Direct Implementation with Minimally Honest Individuals

Juan M. Ortner*
Princeton University
December 2010

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.

1 Introduction

Consider the problem of a planner who wishes to implement the alternative prescribed by some social choice function $f : \Theta \rightarrow A$, where $\Theta$ is the set of possible states of nature and $A$ is the set of 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.

*I am especially indebted to Faruk Gul for his advice and guidance, and to Satoru Takahashi for his encouragement and support at the early stages of this project. I would also like to thank Avidit Acharya, Benjamin Brooks, Edoardo Grillo and Stephen Morris for helpful feedback. All remaining errors are my own. Address: Department of Economics, Fisher Hall, Princeton University, Princeton NJ 08544-1021. E-mail: jortner@princeton.edu.
The most natural class of mechanisms discussed in the literature are direct mechanisms, under which the planner asks each player 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. Given that the planner has no control over which equilibrium obtains, she cannot rely on direct mechanisms to implement a given social choice function.

The mechanism design 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. Probably the most notable example of this sort of 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 generate 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 take an alternative approach to the implementation problem. I assume that agents are *white lie averse*, in the sense that they strictly prefer to tell the truth whenever lying has no effect on the implemented alternative. Put differently, I assume that agents have a minimal degree of honesty, since they prefer to make a truthful announcement than to tell a white lie. 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 equilibrium*. The idea behind fault tolerant equilibrium is that players may not know whether all of their opponents are rational.\(^2\) If a player believes that some of her opponents may fail to act optimally, she may want to reduce the sensitivity of her payoffs to changes in the behavior of irrational players. Thus, for a strategy profile \(s^*\) to be a fault tolerant equilibrium, it must be the case that each player has a strict incentive to play her equilibrium action, even when a fraction of her opponents deviate from equilibrium behavior. A strategy profile \(s^*\) must also satisfy a second condition

---

1See Jackson (1992) for an elaboration of this and related points.
2Fault tolerant equilibrium is closely related to \(k\)-fault tolerant Nash equilibrium, introduced by Eliaz (2002). See Section 3 for more on the connection between these two equilibrium notions.
to be a fault tolerant equilibrium: for every strategy profile $s \neq s^*$, there must be at least one player who would have a strict incentive to change her action if she believed that there was one single irrational agent among her opponents.

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 play 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 construct 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 $\theta$, the mechanism’s outcome is $f(\theta)$. In any other case, the outcome is $a^*$, the status quo. The strategic game that this majoritarian mechanism induces has multiple Nash 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. Importantly, in the appendix 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 one.

This paper provides two distinct justifications for the use of simple direct mechanisms, each based on a different equilibrium concept. Suppose first that each agent believes that some of her opponents may fail to behave optimally, but she knows neither the number of irrational players, nor their identity, 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. Fault tolerant equilibrium selects the strategy profile that is most robust to these deviations, and as such it is good prediction of how rational players will behave in the presence of irrational agents. Theorem 1 tells us that, in this setting, a social planner can use a majoritarian direct mechanism to implement the desired alternative, provided there are at least five white lie averse agents.

The refinement of stochastic stability provides an evolutionary justification for simple majoritarian mechanisms. Suppose a group of agents will repeatedly play the 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. That is, the stochastically stable equilibria are the patterns of behavior that would arise in the long run. 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 next subsection presents an overview of the related literature. Section 2 formally describes the implementation problem and introduces the majoritarian mechanism. Section 3 introduces fault tolerant equilibrium and proves that the majoritarian mechanism implements any social choice function in fault tolerant equilibrium when players are white lie averse. Section 4 shows that under white lie aversion, the majoritarian mechanism also implements any social choice function in stochastically stable equilibrium. Section 5 presents some concluding remarks. Some proofs appear in the Appendix.

1.1 Related literature

The idea of studying implementation when agents have a slight preference for being honest 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 if the planner can impose small fines on the agents, then 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 making honest announcements and shows that any incentive compatible social choice function can be fully implemented. Dutta and Sen (2009) 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.

Eliaz (2002) studies complete information implementation when each player believes that at most $k$ of her opponents may fail to act rationally. He introduces $k$-fault tolerant Nash equilibrium ($k$-FTNE), which explicitly incorporates these beliefs. Indeed, at a $k$-FTNE players have no incentive to change their actions even when irrational players deviate from equilibrium behavior. The paper shows that any social choice correspondence satisfying $k$-monotonicity (an adaptation of Maskin monotonicity to this environment) and no veto power can be implemented in $k$-FTNE.

The present paper uses tools on stochastic dynamic systems developed by Freidlin and Wentzell (1984). Kandori, Mailath and Rob (1993) and Young (1993) introduced these techniques to the economics literature. These papers consider a situation in which a finite
group of players repeatedly plays a 2 × 2 coordination game. For the most part, players choose their strategies as best responses to the action profile played last period (or, in the case of Young (1993), to the action profiles played in a sample of the recent periods), but they occasionally make mistakes. The authors show that players will coordinate their actions most of the time on the risk-dominant equilibrium when the probability of these mistakes is low.

Sandholm (2007) appears to be the first paper to study implementation in stochastically stable equilibrium. He focuses on a simple environment with externalities in which a planner wants the agents to choose an utilitarian action profile (i.e., an action profile that maximizes the sum of their utilities) and shows that the planner can achieve this objective by introducing a simple tax scheme under which each agent pays for the externalities she creates. I stress, however, that Sandholm (2007) uses techniques that are different from the ones I use here. First, Sandholm (2007) considers a logit best-reply dynamic, whereas I consider a perturbed best-reply dynamic. Moreover, he shows that the game that results from introducing the tax scheme is a potential game (see Monderer and Shapley (1996)), with potential function equal to the sum of the players’ payoffs. He then appeals to a result by Blume (1997) to show that the stochastically stable equilibria of this game are the action profiles that maximize the total payoff.

2 Model

Let \( N = \{1, \ldots, n\} \) be a finite set of agents and let \( A \) be a set of possible alternatives. Let \( \Theta \) 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 social planner is interested in implementing a social choice function \( f : \Theta \to A \). Thus, the social planner’s objective is to design a strategic game form \( \langle N, (S_i)_{i \in N}, g \rangle \) that implements \( f \), 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 social 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 \( m \in M \). Most of the complete information implementation literature uses mechanisms that fall into

---

3Cabrales and Serrano (2008) 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.
this category. A mechanism $g$ is a direct mechanism if $S_i = \Theta$ for all $i \in N$ (i.e., if $M = \emptyset$).

I assume that agents are white lie averse. Let $g : S \to A$ be a mechanism, where $S = (\Theta \times M)^n$ is the set of possible announcement profiles. To capture white lie aversion, I assume that $u_i : A \times \Theta \times S \to \mathbb{R}$ for every $i \in N$, where $u_i$ is agent $i$’s utility function. In words, I assume that the agents’ utility depends not only on the implemented alternative and the state of nature, but also on the announced messages.

For each $i$, let $\tilde{u}_i : A \times \Theta \to \mathbb{R}$ and let $\eta > 0$ be a small number. The function $\tilde{u}_i$ will reflect the material payoffs of player $i$, while $\eta$ will represent the cost of lying. For any announcement profile $s \in (\Theta \times M)^n$, let $s_{-i}$ denote the announcements of all players in $N \setminus \{i\}$. I incorporate white lie aversion as a utility perturbation:

**Assumption 1 (white lie aversion)** Suppose $\theta \in \Theta$ is the true state of nature and let $g$ be a mechanism. For every agent $i \in N$, if $m_i$ and $s_{-i}$ are such that $g((\theta', m_i), s_{-i}) = a \forall \theta' \in \Theta$ for some $a \in A$, then

$$u_i(a, \theta, (s_i, s_{-i})) = \begin{cases} \tilde{u}_i(a, \theta) & \text{if } s_i = (\theta, m_i), \\ \tilde{u}_i(a, \theta) - \eta & \text{if } s_i = (\theta', m_i) \text{ with } \theta' \neq \theta. \end{cases}$$

For any other $m_i, s_{-i}$, $u_i(g(s_i, s_{-i}), \theta, (s_i, s_{-i})) = \tilde{u}_i(g(s_i, s_{-i}), \theta) \forall s_i \in \Theta \times \{m_i\}$.

Assumption 1 states that every agent $i \in N$ strictly prefers to tell the truth than to lie if $m_i$ and $s_{-i}$ are such that she cannot change the implemented outcome by changing her announcement of the state of nature. That is, given such $m_i, s_{-i}$, player $i$ incurs a small cost $\eta$ if she lies and announces some $\theta' \neq \theta$. Assumption 1 implies that agents have a minimal degree of honesty, since they dislike telling lies whenever those lies do not benefit them.

The following assumption imposes a very mild condition on the set of alternatives:

**Assumption 2** There exists $a^* \in A$ such that $f(\phi) \neq a^*$ for all $\phi \in \Theta$.

Assumption 2 states that there exists at least one element in $A$ that is never prescribed by the social choice function $f$. For instance, Assumption 2 will always hold if the cardinality of $A$ is strictly larger than the cardinality of $\Theta$.

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 $\phi \in \Theta$ and for every $s = (s_1, \ldots, s_n) \in \Theta^n$, define

$$R^* (\phi | s) := \{ i \in N \text{ s.t. } s_i = \phi \}.$$
In words, $R^* (\phi \mid s)$ is the set of agents who reported state $\phi$ in $s$. The following condition characterizes mechanism $g_M$:

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

Mechanism $g_M$ is a majoritarian mechanism: if strictly more than half the agents announce a state $\phi \in \Theta$, the mechanism implements alternative $f (\phi)$. Otherwise, the mechanism implements alternative $a^*$, which we can interpret as the status quo.

## 3 Implementation in fault tolerant equilibrium

In this section, I introduce 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.

The idea behind fault tolerant equilibrium is that players may not know whether all of their opponents are rational. Solution concepts like Nash equilibrium require players to respond optimally to the equilibrium strategies of their opponents. However, the predictions of these equilibrium notions will in general not be robust to the presence of irrational players. Suppose an agent believes that some of her opponents may fail to act rationally, but knows neither the number of irrational players, nor their identity, nor how irrational players behave. Given her beliefs, she may want to adjust her action to reduce the sensitivity of her payoffs to possible changes in the behavior of irrational players. Fault tolerant equilibrium selects the strategy profile that is most robust to these deviations, and as such it is a good prediction of how rational players will behave in the presence of irrational agents.

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

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

In words, the distance $d (s, s')$ between $s$ and $s'$ is given by the number of players who choose different actions under both strategy profiles.

For any $b \in \mathbb{R}$, let $\lfloor b \rfloor_-$ denote the largest integer strictly smaller than $b$. Given a finite set of players $N = \{1, 2, \ldots, n\}$, define $k^* (N) := \max\{\lfloor \frac{n}{2} \rfloor_-, 1\}$. Note that $k^* (N) =
\[ \left\lfloor \frac{n}{2} - 1 \right\rfloor \geq 1 \] whenever \( n \geq 5 \). Let \( S := \Pi_{i \in N} S_i \), and for every \( i \in N \) let \( S_{-i} := \Pi_{i \in N \setminus \{i\}} S_i \).

**Definition 1** A strategy profile \( s^* = (s_1^*, \ldots, s_n^*) \in S \) is a fault tolerant equilibrium of the strategic game \( \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle \) if:

1. \( \forall i \in N \), \( u_i(s_i^*, s_{-i}) > u_i(s_i', s_{-i}) \) for all \( s_{-i} \in \{s_{-i} : \forall i \neq k : (s_i, s_{-i}) \leq k^*(N)\} \) and for all \( s_i' \in S_i \setminus \{s_i^*\} \).

2. For every \( s = (s_1, \ldots, s_n) \in S \setminus \{s^*\} \), there exists \( i \in N \), \( s_{-i}' \in S_{-i} \) with \( d((s_i, s_{-i}'), s) \leq 1 \) and \( \tilde{s}_i \in S_i \setminus \{s_i\} \) such that \( u_i(s_i, s_{-i}') > u_i(s_i, s_{-i}', \tilde{s}_i, \{s_i\}) \).

The first condition in the definition of fault tolerant equilibrium requires each agent to play a strictly optimal strategy against any action profile of her opponents with the property that at least \( n − k^*(N) − 1 \) are playing their equilibrium actions. Put differently, in a fault tolerant equilibrium each player has a strict incentive to play her equilibrium action even in the presence of \( k^*(N) \) irrational agents.

Besides this, a fault tolerant equilibrium must also satisfy a second condition. Let \( s^* = (s_1^*, \ldots, s_n^*) \) be a fault tolerant equilibrium and let \( s = (s_1, \ldots, s_n) \) be any other strategy profile. Then, there must exist some player \( i \in N \) for which the following holds: even if player \( i \) believes that \( n − 2 \) of her opponents are playing according to \( s \), there must exist a strategy profile \( s_{-i}' \) of her opponents that is consistent with these beliefs, against which player \( i \) strictly prefers to take some action \( \tilde{s}_i \neq s_i \). In other words, if \( s^* \) is a fault tolerant equilibrium and agents are playing according to \( s \neq s^* \), then there must exist at least one player who would have a strict incentive to change her behavior if she believed that one of her opponents is irrational. This condition implies that the fault tolerant equilibrium \( s^* \) is more robust to possible deviations of behavior than any other strategy profile. Indeed, the presence of one single irrational agent is enough to upset any strategy profile \( s \neq s^* \).

The two conditions in definition 1 treat the different strategy profiles in a very asymmetric way. The first condition imposes a strong robustness requirement: a fault tolerant strategy profile should remain a strict equilibrium even in the presence of \( k^*(N) \) irrational agents. The second condition, on the other hand, postulates that strategy profiles different from the fault tolerant equilibrium are not stable when players believe that there is one single irrational agent among them. Taken together, these two conditions imply that a game can have at most one fault tolerant equilibrium: if \( s^* \) is a fault tolerant equilibrium, then no strategy profile \( s \neq s^* \) can satisfy the first condition in definition 1.

The notion of fault tolerant equilibrium is closely related to the \( k \)-fault tolerant Nash equilibrium (\( k \)-FTNE) introduced by Eliaz (2002). Consider a setup in which each player
believes that at most $k$ of her opponents are irrational. A $k$-FTNE describes a stable pattern of behavior for such a setup: for a strategy profile $s'$ to be a $k$-FTNE, each agent must be playing a weakly optimal strategy against any profile of actions of her opponents such that at least $n - k - 1$ are playing according to $s'$. In other words, for any given $k \geq 0$, a strategy profile $s'$ is a $k$-FTNE if the first condition in definition 1 holds, with $k$ instead of $k^*(N)$ and with weak inequalities instead of strict ones. The concept of fault tolerant equilibrium is a strengthening of $k$-FTNE. Indeed, a strategy profile $s^*$ is a fault tolerant equilibrium if it is a strict $k^*(N)$-FTNE and there is no other $k$-FTNE with $k \geq 1$.

The assumption underlying $k$-FTNE is that rational agents know that there can be at most $k$ irrational players in the population. However, even if rational players believe that some of their opponents may fail to act optimally, they may still be uncertain about the actual number of irrational players. In such a setup, it seems reasonable to assume that rational players will coordinate their actions at the strategy profile that is most robust to the presence of irrational agents; and this is precisely the strategy profile that fault tolerance selects.

Consider next the environment of Section 2, and let $\langle N, (S_i)_{i \in N}, g \rangle$ be a strategic game form. Let $E_F(g, \theta)$ denote the set of fault tolerant equilibria of the strategic game that mechanism $g$ induces when the state is $\theta$. Note that $E_F(g, \theta)$ is either empty or has a unique element, since a game can have at most one fault tolerant equilibrium.

**Definition 2** A mechanism $g$ implements the social choice function $f : \Theta \rightarrow A$ in fault tolerant equilibrium if for all $\theta \in \Theta$, (i) $E_F(g, \theta)$ is non-empty and (ii) $g(E_F(g, \theta)) = f(\theta)$.

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

**Theorem 1** Let $f : \Theta \rightarrow A$ be a social choice function and suppose that Assumptions 1 and 2 hold. If $n \geq 5$, mechanism $g_M$ implements $f$ in fault tolerant equilibrium. Moreover, $E_F(g_M, \theta) = \{s^\phi\}$ for every $\theta \in \Theta$.

**Proof.** Suppose that $\theta$ is the true state of nature. If there are at least five white lie averse agents, then the announcement profile $s^\phi$ satisfies the first condition in definition 1. Moreover, it is clear that $g_M(s^\phi) = f(\theta)$ for every $\theta \in \Theta$.

To check that the second condition in definition 1 also holds, consider first announcement profiles $s \neq s^\phi$ such that $|R^*(\theta | s)| \geq \frac{n}{2}$. Note that in this case there always exists
an agent $i \in N$ with $s_i \neq \theta$ and an announcement profile $s'_{-i}$ of $i$'s opponents such that $|R^* (\theta \mid s_i, s'_{-i})| \geq \frac{n}{2} + 1$ and $d ((s_i, s'_{-i}), s) \leq 1$. Under Assumption 1, player $i$ strictly prefers to announce $\theta$ than to announce any $\theta' \neq \theta$ when her opponents are playing according to $s'_{-i}$.

Next, consider announcement profiles $s$ such that there exists $\phi \neq \theta$ with $|R^* (\phi \mid s)| > \frac{n}{2}$. In this case, there always exists an agent $i \in N$ with $s_i = \phi$ and a strategy profile $s'_{-i}$ of $i$'s opponents such that $|R^* (\phi \mid s_i, s'_{-i})| > \frac{n}{2} + 1$ and $d ((s_i, s'_{-i}), s) \leq 1$. If player $i$'s opponents are playing $s'_{-i}$, then player $i$ strictly prefers to announce $\theta$ than to continue announcing $\phi$ (this again follows from white lie aversion).

Consider next message profiles $s$ such that $|R^* (\theta' \mid s)| = |R^* (\theta'' \mid s)| = \frac{n}{2}$ for $\theta', \theta'' \in \Theta$. I already considered the case with $|R^* (\theta \mid s)| = \frac{n}{2}$ 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 $s'_{-i}$ with $d ((s_i, s'_{-i}), s) = 1$ such that $|R^* (\theta'' \mid s_i, s'_{-i})| = \frac{n}{2} + 1$ and $|R^* (\theta' \mid s_i, s'_{-i})| = \frac{n}{2} - 1$ for $\theta', \theta'' \neq \theta$. Note that in $(s_i, s'_{-i})$ 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 $s'_{-i}$, so under white lie aversion she strictly prefers to announce $\theta$ than any other state.

Finally, consider announcement profiles $s$ such that $|R^* (\theta' \mid s)| \leq \frac{n}{2}$ for all $\theta' \in \Theta$, but with at most one $\theta' \neq \theta$ such that $|R^* (\theta' \mid s)| = \frac{n}{2}$ and with $|R^* (\theta \mid s)| < \frac{n}{2}$. There are two cases to consider: (i) there exists $\phi \in \Theta$ such that $|R^* (\phi \mid s)| + 1 > \frac{n}{2}$, and (ii) no such $\phi$ 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 $\theta$ than any $\theta' \in \Theta \setminus \{\theta\}$.

Consider next case (i). Note that at least three states are being announced in $s$. Define $Y (s) := \{ \phi \in \Theta : |R^* (\phi \mid s)| + 1 > \frac{n}{2} \}$. Since at least three states are being announced in $s$, there always exists $i \in N$ with $s_i \neq \theta$ such that $Y (s) \setminus \{s_i\}$ is non-empty (so player $i$ is not announcing the true state, and there exists a state in $Y (s)$ different from the state that player $i$ is announcing). Moreover, there exists an announcement profile $s'_{-i}$ of $i$'s opponents with $d ((s_i, s'_{-i}), s) = 1$ such that $|R^* (\phi \mid s_i, s'_{-i})| > \frac{n}{2}$ for some $\phi \in Y (s) \setminus \{s_i\}$. Given $s'_{-i}$, 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 $\theta$ than any $\theta' \in \Theta \setminus \{\theta\}$.

Theorem 1 states that mechanism $g_M$ implements any social choice function in fault tolerant equilibrium, provided there are at least five white lie averse agents. In other words, in this setup a planner can dispense with any augmentation of the message space and focus
exclusively on the more natural direct mechanisms to implement any social choice function. I stress that, besides its simplicity, mechanism $g_M$ has the attractive feature that it will still implement the desired alternative even in the presence of at most $\left\lfloor \frac{n}{2} - 1 \right\rfloor$ irrational agents. That is, $g_M(s) = f(\theta)$ for all $s \in \{s' \in S : d(s', s^\theta) \leq \left\lfloor \frac{n}{2} - 1 \right\rfloor\}$.

To see why I need $n \geq 5$ in the proof of Theorem 1, suppose $\theta$ is the true state of nature and assume $N = \{1, 2, 3\}$ (similar examples are available when $n = 4$). Suppose further that there exists $\phi \in \Theta \setminus \{\theta\}$ such that, for $i = 2, 3$, $\widetilde{u}_i(f(\phi), \theta) > \widetilde{u}_i(b, \theta)$ for all $b \in A \setminus \{f(\phi)\}$. In this example, the truthful announcement $s^\theta$ is not a fault tolerant equilibrium. To see this, let $s' = (\theta, \theta, \phi)$ and note that $d(s', s^\theta) = 1 = k^*(N)$. The assumptions on preferences imply that $u_2(g(\theta, \phi, \theta, \theta, \phi, \phi)) > u_2(g(s'), \theta, s')$. Therefore, strategy profile $s^\theta$ doesn’t satisfy the first condition in the definition of fault tolerant equilibrium.

Finally, note that the assumption that players are white lie averse is crucial for Theorem 1. This assumption guarantees that truth-telling is a strict Nash equilibrium of the strategic game that the majoritarian mechanism induces, as it implies that players have a strict incentive to tell the truth when most of the other players are also making truthful announcements. If this assumption did not hold, then the majoritarian mechanism would not implement any social choice function in fault tolerant equilibrium. Indeed, if Assumption 1 did not hold then in the truthful announcement profile all agents would be indifferent between telling the truth and announcing any other state of nature.

## 4 Implementation in stochastically stable equilibrium

In this section, I first formally present 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 white lie averse.

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 interact among each other.

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 $i \in N$, and assume that players in $N$ repeatedly play $\langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle$ at $t = 0, 1, 2, \ldots$. Let $s_{i,t}$ denote the strategy played by agent $i$ at period $t$ and $s_t = (s_{1,t}, \ldots, s_{n,t})$ the strategy profile played at period $t$. For every $i \in N$, let $s_{-i,t}$ denote the strategy profile played by all agents but $i$ in period $t$. 11
I now present the assumptions on how behavior evolves in this setup. At \( t = 0 \), agents play according to some \( s_0 \in \Pi_{i=1}^n S_i \). 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.\(^4\) 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 \( \max_{s_i \in S_i} u_i (s_i, s_{-i,t-1}) \). With probability \( \varepsilon \) she trembles/mutates and plays any \( s_i \in S_i \) with positive probability.

Given an initial strategy profile \( s_0 \), 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 \( P^\varepsilon \) denote the transition matrix of this Markov process. Following the literature on stochastic evolutionary game theory, in this section I will refer to each \( s \in S \) as a "state" of the Markov process. The state \( s \in S \) of the Markov process should not be confused with the "state of nature" \( \theta \in \Theta \).

Let \( \tilde{s} \in S \) 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 \( P^\varepsilon \) has an invariant distribution \( \mu^\varepsilon \) such that \( \mu^\varepsilon (\tilde{s}) = 1 \) (i.e., \( \mu^\varepsilon \) puts all its mass on the strict Nash equilibrium \( \tilde{s} \)). Therefore, if \( \varepsilon = 0 \) the matrix \( P^\varepsilon \) 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 \( P^\varepsilon \) is aperiodic and irreducible and therefore has a unique invariant distribution \( \mu^\varepsilon \). One can show that \( \mu^* := \lim_{\varepsilon \to 0} \mu^\varepsilon \) exists.\(^5\)

**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 \( s \in S \) is a stochastically stable equilibrium of \( \langle N, (S_i)_{i \in N}, (u_i)_{i \in N} \rangle \) if \( s \in \text{supp} \mu^* \).

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. I again assume that the realization of \( \theta \) 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 \( P^\varepsilon (\theta) \) denote the transition matrix over \( S \) when the state of nature is \( \theta \), and let \( \mu^\varepsilon (\theta) \) denote its invariant distribution (which is unique for every \( \varepsilon > 0 \)). For every \( \theta \in \Theta \), let \( \mu^* (\theta) := \lim_{\varepsilon \to 0} \mu^\varepsilon (\theta) \).

---

\(^4\)This assumption implies that at any given period and for any \( Y \subseteq N \) there is positive probability that only agents in \( Y \) receive an opportunity to revise their strategies.

\(^5\)See Young (1993) for a proof of this result.
Definition 4 A mechanism $g$ implements the social choice function $f : \Theta \rightarrow A$ in stochastically stable equilibrium if for all $\theta \in \Theta$, $s \in \text{supp}\mu^*_g(\theta)$ implies $g(s) = f(\theta)$.

Given a mechanism $g$ and a state of nature $\theta \in \Theta$, the Markov process over $S$ is ergodic when $\varepsilon > 0$. Therefore, the invariant distribution $\mu^*_g(\theta)$ describes with probability 1 the fraction of time that the Markov process spends at each $s \in S$. This implies that if a mechanism $g$ 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 $\varepsilon$ of mutations is small.

I now present the main result of this section:

Theorem 2 Let $f : \Theta \rightarrow A$ be a social choice function and suppose that Assumptions 1 and 2 hold. If $n \geq 5$, mechanism $g_m$ implements $f$ in stochastically stable equilibrium. Moreover, $\text{supp}\mu^*_{g_m}(\theta) = \{s^\theta\}$ 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 important and useful properties of the set of stochastically stable equilibria. I now present a brief overview of these methods.

For two states $s$ and $s'$ of the Markov process, define the resistance/cost of the transition $s \rightarrow s'$ as the number of mutations needed to complete the transition. Similarly, define a path between states $s$ and $s'$ as a sequence of states $(s^1, s^2, ..., s^k)$ such that $s^r$ is an immediate predecessor of $s^{r+1}$ and such that $s^1 = s$ and $s^k = s'$. The cost of the path $(s^1, s^2, ..., s^k)$ is the sum of the resistance of its transitions. Let $c(s^1, s^2, ..., s^k)$ denote the cost of path $(s^1, s^2, ..., s^k)$.

Formally, for any $\theta \in \Theta$, any initial $s_0$ and any $s \in S$,

$$T^{-1} \sum_{t=0}^{T} 1_s(s_t) \rightarrow \mu^*_g(\theta)(s) \text{ almost surely as } T \rightarrow \infty,$$

where $1_s$ is the indicator function for $s$.

---

6Formally, for any $\theta \in \Theta$, any initial $s_0$ and any $s \in S$,
Let $X = \{X_1, X_2, ..., X_J\}$ 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 $s \in S$ there is a path of zero resistance to at least one of the recurrent classes in $X$. Further, within each recurrent class $X_k$ there is a path of zero resistance from every state to every other. Finally, every path starting at one recurrent class and ending outside that class has positive resistance.

Let $\Omega$ be a union of one or more recurrent classes. The basin of attraction $D(\Omega)$ of $\Omega$ is the set of initial states from which the Markov process converges to $\Omega$ with probability 1 when $\varepsilon = 0$. Define the radius $Ra(\Omega)$ of $D(\Omega)$ as the number of mutations needed to leave $D(\Omega)$ when play begins at $\Omega$. 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

$$Ra(\Omega) := \min_{(s^1, s^2, ..., s^k) \in P(\Omega, S \setminus D(\Omega))} c(s^1, s^2, ..., s^k).$$

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 mutations needed to get into $D(\Omega)$. Formally,

$$CRa(\Omega) := \max_{s \notin D(\Omega)} \min_{(s^1, s^2, ..., s^k) \in P(\Omega, \Omega)} c(s^1, s^2, ..., s^k).$$

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 $\Omega$ that is a union of recurrent classes such that $Ra(\Omega) > CRa(\Omega)$, then the stochastically stable equilibria are all contained in $\Omega$. 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 $\{s^\theta\}$ is a recurrent class and that $Ra(\{s^\theta\}) > CRa(\{s^\theta\})$.

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

**Lemma 1** Let $\theta$ be the true state of nature and suppose that Assumptions 1 and 2 hold. If $n \geq 5$, then $\{s^\theta\}$ is a recurrent class of the game that $g_M$ induces, with $Ra(\{s^\theta\}) \geq 2$.

**Proof.** See Appendix 6.1. □

**Lemma 2** Let $\theta$ be the true state of nature and suppose that Assumptions 1 and 2 hold. If $n \geq 5$, then $CRa(\{s^\theta\}) \leq 1$.
Proof. See Appendix 6.1. ■

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

To see why \( n \geq 5 \) is needed in the proof of Theorem 2, consider the example I used after the proof of Theorem 1. In that example, \( N = \{ 1, 2, 3 \} \) and there exists \( \phi \in \Theta \setminus \{ \theta \} \) such that, for \( i = 1, 2, \), \( u_i ( f ( \phi ) , \theta ) > u_i ( b , \theta ) \) for all \( b \in A \setminus \{ f ( \phi ) \} \). Let \( \bar{s} \) be the announcement profile \( ( \theta , \phi , \phi ) \). Note first that \( \bar{s} \) is a strict Nash equilibrium of the strategic game that mechanism \( g_M \) induces when the state of nature is \( \theta \). Indeed, by white lie aversion, player 1 strictly prefers to announce \( \theta \) than any \( \theta' \neq \theta \) when players 2 and 3 are announcing \( \phi \) (since mechanism \( g_M \) will implement \( f ( \phi ) \) regardless of player 1’s announcement). On the other hand, agents 2 and 3 strictly prefer to announce \( \phi \) than any other \( \theta' \neq \phi \), as they strictly prefer alternative \( f ( \phi ) \) to any \( b \in A \setminus \{ f ( \phi ) \} \).

Since \( \bar{s} \) is a strict Nash equilibrium of the game that \( g_M \) induces when the state of nature is \( \theta \), then \( \{ \bar{s} \} \) is also a recurrent class. In particular, \( \bar{s} \notin D ( \{ s^\theta \} ) \). Moreover, with only one mutation the system can move from \( s^\theta \) to \( \bar{s} \). To see this, suppose the system starts at \( s^\theta \). If agent 2 mutates and announces \( \phi \) instead of \( \theta \), then agent 3 will change her announcement to \( \phi \) if she gets a revision opportunity. Thus, there is a path with total resistance of 1 from \( s^\theta \) to \( \bar{s} \), so \( Ra ( \{ s^\theta \} ) = 1 \). In this case Theorem 2 will not hold. Indeed, one can show that in this example \( s^\theta \in \text{supp} \mu_{2M}^* ( \theta ) \) if and only if \( \bar{s} \in \text{supp} \mu_{gM}^* ( \theta ) \).

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 mutations needed to bring the system 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. However, in the setup I analyzed, from every \( s \neq s^\theta \) there is a path to \( s^\theta \) involving at most one mutation. Therefore, if \( W ( s_0 , s^\theta , \varepsilon ) \) denotes the expected waiting time until the Markov process first reaches \( s^\theta \) when the probability of mistakes is \( \varepsilon \) and the system starts at \( s_0 \), one can apply a result in Ellison (2000) to show that \( W ( s_0 , s^\theta , \varepsilon ) = O ( \varepsilon^{-1} ) \) as \( \varepsilon \to 0 \) for any \( s_0 \neq s^\theta \).

Finally, note that the assumption that players are white lie averse is again crucial for Theorem 2. If this assumption did not hold then truthtelling would not be the unique sto-
chastically stable equilibrium of the game that mechanism \( g_M \) induces, since in this case there would always be a path from \( s^\theta \) to any other announcement profile involving no mutations.

5 Conclusion

I studied a classic implementation problem with complete information under the assumption that agents have a minimal degree of honesty. In particular, I assumed that agents are *white lie averse*: they strictly prefer to tell the truth than to lie whenever lying has no effect on their material payoff. I showed that if agents are white lie averse and if I impose either fault tolerance or stochastic stability, then any social choice function can be fully implemented with a simple direct mechanism. Put differently, under the assumption of white lie aversion and either of these two refinements, a social planner can focus on the more appealing direct mechanisms to implement any social choice function, dispensing with any augmentation of the message space.

6 Appendix

6.1 Proofs of Lemmas 1 and 2

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

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

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

Consider next states \( s^1 \) such that \( |R^*(\theta' | s^1)| = |R^*(\theta'' | s^1)| = \frac{n}{2} \) for \( \theta', \theta'' \in \Theta \). I already considered the case with \( |R^*(\theta | s^1)| = \frac{n}{2} \) 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 \( s^1_i = \theta' \) and such that \( \bar{u}_i( f(\theta''), \theta ) \geq \bar{u}_i( a^*, \theta ) \) (so for agent \( i \) it is a best response to change her announcement to \( \theta'' \), given the announcement profile \( s^1_{-i} \) 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 system can move without any mutations to a state \( s^2 \) such that \( |R^*(\theta'' | s^2)| > \frac{n}{2} \), with \( \theta'' \neq \theta \) (this occurs if only player \( i \) gets a revision opportunity and changes her announcement from \( \theta' \) to \( \theta'' \)). It follows from the previous paragraph that there is a path from \( s^2 \) to \( s^\theta \) involving one single mutation. In case (ii), given \( s^1_{-i} \) 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 \( \theta \). Therefore, in this case the system can move to \( s^\theta \) without any mutations.

Finally, consider announcement profiles \( s^1 \) such that \( |R^*(\theta' | s^1)| \leq \frac{n}{2} \) for every \( \theta' \in \Theta \), but with at most one \( \theta' \neq \theta \) such that \( |R^*(\theta' | s)| = \frac{n}{2} \) and with \( |R^*(\theta | s)| < \frac{n}{2} \). There are two cases to consider: (i) there exists \( \phi \in \Theta \) such that \( |R^*(\phi | s^1)| + 1 > \frac{n}{2} \), and (ii) no such \( \phi \) 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 \( s^1_{-i} \), each agent \( i \) strictly prefers to announce \( \theta \) than any \( \theta' \in \Theta \setminus \{\theta\} \), so the system can move to \( s^\theta \) without any mutations.

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

### 6.2 Fault tolerance and stochastic stability

Theorems 1 and 2 together imply that, under white lie aversion, the solution concepts of fault tolerant 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 one.

To see this, consider the following example from Ellison (2000). There is a finite set of players \( N = \{1, 2, ..., n\} \). 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:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1, 1</td>
<td>0, 0</td>
<td>0, 0</td>
</tr>
<tr>
<td>B</td>
<td>0, 0</td>
<td>−4, −4</td>
<td>3, 3</td>
</tr>
<tr>
<td>C</td>
<td>0, 0</td>
<td>3, 3</td>
<td>−4, −4</td>
</tr>
</tbody>
</table>

Given \( s_{-i} \), 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, when the number of players is large enough, the stochastically stable equilibria of this game are \( s^B \) and \( s^C \) (where \( s^B = (B, B, ..., B) \), and similarly for \( s^C \)). That is, in the long run we should expect to see agents alternating between playing \( B \) and \( C \). However, neither \( s^B \) nor \( s^C \) are fault tolerant equilibria of the static random matching game. In fact, one can check that the fault tolerant equilibrium of the static random matching game is \( s^A \).
References


