Direct Implementation with Minimally Honest Individuals

Juan Ortner*
Boston University

January 13, 2015

Abstract

I consider a standard implementation problem under complete information when agents have a minimal degree of honesty. In particular, I assume that agents are *white lie averse*: they strictly prefer to tell the truth whenever lying has no effect on their material payoff. I show that if there are at least five agents who are all white lie averse and if I impose either of two refinements of Nash equilibrium, then a simple direct mechanism fully implements any social choice function.

JEL Classification codes: C72, C73, D71, D78.

Keywords: implementation, mechanism design, white lie aversion, direct mechanisms.

*I am indebted to Faruk Gul and Satoru Takahashi for their advice and guidance, and to Avidit Acharya, Ben Brooks, Edoardo Grillo, Navin Kartik and Stephen Morris for useful comments. I also thank the editor and an anonymous referee for very thoughtful suggestions. All remaining errors are my own. Address: Boston University, Department of Economics, 270 Bay State Road, Boston, MA 02215. E-mail: jortner@bu.edu.

1 Introduction

Consider the problem of a planner who wishes to implement the alternative prescribed by a social choice function $f : \Theta \to A$, where Θ is the set of possible states of nature and A is the set of possible alternatives. A state determines the preferences of the agents over the elements of A. The social choice function assigns an alternative to each possible state. The state is common knowledge among the agents but is unknown to the planner. The problem of the planner is to design a mechanism to implement the social choice function.

The most natural class of mechanisms discussed in the literature are direct mechanisms, under which the planner asks each agent to announce the state of nature. If there are at least three agents one can easily construct direct mechanisms such that truth-telling is a Nash equilibrium. However, direct mechanisms will in general yield other non-truthful equilibria. If the planner has no control over which equilibrium obtains, she cannot rely on direct mechanisms to implement a given social choice function.

The implementation literature addresses the issue of multiple equilibria by seeking more complicated mechanisms with richer message spaces. In other words, the literature focuses on mechanisms that require players to make additional announcements besides the information that is directly relevant to the environment. The most notable example of this augmentation of the message space is Maskin's (1999) integer game. Despite the success that the theory has had in characterizing implementable social choice functions, the complex message spaces and game forms required to achieve full implementation have been criticized in the literature for their implausibility. Some researchers have also expressed concerns about the appropriateness of Nash equilibrium as a solution concept for the games that these mechanisms induce.[1]

In this paper I assume that agents are *white lie averse*: they strictly prefer to tell the truth whenever lying has no effect on the implemented alternative. I show that if there are at least five agents who are white lie averse and if I impose either of two refinements of Nash equilibrium, then a simple direct mechanism fully implements any social choice function. Therefore, under these conditions a planner can achieve full implementation using a simple direct mechanism, and can thus dispense with any augmentation of the message space.

The first refinement I consider is *fault tolerant Nash equilibrium*, introduced by Eliaz (2002). The idea behind fault tolerant Nash equilibrium is that players may not know whether all of their opponents are rational. Suppose that players believe that there are at most k irrational agents in the population, but that they know neither the identity of the irrational players, nor how irrational players behave. A *k-fault tolerant Nash equilibrium*

[1] See Jackson (1992) for an elaboration of this and related points.

(k-FTNE) is a strategy profile that is robust to the presence of k irrational agents: under a k-FTNE each agent has an incentive to play her equilibrium action, regardless of the identity and actions of irrational players, as long as $n-k-1$ of her opponents adhere to equilibrium behavior.

The second refinement I consider is *stochastically stable equilibrium*, introduced by Kandori, Mailath and Rob (1993) and Young (1993). This equilibrium concept was proposed as a way of studying which outcomes are more likely to arise in the long-run. Suppose that a group of agents plays a strategic game infinitely many times. Assume also that players follow a myopic behavioral rule whenever they have an opportunity to revise their strategies, and that they occasionally make mistakes. The stochastically stable equilibria are those strategy profiles at which players will coordinate their actions most of the time in the long run when the probability with which they make mistakes is low.

The direct mechanism I use is a majoritarian aggregation rule. The planner asks each agent to announce the state of nature. If more than half the population announces the same state θ, the mechanism's outcome is $f(\theta)$. In any other case, the outcome is some fixed alternative a^*. Importantly, a^* need not be a particularly bad outcome, and it may even be a Pareto optimal alternative. Besides its simplicity, from a practical point of view this mechanism has the attractive feature of being anonymous (i.e., the outcome is unchanged if agents are permuted) and completely independent of the preferences of the agents. The strategic game that this mechanism induces may have multiple equilibria. However, if there are at least five players and they are all white lie averse, then both fault tolerance and stochastic stability yield the same unique prediction: all agents make truthful announcements and the planner is able to implement the desired alternative.[2]

Other papers have studied implementation under the refinements that I consider. Eliaz (2002) studies complete information implementation in k-FTNE. Adapting Maskin's (1999) canonical mechanism, he shows that any social choice correspondence satisfying k-*monotonicity* and no veto power can be implemented in k-FTNE. In contrast, the current paper shows that any social choice function can be implemented in k-FTNE with a simple direct mechanism when players have a small preference for honesty. Sandholm (2007) studies implementation in stochastically stable equilibrium in an environment with externalities in which a planner wants the agents to choose an utilitarian action profile. Sandholm (2007) shows that the planner can achieve this objective by introducing a simple tax scheme under which each

[2]In Appendix A.4 I give an example of a game in which fault tolerance and stochastic stability yield different predictions. Therefore, these solution concepts are logically independent, and neither of them implies the other.

agent pays for the externalities she creates. In contrast to Sandholm (2007), the current paper studies a general implementation problem in which transfers among individuals may not be possible.[3] Neither Eliaz (2002) nor Sandholm (2007) study implementation with white lie averse agents.

The results in this paper provide two distinct justifications for the use of simple direct mechanisms, each based on a different equilibrium concept. Suppose first that agents believe that a fraction of their opponents may fail to behave optimally, but they know neither the identity of irrational players, nor how irrational players behave. Rational players will likely coordinate their actions at a fault tolerant Nash equilibrium in such an environment, as this is a strategy profile that is robust to the presence of irrational agents. The results in this paper then show that a social planner can use a majoritarian direct mechanism to implement the desired alternative, provided there are at least five agents and they are all white lie averse.

The refinement of stochastic stability provides an evolutionary justification for simple majoritarian mechanisms. Suppose a group of agents will repeatedly play the strategic game induced by the mechanism that the planner puts in place. Kandori, Mailath and Rob (1993) and Young (1993) introduced the notion of stochastic stability to predict long run behavior in such an environment. If there are at least five agents and they all have a minimal degree of honesty, then the results in this paper tell us that a social planner can use a majoritarian direct mechanism to achieve full implementation in the long run.

The idea of studying implementation when agents have a minimal degree of honesty is due to Matsushima (2008a), who considers the problem of implementing a social choice function in a complete information setup with three white lie averse agents. He shows that, in this environment, any social choice function can be exactly implemented in iteratively undominated strategies with a mechanism similar to the one in Abreu and Matsushima (1992). Matsushima (2008b) considers implementation in Bayesian environments when agents have an intrinsic preference for being honest and shows that any incentive compatible social choice function can be fully implemented. Dutta and Sen (2012) study Nash implementation in settings in which individuals may be partially honest. They show that any social choice correspondence satisfying no veto power can be implemented in Nash equilibrium if there is at least one partially honest individual in the population.

Kartik and Tercieux (2012) study implementation in Nash equilibrium when agents have

[3]Cabrales and Serrano (2011b) also study implementation in stochastically stable equilibrium. Focusing on economic environments, they find sufficient conditions for implementation in stochastically stable equilibrium of strongly Pareto efficient social choice functions. See also the results in Cabrales and Serrano (2011a).

to provide evidence along with their messages. They show that, in settings in which players can fabricate evidence at a cost, any social choice correspondence satisfying *cost-monotonicity* (a condition weaker than Maskin-monotonicity) and no veto power is implementable in Nash equilibrium. The model in which individuals have to pay a small cost to produce false evidence is equivalent to one in which individuals have a small preference for honesty. The authors show that under this cost structure any social choice correspondence satisfies cost-monotonicity. Thus, in this case any social choice correspondence satisfying no veto power is implementable in Nash equilibrium.

The main message of these papers is that the set of implementable social choice functions becomes significantly larger when agents are minimally honest. These papers derive permissive results using complex augmented mechanisms. For instance, Dutta and Sen (2012) and Kartik and Tercieux (2012) derive their results using augmented mechanisms which are modifications of Maskin's canonical mechanism. The contribution of the current paper is to show that, under either of two refinements, these permissive results continue to hold even when the planner uses simple direct mechanisms. Moreover, the current paper also shows by example that there are social choice functions satisfying no veto power that cannot be implemented in Nash equilibrium with a direct mechanism when players are white lie averse. This shows that the use of refinements is necessary for the results in the current paper, and also clarifies that augmented mechanisms are needed for the permissive results in Dutta and Sen (2012) and Kartik and Tercieux (2012).

Dutta and Sen (2012) also show that, if the environment is *separable*, the planner can implement any social choice function in Nash equilibrium with a direct mechanism when all agents are minimally honest. Holden, Kartik and Tercieux (2014) show that a planner can implement any social choice function in two rounds of iterated deletion of strictly dominated strategies using a simple mechanism, provided that agents have a preference for honesty and the environment satisfies a *separable punishment* condition. In contrast to Dutta and Sen (2012) and Holden, Kartik and Tercieux (2014), the results in the current paper don't require restrictions on the environment. On the other hand, the results in the current paper demand stronger solution concepts.

Finally, there are other papers showing that the implementation problem becomes easier when considering refinements of Nash equilibrium; e.g., Moore and Repullo (1988) and Palfrey and Srivastava (1991). The permissive results in this literature are obtained using augmented

mechanisms.[4] [5] This paper adds to this strand of literature by showing that two particular refinements of Nash equilibrium allow a planner to implement any social choice function with a simple direct mechanism when agents have a minimum degree of honesty.

2 Model

Let $N = \{1, .., n\}$ be a finite set of agents and let A be a set of possible alternatives. Let Θ be a finite set of possible states of the world. Each state $\theta \in \Theta$ specifies the preferences of the agents over the elements in A. The realization of the state is common knowledge among the agents, but is unknown to the planner.

The planner's objective is to implement a social choice function $f : \Theta \to A$. To achieve this objective the planner designs a strategic game form $\langle N, (S_i)_{i \in N}, g \rangle$, where S_i is player i's action set and $g : \prod_{i=1}^{n} S_i \to A$ is a mechanism. I restrict the planner to use mechanisms g with $S_i = \Theta \times M$ for all $i \in N$, where M is a (possibly empty) set of messages. That is, the planner can only use mechanisms under which each player announces a state of nature plus (possibly) some other message $m \in M$. The restriction to this class of mechanisms allows me to interpret a message $s = (\theta, m) \in S_i$ as being truthful if θ is the true state of nature. Note that most of the complete information implementation literature uses mechanisms that fall into this category. A mechanism g is a *direct mechanism* if $S_i = \Theta$ for all $i \in N$ (i.e., if $M = \emptyset$).

Let $g : S \to A$ be a mechanism, where $S = (\Theta \times M)^n$ is the set of possible announcement profiles. For each $i \in N$, let $u_i : A \times \Theta \times S \to \mathbb{R}$ be agent i's utility function. Note that the agents' utility depends on the implemented alternative, the state of nature *and* the announced messages. For each i, let $\widetilde{u}_i : A \times \Theta \to \mathbb{R}$ and let $\eta > 0$ be a small number. The function $\widetilde{u}_i$ represents the material payoff of player i, while η represents the cost of lying. For any announcement profile $\mathbf{s} \in (\Theta \times M)^n$, let $\mathbf{s}_{-i}$ denote the announcements of all players in $N \setminus \{i\}$. I incorporate white lie aversion as a utility perturbation.

Assumption 1 (white lie aversion) *Let g be a mechanism. For all $i \in N$, if $\mathbf{s}_{-i}$ is such that there exists $s_i' \in \theta \times M$ and $s_i'' \notin \theta \times M$ with the property that $\widetilde{u}_i(g(s_i', \mathbf{s}_{-i}), \theta) =$*

[4] Moreover, as Chung and Ely (2003) and Aghion et al. (2012) show, the implementation results in these papers are not robust to small perturbations in the information structure.

[5] For settings in which transfers among players are possible (or, more generally, if it is possible to punish an agent while simultaneously rewarding the others), Moore and Repullo (1988) construct a simple mechanism that implements any social choice function in Subgame Perfect equilibrium.

$\widetilde{u}_i(g(s_i'', \mathbf{s}_{-i}), \theta) \geq \widetilde{u}_i(g(s_i, \mathbf{s}_{-i}), \theta)\ \forall s_i \in \Theta \times M$, *then*

$$u_i(g(s_i', \mathbf{s}_{-i}), \theta, (s_i', \mathbf{s}_{-i})) = \widetilde{u}_i(g(s_i', \mathbf{s}_{-i}), \theta) > u_i(g(s_i'', \mathbf{s}_{-i}), \theta, (s_i'', \mathbf{s}_{-i})) = \widetilde{u}_i(g(s_i'', \mathbf{s}_{-i}), \theta) - \eta.$$

For any other $\mathbf{s}_{-i}$, $u_i(g(s_i', \mathbf{s}_{-i}), \theta, (s_i', \mathbf{s}_{-i})) = \widetilde{u}_i(g(s_i', \mathbf{s}_{-i}), \theta)\ \ \forall s_i' \in \Theta \times M$.

Assumption 1 states that every agent $i \in N$ strictly prefers to tell the truth than to lie if $\mathbf{s}_{-i}$ is such that she cannot gain by sending a false message. Assumption 1 implies that agents have a minimal degree of honesty, since they dislike telling lies whenever those lies do not benefit them.

I now present the strategic form $\langle N, (S_i)_{i \in N}, g_M \rangle$ I will use throughout the paper. Let $S_i = \Theta$ for all $i \in N$. For every $\theta \in \Theta$ and for every $\mathbf{s} = (s_1, ..., s_n) \in \Theta^n$, let

$$R^*(\theta \mid \mathbf{s}) = \{i \in N \text{ s.t. } s_i = \theta\}$$

be the set of agents who reported state θ in $\mathbf{s}$.

Fix any $a^* \in A$. The following condition characterizes mechanism g_M:

$$g_M(\mathbf{s}) = \begin{cases} f(\theta) & \text{if } |R^*(\theta \mid \mathbf{s})| > \frac{n}{2} \text{ for some } \theta \in \Theta, \\ a^* & \text{otherwise.} \end{cases}$$

Mechanism g_M is a majoritarian mechanism: if strictly more than half the agents announce a state $\theta \in \Theta$, the mechanism implements alternative $f(\theta)$. Otherwise, the mechanism implements alternative a^*, which can be interpreted as the status quo. Note that a^* need not be a particularly bad outcome for the agents; indeed, a^* could even be a Pareto optimal alternative. Moreover, note that besides its simplicity, from a practical point of view mechanism g_M has the attractive property of being anonymous (i.e., the outcome is unchanged if agents are permuted) and completely independent of the agents' preferences.

3 Implementation in fault tolerant Nash equilibrium

In this section I present the solution concept of fault tolerant equilibrium and show that the majoritarian mechanism g_M fully implements any social choice function in fault tolerant equilibrium when agents are white lie averse. For clarity of exposition, this section focuses exclusively on implementation in pure strategies. Appendix A.3 shows that the results in this section continue to hold even if we allow for mixed strategies.

Fault tolerant Nash equilibrium was introduced by Eliaz (2002). The idea behind this equilibrium concept is that players may not know whether all of their opponents are rational. In particular, suppose that each agent believes that at most k of her opponents may fail to act rationally, but knows neither the identity of irrational players, nor how irrational players behave. Standard solution concepts will in general not provide a robust prediction for this environment, as rational players might choose to adjust their actions to take into account the possibility that irrational agents deviate from equilibrium behavior. A fault tolerant Nash equilibrium is a strategy profile that is robust to the presence of irrational agents, and as such it is a good prediction of how rational players will behave in this environment.

Let $\left\langle N, (S_i)_{i\in N}, (u_i)_{i\in N} \right\rangle$ be a strategic game, where $N = \{1, 2, ..., n\}$ is a finite set of players, S_i is the set of actions of player i and $u_i : \Pi_{i\in N} S_i \to \mathbb{R}$ is the utility function of player i. For any pair of strategy profiles $\mathbf{s}, \mathbf{s}' \in \Pi_{i\in N} S_i$, let the distance between $\mathbf{s}$ and $\mathbf{s}'$ be

$$d(\mathbf{s}, \mathbf{s}') = |\{i \in N : s_i \neq s_i'\}|.$$

Definition 1 *A strategy profile $\mathbf{s}^* = (s_1^*, ..., s_n^*) \in S$ is a k-fault tolerant Nash equilibrium of the strategic game $\left\langle N, (S_i)_{i\in N}, (u_i)_{i\in N} \right\rangle$ if, $\forall i \in N$, $u_i(s_i^*, \mathbf{s}_{-i}) \geq u_i(s_i', \mathbf{s}_{-i})$ for all $\mathbf{s}_{-i} \in \{\widetilde{\mathbf{s}}_{-i} \in S_{-i} : d((s_i^*, \widetilde{\mathbf{s}}_{-i}), \mathbf{s}^*) \leq k\}$ and for all $s_i' \in S_i \setminus \{s_i^*\}$.*

Definition 1 coincides with the equilibrium notion of fault tolerant Nash equilibrium introduced by Eliaz (2002). For a strategy profile $\mathbf{s}^*$ to be a k-fault tolerant Nash equilibrium (k-FTNE), each agent must play an optimal strategy against any action profile of her opponents with the property that at least $n - k - 1$ are playing their equilibrium action. Put differently, in a fault tolerant Nash equilibrium each player has an incentive to play her equilibrium action even in the presence of k irrational agents, regardless of the identity of the irrational agents and of the actions that irrational agents take. Note that 0-FTNE coincides with Nash equilibrium. On the other hand, $(n-1)$-FTNE coincides with weakly dominant strategy equilibrium. For $k \in (0, n-1)$, k-FTNE lies between these two solution concepts. Finally, note that k-FTNE becomes more demanding as k increases: for any k, k' with $n - 1 \geq k' > k \geq 0$, the set of k'-FTNE is a (possibly empty) subset of the set of k-FTNE.

Consider next the environment of Section 2 and let $\left\langle N, (S_i)_{i\in N}, g \right\rangle$ be a strategic game form. For any $k \geq 0$, let $E_k(g, \theta)$ denote the set of k-FTNE of the strategic game that mechanism g induces when the state is θ. For any $\mathbf{s} \in S$, let $B(\mathbf{s}, k)$ denote the set of

strategy profiles whose distance from $\mathbf{s}$ is no more than k:

$$B(\mathbf{s},k)=\{\mathbf{s}'\in S:d(\mathbf{s},\mathbf{s}')\leq k\}.$$

Definition 2 *A mechanism g implements the social choice function $f:\Theta\to A$ in k-FTNE if for all $\theta\in\Theta$, $g(B(\mathbf{s},k))=f(\theta)$ for all $\mathbf{s}\in E_k(g,\theta)$.*

By definition 2, a mechanism g implements a social choice function f in k-FTNE if it selects the correct alternative even when irrational players deviate from equilibrium behavior.

For each $\theta\in\Theta$, let $\mathbf{s}^\theta$ denote the strategy profile under which every player announces state θ, i.e., $\mathbf{s}^\theta=(\theta,\theta,...,\theta)$. I now present the main result of this section:

Theorem 1 *Let $f:\Theta\to A$ be a social choice function and suppose that Assumption 1 holds. If $n\geq 5$, then mechanism g_M implements f in 1-FTNE. Moreover, $E_1(g_M,\theta)=\left\{\mathbf{s}^\theta\right\}$ for every $\theta\in\Theta$.*

Proof. Suppose that θ is the true state of nature. If $n\geq 5$, then the announcement profile $\mathbf{s}^\theta$ is a 1-FTNE. Moreover, it is clear that $g_M\left(B\left(\mathbf{s}^\theta,1\right)\right)=f(\theta)$ for every $\theta\in\Theta$. Therefore, to prove Theorem 1 it suffices to show that, for every $\mathbf{s}=(s_1,...,s_n)\neq\mathbf{s}^\theta$, there exists $i\in N$, $\mathbf{s}'_{-i}\in\Theta^{n-1}$ with $d\left(\left(s_i,\mathbf{s}'_{-i}\right),\mathbf{s}\right)\leq 1$ and $\widetilde{s}_i\in\Theta\backslash\{s_i\}$ such that player i strictly prefers to announce $\widetilde{s}_i$ than to announce s_i whenever her opponents announce $\mathbf{s}'_{-i}$.

Consider first announcement profiles $\mathbf{s}\neq\mathbf{s}^\theta$ such that $\left|R^*(\theta\mid\mathbf{s})\right|\geq\frac{n}{2}$. Note that in this case there always exists an agent $i\in N$ with $s_i\neq\theta$ and an announcement profile $\mathbf{s}'_{-i}$ of i's opponents such that $\left|R^*\left(\theta\mid s_i,\mathbf{s}'_{-i}\right)\right|\geq\frac{n}{2}+1$ and $d\left(\left(s_i,\mathbf{s}'_{-i}\right),\mathbf{s}\right)\leq 1$. Note that player i is not pivotal when her opponents announce $\mathbf{s}'_{-i}$. Therefore, by Assumption 1 player i strictly prefers to announce θ than any $\theta'\neq\theta$ when her opponents are playing $\mathbf{s}'_{-i}$, so $\mathbf{s}$ cannot be a 1-FTNE.

Next, consider announcement profiles $\mathbf{s}\neq\mathbf{s}^\theta$ such that there exists $\phi\neq\theta$ with $\left|R^*(\phi\mid\mathbf{s})\right|>\frac{n}{2}$. In this case, there always exists an agent $i\in N$ with $s_i=\phi$ and a strategy profile $\mathbf{s}'_{-i}$ of i's opponents such that $\left|R^*\left(\phi\mid s_i,\mathbf{s}'_{-i}\right)\right|>\frac{n}{2}+1$ and $d\left(\left(s_i,\mathbf{s}'_{-i}\right),\mathbf{s}\right)\leq 1$. Note that player i is not pivotal if her opponents announce $\mathbf{s}'_{-i}$. Thus, by Assumption 1, player i strictly prefers to announce θ than to announce ϕ, so $\mathbf{s}$ cannot be a 1-FTNE.

Consider next message profiles $\mathbf{s}\neq\mathbf{s}^\theta$ such that $\left|R^*(\theta'\mid\mathbf{s})\right|=\left|R^*(\theta''\mid\mathbf{s})\right|=\frac{n}{2}$ for $\theta',\theta''\in\Theta$. I already considered the case with $\left|R^*(\theta\mid\mathbf{s})\right|=\frac{n}{2}$ above. Therefore, I now focus on the case with $\theta',\theta''\neq\theta$. In this case, there exists $i\in N$ with $s_i=\theta'$ and an announcement profile of i's opponents $\mathbf{s}'_{-i}$ with $d\left(\left(s_i,\mathbf{s}'_{-i}\right),\mathbf{s}\right)=1$ such that $\left|R^*\left(\theta''\mid s_i,\mathbf{s}'_{-i}\right)\right|=\frac{n}{2}+1$ and $\left|R^*\left(\theta'\mid s_i,\mathbf{s}'_{-i}\right)\right|=\frac{n}{2}-1$ for $\theta',\theta''\neq\theta$. Note that in $\left(s_i,\mathbf{s}'_{-i}\right)$ player i is announcing the state

of nature that is announced the least. Moreover, player i cannot change the implemented outcome by changing her announcement if her opponents are playing according to $\mathbf{s}'_{-i}$. It follows from Assumption 1 that player i strictly prefers to announce θ than any other state, so $\mathbf{s}$ cannot be a 1-FTNE.

Finally, consider announcement profiles $\mathbf{s}$ such that $|R^*(\theta' \mid \mathbf{s})| \leq \frac{n}{2}$ for all $\theta' \in \Theta$, but with at most one $\theta' \neq \theta$ such that $|R^*(\theta' \mid \mathbf{s})| = \frac{n}{2}$ and with $|R^*(\theta \mid \mathbf{s})| < \frac{n}{2}$. There are two cases to consider: (i) there exists $\phi \in \Theta$ such that $|R^*(\phi \mid \mathbf{s})| + 1 > \frac{n}{2}$, and (ii) no such ϕ exists. In case (ii), no agent can change the implemented outcome by changing her announcement (i.e., the implemented alternative will still be a^*), so under Assumption 1 all agents strictly prefer to announce θ than any $\theta' \in \Theta \backslash \{\theta\}$. Thus, such announcement profiles cannot be a 1-FTNE.

Consider next case (i). Note that at least three states are being announced in $\mathbf{s}$. Define $Y(\mathbf{s}) = \{\phi \in \Theta : |R^*(\phi \mid \mathbf{s})| + 1 > \frac{n}{2}\}$. Since at least three states are being announced in $\mathbf{s}$, there always exists $i \in N$ with $s_i \neq \theta$ such that $Y(\mathbf{s}) \setminus \{s_i\}$ is non-empty (so player i is not announcing the true state, and there exists a state in $Y(\mathbf{s})$ different from the state that player i is announcing). Moreover, there exists an announcement profile $\mathbf{s}'_{-i}$ of i's opponents with $d\left(\left(s_i, \mathbf{s}'_{-i}\right), \mathbf{s}\right) = 1$ such that $\left|R^*\left(\phi \mid s_i, \mathbf{s}'_{-i}\right)\right| > \frac{n}{2}$ for some $\phi \in Y(\mathbf{s}) \setminus \{s_i\}$. Given $\mathbf{s}'_{-i}$, mechanism g_M will implement $f(\phi)$ regardless of player i's announcement, so under white lie aversion player i strictly prefers to announce the true state θ than any $\theta' \in \Theta \backslash \{\theta\}$. Hence, such announcement profiles cannot be a 1-FTNE. ■

Theorem 1 states that mechanism g_M fully implements any social choice function in 1-FTNE: for any $\theta \in \Theta$, the truthful announcement profile $\mathbf{s}^\theta$ is the unique 1-FTNE of the game induced by g_M at state θ. For any $k > 1$, the set of k-FTNE is a subset of the set of 1-FTNE. It then follows from Theorem 1 that, for any $k > 1$, no strategy profile different from the truthful announcement $\mathbf{s}^\theta$ can be k-FTNE of the game that mechanism g_M induces at state θ. Moreover, it is easy to see that when there are at least five agents and they are all white lie averse, $\mathbf{s}^\theta$ is a k-FTNE of the game that mechanism g_M induces at state θ for all $k < \frac{n}{2} - 1$. Finally, note also that $g_M(B(\mathbf{s}^\theta, k)) = f(\theta)$ for all states θ and for all $k < \frac{n}{2} - 1$. Therefore, g_M implements f in k-FTNE for all $k < \frac{n}{2} - 1$ when $n \geq 5$ and Assumption 1 holds. The following corollary summarizes this discussion.[6]

Corollary 1 *Let $f : \Theta \to A$ be a social choice function and suppose that Assumption 1 holds. If $n \geq 5$, then mechanism g_M implements f in k-FTNE for all $k < \frac{n}{2} - 1$. Moreover,*

[6]I state Theorem 1 using 1-FTNE, instead of the more demanding k-FTNE, to conform with the general philosophy in implementation theory of using the minimal refinement possible.

$E_k(g_M, \theta) = \{\mathbf{s}^\theta\}$ *for every* $\theta \in \Theta$.

The intuition behind Theorem 1 and Corollary 1 is as follows. White lie averse agents have a strict incentive to be honest whenever they are not pivotal. This implies that truth-telling is a k-FTNE (with $k < \frac{n}{2} - 1$) under mechanism g_M if there are at least five players and they are all white lie averse, since under this mechanism no player can change the implemented outcome by changing her announcement when all but $k < \frac{n}{2} - 1$ of her opponents announce the true state. Consider next any strategy profile $\mathbf{s} \neq \mathbf{s}^\theta$. For $\mathbf{s}$ to be a k-FTNE all players who don't make truthful announcements under $\mathbf{s}$ must be pivotal, since otherwise they would have a strict incentive to be honest. Under the majority mechanism g_M, if a non-truthful agent i is pivotal under announcement profile $\mathbf{s}$, there is always a deviation by an agent $j \neq i$ that will make player i no longer pivotal. By Assumption 1, player i would strictly prefer to announce the true state if j deviated in such way, so $\mathbf{s}$ cannot be a k-FTNE with $k \geq 1$.

To see why I need $n \geq 5$ in the proof of Theorem 1, suppose $\theta' \in \Theta$ is the true state of nature and assume $N = \{1, 2, 3\}$ (similar examples can be constructed for the case with $n = 4$). Suppose further that there exists $\theta'' \in \Theta \backslash \{\theta'\}$ with $f(\theta'') \neq f(\theta')$ such that $\widetilde{u}_2(f(\theta''), \theta') > \widetilde{u}_2(b, \theta')$ for all $b \in A \backslash \{f(\theta'')\}$. In this example, the truthful announcement $\mathbf{s}^{\theta'}$ is not a 1-FTNE at state θ'. To see this, let $\mathbf{s}' = (\theta', \theta', \theta'')$ and note that $d\left(\mathbf{s}', \mathbf{s}^{\theta'}\right) = 1$. The assumptions on preferences imply that $u_2(g_M(\theta', \theta'', \theta''), \theta', (\theta', \theta'', \theta'')) > u_2(g_M(\mathbf{s}'), \theta', \mathbf{s}')$. Hence, $\mathbf{s}^{\theta'}$ is not a 1-FTNE.

Theorem 1 shows that any social choice function is implementable with a direct mechanism in 1-FTNE whenever there are at least five agents and they are all white lie averse. I now argue that neither 1-FTNE nor white lie aversion by themselves are sufficient to obtain such a permissive result. Therefore, the combination of these two is needed for Theorem 1.

Consider first the problem of implementation in k-FTNE, without assuming that agents are white lie averse. This problem was studied by Eliaz (2002), who showed that any social choice function that is implementable in k-FTNE must be *k-monotonic*.[7] Since 1-monotonicity is a stronger condition than Maskin-monotonicity, it follows that there is a wide range of social choice functions that cannot be implemented in 1-FTNE when players are not white lie averse.

Consider next the problem of implementation when all agents are white lie averse, but using Nash equilibrium as a solution concept instead of the more demanding 1-FTNE. The following example shows that, in such a setting, it is not possible to obtain a permissive

[7] A social choice function f is k-monotonic if whenever $f(\theta') = a$ and $f(\theta'') \neq a$, there exists $M \subset N$ and $b \in A$ such that $|M| \geq k+1$, every $i \in M$ satisfies $u_i(a, \theta') \geq u_i(b, \theta')$, and at least one player $j \in M$ satisfies $u_j(b, \theta'') > u_j(a, \theta'')$.

result as in Theorem 1.

Example 1 Let $f : \Theta \to A$ be a social choice function with $f(\theta') = a$ and $f(\theta'') = b \neq a$ for some $\theta', \theta'' \in \Theta$. Let $N = \{1, ..., n\}$ be the set of agents in the population. Assume that all players in N are white lie averse. The agents' material preferences are such that $\widetilde{u}_i(a, \theta'') > \widetilde{u}_i(c, \theta'')$ for all $c \neq a$ and all $i \in N$; that is, at state θ'' all players strictly prefer alternative a to any other alternative. Appendix A.1 shows that, in this setting, the social choice function f cannot be implemented in Nash equilibrium even when all agents are white lie averse (regardless of the number of agents in the population).

Example 1 shows that white lie aversion by itself is not sufficient for the permissive results in Theorem 1: in this setting, there exists no mechanism (either direct or augmented) that implements social choice function f in Nash equilibrium when players are white lie averse.

A weakness of Example 1 is that the social choice function f violates no veto power.[8] The following example presents a social choice function satisfying no veto power that cannot be implemented in Nash equilibrium *with a direct mechanism* when players are white lie averse.

Example 2 Consider a setting with $N = \{1, 2, 3, 4, 5\}$, $A = \{a, b, c, d\}$ and $\Theta = \{\theta', \theta''\}$. The social choice function that the planner wants to implement is $f(\theta') = a$ and $f(\theta'') = d$.

The table below shows the agents' material preferences at each state in Θ. For instance, at state θ' agent 1 is materially indifferent between the four alternatives in A, while at state θ'' her material preferences are such that $\widetilde{u}_1(a, \theta'') > \widetilde{u}_1(b, \theta'') > \widetilde{u}_1(c, \theta'') > \widetilde{u}_1(d, \theta'')$.

	1	2	3	4	5
θ'	a b c d	a b c d	b	c	d
			c	d	b
			d	b	c
			a	a	a
θ''	a	b	c	a b c d	a b c d
	b	c	a		
	c	a	b		
	d	d	d		

Appendix A.1 shows that, in this setting, there is no direct mechanism that implements the social choice function f in Nash equilibrium when players are white lie averse.

[8] A social choice function f satisfies no veto power if $f(\theta) = a$ whenever for at least $n-1$ players we have that $\widetilde{u}_i(a, \theta) \geq \widetilde{u}_i(b, \theta)$ for all $b \in A$.

Dutta and Sen (2012) show that any social choice function satisfying no veto power can be implemented in Nash equilibrium when agents are partially honest (see Kartik and Tercieux (2012) for a related result). The mechanism that Dutta and Sen (2012) use to establish this result is an augmented mechanism. Example 2, on the other hand, presents a social choice function that satisfies no veto power but cannot be implemented with a direct mechanism in Nash equilibrium when players are partially honest. This example clarifies that augmented mechanisms are needed in order to obtain the permissive results in Dutta and Sen (2012).

4 Implementation in stochastically stable equilibrium

In this section I first present the solution concept of stochastically stable equilibrium. I then show that the majoritarian mechanism g_M implements any social choice function in stochastically stable equilibrium if there are at least five agents who are all white lie averse. I restrict attention to pure strategies.[9]

The solution concept of stochastically stable equilibrium was first introduced into the economics literature to study which outcomes are more likely to arise in the long-run in evolutionary settings. Therefore, I present it by thinking about an evolutionary setup in which a group of players repeatedly interacts among each other.

Let $\left\langle N, (S_i)_{i\in N}, (u_i)_{i\in N}\right\rangle$ be a strategic game with $|N| < \infty$ and $|S_i| < \infty$ for all $i \in N$, and assume that players in N repeatedly play $\left\langle N, (S_i)_{i\in N}, (u_i)_{i\in N}\right\rangle$ in periods $t = 0, 1, 2, ...$ Let $s_{i,t}$ denote the strategy played by agent i in period t and let $\mathbf{s}_t = (s_{1,t}, ..., s_{n,t})$ the strategy profile played in period t. For every $i \in N$, let $\mathbf{s}_{-i,t}$ denote the strategy profile played by all agents but i in period t.

I now present the assumptions on how behavior evolves in this setup. At $t = 0$, agents play according to some $\mathbf{s}_0 \in \Pi_{i=1}^n S_i$. Then, at each date $t \geq 1$, every agent faces an independent probability $p \in (0, 1)$ of getting an opportunity to revise the strategy she played last period. Suppose agent i gets a revision opportunity at date t. With probability $1-\varepsilon$ (with $\varepsilon \in [0, 1]$) she randomizes among the strategies that solve $\max_{s_i\in S_i} u_i(s_i, \mathbf{s}_{-i,t-1})$. With probability ε she makes a mistake and plays any $s_i \in S_i$ with positive probability. That is, with probability $1-\varepsilon$ a player who gets a revision opportunity chooses a strategy that maximizes her payoff against $\mathbf{s}_{-i,t-1}$; with probability ε she "makes a mistake" and chooses randomly from her set of available strategies.

[9] As I explain in Appendix A.3, there no natural definition of mixed strategy stochastically stable equilibrium.

Given an initial strategy profile $\mathbf{s}_0$, for each $\varepsilon \in [0,1]$ the behavioral rule outlined above defines a Markov process over the (finite) set of strategy profiles $S = \Pi_{i=1}^{n} S_i$. Let P^{ε} denote the transition matrix of this Markov process. Following the literature on stochastic evolutionary game theory, I will refer to each $\mathbf{s} \in S$ as a "state" of the Markov process. The state $\mathbf{s} \in S$ should not be confused with the "state of nature" $\theta \in \Theta$.

Let $\widetilde{\mathbf{s}} \in S$ be a *strict* Nash equilibrium of $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$. When $\varepsilon = 0$ (so players do not make mistakes), the transition matrix P^0 has an invariant distribution $\mu^{\widetilde{\mathbf{s}}}$ such that $\mu^{\widetilde{\mathbf{s}}}(\widetilde{\mathbf{s}}) = 1$ (i.e., $\mu^{\widetilde{\mathbf{s}}}$ puts all its mass on the strict Nash equilibrium $\widetilde{\mathbf{s}}$). Therefore, the matrix P^0 will have multiple invariant distributions whenever the strategic game $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ has multiple strict equilibria. However, for every $\varepsilon > 0$ the matrix P^{ε} is aperiodic and irreducible and therefore has a unique invariant distribution μ^{ε}. One can show that $\mu^* = \lim_{\varepsilon \downarrow 0} \mu^{\varepsilon}$ exists (see Young (1993) for a proof of this result). Moreover, μ^* is one of the invariant distributions of the unperturbed process with transition matrix P^0.

Definition 3 *Let $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic game with $|N| < \infty$ and $|S_i| < \infty$ for all $i \in N$. A strategy profile $\mathbf{s} \in S$ is a stochastically stable equilibrium of $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ if $\mathbf{s} \in \text{supp}\mu^*$.*

The Markov process over S is ergodic for every $\varepsilon > 0$. Therefore, the invariant distribution μ^{ε} describes with probability 1 the fraction of time that the Markov process spends at each $\mathbf{s} \in S$.[10] This implies that the set of states in the support of μ^* (i.e., the stochastically stable equilibria) are the action profiles at which players will coordinate most of the time in the long run when the probability with which they make mistakes is small.

Consider next the implementation environment of Section 2. At $t = 0$ nature chooses $\theta \in \Theta$, which determines the preference profile of the agents and which remains fixed forever. The realization of θ is common knowledge among all agents, but is not known by the planner. The planner's objective is to design a strategic game form $\langle N, (S_i)_{i \in N}, g \rangle$ with the property that $|S_i| < \infty$ for all $i \in N$ to implement a given social choice function f.

Given a strategic game form $\langle N, (S_i)_{i \in N}, g \rangle$, let $P_g^{\varepsilon}(\theta)$ denote the transition matrix over $S = \Pi_{i=1}^{n} S_i$ when the state of nature is θ, and let $\mu_g^{\varepsilon}(\theta)$ denote its invariant distribution (which is unique for every $\varepsilon > 0$). For every $\theta \in \Theta$, let $\mu_g^*(\theta) = \lim_{\varepsilon \downarrow 0} \mu_g^{\varepsilon}(\theta)$.

Definition 4 *A mechanism g implements the social choice function $f : \Theta \to A$ in stochastically stable equilibrium if for all $\theta \in \Theta$, $\mathbf{s} \in \text{supp}\mu_g^*(\theta)$ implies $g(\mathbf{s}) = f(\theta)$.*

[10] Formally, for any initial $\mathbf{s}_0$ and any $\mathbf{s} \in S$, $T^{-1} \sum_{t=0}^{T} \mathbf{1}_{\mathbf{s}}(\mathbf{s}_t) \to \mu^{\varepsilon}(\mathbf{s})$ almost surely as $T \to \infty$, where $\mathbf{1}_{\mathbf{s}}$ is the indicator function for $\mathbf{s}$.

As before, the invariant distribution $\mu_g^{\varepsilon}(\theta)$ describes with probability 1 the fraction of time that the Markov process spends at each $\mathbf{s} \in S$. Therefore, if a mechanism stochastically implements a social choice function f, then the planner knows that the implemented outcome will be the correct one most of the time, provided the probability of mistakes is small.

I now present the main result of this section:

Theorem 2 *Let $f : \Theta \to A$ be a social choice function and suppose that Assumption 1 holds. If $n \geq 5$, mechanism g_M implements f in stochastically stable equilibrium. Moreover,* $\text{supp}\mu_{g_M}^*(\theta) = \{\mathbf{s}^\theta\}$ *for every $\theta \in \Theta$.*

Theorem 2 states that if agents have a minimal degree of honesty in the form of white lie aversion, then a social planner can use a majoritarian mechanism to implement any social choice function in stochastically stable equilibrium. In other words, under these conditions a planner can use a simple direct mechanism to achieve full implementation and can thus dispense with any augmentation of the message space.

The proof of Theorem 2 uses tools on stochastic dynamic systems developed by Freidlin and Wentzell (1984). Foster and Young (1990) were the first to apply these tools to evolutionary biology, while Kandori, Mailath and Rob (1993) and Young (1993) introduced them to the economics literature. Ellison (2000) extended these techniques, uncovering useful properties of the set of stochastically stable equilibria. I now present a brief overview of these methods.

For two states $\mathbf{s}$ and $\mathbf{s}'$ of the Markov process, define the cost of the transition $\mathbf{s} \to \mathbf{s}'$ as the number of mistakes needed to complete the transition. Define a path between states $\mathbf{s}$ and $\mathbf{s}'$ as a sequence of states $(\mathbf{s}^1, ..., \mathbf{s}^k)$ such that $\mathbf{s}^r$ is an immediate precessor of $\mathbf{s}^{r+1}$ and such that $\mathbf{s}^1 = \mathbf{s}$ and $\mathbf{s}^k = \mathbf{s}'$. The cost of the path $(\mathbf{s}^1, ..., \mathbf{s}^k)$ is the sum of the cost of its transitions. Let $c(\mathbf{s}^1, ..., \mathbf{s}^k)$ denote the cost of path $(\mathbf{s}^1, ..., \mathbf{s}^k)$.

Let $X = \{X_1, ..., X_J\}$ be the *recurrent classes* of the Markov process, with $X_k \subseteq S$ for all $X_k \in X$. These classes are disjoint and satisfy the following three properties. First, from every $\mathbf{s} \in S$ there is a path of zero cost to at least one of the recurrent classes. Second, within each recurrent class there is a path of zero cost from every state to every other. Third, every path starting at one recurrent class and ending outside that class has positive cost.

Let Ω be a union of one or more recurrent classes. The *basin of attraction* $D(\Omega)$ of Ω is the set of initial states from which the Markov process converges to Ω with probability 1 when $\varepsilon = 0$. Define the *radius* $Ra(\Omega)$ of $D(\Omega)$ as the number of mistakes needed to leave $D(\Omega)$ when play begins at a state in Ω. For every sets of states Z and Y, let $P(Z, Y)$ denote

the set of all paths starting in Z and ending in Y. Then $Ra(\Omega)$ is given by

$$Ra(\Omega) = \min_{(\mathbf{s}^1,...,\mathbf{s}^k)\in P(\Omega,S\setminus D(\Omega))} c\left(\mathbf{s}^1,...,\mathbf{s}^k\right).$$

In words, the radius is a measure of how hard it is to leave a given basin of attraction. The *coradius* $CRa(\Omega)$ of $D(\Omega)$ is the maximum number of mistakes needed to get into $D(\Omega)$:

$$CRa(\Omega) = \max_{\mathbf{s}\notin D(\Omega)} \min_{(\mathbf{s}^1,...,\mathbf{s}^k)\in P(\{\mathbf{s}\},\Omega)} c\left(\mathbf{s}^1,...,\mathbf{s}^k\right).$$

The coradius is then a measure of how easy it is to enter a given basin of attraction.

Ellison (2000) showed that if there is a set Ω that is a union of recurrent classes such that $Ra(\Omega) > CRa(\Omega)$, then the stochastically stable equilibria are all contained in Ω. In what follows, I will refer to this result as Ellison's Theorem. With this result in hand, to prove Theorem 2 it suffices to show that $\left\{\mathbf{s}^\theta\right\}$ is a recurrent class and that $Ra\left(\left\{\mathbf{s}^\theta\right\}\right) > CRa\left(\left\{\mathbf{s}^\theta\right\}\right)$.

For clarity of exposition, I divide the proof of Theorem 2 into two Lemmas. The first Lemma shows that $\left\{\mathbf{s}^\theta\right\}$ is a recurrent class and finds a lower bound on $Ra\left(\left\{\mathbf{s}^\theta\right\}\right)$, while the second Lemma finds an upper bound on $CRa\left(\left\{\mathbf{s}^\theta\right\}\right)$. These two Lemmas together with Ellison's Theorem will immediately imply Theorem 2.

Lemma 1 *Let θ be the true state of nature and suppose that Assumption 1 holds. If $n \geq 5$, then $\left\{\mathbf{s}^\theta\right\}$ is a recurrent class for the game that g_M induces at state θ, with $Ra\left(\left\{\mathbf{s}^\theta\right\}\right) \geq 2$.*

Proof. See Appendix A.2. ■

Lemma 2 *Let θ be the true state of nature and suppose that Assumption 1 holds. If $n \geq 5$, then $CRa\left(\left\{\mathbf{s}^\theta\right\}\right) \leq 1$.*

Proof. See Appendix A.2. ■

Proof of Theorem 2. By Lemma 1, $\left\{\mathbf{s}^\theta\right\}$ is a recurrent class with $Ra\left(\left\{\mathbf{s}^\theta\right\}\right) \geq 2$. By Lemma 2, $CRa\left(\left\{\mathbf{s}^\theta\right\}\right) \leq 1 < 2 \leq Ra\left(\left\{\mathbf{s}^\theta\right\}\right)$. Therefore, Ellison's Theorem implies that $\mathbf{s}^\theta$ is the unique stochastically stable equilibrium. ■

To see the intuition behind Theorem 2, note that white lie averse agents have a strict incentive to be honest whenever they are not pivotal. Since under mechanism g_M no agent is pivotal at state $\mathbf{s}^\theta$, it follows that truth-telling is always a recurrent class of the Markov process $P^\varepsilon_{g_M}(\theta)$. Moreover, this also implies that the radius of the truthful announcement $\mathbf{s}^\theta$ is at least 2 whenever there are five or more agents and they are all white lie averse. Indeed,

starting from the truthful announcement $\mathbf{s}^\theta$, under mechanism g_M no agent is pivotal after a single mistake if $n \geq 5$. Therefore, if play starts at $\mathbf{s}^\theta$ and there is a single mistake, all players will have a strict incentive to tell the truth, and so play will revert to $\mathbf{s}^\theta$ if there are no further mistakes.

On the other hand, Lemma 2 shows that from any strategy profile $\mathbf{s} \neq \mathbf{s}^\theta$ there is a path to $\mathbf{s}^\theta$ involving at most one mistake (so that the coradius of $\mathbf{s}^\theta$ is at most 1). Intuitively, for any strategy profile different from $\mathbf{s} \neq \mathbf{s}^\theta$, there is a mistake that will make the non-truthful agents non-pivotal, so that they will then have a strict incentive to be honest. These two results, together with Ellison's Theorem, imply that the truthful announcement is the unique stochastically stable equilibrium.

To see why $n \geq 5$ is needed in the proof of Theorem 2, consider the following example: $N = \{1, 2, 3\}$ and there exists $\theta', \theta'' \in \Theta$ such that, for $i = 2, 3$, $\widetilde{u}_i\left(f\left(\theta''\right), \theta'\right) > \widetilde{u}_i\left(b, \theta'\right)$ for all $b \in A \backslash \left\{f\left(\theta''\right)\right\}$ (similar examples can be constructed for the case with $n = 4$). Suppose further that $f(\theta'') \neq a^*$ and $f(\theta'') \neq f(\theta')$. Let $\overline{\mathbf{s}}$ be the announcement profile $(\theta', \theta'', \theta'')$. Note first that $\overline{\mathbf{s}}$ is a strict Nash equilibrium of the strategic game that mechanism g_M induces when the state of nature is θ'. Indeed, by white lie aversion, player 1 strictly prefers to announce θ' than any $\phi \neq \theta'$ when players 2 and 3 are announcing θ'' (since mechanism g_M will implement $f\left(\theta''\right)$ regardless of player 1's announcement). On the other hand, agents 2 and 3 strictly prefer to announce θ'' than any other $\phi \neq \theta''$, as they strictly prefer alternative $f\left(\theta''\right)$ to any $b \in A \backslash \left\{f\left(\theta''\right)\right\}$.

Since $\overline{\mathbf{s}}$ is a strict Nash equilibrium of the game that g_M induces when the state of nature is θ', then $\{\overline{\mathbf{s}}\}$ is also a recurrent class. In particular, $\overline{\mathbf{s}} \notin D\left(\left\{\mathbf{s}^{\theta'}\right\}\right)$. Moreover, with only one mistake the Markov process can move from $\mathbf{s}^{\theta'}$ to $\overline{\mathbf{s}}$. To see this, suppose the Markov process starts at $\mathbf{s}^{\theta'}$. If agent 2 makes a mistake and announces θ'' instead of θ', then agent 3 will strictly prefer to change her announcement to θ'' if she gets a revision opportunity. Thus, there is a path with total cost of 1 from $\mathbf{s}^{\theta'}$ to $\overline{\mathbf{s}}$, so $Ra\left(\left\{\mathbf{s}^{\theta'}\right\}\right) = 1$. In this case Theorem 2 will not hold. Indeed, one can show that in this example $\mathbf{s}^{\theta'} \in \text{supp}\mu^*_{g_M}\left(\theta'\right)$ if and only if $\overline{\mathbf{s}} \in \text{supp}\mu^*_{g_M}\left(\theta'\right)$.

Theorem 2 shows that, when there are five or more players and they are all white lie averse, any social choice function can be implemented in stochastically stable equilibrium with a simple direct mechanism. I now argue that the combination of stochastic stability and white lie aversion is needed for this positive result. Examples 1 and 2 show that white lie aversion by itself is not enough to obtain a positive result as in Theorem 2. The following example shows that there are social choice functions that cannot be implemented in stochastically stable equilibrium when players are not white lie averse.

Example 3 Suppose there exist states $\theta', \theta'' \in \Theta$ such that $\widetilde{u}_i(\cdot, \theta') = \widetilde{u}_i(\cdot, \theta'')$ for all $i \in N$; that is, players have the same material preferences over the alternatives in A under states θ' and θ''. Suppose further that the social choice function f that the planner wants to implement is such that $f(\theta') \neq f(\theta'')$. In this setting, for any mechanism g (either direct or augmented), the set of stochastically stable equilibria under states θ' and θ'' will coincide if players are not white lie averse. Therefore, f cannot be implemented in stochastically stable equilibrium when players are not white lie averse.

Finally, the analysis I presented in this section is immune to the critique made by Ellison (1993) to models of stochastic evolution. Ellison (1993) showed that convergence to the stochastically stable equilibrium can take an extremely long period of time in models like the one in Kandori, Mailath and Rob (1993), especially when the number of agents is large. The reason for this is that, in these models, the number of mistakes needed to bring the Markov process into the basin of attraction of the stochastically stable equilibrium increases with the number of players. In this case, the predictions of the solution concept of stochastic stability are weak, as those predictions will only materialize in the very long run. In contrast, in the current paper's setting, from every $\mathbf{s} \neq \mathbf{s}^\theta$ there is a path to $\mathbf{s}^\theta$ involving at most one mistake, regardless of the number of players. Therefore, if $W\left(\mathbf{s}_0, \mathbf{s}^\theta, \varepsilon\right)$ denotes the expected waiting time until the Markov process first reaches $\mathbf{s}^\theta$ when the probability of mistakes is ε and the Markov process starts at $\mathbf{s}_0$, then Theorem 1 in Ellison (2000) implies that $W\left(\mathbf{s}_0, \mathbf{s}^\theta, \varepsilon\right) = O\left(\varepsilon^{-1}\right)$ as $\varepsilon \to 0$ for any $\mathbf{s}_0 \neq \mathbf{s}^\theta$.

A Appendix

A.1 Examples 1 and 2

A.1.1 Example 1

In this appendix, I show that the social choice function in Example 1 cannot be implemented in Nash equilibrium when players are white lie averse.

Towards a contradiction, suppose that there exists a game form $\left\langle N, (S_i)_{i \in N}, g \right\rangle$ with $S_i = \Theta \times M$ for all $i \in N$ that implements f. Then, there exists a strategy profile $\mathbf{s} = (s_1, ..., s_n) \in (\Theta \times M)^n$ such that $g(\mathbf{s}) = a$. Let $N_1 = \{i \in N : s_i \notin \theta'' \times M\}$ be the set of players who announce a state of nature different from θ'' under strategy profile $\mathbf{s}$, and let $N_1^{np} = \{i \in N_1 : i \text{ is not pivotal at } \mathbf{s}\}$ be the set of players in N_1 that are not pivotal at $\mathbf{s}$ (i.e., if $i \in N_1^{np}$, then $g(s_i', \mathbf{s}_{-i}) = a$ for all s_i'). If $N_1^{np} = \emptyset$, then the announcement profile $\mathbf{s}$ is a

Nash equilibrium at state θ'', since all players are getting their preferred alternative, and those players who are not announcing the true state are pivotal (so, given their preferences, they strictly prefer to make an untruthful announcement than to tell the truth). This contradicts the assumption that g implements f, so it must be that $N_1^{np} \neq \emptyset$. Let $j_1 = \min\{i \in N_1^{np}\}$, and let $\mathbf{s}^1 = (s_1^1, ..., s_n^1) \in (\Theta \times M)^n$ be the announcement profile such that $s_{j_1}^1 = (\theta'', m_{j_1})$ (where m_{j_1} is the message in M that j_1 was announcing under $\mathbf{s}$) and $s_i^1 = s_i$ for all $i \neq j_1$. Since j_1 is not pivotal at $\mathbf{s}$, $g(\mathbf{s}^1) = a$.

Next, let $N_2 = \{i \in N : s_i^1 \notin \theta'' \times M\}$, and let $N_2^{np} = \{i \in N_2 : i \text{ is not pivotal at } \mathbf{s}^1\}$. If $N_2^{np} = \emptyset$, then the announcement profile $\mathbf{s}^1$ is a Nash equilibrium at state θ'', since all players are getting their preferred alternative, and those players who are not announcing the true state are pivotal. This contradicts the assumption that g implements f, so it must be that $N_2^{np} \neq \emptyset$. Let $j_2 = \min\{i \in N_2^{np}\}$, and let $\mathbf{s}^2 = (s_1^2, ..., s_n^2) \in (\Theta \times M)^n$ be such that $s_{j_2}^2 = (\theta'', m_{j_2})$ and $s_i^2 = s_i^1$ for all $i \neq j_2$. Since j_2 is not pivotal at $\mathbf{s}^1$, $g(\mathbf{s}^2) = a$.

If we continue with this procedure, there will be an iteration $r \geq 1$ such that either: (a) $\mathbf{s}^r$ is such that $s_i^r = (\theta'', m_i)$ for all i, or (b) $\mathbf{s}^r$ is such that all i with $s_i^r \notin \theta'' \times M$ are pivotal at $\mathbf{s}^r$. Moreover, in either case $g(\mathbf{s}^r) = a$. Note that $\mathbf{s}^r$ is a Nash equilibrium at state θ'', regardless of whether the relevant case is (a) or (b): in case (a) all players are announcing the truth and obtaining their most preferred alternative, while in case (b) all players are obtaining their preferred alternative and those who are not announcing state θ'' are pivotal at $\mathbf{s}^r$. Therefore, g does not implement f in Nash equilibrium.

A.1.2 Example 2

In this appendix I show that there is no direct mechanism that implements the social choice function in Example 2. For ease of exposition, I divide the argument into two claims.

Claim A1 *Consider the setting in Example 2 and suppose that Assumption 1 holds. Suppose further that there exists a direct mechanism $g : \Theta^5 \to A$ that implements f in Nash equilibrium. Then it must be that, for $\theta = \theta', \theta''$, $\mathbf{s}^\theta$ is a Nash equilibrium of the strategic game that g induces at state θ, with $g(\mathbf{s}^\theta) = f(\theta)$.*

Proof. Since g implements f in Nash equilibrium, there must exist a strategy profile $\mathbf{s}^* \in \Theta^n$ such that $\mathbf{s}^*$ is a Nash equilibrium of the strategic game that mechanism g induces at state θ', with $g(\mathbf{s}^*) = f(\theta') = a$. I now show that, under the setting in Example 2, it must be that $\mathbf{s}^* = \mathbf{s}^{\theta'}$.

Note first that, since players 1 and 2 are materially indifferent between all the alternatives in A, for $\mathbf{s}^*$ to be a Nash equilibrium it must be that $s_1^* = s_2^* = \theta'$; otherwise, if $s_i^* \neq \theta'$ for $i \in \{1, 2\}$, by white lie aversion player i would have a strict preference to deviate and announce θ' instead.

Note next that at state θ', players 3, 4 and 5 are getting their worse alternative by playing according to strategy profile $\mathbf{s}^*$. Therefore, for $\mathbf{s}^*$ to be a Nash equilibrium at state θ' it must be that, for all $i \in \{3, 4, 5\}$, $g(s_i, \mathbf{s}_{-i}^*) = a$ for all $s_i \neq s_i^*$; otherwise, player $i \in \{3, 4, 5\}$ would find it strictly optimal to change her announcement when other players are announcing $\mathbf{s}_{-i}^*$. Assumption 1 then implies that $s_i^* = \theta'$, since otherwise player i would have a strict incentive to announce θ' instead of s_i^* when her opponents are announcing $\mathbf{s}_{-i}^*$ and the state is θ'. This, together with the arguments in the previous paragraph, implies that $\mathbf{s}^* = \mathbf{s}^{\theta'}$.

Given the symmetry of the environment at states θ' and θ'', a symmetric argument can be used to establish that if $\mathbf{s}^{**}$ is a Nash equilibrium of the strategic game that mechanism g induces at state θ'' with $g(\mathbf{s}^{**}) = f(\theta'') = d$, then it must be that $\mathbf{s}^{**} = \mathbf{s}^{\theta''}$ (such a Nash equilibrium $\mathbf{s}^{**}$ must exist if g implements f). ■

Claim A2 *Consider the setting in Example 2 and suppose that Assumption 1 holds. Suppose further that there exists a direct mechanism $g : \Theta^5 \to A$ that implements f in Nash equilibrium. Then, the mechanism g must satisfy the following conditions:*

(i) $g(\mathbf{s}) = a$ *for all* $\mathbf{s} = (s_1, ..., s_5) \in \Theta^5$ *such that* $s_1 = s_2 = \theta'$.

(ii) $g(\mathbf{s}) = d$ *for all* $\mathbf{s} = (s_1, ..., s_5) \in \Theta^5$ *such that* $s_4 = s_5 = \theta''$.

Before proceeding to its proof, note that Claim A2 implies that there is no direct mechanism that implements f in Nash equilibrium, since the two conditions in the claim cannot be satisfied simultaneously.

Proof of Claim A2. Let g be a direct mechanism that implements f in Nash equilibrium. By Claim A1, $\mathbf{s}^{\theta'}$ is a Nash equilibrium of the strategic game that mechanism g induces at state θ', with $g(\mathbf{s}^{\theta'}) = a$. This implies that $g(\mathbf{s}_{-i}^{\theta'}, s_i) = a$ for all $i \in \{3, 4, 5\}$ and all $s_i \in \Theta = \{\theta', \theta''\}$: otherwise, if $g(\mathbf{s}_{-i}^{\theta'}, \theta'') \neq a$ for some $i \in \{3, 4, 5\}$, then $\mathbf{s}^{\theta'}$ would not be a Nash equilibrium at state θ' since player i would have a strict incentive to announce θ'' when her opponents are announcing $\mathbf{s}_{-i}^{\theta'}$.

Next, let $\tilde{\mathbf{s}} = (\theta', \theta', \theta'', \theta'', \theta'')$; i.e., $\tilde{\mathbf{s}}$ is such that agents 1 and 2 announce θ', and agents 3, 4 and 5 announce θ''. I now show that, if g implements f in Nash equilibrium, it must be that $g(\tilde{\mathbf{s}}) = a$. Suppose by contradiction that $g(\tilde{\mathbf{s}}) \neq a$. Note that by Assumption 1, at state

θ' players $i = 1, 2$ find it optimal to announce $\tilde{s}_i = \theta'$ when their opponents are announcing $\tilde{\mathbf{s}}_{-i}$. Therefore, at state θ' there must exist a player $j \in \{3, 4, 5\}$ who finds it optimal to announce θ' when her opponents are announcing $\tilde{\mathbf{s}}_{-j}$; if no such j existed, $\tilde{\mathbf{s}}$ would be a Nash equilibrium at state θ', which would contradict the assumption that g implements f.

Let $\mathbf{s}' = (\tilde{\mathbf{s}}_{-j}, \theta')$, where $j \in \{3, 4, 5\}$ is an agent who finds it optimal to announce θ' when her opponents are announcing $\tilde{\mathbf{s}}_{-j}$ and the true state of nature is θ'. Strategy profile $\mathbf{s}'$ is such that players 1, 2 and j are announcing θ' and players $k \in \{3, 4, 5\}\backslash\{j\}$ are announcing θ''. Note that player j's material preferences must be such that she weakly prefers $g(\mathbf{s}')$ to $g(\tilde{\mathbf{s}})$ at state θ'; otherwise it would not be optimal for her to announce θ' when her opponents announce $\tilde{\mathbf{s}}_{-j}$. Since $g(\tilde{\mathbf{s}}) \neq a$ and since a is player j's worse alternative at state θ', it must be that $g(\mathbf{s}') \neq a$. Note further that by Assumption 1, at state θ' players $i = 1, 2$ find it optimal to announce $s'_i = \theta'$ when their opponents are announcing $\mathbf{s}'_{-i}$. Finally, at state θ' players $k \in \{3, 4, 5\}\backslash\{j\}$ also find it optimal to announce $s'_k = \theta''$ when their opponents are announcing $\mathbf{s}'_{-k}$: the mechanism's outcome is $g(\mathbf{s}') \neq a$ if $k \in \{3, 4, 5\}\backslash\{j\}$ announces $s'_k = \theta''$ when her opponents are announcing $\mathbf{s}'_{-k}$, while the mechanism's outcome would be a if she announces θ' when her opponents announce $\mathbf{s}'_{-k}$ (i.e., $g(\mathbf{s}'_{-k}, \theta') = a$).[11] But this implies that $\mathbf{s}'$ is a Nash equilibrium at state θ' with $g(\mathbf{s}') \neq a$, a contradiction to the assumption that g implements f. Hence, it must be that $g(\tilde{\mathbf{s}}) = a$.

Finally, suppose that there exists $i, j \in \{3, 4, 5\}$ such that $g(\mathbf{s}^{\theta'}_{-ij}, \theta'', \theta'') \neq a$, where $(\mathbf{s}^{\theta'}_{-ij}, \theta'', \theta'')$ is the strategy profile under which players i and j announce θ'' and the other three players announce θ'. Let $\hat{\mathbf{s}} = (\mathbf{s}^{\theta'}_{-ij}, \theta'', \theta'')$. Note that players $k = i, j$ find it optimal to announce θ'' when their opponents are announcing $\hat{\mathbf{s}}_{-k}$: if k changes her announcement to θ' the mechanism's outcome changes to a, which is k's worse alternative.[12] Moreover, by white lie aversion players $k' = 1, 2$ also find it strictly optimal to announce θ' when their opponents are announcing according $\hat{\mathbf{s}}_{-k'}$. Since $g(\hat{\mathbf{s}}) \neq a$ and g implements f in Nash equilibrium, player $\hat{i} = \{3, 4, 5\}\backslash\{i, j\}$ must find it optimal to announce θ'' when her opponents are announcing $\hat{\mathbf{s}}_{-\hat{i}}$ (otherwise, $\hat{\mathbf{s}}$ would be a Nash equilibrium at state θ', and so g would not implement f). Given the preferences of $\hat{i}$, it must be that $g(\hat{\mathbf{s}}_{-\hat{i}}, \theta'') = g(\tilde{\mathbf{s}}) \neq a$, since otherwise $\hat{i}$ would not find it optimal to announce θ'' when her opponents are announcing according to $\hat{\mathbf{s}}_{-\hat{i}}$ (recall that $\tilde{\mathbf{s}} = (\theta', \theta', \theta'', \theta'', \theta'') = (\hat{\mathbf{s}}_{-\hat{i}}, \theta'')$). But this contradicts the fact that $g(\tilde{\mathbf{s}}) = a$, which was established in the previous paragraph. Hence, it must be that $g(\mathbf{s}^{\theta'}_{-ij}, \theta'', \theta'') = a$ for all

[11] To see this, let $\mathbf{s}'' = (\mathbf{s}'_{-k}, \theta')$ and note that $\mathbf{s}''$ is such that all players but $i = \{3, 4, 5\}\backslash\{j, k\}$ are announcing θ'; i.e., $\mathbf{s}'' = (\mathbf{s}^{\theta'}_{-i}, \theta'')$. In the first paragraph of this proof I had established that $g(\mathbf{s}'') = a$ for all such announcement profiles.

[12] This follows because any strategy profile of the form $(\mathbf{s}^{\theta'}_{-i}, \theta'')$ with $i \in \{3, 4, 5\}$ is such that $g(\mathbf{s}^{\theta'}_{-i}, \theta'') = a$; see the first paragraph of the current proof.

$i, j \in \{3, 4, 5\}$.

The arguments above establish part (i) of Claim A2. Given the symmetry of the environment at states θ' and θ'', a symmetric argument can be used to establish that $g(\mathbf{s}) = d$ for all $\mathbf{s} = (s_1, ..., s_5) \in \Theta^5$ such that $s_4 = s_5 = \theta''$. ■

A.2 Proofs of Lemmas 1 and 2

Proof of Lemma 1. Under Assumption 1, $\mathbf{s}^\theta$ is a strict Nash equilibrium of the game that mechanism g_M induces. Therefore, $\left\{\mathbf{s}^\theta\right\}$ is a recurrent class of the stochastic process, since the Markov process can only leave this state with the aid of a mistake. When $n \geq 5$, Assumption 1 implies that every $\mathbf{s}'$ with $|R^*(\theta \mid \mathbf{s}')| = n - 1$ belongs to $D\left(\left\{\mathbf{s}^\theta\right\}\right)$. Indeed, at any such $\mathbf{s}'$ no agent can change the implemented outcome by changing her announcement. Thus, if $\varepsilon = 0$ and the Markov process starts at such $\mathbf{s}'$, eventually the agent who was announcing a state different from θ will get a revision opportunity and will change her announcement to θ. This implies that every path starting at $\mathbf{s}^\theta$ and ending in some state $\mathbf{s} \notin D\left(\left\{\mathbf{s}^\theta\right\}\right)$ must involve at least two mistakes. Therefore, $Ra\left(\left\{\mathbf{s}^\theta\right\}\right) \geq 2$. ■

Proof of Lemma 2. To prove Lemma 2, I need to show that from every $\mathbf{s} \in S \backslash \left\{\mathbf{s}^\theta\right\}$ there exists a path involving at most 1 mistake leading to $\mathbf{s}^\theta$. Consider first states $\mathbf{s}^1 \neq \mathbf{s}^\theta$ such that $|R^*(\theta \mid \mathbf{s}^1)| \geq \frac{n}{2}$. With one mistake the Markov process can move to a state $\mathbf{s}^2$ such that $|R^*(\theta \mid \mathbf{s}^2)| > \frac{n}{2}$. At $\mathbf{s}^2$, agents announcing a state different from θ cannot change the implemented outcome by changing their announcements, so under Assumption 1 they all strictly prefer to announce θ than to continue with their announcements. Therefore, from $\mathbf{s}^2$ the Markov process can move to $\mathbf{s}^\theta$ without any further mistakes.

Consider next states $\mathbf{s}^1$ such that $|R^*(\phi \mid \mathbf{s}^1)| > \frac{n}{2}$ for some $\phi \neq \theta$. From $\mathbf{s}^1$ the Markov process can move to a state $\mathbf{s}^2$ such that $|R^*(\phi \mid \mathbf{s}^2)| > \frac{n}{2} + 1$ with one mistake. At $\mathbf{s}^2$, no agent can change the implemented outcome by changing her announcement. Assumption 1 then implies that, given $\mathbf{s}^2_{-i}$, each agent i strictly prefers to announce θ than any $\theta' \in \Theta \backslash \{\theta\}$, so the Markov process can move to $\mathbf{s}^\theta$ without any further mistakes.

Consider next states $\mathbf{s}^1$ such that $|R^*(\theta' \mid \mathbf{s}^1)| = |R^*(\theta'' \mid \mathbf{s}^1)| = \frac{n}{2}$ for $\theta', \theta'' \in \Theta$. I already considered the case with $|R^*(\theta \mid \mathbf{s}^1)| = \frac{n}{2}$ above. Therefore, I now focus on the case with $\theta', \theta'' \neq \theta$. There are two possibilities: (i) there exists $i \in N$ such that $s_i^1 = \theta'$ and such that $\widetilde{u}_i(f(\theta''), \theta) \geq \widetilde{u}_i(a^*, \theta)$ (so for agent i it is a best response to change her announcement to θ'', given the announcement profile $\mathbf{s}^1_{-i}$ of her opponents), and (ii) no such i exists (so that every player strictly prefers alternative a^* to any other alternative that mechanism g_M

would implement if she changed her announcement). In case (i), the Markov process can move without any mistakes to a state $\mathbf{s}^2$ such that $|R^*(\theta'' \mid \mathbf{s}^2)| > \frac{n}{2}$, with $\theta'' \neq \theta$ (this occurs if only player i gets a revision opportunity and changes her announcement from θ' to θ''). It follows from the previous paragraph that there is a path from $\mathbf{s}^2$ to $\mathbf{s}^\theta$ involving one single mistake. In case (ii), given $\mathbf{s}^1_{-i}$ it is weakly optimal for each agent i to announce any $\phi \notin \{\theta', \theta''\}$. In particular, for every agent it is a weak best reply to announce θ. Therefore, in this case the Markov process can move to $\mathbf{s}^\theta$ without any mistakes.

Finally, consider announcement profiles $\mathbf{s}^1$ such that $|R^*(\theta' \mid \mathbf{s}^1)| \leq \frac{n}{2}$ for every $\theta' \in \Theta$, but with at most one $\theta' \neq \theta$ such that $|R^*(\theta' \mid \mathbf{s})| = \frac{n}{2}$ and with $|R^*(\theta \mid \mathbf{s})| < \frac{n}{2}$. There are two cases to consider: (i) there exists $\phi \in \Theta$ such that $|R^*(\phi \mid \mathbf{s}^1)| + 1 > \frac{n}{2}$, and (ii) no such ϕ exists. In case (ii), no agent can change the implemented outcome by changing her announcement (i.e., mechanism g_M would still implement alternative a^*). Assumption 1 then implies that, given $\mathbf{s}^1_{-i}$, each agent i strictly prefers to announce θ than any $\theta' \in \Theta \backslash \{\theta\}$, so the Markov process can move to $\mathbf{s}^\theta$ without any mistakes.

Consider next case (i), and let $Y(\mathbf{s}^1) = \{\phi \in \Theta : |R^*(\phi \mid \mathbf{s}^1)| + 1 > \frac{n}{2}\}$. Since $n \geq 5$, the cardinality of $Y(\mathbf{s}^1)$ can be at most two. Note also that at least three states of nature are being announced in $\mathbf{s}^1$, so there exists at least one state of nature $\theta' \in \Theta$ with $|R^*(\theta' \mid \mathbf{s}^1)| \leq \frac{n}{2} - 1$ (i.e., $\theta' \notin Y(\mathbf{s}^1)$). There are two subcases to consider: (i.a) there exists $i \in N$ who finds it strictly optimal to change her announcement given $\mathbf{s}^1_{-i}$ (note that this implies that player i's best reply to $\mathbf{s}^1_{-i}$ is to announce some $\phi \in Y(\mathbf{s}^1)$), and (i.b) no such i exists. In case (i.a), the Markov process can move to a state $\mathbf{s}^2$ such that $|R^*(\phi \mid \mathbf{s}^2)| > \frac{n}{2}$ without any mistakes (this happens if only player i gets a revision opportunity and announces $\phi \in Y(\mathbf{s}^1)$), and from such a state there is a path involving one mistake leading to $\mathbf{s}^\theta$. On the other hand, in case (i.b) it is weakly optimal for every agent i to announce θ' when her opponents are announcing $\mathbf{s}^1_{-i}$ (where $\theta' \in \Theta$ is such that $|R^*(\theta' \mid \mathbf{s}^1)| \leq \frac{n}{2} - 1$), since all agents weakly prefer alternative a^* to any other alternative that the mechanism would implement if they changed their announcements. Therefore, the Markov process can move to a state $\mathbf{s}^2$ with $|R^*(\theta' \mid \mathbf{s}^2)| > \frac{n}{2}$ without any mistakes, and from such a state there is a path to $\mathbf{s}^\theta$ involving at most one mistake. ■

A.3 Mixed strategies

The body of the paper focuses exclusively on pure strategy equilibria. This appendix shows that the results in Section 3 continue to hold even if we allow for mixed strategies: if there are at least five agents and they are all white lie averse, then for generic payoffs the mechanism

g_M does not have undesirable mixed strategy 1-FTNE. This appendix, however, does not extend the results in Section 4 to the case of mixed strategies. The reason for this is that, to the best of my knowledge, there is no natural extension of stochastically stable equilibrium to mixed strategies.[13]

I start by extending the definition of white lie aversion to the case of mixed strategies. Assume that, for all $i \in N$, agent i's material preferences satisfy the axioms of expected utility, and let $\widetilde{u}_i : A \times \Theta \to \mathbb{R}$ be agent i's utility index over alternatives in A at each state state of nature in Θ. Let $\Delta(A)$ denote the set of lotteries over alternatives, and let $\widetilde{U}_i : \Delta(A) \times \Theta \to \mathbb{R}$ denote agent i's material utility over lotteries in $\Delta(A)$ at each state of nature: for any $p \in \Delta(A)$ and $\theta \in \Theta$, $\widetilde{U}_i(p, \theta) = \mathbb{E}_p[\widetilde{u}_i(a, \theta)]$, where $\mathbb{E}_p$ is the expectation with respect to lottery p.

For any mechanism $g : (\Theta \times M)^n \to A$, let $\Sigma_i = \Delta(\Theta \times M)$ denote the set of agent i's mixed strategies and let $\Sigma = \Pi_{i \in N} \Sigma_i$ denote the set of mixed strategy profiles. For any profile of mixed strategies $\sigma \in \Sigma$, let σ_{-i} denote the mixed strategies of all players in $N \backslash \{i\}$. For any state of nature $\theta \in \Theta$, agent i's material utility from a mixed strategy profile $\sigma = (\sigma_i, \sigma_{-i})$ is $\widetilde{U}_i(g(\sigma_i, \sigma_{-i}), \theta) = \mathbb{E}_\sigma[\widetilde{u}_i(g(\sigma_i, \sigma_{-i}), \theta)]$, where $\mathbb{E}_\sigma$ denotes the expectation with respect to the mixed strategy profile $\sigma = (\sigma_i, \sigma_{-i})$.

Let $U_i : \Delta(A) \times \Theta \times \Sigma \to \mathbb{R}$ be agent i's utility. For all $\theta \in \Theta$, let $T(\theta) = \{\sigma \in \Delta(\Theta \times M) : \text{supp}\sigma \subset \theta \times M\}$ be the set of mixed strategies under which the player announces state θ with probability 1. The following definition extends Assumption 1 to the case of mixed strategies.

Assumption A1 (white lie aversion - mixed strategies) *Let g be a mechanism. For all $i \in N$, if σ_{-i} is such that there exists $\sigma_i' \in T(\theta)$ and $\sigma_i'' \notin T(\theta)$ with the property that $\widetilde{U}_i(g(\sigma_i', \sigma_{-i}), \theta) = \widetilde{U}_i(g(\sigma_i'', \sigma_{-i}), \theta) \geq \widetilde{U}_i(g(\sigma_i, \sigma_{-i}), \theta)$ $\forall \sigma_i \in \Delta(\Theta \times M)$, then*

$$U_i(g(\sigma_i', \sigma_{-i}), \theta, (\sigma_i', \sigma_{-i})) = \widetilde{U}_i(g(\sigma_i', \sigma_{-i}), \theta) > U_i(g(\sigma_i'', \sigma_{-i}), \theta, (\sigma_i'', \sigma_{-i})) = \widetilde{U}_i(g(\sigma_i'', \sigma_{-i}), \theta) - \eta.$$

For any other σ_{-i}, $U_i(g(\sigma_i', \sigma_{-i}), \theta, (\sigma_i', \sigma_{-i})) = \widetilde{U}_i(g(\sigma_i', \sigma_{-i}), \theta)$ $\forall \sigma_i' \in \Delta(\Theta \times M)$.

Next, I extend the definition of k-FTNE to mixed strategies. Let $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ be a strategic game, where $N = \{1, 2, ..., n\}$ is a finite set of players, S_i is the set of actions of player i and $u_i : \Pi_{i \in N} S_i \to \mathbb{R}$ is the utility index of player i. All players in N are expected

[13]The tools developed in the literature to find the set of stochastically stable equilibria (i.e., Kandori, Mailath and Rob (1993), Young (1993) and Ellison (2000)) require that the set of strategy profiles be finite. Having a finite set of strategy profiles guarantees that the Markov chain defined by the behavioral rule that players use to update their strategies has a finite set of states. Clearly, with mixed strategies the set of strategy profiles would not be finite.

utility maximizers. Let $\Sigma_i = \Delta(S_i)$ be the set of mixed strategies of player i and $\Sigma = \Pi_{i \in N} \Sigma_i$ be the set of mixed strategy profiles. For any $\sigma = (\sigma_i, \sigma_{-i}) \in \Sigma$, let $U_i(\sigma_i, \sigma_{-i})$ denote player i's expected utility from mixed strategy profile σ.

For any $\sigma, \sigma' \in \Sigma$, let the distance between σ and σ' be

$$d(\sigma, \sigma') = |\{i \in N : \sigma_i \neq \sigma'_i\}|.$$

Definition A1 *A mixed strategy profile $\sigma^* = (\sigma_1^*, ..., \sigma_n^*) \in \Sigma$ is a k-fault tolerant Nash equilibrium of the strategic game $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ if, $\forall i \in N$, $U_i(\sigma_i^*, \sigma_{-i}) \geq U_i(\sigma'_i, \sigma_{-i})$ for all $\sigma_{-i} \in \{\widetilde{\sigma}_{-i} \in \Sigma_{-i} : d((\sigma_i^*, \widetilde{\sigma}_{-i}), \sigma^*) \leq k\}$ and for all $\sigma'_i \in \Sigma_i \backslash \{\sigma_i^*\}$.*

For any strategic game form $\langle N, (S_i)_{i \in N}, g \rangle$ and for any integer $k \geq 0$, let $\tilde{E}_k(g, \theta)$ be the set of mixed strategy k-FTNE of the strategic game that mechanism g induces when the state is θ. For any $\sigma \in \Sigma$, let

$$B(\sigma, k) = \{\sigma' \in \Sigma : d(\sigma, \sigma') \leq k\}.$$

Definition A2 *A mechanism implements the social choice function $f : \Theta \to A$ in mixed-strategy k-FTNE if for all $\theta \in \Theta$, $g(B(\sigma, k)) = f(\theta)$ for all $\sigma \in \tilde{E}_k(g, \theta)$.*

Theorem A1 *Let $f : \Theta \to A$ be a social choice function and suppose that Assumption A1 holds. If $n \geq 5$, then for generic material payoffs the mechanism g_M implements f in mixed strategy 1-FTNE. Moreover, $\tilde{E}_1(g_M, \theta) = \{\mathbf{s}^\theta\}$ for every $\theta \in \Theta$.*

Proof. Suppose the state of nature is θ. Theorem 1 shows that the unique pure strategy 1-FTNE of the strategic game that mechanism g_M induces at state θ is the truthful announcement profile $\mathbf{s}^\theta$. To establish the result I show that, for generic payoffs, such strategic game has no mixed strategy 1-FTNE.

Towards a contradiction, suppose that at state θ there exists a mixed strategy profile $\sigma^* \neq \mathbf{s}^\theta$ such that $\sigma^* \in \tilde{E}_1(g_M, \theta)$. By Theorem 1, it must be that at least one player is making at least two announcements with positive probability under σ^*. Fix a player $i \in N$ who randomizes under σ^*. Since $\sigma^* \in \tilde{E}_1(g_M, \theta)$, it must be that player i is pivotal with positive probability when her opponents play according to σ^*_{-i}; if she were not pivotal with positive probability, by white lie aversion she would strictly prefer to announce the true state with probability 1 than to randomize.

Note that for any $\theta' \in \text{supp}\sigma_i^*$, player i's material payoff from announcing θ' is

$$\begin{aligned}\widetilde{U}_i(g(\theta', \sigma_{-i}^*), \theta) &= \mathbb{E}_{\sigma_{-i}^*}[\widetilde{u}_i(g_M(\theta', \sigma_{-i}^*), \theta)] \\ &= \Pr(i \text{ is pivotal}|\sigma_{-i}^*)\mathbb{E}_{\sigma_{-i}^*}[\widetilde{u}_i(g_M(\theta', \sigma_{-i}^*), \theta)|i \text{ is pivotal}] + \\ &\quad \Pr(i \text{ is not pivotal}|\sigma_{-i}^*)\mathbb{E}_{\sigma_{-i}^*}[\widetilde{u}_i(g_M(\theta', \sigma_{-i}^*), \theta)|i \text{ is not pivotal}]. \quad \text{(A.1)}\end{aligned}$$

Note further that, under mechanism g_M, player i is only pivotal when the announcement of i's opponent $\mathbf{s}_{-i}$ is such that

$$|R(\phi|(\mathbf{s}_{-i}, \phi))| > \frac{n}{2} \geq |R(\phi|(\mathbf{s}_{-i}, \phi))| - 1$$

for some state $\phi \in \Theta$. Therefore, one of two things can happen if i is pivotal and announces θ': (i) i's announcement of θ' leads to alternative $f(\theta')$ being implemented, or (ii) i's announcement of θ' leads to alternative a^* being implemented.[14]

Given the strategy of i's opponents σ_{-i}^*, let $q(\sigma_{-i}^*, \theta') \in [0, 1]$ be the probability that, conditional on i being pivotal, i's announcement of θ' leads to alternative $f(\theta')$ being implemented.[15] Using this notation in equation (A.1),

$$\begin{aligned}\widetilde{U}_i(g(\theta', \sigma_{-i}^*), \theta) &= \Pr(i \text{ is pivotal}|\sigma_{-i}^*)[q(\sigma_{-i}^*, \theta')\widetilde{u}_i(f(\theta'), \theta) + (1 - q(\sigma_{-i}^*, \theta'))\widetilde{u}_i(a^*, \theta)] + \\ &\quad \Pr(i \text{ is not pivotal}|\sigma_{-i}^*)\mathbb{E}_{\sigma_{-i}^*}[\widetilde{u}_i(g_M(\theta', \sigma_{-i}^*), \theta)|i \text{ is not pivotal}]. \quad \text{(A.2)}\end{aligned}$$

Since σ^* is a 1-FTNE at state θ, it must be that for all $\theta' \in \text{supp}\sigma_i^*$ and for all σ_{-i} such that $d((\sigma_i^*, \sigma_{-i}), \sigma^*) \leq 1$, $\widetilde{U}_i(g(\theta', \sigma_{-i}), \theta) \geq \widetilde{U}_i(g(\tilde{\sigma}_i, \sigma_{-i}), \theta)$ for all $\tilde{\sigma}_i \in \Sigma_i$. This implies that, for $\theta', \theta'' \in \text{supp}\sigma_i^*$ and for all σ_{-i} such that $d((\sigma_i^*, \sigma_{-i}), \sigma^*) \leq 1$, it must be that

$$\begin{aligned}\widetilde{U}_i(g(\theta', \sigma_{-i}), \theta) &= \widetilde{U}_i(g(\theta'', \sigma_{-i}), \theta) \iff \\ q(\sigma_{-i}, \theta')(\widetilde{u}_i(f(\theta'), \theta) - \widetilde{u}_i(a^*, \theta)) &= q(\sigma_{-i}, \theta'')(\widetilde{u}_i(f(\theta''), \theta) - \widetilde{u}_i(a^*, \theta)), \quad \text{(A.3)}\end{aligned}$$

where the equality in (A.3) follows after using equation (A.2) and noting that when i is not pivotal, i's announcement does not influence the mechanism's outcome and thus i's material payoff is the same regardless of whether she announces θ' or θ''. The rest of proof shows

[14] The first case arises when the announcement $\mathbf{s}_{-i}$ of i's opponents is such that $|R(\theta'|(\mathbf{s}_{-i}, \theta'))| > \frac{n}{2} \geq |R(\theta'|(\mathbf{s}_{-i}, \theta'))| - 1$. The second case arises when $|R(\phi|(\mathbf{s}_{-i}, \phi))| > \frac{n}{2} \geq |R(\phi|(\mathbf{s}_{-i}, \phi))| - 1$ for some $\phi \neq \theta'$.

[15] Let $p(\sigma_{-i}^*, \theta')$ be the probability with which the announcement of i's opponents $\mathbf{s}_{-i}$ is such that $|R(\theta'|(\mathbf{s}_{-i}, \theta'))| > \frac{n}{2} \geq |R(\theta'|(\mathbf{s}_{-i}, \theta'))| - 1$ when i's opponents use the mixed strategy profile σ_{-i}^*. Then, $q(\sigma_{-i}^*, \theta') = \frac{p(\sigma_{-i}^*, \theta')}{\Pr(i \text{ is pivotal}|\sigma_{-i}^*)}$.

that, generically, the equalities in (A.3) can never be satisfied and hence there cannot be a 1-FTNE in which some players use mixed strategies.

Suppose first that $\widetilde{u}_i$ and $q(\sigma^*_{-i}, \theta')$ are such that $q(\sigma^*_{-i}, \theta')(\widetilde{u}_i(f(\theta'), \theta) - \widetilde{u}_i(a^*, \theta)) \neq 0$ for some $\theta' \in \text{supp}\sigma^*_i$. In this case, the equality in (A.3) cannot hold for all σ_{-i} such that $d((\sigma^*_i, \sigma_{-i}), \sigma^*) \leq 1$: if (A.3) holds for σ^*_{-i}, then for generic payoffs and under mechanism g_M it is always possible to find a strategy profile of i's opponents σ_{-i} with $d((\sigma^*_i, \sigma_{-i}), \sigma^*) = 1$ such that either $q(\sigma_{-i}, \theta'') \leq q(\sigma^*_{-i}, \theta'')$ and $q(\sigma_{-i}, \theta') > q(\sigma^*_{-i}, \theta')$ or $q(\sigma_{-i}, \theta'') < q(\sigma^*_{-i}, \theta'')$ and $q(\sigma_{-i}, \theta') \geq q(\sigma^*_{-i}, \theta')$.

The arguments in the previous paragraphs imply that, for σ^* to be a 1-FTNE, it must be that $q(\sigma^*_{-i}, \theta')(\widetilde{u}_i(f(\theta'), \theta) - \widetilde{u}_i(a^*, \theta)) = 0$ for all $\theta' \in \text{supp}\sigma^*_i$. The rest of the proof establishes that under Assumption A1 this cannot happen in a 1-FTNE either. Note first that since player i is white lie averse, for i to find it optimal to use strategy σ^*_i when her opponents are using σ^*_{-i} it must be that, for all $\theta' \in \text{supp}\sigma^*_i$,

$$\begin{aligned} \widetilde{U}_i(g(\theta', \sigma^*_{-i}), \theta) &> \widetilde{U}_i(g(\theta, \sigma^*_{-i}), \theta) \iff \\ q(\sigma^*_{-i}, \theta')(\widetilde{u}_i(f(\theta'), \theta) - \widetilde{u}_i(a^*, \theta)) &> q(\sigma^*_{-i}, \theta)(\widetilde{u}_i(f(\theta), \theta) - \widetilde{u}_i(a^*, \theta)). \end{aligned} \tag{A.4}$$

If (A.4) did not hold, then under Assumption A1 player i would have a strict incentive to make a truthful announcement with probability 1 when her opponents are using strategy profile σ^*_{-i}. An implication of (A.4) is that, if $\sigma^* = (\sigma^*_1, ..., \sigma^*_n) \in \tilde{E}_1(g_M, \theta)$ and σ^*_i is a non-trivial mixed strategy, then σ^*_i puts probability zero on the truthful announcement θ.

Let $N(\theta) = \{j \in N : \sigma^*_j(\theta) = 1\}$ be the set of players who announce state θ with probability 1 under strategy profile σ^*. I now show that, if σ^* is a 1-FTNE, then it must be that $|N(\theta)| \leq \frac{n}{2} - 1$. To see this, note that if $|N(\theta)| > \frac{n}{2} - 1$, then for every $i \notin N(\theta)$ there would exist a strategy profile σ_{-i} of i's opponents with $d((\sigma^*_i, \sigma_{-i}), \sigma^*) \leq 1$ such that the number of players announcing θ with probability 1 under strategy profile $(\sigma^*_i, \sigma_{-i})$ is strictly greater than $\frac{n}{2}$. Note that agent i is never pivotal when her opponents play according to strategy σ_{-i}: the outcome of mechanism g_M would be $f(\theta)$ regardless of i's announcement. This implies that σ^* cannot be a 1-FTNE at state θ, since under Assumption A1 player i would find it strictly optimal to announce θ with probability 1 when her opponents play according to σ_{-i} than to play according to the mixed strategy σ^*_i. Hence, it must be that $|N(\theta)| \leq \frac{n}{2} - 1$.

Next, I show that if σ^* is a 1-FTNE and if player i uses a non-trivial mixed strategy under σ^*, then it must be that $q(\sigma^*_{-i}, \theta) = 0$. To see this, note that if i is pivotal, i's announcement of θ will lead to alternative $f(\theta)$ being implemented only in situations in which

the announcement of i's opponents $\mathbf{s}_{-i}$ is such that $|R(\theta|(\mathbf{s}_{-i},\theta))| > \frac{n}{2} \geq |R(\theta|(\mathbf{s}_{-i},\theta))| - 1$. By the previous paragraph, the number of players that announce θ with probability 1 under σ^* is weakly lower than $\frac{n}{2} - 1$. Moreover, equation (A.4) implies that those players who use a non-trivial mixed strategy under σ^* will make a truthful announcement θ with probability zero. Therefore, if i's opponent's play according to σ^*_{-i}, the probability that $\mathbf{s}_{-i}$ is such that $|R(\theta|(\mathbf{s}_{-i},\theta))| > \frac{n}{2} \geq |R(\theta|(\mathbf{s}_{-i},\theta))| - 1$ is zero. Hence, it must be that $q(\sigma^*_{-i},\theta) = 0$.

The fact that $q(\sigma^*_{-i},\theta) = 0$, together with equation (A.4), implies that for all $\theta' \in \text{supp}\sigma^*_i$, $q(\sigma^*_{-i},\theta')(\widetilde{u}_i(f(\theta'),\theta) - \widetilde{u}_i(a^*,\theta)) > 0$. ■

Theorem A1 shows that, for generic payoffs, mechanism g_M implements any social choice function in 1-FTNE even when we consider mixed strategies. The following example clarifies why the genericity of payoffs is required for the result.

Example A1 Consider the following setting: $N = \{1,2,3,4,5\}$, $\Theta = \{\theta', \theta'', \theta'''\}$ and $A = \{a, b, a^*\}$. For all $i \in N$, $\widetilde{u}_i(a,\theta''') = \widetilde{u}_i(b,\theta''') = 1$ and $\widetilde{u}_i(a^*,\theta''') = 0$. The social choice function that the planner wishes to implement is such that $f(\theta''') = a^*$, $f(\theta') = a$ and $f(\theta'') = b$.

Suppose the planner uses mechanism g_M and let $a^* = f(\theta''')$ be the alternative that this mechanism implements when no state is announced by more than half the population. Consider the mixed strategy profile $\sigma^* = (\sigma^*_1, \sigma^*_2, \sigma^*_3, \sigma^*_4, \sigma^*_5)$ such that, for all $i \in N$, $\sigma^*_i(\theta') = \sigma^*_i(\theta'') = 1/2$. I now show that, for these specific payoffs, σ^* is a 1-FTNE of the strategic game that mechanism g_M induces at state θ'''. However, any minimal variation in the material payoffs of any agent is enough for the strategic game that mechanism g_M induces at state θ''' to have no mixed strategy 1-FTNE.

Note that under strategy profile σ^*, $q(\sigma^*_{-i},\theta') = q(\sigma^*_{-i},\theta'')$ for all $i \in N$ (recall that for each $\theta \in \Theta$, $q(\sigma^*_{-i},\theta)$ is the probability with which, conditional on i being pivotal, i's announcement of θ leads to alternative $f(\theta)$ being implemented – see the proof of Theorem A1 for more details). This, together with the fact that $\widetilde{u}_i(a,\theta''') = \widetilde{u}_i(b,\theta''') = 1$, implies that for all $i \in N$,

$$q(\sigma^*_{-i},\theta')(\widetilde{u}_i(f(\theta'),\theta''') - \widetilde{u}_i(a^*,\theta''')) = q(\sigma^*_{-i},\theta'')(\widetilde{u}_i(f(\theta''),\theta''') - \widetilde{u}_i(a^*,\theta''')) > 0. \qquad \text{(A.5)}$$

Equation (A.5) in turn implies that, given σ^*_{-i}, player i is materially indifferent between announcing θ' or θ'' when the true state of nature is θ'''. Moreover, given σ^*_{-i} player i strictly prefers to announce either of these than to announce the truthful state θ'''.

For each $i \in N$, let $\mathcal{F}_i$ be the set of all subsets of $N\backslash\{i\}$ that have cardinality 2. The

probability $q(\sigma_{-i}, \theta')$ is proportional to $\sum_{B \in \mathcal{F}_i} [\Pi_{j \in B} \sigma_j(\theta') \Pi_{j \notin B}(1 - \sigma_j(\theta'))]$.[16] For each $i \in N$, let $p_i = \sigma_i(\theta')$. Then, $q(\sigma_{-1}, \theta')$ is proportional to:

$$\begin{aligned} & \sum_{B \in \mathcal{F}_1} [\Pi_{j \in B} \sigma_j(\theta') \Pi_{j \notin B}(1 - \sigma_j(\theta'))] \\ = \; & p_2 p_3 (1-p_4)(1-p_5) + p_2 p_4 (1-p_3)(1-p_5) + p_2 p_5 (1-p_3)(1-p_4) + \\ & p_3 p_4 (1-p_2)(1-p_5) + p_3 p_5 (1-p_2)(1-p_4) + p_4 p_5 (1-p_2)(1-p_3). \end{aligned} \tag{A.6}$$

The probabilities $q(\sigma_{-i}, \theta')$ for $i \neq 1$ can be computed similarly.

Using equation (A.6) one can check that, for all $j \in N \backslash \{1\}$, the derivative of $q(\sigma_{-1}, \theta')$ with respect to p_j is equal to zero when all players $i \in N \backslash \{1, j\}$ are using strategy $\sigma_i^*(\theta') = p_i = 1/2$. That is, changing the strategy of any $j \neq 1$ has no effect on $q(\sigma_{-1}^*, \theta')$ when every player $i \in N \backslash \{1, j\}$ uses strategy σ_i^*. Given the symmetry in the players' strategies, the same is true for $q(\sigma_{-1}^*, \theta'')$. Therefore, for any σ_{-1} such that $d((\sigma_{-1}, \sigma_1^*), \sigma^*) \leq 1$, player 1 finds it optimal to announce either θ' or θ'' when the other players are announcing σ_{-1}. The symmetry of the environment implies that the same is true for all players $i \neq 1$, and so $\sigma^* \in \tilde{E}_1(g_M, \theta''')$.

The crucial property of σ^* that makes it a 1-FTNE is that, under this strategy profile, the probabilities $q(\sigma_{-i}^*, \theta')$ and $q(\sigma_{-i}^*, \theta'')$ remain constant as we change the strategy of any single agent $j \neq i$. Note that finding a strategy profile σ^* such that for all $i \in N$ and for all $j \in N \backslash \{i\}$, $q(\sigma_{-i}^*, \theta')$ is constant on $\sigma_j^*(\theta')$, involves finding mixing probabilities $\sigma_k^*(\theta')$ for all $k \in N$ that solve a system of equations (the equations of this system are, for each $i \in N$ and each $j \in N \backslash \{i\}$, the derivative of $q(\sigma_{-i}^*, \theta')$ with respect to $\sigma_j^*(\theta')$).[17]

The agents' preferences in the current example were constructed as follows: (i) find a strategy profile σ^* such that $q(\sigma_{-i}^*, \theta') > 0$ and $q(\sigma_{-i}^*, \theta'') > 0$ remain constant as we change the strategy of any single agent $j \neq i$; (ii) find material preferences such that all players are indifferent between announcing θ' and θ'' given $q(\sigma_{-i}^*, \theta')$ and $q(\sigma_{-i}^*, \theta'')$, and strictly prefer to announce either of these than any other state. Clearly, for any strategy profile σ^* that satisfies (i), the set of material preferences that satisfy (ii) is non-generic.

[16] Indeed, $\sum_{B \in \mathcal{F}_i} \left[\Pi_{j \in B} \sigma_j(\theta') \Pi_{j \notin B}(1 - \sigma_j(\theta'))\right]$ is the probability with which exactly two of i's opponents announce θ'. In this case, i is pivotal and her announcement of θ' leads to $f(\theta')$ being implemented.

[17] It can be shown that, in this setting with five players, the only real solutions to this system of equations satisfy $\sigma_k(\theta) = p \in \{0, 1/2, 1\}$ for all $k \in N$ and all $\theta \in \Theta$.

A.4 Fault tolerance and stochastic stability

Theorems 1 and 2 together imply that, under white lie aversion, the solution concepts of fault tolerant Nash equilibrium and stochastically stable equilibrium give the same unique prediction for the game that mechanism g_M induces. In this appendix, I give an example of a game in which fault tolerance and stochastic stability yield different predictions. Therefore, these two solution concepts are logically independent, and neither of them implies the other.

To see this, consider the following example from Ellison (2000). There is a finite set of players $N = \{1, 2, ..., n\}$, with n odd. At every period $t = 0, 1, 2, ...$, each player $i \in N$ is randomly matched (with uniform probabilities) with some other player to play the following symmetric strategic game:

	A	B	C
A	$1, 1$	$0, 0$	$0, 0$
B	$0, 0$	$-4, -4$	$3, 3$
C	$0, 0$	$3, 3$	$-4, -4$

Given $\mathbf{s}_{-i}$, player i's payoff from playing s_i is $\frac{1}{n-1} \sum_{j \in N \setminus \{i\}} u(s_i, s_j)$. Players can revise their strategies in every period, so that in every t each player chooses a best reply to the strategy profile played $t - 1$. Ellison (2000) showed that the stochastically stable equilibria of this game are $\mathbf{s}^B$ and $\mathbf{s}^C$ (where $\mathbf{s}^B = (B, B, ..., B)$, and similarly for $\mathbf{s}^A$ and $\mathbf{s}^C$), provided the number of players is large enough. That is, in the long run we should expect to see agents alternating between playing B and C. However, neither $\mathbf{s}^B$ nor $\mathbf{s}^C$ are fault tolerant Nash equilibria of the static random matching game. In fact, one can check that for k small enough, $\mathbf{s}^A$ is a k-FTNE of the static random matching game.

References

[1] Abreu, D. and H. Matsushima (1992): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica* Vol. 60, No. 5, pp. 993-1008.

[2] Aghion, P., D. Fudenberg, R. Holden, T. Kunimoto and O. Tercieux (2012): "Subgame-Perfect Implementation under Information Perturbations," *Quarterly Journal of Economics*, Vol. 127, pp. 1843-1881.

[3] Cabrales, A. and R. Serrano (2011a): "Implementation in Adaptive Better-response Dynamics: Towards a General Theory of Bounded Rationality in Mechanisms," *Games and Economic Behavior*, Vol. 73, pp. 360-374.

[4] Cabrales, A. and R. Serrano (2011b): "Stochastically Stable Implementation," unpublished manuscript.

[5] Chung, K. S, and J. C. Ely (2003): "Implementation with Near-Complete Information," *Econometrica*, Vol. 71, pp. 857-871.

[6] Dutta, B. and A. Sen (2012): "Nash Implementation with Partially Honest Individuals," *Games and Economic Behavior*, Vol. 74, pp. 154-169.

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

[8] Ellison, G. (1993): "Learning, Local Interaction and Coordination," *Econometrica* Vol. 61, pp. 1047-1071.

[9] Ellison, G. (2000): "Basins of Attraction, Long-Run Stochastic Stability, and the Speed of Step-by-Step Evolution," *Review of Economic Studies* Vol. 67, pp. 17-45.

[10] Foster, D. and P. Young (1990): "Stochastic Evolutionary Game Dynamics," *Theoretical Population Biology*, Vol. 38, pp. 219-232.

[11] Freidlin, M. I. and A. D. Wentzell (1984): *Random Perturbations of Dynamical Systems*. New York: Springer Verlag.

[12] Holden, R., N. Kartik, and O. Tercieux (2014): "Simple Mechanisms and Preferences for Honesty," *Games and Economic Behavior*, Vol. 83, pp. 284-290.

[13] Kandori, M., G. Mailath, and R. Rob (1993): "Learning, Mutations and Long Run Equilibria in Games," *Econometrica* Vol. 61, pp. 29-56.

[14] Kartik, N., and O. Tercieux (2012): "Implementation with Evidence," *Theoretical Economics* , Vol. 7, pp. 323-355.

[15] Jackson, M. (1992): "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *Review of Economic Studies*, Vol. 59, pp. 757– 775.

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

[17] Matsushima, H. (2008a): "Behavioral Aspects of Implementation Theory," *Economic Letters* Vol. 100, pp. 161-164.

[18] Matsushima, H. (2008b): "Role of Honesty in Full Implementation," *Journal of Economic Theory* Vol. 139, pp. 353-359.

[19] Moore, J., and R. Repullo (1988): "Subgame Perfect Implementation," *Econometrica*, Vol. 56, pp. 1191-1220.

[20] Palfrey, T. R. and S. Srivastava (1991): "Nash Implementation Using Undominated Strategies," *Econometrica*, Vol. 59, pp. 479-501.

[21] Sandholm, W. (2007): "Pigouvian Pricing and Stochastic Evolutionary Implementation," *Journal of Economic Theory* Vol. 132, pp. 367-382.

[22] Young, P. (1993): "The Evolution of Conventions," *Econometrica* Vol. 61, pp. 57-84.