid E, t) = \tilde{f}(\cdot \mid E, t)$ for all $t \in \prod_{i=1}^{L} T_i$ and note that for any $t \in T$ the distribution $\tilde{f}(x \mid E, t)$ puts weight only on $x = f(t)$. The choice of the outcome function $g(\cdot \mid E, t)$ implies that if all agent-types follow truthful and obedient strategies, then $f(\cdot)$ is implemented. Finally, suppose the mechanism recommends $E^r$ (i.e. the realization of $\mu(\cdot|t)$ equals $E^r$), yet the agents report $E \neq E^r$. Then we let the punishment $g^p(\cdot \mid E, t)$ assign probability one to the outcome $\underline{x}_{\hat{i}}$, where $\hat{i} = \min\{i \mid E_i^r \neq E_i\}$.

Let us now demonstrate that $(\mu, g, g^p)$ has a Perfect Bayesian equilibrium in which all agent-types follow truthful and obedient strategies. First, given any realized profile of types $t$ and any first-stage profile of menu choices $(c_1(\tilde{t}_1), \ldots, c_L(\tilde{t}_L))$, it is optimal for agent $i$ to obey the mechanism's recommendation whenever possible, i.e. whenever the mechanism recommends $E_i^r \in \mathcal{E}_i(t_i)$, irrespective of her beliefs.[20] Indeed, sending $E_i \neq E_i^r$ either results in $i = \hat{i}$ (yielding

[20] This implies that our equilibrium is Perfect Bayesian, and not just Bayesian.

$i$ the worst possible outcome $\underline{x}_i$), or does not change the outcome if some $j$, $j < i$, does not obey the recommendation, i.e. $E_j \neq E_j^r$.

Next, let us show that agent $i$ of type $t_i$ cannot gain by selecting $c_i(\tilde{t}_i) \neq c_i(t_i)$ from the menu when every agent $j \neq i$ of type $t_j$ chooses element $c_j(t_j)$ from her menu and obeys the mechanism's recommendation (in this case, by construction, $E_j^r \in \mathcal{E}_j(t_j)$ for every $j \neq i$).

By construction of mechanism $R$, the expected payoff which type $t_i$ obtains after choosing $c_i(\tilde{t}_i)$ in $R$ does not exceed the expected payoff that she would obtain in mechanism $\Gamma$ when other agents use $\sigma_{-i}^*(t_{-i})$ and agent $i$ employs a strategy $\hat{\sigma}_i^D$ which: (i) coincides with $\sigma_i^*(\tilde{t}_i)$ at information sets where this is feasible, (ii) at other information sets (those where with positive probability $\sigma_i^*(\tilde{t}_i)$ prescribes that $i$ send collections of messages that do not belong to $\mathcal{E}_i(t_i)$) prescribes that $i$ send collections of messages from $\mathcal{E}_i(t_i)$ with the same probability as prescribed by $\sigma_i^*(\tilde{t}_i)$, and send arbitrary collections of messages with complementary probability.

Indeed, in mechanism $R$ any realization $E_i$ of $\mu_i(\tilde{t}_i, t_{-i})$ that belongs to $\mathcal{E}_i(t_i)$ occurs with the same probability as on paths of the game $\Gamma$ along which agent $i$ sends the collection of verifying messages $E_i$ when agents use the strategy profile $(\sigma_i^*(\tilde{t}_i), \sigma_{-i}^*(t_{-i}))$. Conditional on these events in the respective games, the outcomes of $R$ and $\Gamma$ coincide and are distributed according to $g(\cdot \mid E, (\tilde{t}_i, t_{-i}))$.

Similarly, in mechanism $R$ realizations $E_i$ of $\mu_i(\tilde{t}_i, t_{-i})$ that *do not* belong to $\mathcal{E}_i(t_i)$ occur with the same probability as on paths of the game $\Gamma$ in which agent $i$ sends the collection of verifying messages $E_i$ when agents use the strategy profile $(\sigma_i^*(\tilde{t}_i), \sigma_{-i}^*(t_{-i}))$. Conditional on these events in the respective games, the outcome of $R$ (given by $\underline{x}_i$) yields agent $i$ a payoff no higher than in the corresponding outcome of $\Gamma$. Hence, after choosing element $c_i(\tilde{t}_i)$ from the menu of mechanism $R$, agent $i$ of type $t_i$ gets an expected payoff that does not exceed her expected payoff in mechanism $\Gamma$ when she uses strategy $\hat{\sigma}_i^D$.

But by choosing menu element $c_i(t_i)$ in mechanism $R$ agent type $t_i$ gets an expected payoff that coincides with her expected payoff in $\Gamma$ when the strategy profile $\sigma^*(t)$ is followed. Since $\sigma^*(t)$ is an equilibrium of $\Gamma$, a deviation from $\sigma_i^*(t_i)$ to $\hat{\sigma}_i^D$ is not profitable for agent $i$ in $\Gamma$. Consequently, the deviation to choosing menu element $c_i(\tilde{t}_i)$, $\tilde{t}_i \neq t_i$, in mechanism $R$ is suboptimal. Hence, in mechanism $R$ it is optimal for type $t_i$ to choose menu element $c_i(t_i)$ and follow the principal's recommendation. □

**Proof of Theorem 4.** To prove the theorem, consider the following two-agent example. Agent 2 has two possible types: $t_2^1 = (\theta_2^1, \tilde{m}_1)$ and $t_2^2 = (\theta_2^2, \tilde{m}_2)$. Agent 1 has four possible types: $t_1^1 = (\theta_1^1, \{\{m_1, m_2\}, \{m_1, m_3\}\})$, $t_1^2 = (\theta_1^2, \{m_1, m_2\})$, $t_1^3 = (\theta_1^3, \{m_1, m_3\})$ and $t_1^4 = (\theta_1^4, \{m_2, m_3, m_4\})$.[21]

The type combinations $(t_1^2, t_2^2)$ and $(t_1^3, t_2^1)$ are infeasible, i.e. $F(t_1^2, t_2^2) = F(t_1^3, t_2^1) = 0$. Additionally, let $F(t_1^1|t_2^1) = F(t_1^1|t_2^2) = p_1^1 > 0$, $F(t_1^4|t_2^1) = F(t_1^4|t_2^2) = p_1^4 > 0$, and $F(t_1^2|t_2^1) = F(t_1^3|t_2^2) = \hat{p}_1 = 1 - p_1^1 - p_1^4 > 0$. Finally, letting $F_2(t_2^1) = p_2^1$ and $F_2(t_2^2) = p_2^2$, we assume that $0.4 \leqslant p_2^1 < 1/2$.

Let $X = \{x_0, x_{11}, x_{12}, x_2, x_3, x_4\}$. The social choice function $f$ we wish to implement is $f(t_1^1, t_2^1) = x_{11}$, $f(t_1^1, t_2^2) = x_{12}$, $f(t_1^2, t_2^1) = x_2$, $f(t_1^3, t_2^2) = x_3$, and $f(t_1^4, t_2^1) = f(t_1^4, t_2^2) = x_4$. The utility function $u_1(x, t_1, t_2)$ of agent 1 is given by:

[21] To achieve some economy of notation in the description of agent 1's types, we list only the largest collections of verifying messages she can send; we do allow her to send any subset of those.

| | $x_{11}$ | $x_{12}$ | $x_2$ | $x_3$ | $x_4$ | $x_0$ |
|---|---|---|---|---|---|---|
| $t_1^1$ | 1 | 2 | 0.9 | 0.9 | 10 | 1.5 |
| $t_1^2, t_1^3$ | 10 | 10 | 2 | 2 | 10 | 0 |
| $t_1^4$ | 200 | 200 | 200 | 200 | 2 | 1 |

Note that this preference structure violates Assumption 1. The preferences of agent 2 are given by $u_2(x_0, t_2^i, t_1^j) = -K$ for some large $K > 0$, and $u_2(x, t_2^i, t_1^j) = 0$ for all $i, j$ and $x \in X \setminus x_0$.[22]

The incentive problem can intuitively be described as follows. Agent 2 can be easily induced to submit her verifying message, and thereby reveal her type, if the mechanism implements $x_0$ whenever agent 2 sends nothing. For agent 1, type $t_1^4$ prefers the allocation designed for any other type of agent 1 to $x_4$. In order to prevent $t_1^4$ from having access to these other allocations, types $t_1^j$, $j \in \{1, 2, 3\}$, must present evidence unavailable to $t_1^4$, i.e. send $m_1$. Similarly, type $t_1^j$, $j \in \{1, 2, 3\}$, prefers $x_4$ to her own allocation, so type $t_1^4$ must present $m_4$. Finally, since types $t_1^2$ and $t_1^3$ prefer the allocations designed for type $t_1^1$ to their own allocations, type $t_1^1$ must distinguish itself from those types. This can be accomplished with the help of agent 2: when agent 2 sends $\tilde{m}_2$, then type $t_1^1$ must send $m_2$ (which is unavailable to $t_1^3$); when agent 2 sends $\tilde{m}_1$, then type $t_1^1$ must send $m_3$ (which is unavailable to $t_1^2$).

Consider the following extensive form mechanism:

*Stage 1.* Agent 1 is given an opportunity to send verifying messages.
If she sends message $m_4$, then implement $x_4$. If she sends message $m_1$, then go to stage 2. Otherwise implement $x_0$.

*Stage 2.* Agent 2 is given an opportunity to send verifying messages.
If she sends $\tilde{m}_1$ or $\tilde{m}_2$, then go to stage 3. Otherwise implement $x_0$.

*Stage 3.* Agent 1 is given a second opportunity to send verifying messages.
(i) If agent 1 sends no message, then implement $x_2$.
(ii) If agent 2 has sent $\tilde{m}_1$ in stage 2, and agent 1 sends $m_2$ in stage 3, then implement $x_2$.
(iii) If agent 2 has sent $\tilde{m}_1$ in stage 2, and agent 1 sends $m_3$ in stage 3, then implement $x_{11}$.
(iv) If agent 2 has sent $\tilde{m}_2$ in stage 2, and agent 1 sends $m_2$ in stage 3, then implement $x_{12}$
(v) If agent 2 has sent $\tilde{m}_2$ in stage 2, and agent 1 sends $m_3$ in stage 3, then implement $x_3$.

It is straightforward to verify that the following strategies form a perfect Bayesian equilibrium that implements $f(\cdot)$. Agent 2 of either type sends her available message. Agent 1 of type $t_1^4$ sends message $m_4$ in stage 1, and agent 1 of type $t_1^2$ ($t_1^3$) sends message $m_1$ in the first stage and her other message in stage 3. Agent type $t_1^1$ sends message $m_1$ in the first stage and, in the third stage sends message $m_2$ ($m_3$) if agent 2 has sent $\tilde{m}_2$ ($\tilde{m}_1$) in the second stage. □

*Impossibility of implementing $f(\cdot)$ in the Revelation Mechanism*

Suppose to the contrary that there exists a Revelation Mechanism $R = (\mu, g, g^p)$ that implements $f(\cdot)$ in Bayesian equilibrium.

For brevity, we will refer to agents' choices in stage 1 of this mechanism as type announcements. Recall that such cheap-talk announcements are equivalent to choices from menus. Also, we need to introduce some additional notation. Let $g^{*e}(\cdot \mid E_1, t_1, \hat{t}_1)$ be the expected probability

[22] Slightly modifying the preferences of agent 2 would ensure that s.c.f. $f$ is ex post efficient, and is still implementable via our mechanism.

distribution over outcomes assigned when all of the following happen: (i) agent 1 has type $t_1$ and announces type $\hat{t}_1$, (ii) agent 2 announces her type truthfully, whatever it is, and sends the requested verifying message, (iii) in stage 3 agent 1 is requested to send the set of verifying messages $E_1$ and does so. Note that the expectation is taken with respect to type of agent 2, given type $t_1$ of agent 1.

Further, let $g^p(\cdot \mid t_1, \hat{t}_1, E_1, \hat{E}_1)$ be expected probability distribution over outcomes assigned when all of the following happen: (i) agent 1 has type $t_1$ and announces type $\hat{t}_1$ in stage 1, (ii) agent 2 announces her type truthfully in stage 1 and sends the requested verifying message; (iii) in stage 3 agent 1 is requested to send the set of verifying message $E_1$ and sends $\hat{E}_1$.

The revelation mechanism has to satisfy the following conditions:

(A) If agent 1 announces $t_1^1$ and agent 2 announces $t_2^2$ in stage 1, then agent 1 must be asked to produce a report including $m_2$ with probability of at least 4/5, i.e.

$$\sum_{E \in \{(m_1,m_2),m_2\}} \mu\big(E \mid t_1^1, t_2^2\big) \geqslant \frac{4}{5}.$$

Otherwise agent $t_1^3$ will announce $t_1^1$ and get $x_{12}$ with a probability that exceeds 1/5—giving $t_1^3$ an expected payoff that would exceed her payoff from her equilibrium allocation $x_3$.

(B) If in the first stage agent 1 announces $t_1^1$ and agent 2 announces $t_2^1$, then agent 1 must be asked to produce a report including $m_3$ with probability of at least 4/5, i.e.

$$\sum_{E \in \{(m_1,m_3),m_3\}} \mu\big(E \mid t_1^1, t_2^1\big) \geqslant \frac{4}{5}.$$

Otherwise agent $t_1^2$ will announce $t_1^1$ and get $x_{11}$ with a probability that exceeds 1/5—giving $t_1^2$ an expected payoff exceeding her payoff from her equilibrium allocation $x_2$.

(C) If agent 1 announces type $t_1^1$ and her true type is $t_1^4$, then she must be asked to produce a report including $m_1$ with probability of at least 0.99, i.e.

$$\sum_{i \in 1,2} \sum_{E \in \{(m_1,m_3),(m_1,m_2),m_1\}} \mu\big(E \mid t_1^1, t_2^i\big) F\big(t_2^i | t_1^4\big) \geqslant 0.99.$$

Otherwise $t_1^4$ will announce $t_1^1$, since $t_1^4$ will then be able to get a payoff of 200 with probability of at least 0.01. This will give her an expected payoff exceeding her payoff from her equilibrium allocation $x_4$.

(A), (B) and (C) imply that:

(D) $\mu((m_1, m_3) \mid t_1^1, t_2^1) \geqslant 0.7$. That is, if agent 1 announces $t_1^1$ and agent 2 announces $t_2^1$, then agent 1 is asked to report $(m_1, m_3)$ with probability of at least 0.7. For suppose not. Then by $B$ agent 1 is asked to produce a report consisting of only $m_3$ with probability of at least 0.1. Then since $p_2^1 > 0.4$, (C) cannot hold.

(E) $\mu_1((m_1, m_2)] \mid t_1^1, t_2^2) \geqslant 0.7$. That is, if agent 1 announces $t_1^1$ and agent 2 announces $t_2^2$, then agent 1 is asked to report $(m_1, m_2)$ with probability of at least 0.7. The proof is similar to that in (D).

(F) Let us compute the following conditional probabilities for $k \in \{1, 4\}$ which will be useful below: $F(t_2^1 \mid t_1^k) = \frac{F(t_2^1, t_1^k)}{F_1(t_1^k)} = \frac{F(t_1^k | t_2^1) F(t_2^1)}{F_1(t_1^k)} = F(t_2^1) = p_2^1$. All equalities here are by definition, except the second one which holds because $F(t_1^k | t_2^1) = F_1(t_1^k)$ for $k \in \{1, 4\}$.

The following steps establish the impossibility of implementing $f(\cdot)$ via a Revelation Mechanism:

(i) *The probability* $Pr^r((m_1, m_3) \mid t_1^k, t_1^1)$ *that agent* 1 *of type* $t_1^k$ $(k \in \{1, 4\})$ *reporting type* $t_1^1$ *is asked to send* $(m_1, m_3)$ *exceeds* $0.7 p_2^1$. This follows from $F$ and $D$.

(ii) $\sum_{x \in \{x_{11}, x_{12}, x_2, x_3\}} g^p(x \mid t_1^4, t_1^1, (m_1, m_3), \phi) \leqslant \frac{2}{0.7 p_2^1 200}$.

Suppose that agent type $t_1^4$ announces $t_1^1$ and sends no verifying messages (i.e. sends $\phi$). By (i) she is asked to report $(m_1, m_3)$ with probability exceeding $0.7 p_2^1$. Then her expected payoff from this deviation is at least $\sum_{x \in \{x_{11}, x_{12}, x_2, x_3\}} g^p(x \mid t_1^4, t_1^1, (m_1, m_3), \phi) \times 0.7 p_2^1 \times 200$. This cannot exceed the payoff 2 that type $t_1^4$ gets from her equilibrium allocation $x_4$.

(iii) $g^p(x \mid t_1^4, t_1^1, (m_1, m_3), \phi) = g^p(x \mid t_1^1, t_1^1, (m_1, m_3), \phi)$. This is so because the distribution of types of agent 2 is the same when the type of agent 1 is either $t_1^1$ or $t_1^4$.

(iv) $\sum_{x \in \{x_0, x_4\}} g^p(x \mid t_1^1, t_1^1, (m_1, m_3), \phi) \geqslant 1 - \frac{2}{0.7 p_2^1 \times 200} = 1 - \frac{1}{70 p_2^1}$. This follows from (ii) and (iii).

(v) Using (iv) we conclude that the expected payoff of type $t_1^1$ who announces her type truthfully, is requested to report $(m_1, m_3)$ but sends no report is at least $u_1(x_0, t_1^1, t_2^k)(1 - \frac{1}{70 p_2^1}) = 1.5(1 - \frac{1}{70 p_2^1}) \geqslant 1.5(1 - \frac{1}{70 \times 0.4}) = \frac{1.5 \times 27}{28} > 1.4$.

(vi) $g^{*e}(x_{11} \mid (m_1, m_3), t_1^1, t_1^1) > 0.6$.

Since by assumption our mechanism implements $f(\cdot)$, $g^{*e}(x_{11} \mid (m_1, m_3), t_1^1, t_1^1)$ must be equal to the probability that agent 2 has type $t_2^1$, given that both agents announce their types truthfully in stage 1, agent 1 has type $t_1^1$, and the mechanism requests that agent 1 send $(m_1, m_3)$ in stage 3. Therefore, by Bayes' rule, we have:

$$g^{*e}\big(x_{11} \mid (m_1, m_3), t_1^1, t_1^1\big) = \frac{\mu((m_1, m_3) \mid t_1^1, t_2^1) p_2^1}{\mu((m_1, m_3) \mid t_1^1, t_2^1) p_2^1 + \mu((m_1, m_3) \mid t_1^1, t_2^2) p_2^2}.$$

Since $\mu((m_1, m_3) \mid t_1^1, t_2^1) \geqslant 0.7$ by (D), $\mu((m_1, m_3) \mid t_1^1, t_2^2) \leqslant 0.3$ by (E), and $0.4 \leqslant p_2^1 < p_2^2 \leqslant 0.6$, it follows that $g^{*e}(x_{11} | (m_1, m_3), t_1^1, t_1^1) > 0.6$.

(vii) Part (vi) implies that the expected payoff of type $t_1^1$ when she announces her type truthfully, is requested to send $(m_1, m_3)$ and sends this report does not exceed $1 \times g^{*e}(x_{11} \mid (m_1, m_3), t_1^1, t_1^1) + 2(1 - g^{*e}(x_{11} \mid (m_1, m_3), t_1^1, t_1^1)) \leqslant 1 \times 0.6 + 2 \times 0.4 = 1.4$.

(viii) In combination, (v) and (vii) imply that when type $t_1^1$ announces her type truthfully and is requested to report $(m_1, m_3)$ (which happens at least with probability 0.7), she prefers to deviate and report nothing rather than $(m_1, m_3)$, because in the latter case she is assigned either $x_0$ or $x_4$ with a probability exceeding $\frac{27}{28}$ and gets an expected payoff exceeding 1.4.

So, the social choice function $f(\cdot)$ is not implementable.

## References

Alger, I., Ma, C.A., 2003. Moral hazard, insurance and some collusion. J. Econ. Behav. Organ. 50, 225–247.

Alger, I., Renault, R., 2006. Screening ethics when honest agents care about fairness. Int. Econ. Rev. 46, 59–85.

Bull, J., Watson, J., 2004a. Evidence disclosure and verifiability. J. Econ. Theory 118, 1–31.

Bull, J., Watson, J., 2004b. Hard evidence and mechanism design. Mimeo. University of California, San Diego.

Che, Y.-K., Gale, I., 2000. The optimal mechanism for selling to a budget constrained buyer. J. Econ. Theory 92, 198–233.

Chen, Y., 2000. Promises, trust, and contracts. J. Law, Econ., Organ. 16, 209–231.

Crocker, K., Morgan, J., 1998. Is honesty the best policy? Curtailing insurance fraud through optimal incentive contracts. J. Polit. Economy 106, 355–375.

Deneckere, R., De Palma, A., 1995. The market for audit services and mandatory rotation. Working paper. H.E.C., University of Geneva.

Deneckere, R., Severinov, S., 2003. Optimal screening with costly misrepresentation. Mimeo, Fuqua School of Business.

Ekman, P., 1973. Darwin and Facial Expression: A Century of Research in Review. Academic Press, New York.

Erard, J., Feinstein, B., 1994. Honesty and evasion in the tax compliance game. RAND J. Econ. 25 (1), 447–456.

Forges, F., Koessler, F., 2005. Communication equilibria with partially verifiable types. J. Math. Econ. 41, 793–811.

Glazer, J., Rubinstein, A., 2001. Debates and decisions: On a rationale of argumentation rules. Games Econ. Behav. 36, 158–173.

Glazer, J., Rubinstein, A., 2004. On optimal rules of persuasion. Econometrica 72, 1715–1736.

Gneezy, U., 2002. Deception: The role of consequences. Mimeo. Graduate School of Business, University of Chicago, pp. 1–41.

Green, J., Laffont, J.-J., 1986. Partially verifiable information and mechanism design. Rev. Econ. Stud. 53, 447–456.

Grossman, S., 1981. The informational role of warranties and private disclosure about product quality. J. Law Econ. 24, 461–483.

Kartik, N., 2004. Information transmission with almost-cheap talk. Mimeo, UC San Diego.

Kartik, N., Ottaviani, M., Squintani, F., 2007. Credulity, lies and costly talk. J. Econ. Theory 134, 93–116.

Lacker, J.M., Weinberg, J.A., 1989. Optimal contracts under costly state falsification. J. Polit. Economy 97 (6), 1345–1363.

Lipman, B., Seppi, D., 1995. Robust inference in communication games with partial provability. J. Econ. Theory 66 (2), 370–405.

Maggi, G., Rodriguez-Clare, A., 1995. Costly distortion of information in agency problems. RAND J. Econ. 26, 675–689.

Maskin, E., 1999. Nash equilibrium and welfare optimality. Rev. Econ. Stud. 66, 23–38.

Milgrom, P., 1981. Good news and bad news: Representation theorems and applications. Bell J. Econ. 12, 380–391.

Postlewaite, A., Schmeidler, D., 1986. Implementation in differential information economies. J. Econ. Theory 39, 14–33.

Sanchirico, C., 2001. Relying on the information of interested—And potentially dishonest—Parties. Amer. Law Econ. Rev. 3, 320–335.

Sanchirico, C., Triantis, G., 2004. Evidence arbitrage: The fabrication of evidence and the verifiability of contract performance. Mimeo, University of Pennsylvania.

Severinov, S., Deneckere, R., 2006. Screening when not all agents are strategic: Does a monopolist need to exclude? RAND J. Econ. 37 (4), 816–841.

Squintani, F., 2004. Contracts, liability restrictions and costly verification. Mimeo, UCL.

Tennyson, S., 1997. Economic institutions and individual ethics: Consumer attitudes towards insurance fraud. J. Econ. Behav. Organ. 32, 247–265.