Journal of Economic Theory **86**, 159–184 (1999)
Article ID jeth.1999.2530, available online at http://www.idealibrary.com on IDEAL®

Adaptive Dynamics and the Implementation Problem with Complete Information*

Antonio Cabrales

Department of Economics, Universitat Pompeu Fabra,
Ramon Trias Fargas 25, E-08005 Barcelona, Spain
antonio.cabrales@econ.upf.es

Received January 28, 1998; revised February 4, 1999

This paper studies the equilibrating process of several implementation mechanisms using naive adaptive dynamics. We show that the dynamics converge and are stable, for the canonical mechanism of implementation in Nash equilibrium. In this way we cast some doubt on the criticism of "complexity" commonly used against this mechanism. For a mechanism that implements using the iterated deletion of dominated strategies, the dynamics converge but are less stable. *Journal of Economic Literature* Classification Numbers: C72, D70, D78. © 1999 Academic Press

Key Words: implementation; bounded rationality; evolutionary dynamics; mechanisms.

1. INTRODUCTION

The theory of implementation tries to address the problem of designing game forms (which in this literature are called *mechanisms*) whose equilibria satisfy certain socially desirable properties but which do not necessitate vast amounts of knowledge by the authorities to put them in place. Instead, these social arrangements should basically self police themselves, and the designer should only make sure that the rules of the game are respected by the players.

* This paper was written while the author was visiting University College London, and he thanks the members of the Economics Department for their hospitality and encouragement. Thanks are also due to the helpful comments of an associate editor, a referee, Ken Binmore, Tilman Börgers, Luis Corchón, Dilip Mookherjee, Larry Samuelson, Roberto Serrano, Joel Sobel, Richard Vaughan, Fernando Vega-Redondo, Jürgen Weibull and seminar audiences at the London School of Economics, Universidad de Alicante and Barcelona Jocs. The financial support of the European Union Human Capital and Mobility Program under Contract ERB CHBGCT 930443 is gratefully acknowledged. Partial support was also obtained from Spain's Ministry of Education under Grant PB96-0302 and from the Generalitat de Catalunya under Grant 1997SGR00138.

In the past few years there have been impressive advances in the theory of implementation. As Sjöström [25] points out, “With enough ingenuity the planner can implement ‘anything.” This “ingenuity” often involves the construction of complicated games and the choice of the solution concept. As is often the case in economics, very little attention has been paid to the issue of how equilibrium is reached, and whether it is stable. This situation is worrisome given the importance of the issues at hand and the fact that the theory makes normative recommendations. It would not be sensible to apply these social engineering recipes without first thinking about whether real people will achieve the desired outcomes.

Some exceptions are the papers of Muench and Walker [20] and De Trenqualye [27] who study the conditions for local stability of the Groves and Ledyard [11] mechanism. Walker [30] proposes a stable mechanism yielding nearly Walrasian allocations in large economies. Jordan [15] shows that for any mechanism which implements the Walrasian correspondence in Nash equilibria with agents that are uninformed about other agents characteristics and any dynamic adjustment process there is an environment for which the equilibria are unstable with respect to the dynamics. Vega-Redondo [29] proposes a mechanism for which a best-response dynamic adjustment process is globally convergent to the Lindahl equilibrium outcome in an economy which has one private good, one public good and a linear production technology for the public good. De Trenqualye [28] proposes a mechanism that is locally stable for the implementation of Lindahl equilibria in an economy with multiple private goods, one public good, a linear production technology for the public good and quasi-linear preferences. Cabrales and Ponti [5] study the convergence and stability properties of Sjöström's [25] mechanism[1] under fictitious play and when one assumes that the dynamics are monotonic in the sense of Samuelson and Zhang [23].[2]

This paper studies first (a slight variation of) the canonical mechanism for implementation in Nash equilibria (see Maskin [17], Repullo [22]). We show that it has good dynamic properties when the assumption of *monotonicity* is replaced by *strict monotonicity*, the possible preference profiles and outcomes of the social choice rule are finite (although outcomes that are not part of the social choice rule can be infinite), and some punishments are possible. The dynamics are such that agents play the game repeatedly and once in a while they get a chance to replace the strategies

[1] Sjöström's [25] mechanism and the one that Jackson, Palfrey and Srivastava [14] study for separable environments are very similar and most of our results would generalize easily for that mechanism as well.

[2] A member of the class of monotonic dynamics is the replicator dynamics of evolutionary game theory, (Taylor and Jonker [26]).

they currently use. When they do it, they put positive probability on strategies that are best responses to the current strategy profile of the other players' and probability zero on strategies that give lower payoff than the one they are currently using.[3] Under these assumptions the dynamics converge to the set of Nash equilibria (so the social choice correspondence is implemented) and once the dynamics converge to an equilibrium, they stay there.

According to Jackson [13] "A nagging criticism of the theory is that the mechanisms used in the general constructive proofs have 'unnatural' features." Moore [18] also complains that the mechanisms for Nash implementation are "highly complex—often employing some unconvincing device such as an integer game." Our result shows that even unsophisticated agents using very simple adjustment rules can reach the set of equilibria of the mechanism. Therefore the criticism is misplaced if by "complex" we mean that the outcome that is desired by the planner will not be achieved by boundedly rational agents. On the other hand, it may be that the critics are right. If "complexity" is associated with the issue of the speed of convergence (which we do not explore) it may be that the canonical mechanism is slower than others.

The structure of the general constructive mechanism is as follows. The agents have to announce a state of the world, an outcome and an integer. If all agents agree on a state and an outcome, the outcome is implemented. If one agent disagrees and proposes an alternative, there is a test that the alternative has to pass. If it passes the test, the alternative outcome is implemented, otherwise it is not. A condition called *monotonicity* (Maskin [17]) ensures that an alternative will be proposed if and only if there is agreement on a lie. The mechanism also specifies what happens when more than one agent disagrees. In these cases, the mechanism gives the agent who proposed the largest integer her favorite outcome given the state of the world she announces. No situation with more than one dissident can be an equilibrium (if the best outcomes for the different agents are different). The reason is that in that case there is always one "loser" and the "loser" could "win" by announcing a high enough integer.[4] This exploitation of the non-existence of an equilibrium is one of the things that appears more worrisome to the critics about the use of integer game constructions.

The intuition for why there is convergence with the canonical mechanism is simple. If in one period all the agents make a coincident announcement

[3] They may (although it is not necessary) give weight to strategies that are not a best response but do better than the one currently used.

[4] One can show that there is no equilibrium, even in mixed strategies, that gives positive weight to strategy profiles where the integers are used to determine a winner.

that is false, some agent has an incentive to change the announcement to the truth by monotonicity. At that point any agent can invoke the integer game and, by choosing a large enough integer, win it. When the integer game is invoked, announcing the truth cannot hurt any player, provided it is accompanied by a high integer, since it only determines the preference profile with respect to which her most preferred outcome is chosen (if she wins the integer game). If everybody (simultaneously) announces the true outcome, the designer's desired outcome is achieved and from that point on, the strengthened version of monotonicity ensures that no player desires to deviate.

We also examine a mechanism that implements the social choice rule in the iterated deletion of dominated strategies. We show that although convergence to the equilibria of these games can be achieved, they are not stable when agents choose improving strategies which are not necessarily best responses. The problem is that drift between strategies that have the same payoff as the equilibrium payoff can destabilize the equilibrium outcome. This result is far from being merely a theoretical curiosity. As Binmore and Samuelson [3] point out, "the experimental evidence is now strong that one cannot rely on predictions that depend on deleting weakly dominated strategies." The mechanism we study, which is the one proposed by Abreu and Matsushima [2], implements the social choice rule in iteratively weakly undominated strategies. Besides being a good example of the literature on implementation with solution concepts different from Nash equilibrium, it has an additional interest because it allows us to discuss the mechanism of Abreu and Matsushima [1]. This mechanism *virtually* implements the social choice rule (that is, it implements with arbitrarily high probability) in strategies that survive the iterative deletion of strictly dominated strategies. This would seem to be a good mechanism from a dynamic perspective, given that iteratively strictly dominated strategies are asymptotically eliminated for most adaptive dynamics (see Nachbar [21], Samuelson and Zhang [23] or Cabrales and Sobel [6]). The problem is that if the mechanism implements with very high probability the social choice rule, then it will do so in iteratively strictly ε-*undominated* strategies, for ε very small. This implies that as the mechanism becomes more effective in doing its job, it becomes closer to the one in Abreu and Matsushima [2] and thus it becomes open to the sort of instability problems which that mechanism has. We think that this trade-off between close implementability and stability needs to be pointed out and we formalize it.

Section 2 describes the model and the dynamics we use. Section 3 studies the problem of Nash implementation with adaptive dynamics. Section 4 studies the dynamics of the mechanisms of Abreu and Matsushima [2] and Abreu and Matsushima [1]. An appendix gathers the proofs.

2. THE MODEL AND THE DYNAMICS

There is a set $I = 1, ..., n$ of agents, and the preferences of agent $i \in I$ are represented with a (Von Neumann–Morgenstern) utility function $v_i: A \times \Phi_i \to R$, where A is a set of alternatives and Φ_i specifies a finite set of possible utility functions. An element ϕ_i of Φ_i is meant to represent the preferences of agent i over A. A *preference profile* is a vector $\phi = (\phi_1, ..., \phi_n)$, where $\phi_i \in \Phi_i$. The set of possible preference profiles, denoted by Γ, is a subset of $\Phi = \times_{i \in N} \Phi_i$. Since we are concerned with environments with complete information, the preference profiles will be common knowledge among the agents.

A *social choice rule* is a (possibly multi-valued) mapping $F: \Gamma \to A$, where $\Gamma \subset \Phi$ is the set of possible preference profiles. A *mechanism* is a pair (M, g), where $M = M_1 \times \cdots \times M_n$ and $g: M \to A$. M_i is the *message* space of agent i and g is the *outcome function.* A *mechanism* and a *preference profile* define a game.

Let $M_{-i} = M_1 \times \cdots \times M_{i-1} \times M_{i+1} \cdots \times M_n$. Given a *mechanism* (M, g) and a *preference profile* ϕ, we will say that m_i is a best response for player i, to $m_{-i} \in M_{-i}$ if $v_i(g(m_i, m_{-i}), \phi_i) \geqslant v_i(g(m'_i, m_{-i}), \phi_i)$ for all $m'_i \in M_i$. A message profile m is a *Nash equilibrium* (NE) if m_i is a best response to m_{-i} for all $i \in N$. Let $NE(\phi) = \{g(m) \mid m \text{ is a NE at } \phi\}$.

We say that a mechanism (M, g) *implements* a social choice rule F in *Nash equilibrium* if for all $\phi \in \Gamma$, $F(\phi) = NE(\phi)$.

We will assume now that the implementation game is played repeatedly by the agents and that they can use the information obtained in previous periods to modify their behavior in subsequent rounds of play. The underlying set-up that one can keep in mind is that of a population of agents who play repeatedly implementation games for the provision of public goods in different groups of a society over time. An agent lives in an apartment building where the owners have to decide whether to purchase an elevator and of what quality, or whether to paint the external surface of the building and in what color. The agent is also the member of a sports club where members have to decide on the level of upkeep of the tennis courts and the putting green. She is also working in an office where workers decide on the regulation of the temperature in the (shared) working space. Besides, many of these decisions have to be taken repeatedly. If they were all taken using the same implementation mechanism, there would be ample opportunities for learning and adjusting play between repetitions.[5]

[5] To be completely consistent with this story, we would need to have (among other things) a population with several individuals playing the role of each agent i, and some kind of matching process. This modification is conceptually easy to do and the results still follow, but it involves substantial notational complication and we omit it for expositional simplicity.

To keep the problem tractable we will make some assumptions about the way in which the play and the updating takes place.

We will say that message $m_i \in M_i$ improves upon message m'_i given the message profile m if

$$v_i(g(m_i, m_{-i}), \phi_i) \geqslant v_i(g(m'_i, m_{-i}), \phi_i).$$

Let $\varepsilon > 0$. We say that m_i is an *εU-improvement* upon m'_i if

$$v_i(g(m_i, m_{-i}), \phi_i) - v_i(g(m'_i, m_{-i}), \phi_i) > -\varepsilon U.$$

We assume that agents play the game repeatedly (maintaining the role i and preference index ϕ_i). Each individual starts by playing some arbitrary (pure) strategy and before each repetition of the game they have an opportunity to change their (pure) strategy with some probability. The dynamics will be fully described when one identifies the transition probabilities between strategies. Instead of fully describing the process we enumerate a set of assumptions that are sufficient for the results of the paper.

(Y0) The transition probabilities depend exclusively on the present message profile.

(Y1) All individuals are given the chance to update their strategies with independent probabilities in the interval (0, 1).

(Y2) If the individual is given the chance to update her strategy, any best response to the present message profile is adopted with positive probability. If there are several messages which are best responses to the current one and they announce the same pair ϕ, a (as we will see, this means that they only differ in the integer that the mechanism requires mentioning), one of these messages is chosen with a probability bounded away from zero.

(Y3) A strategy which does not improve upon the strategy currently in use is adopted with zero probability.

Some alternatives to assumption (Y2) will be used in Section 4

(Y4) If the individual is given the chance to update her strategy, any strategy that improves upon the strategy currently in use, given the present message profile, is adopted with positive probability.

(Y5) If the individual is given the chance to update her strategy, any strategy that is εU-improvement upon the strategy currently in use, given the present message profile, is adopted with positive probability.

These assumptions permit us to obtain clear-cut results in a relatively simple fashion. Assumption (Y0) simplifies the analysis by making the strategy profile of a certain period the state variable of the system, but it is not essential for the results. It would suffice if the system had a finite memory, for example.

Assumptions (Y1) and (Y2) are designed to exploit a special characteristic of the mechanisms. For many strategy profiles the agents have lots of alternative strategies that yield the same payoff, and some of them are both a best response to the prevalent strategy profile, and lead to implementation of the social choice rule. If assumptions (Y1) and (Y2) are satisfied there will be convergence to an equilibrium from those states. It turns out that those states are also easily accessible from other states. Notice that assumption (Y2) does not demand that all agents choose best responses all the time, but only that they find them with some probability.

Assumption (Y2) does not say anything about the probabilities of strategies that are *not* best responses. In particular, it does not specify what happens with improving strategies that are not best responses. This implies that (Y2) is a less restrictive assumption than (Y4), so any proposition that is true with (Y2) as an assumption will also be true with (Y4). Specifically, Proposition 1 is true with (Y2) replaced by (Y4). We discuss in Section 4 why assumption (Y4) rather than (Y2) is used in the context of implementation in iteratively undominated strategies.

Both assumptions (Y4) and (Y2) remain silent as to the relative sizes of the probabilities of transitions to best-responses versus improving strategies that are not best responses. All that is needed in the propositions is that there is enough drift away from strategies that are not strict best responses, and no flow out of strategies that are strict best responses. The latter is achieved with assumption (Y3), which makes the equilibria of the canonical mechanism absorbing states. One could even relax (Y3) by adding small probabilities of mutations in all directions, which would make the process ergodic, and then look at the stationary distribution. The limit of that distribution as mutations go to zero would put weight only on the socially desired outcomes for the canonical mechanism. It is unclear how this would affect the results in Section 4.

The statement in assumption (Y2) about best response messages that share the same a, ϕ announcements is used, as we explain in Section 3, to account for the fact that the strategy spaces are infinite but for Proposition 1 some transition probabilities have to be bounded away from zero.

Assumption (Y5) modifies (Y4) in a way that will be suitable to discuss virtual implementation.

Properties (Y0) to (Y4) make our dynamics similar to the ones in Kim and Sobel [16]. The difference here is that they require individual (sequential) adjustments and we assume that there is positive probability of simultaneous adjustments. Assumption (Y2) corresponds to their assumption

(BR), (Y4) corresponds to their assumption (R) and Assumption (Y3) to their assumption (NL). Our dynamics are also closely related to the ones in Hurkens [12] and Gilboa and Matsui [9].

3. NASH IMPLEMENTATION

In the first subsection we describe the mechanism and the second will show that the dynamics described in Section 2 converge and are stable for that mechanism.

3.1. *The Canonical Mechanism* (*Almost*)

We say that F is *monotonic* if for all a, ϕ, ϕ', with $a \in F(\phi)$ and $a \notin F(\phi')$ there is an i and a' such that $v_i(a, \phi) \geqslant v_i(a', \phi')$ and $v_i(a', \phi') > v_i(a, \phi')$.

Monotonicity is a necessary and almost sufficient condition for Nash implementation. We use somewhat stronger assumptions,

(N1) For all a, ϕ, ϕ', with $a \in F(\phi)$ and $a \notin F(\phi')$ there is an i and a' such that $v_i(a, \phi) > v_i(a', \phi)$ and $v_i(a', \phi') > v_i(a, \phi')$.

(N2) For all i, ϕ and $a \in F(\phi)$ there is $a' \in A$ such that $v_i(a, \phi) > v_i(a', \phi)$.

(N3) The set Γ is finite. So are the sets $F(\phi)$ for all $\phi \in \Gamma$.

We denote by $i(\phi, \phi')$ one (arbitrarily chosen) of the agents that satisfy the condition of assumption (N1), and by $a'(\phi, \phi')$ one (arbitrarily chosen) of the outcomes such that $v_{i(\phi, \phi')}(a, \phi) > v_{i(\phi, \phi')}(a'(\phi, \phi'), \phi)$ and $v_{i(\phi, \phi')}(a'(\phi, \phi'), \phi') > v_{i(\phi, \phi')}(a, \phi')$. This agent $i(\phi, \phi')$ is often called the *test agent* and $a'(\phi, \phi')$ the *test outcome* in the implementation literature. Let us also denote by $a'_i(a, \phi)$ one of the outcomes a' in assumption (N2).

Under our dynamics, all best-responding messages are chosen with positive probability. If the Nash equilibrium of the mechanism were such that some agent had more than one best response, it could be easily destabilized. To avoid this we will use two assumptions; (N1), which demands that the test outcome be a strict improvement over the "status quo" and (N2) by which it is always possible to punish a dissident who has no reason to dissent (she is not a test agent). Assumption (N3) is used to guarantee convergence to the desired outcome in finite time.

Assumption (N2) does not seem very restrictive, since it will be sufficient for example that agents have strictly monotonic preferences over a private good over which fines can be levied, or that there is an outcome which is bad for all agents. Assumption (N1) is a slight strengthening of Maskin-monotonicity and would be satisfied if preferences were strictly convex, for example. Assumption (N3) limits the set of allowable preference profiles and the social choice rules to be finite valued, but the set of possible outcomes A might still be infinite, so the Euclidean spaces for outcomes that are common in economics are not excluded. This still leaves a nontrivial set of social choice rules like the Walrasian or Lindahl correspondences which can be implemented under reasonable subsets of preferences.

We will use a slight variation of the canonical mechanism for implementation in Nash equilibria, as described, for example in Repullo [22].

Let $A_F = \{a \in A : \exists \phi \text{ with } a \in F(\phi) \text{ or } \exists \phi, \phi' \text{ with } a = a'(\phi, \phi')\}$.

Let $M_i = A_F \times \Gamma \times N$, so that each individual announces an outcome, a preference profile, and a positive integer; and $M = M_1 \times \cdots \times M_n$, and let members of M_i and M be denoted m_i and m respectively. Let the first component of m_i, that is, the outcome announced by agent i be m_i^1 and the second component, the preference profile announced by agent i, be m_i^2. Let $i(m)$ be the individual who has the lowest index among those who announce the highest integer in the message profile m.

Let $b_i(\phi)$ be such that $v_i(b_i(\phi), \phi) \geqslant v_i(a, \phi)$ for all $a \in A$.

To define g, let's divide M into the following regions,

$$D_1 = \{m \mid \exists \phi \in \Gamma, a \in F(\phi) \text{ such that for all } i, m_i = (a, \phi, n_i), \text{ for some } n_i \in N\}$$

$$D_2 = \{m \mid m_i = (a, \phi, n_i)\ \forall i \neq i(\phi, \phi'), \text{ and } m_{i(\phi, \phi')} = (a'(\phi, \phi'), \phi', n_{i(\phi, \phi')})\}$$

$$D_3^j = \{m \mid m_i = (a, \phi, n_i)\ \forall i \neq j, \text{ and } m \notin D_1 \cup D_2\}$$

$$D_4 = \{m \mid m \notin D_1 \cup D_2 \cup D_3^1 \cdots \cup D_3^n\}$$

$$g(m) = \begin{cases} a & \text{if } m \in D_1 \\ a'(\phi, \phi') & \text{if } m \in D_2 \\ a_j'(a, \phi) & \text{if } m \in D_3^j \\ b_{i(m)}(m_{i(m)}^2) & \text{if } m \in D_4 \end{cases}$$

This mechanism can be described in the following way. If everybody agrees on an outcome and a state, then that outcome is implemented. If all agents but one announce the same outcome, and the dissident is the test agent and she announces the test outcome, then the test outcome is implemented. If there is one dissident but she is not the test agent (or she is the test agent but does not announce the test outcome), then the dissident is punished. If more than one person disagrees, then the outcome is the favorite one (under the preference profile she announces) for the agent who announces the largest integer.

There are a couple of small differences between this mechanism and the one in Repullo [22]. One is that we punish deviations from the equilibrium by agents other than the *test agent*, (and even punish announcements by the *test agent* which are not part of the test pair). As we discussed above, this is done to avoid having multiple best responses in equilibrium. The other difference is that we ask that the "allowable" dissident, the *test agent*, and the *test outcome* must be designated beforehand, and the outcomes that can be announced must be either test-outcomes or outcomes of the social choice rule for some preference profile. We do this so that we can allow a possibly infinite set of outcomes A, while maintaining a relatively small state space. This is important because the agents in this model are not very sophisticated and they find their way to equilibrium by trial and error. The smaller the state space, the faster will convergence be.

Unlike Repullo [22] and other papers in the literature on Nash implementation we do not make the assumption of *absence of veto power*. This assumption says that for all $a \in A$, $\phi \in \Gamma$, if $u_j(a, \phi) \geqslant u_j(a', \phi)$ for all $a' \in A$ and for all $j \neq i$, then $a \in F(\phi)$. Without this assumption we can have, for example, the situation that an outcome a, which is the best in all players' utility functions for some preference profile ϕ, is not selected by the social choice rule, (that is, $a \notin F(\phi)$). Under these conditions, we would have a Nash equilibrium with outcome a, when the true preference profile is ϕ. A message profile in D_4 would deliver such an equilibrium, since no agent would have an incentive to change a strategy that is already delivering the best possible outcome. The assumption of *absence of veto power* is not needed in our case since such kind of Nash equilibria would not be stable under our dynamics. We will show in Proposition 1 that from message profiles in D_4 the dynamics eventually drift into D_1 (for the true ϕ), and by definition $a \notin F(\phi)$.

3.2. *The Dynamics of Nash Implementation*

The main result in this section is that the dynamics defined in Section 2 for the game induced by the mechanism in subsection 3.1 are such that the strategy profile will almost surely lead to one of the outcomes that the designer wants to implement, and that outcome is then implemented forever. In addition, if none of the outcomes that the planner wants to implement are already being implemented, all outcomes in the social choice rule are implemented with positive probability.

Assumption (Y2) requires that when there are several best responses which announce the same pair ϕ, a, that is, they announce different integers, one of these messages is chosen with a probability bounded away from zero. It is then necessary that the sets of allowable pairs ϕ, a are finite. This is true by assumption (N3), the fact that the announced a must belong to A_F and we single out one and only one *test agent* and *test outcome* for every pair ϕ, ϕ'. Without this assumption it would be possible that agents would spend infinite amounts of time cycling around the integer game. Because of this modification, whenever we say in the proof of the propositions that something happens with positive probability it means actually with probability that is bounded away from zero.

Define the set $S_a = \{m \mid \exists \phi, a \in F(\phi)$, such that $\forall i; m_i = (a, \phi, n_i)\}$. The set S_a is the set of message profiles in D_1 where the outcome is a.

PROPOSITION 1. *Let the true preference profile be* ϕ. *Given dynamics that satisfy properties* (Y0), (Y1), (Y2), (Y3), *and given a social choice function that satisfies* (N1), (N2), (N3);

(a) *If* $m(0)$ *is a such that* $m(0) \notin S_a$ *for any* $a \in \phi$, *then for all* $a \in \phi$; $P($*for some* $t', m(t) \in S_a, \forall t \geqslant t') > 0$.

(b) $P(\bigcup_{a \in F(\phi)} \{$*for some* $t', m(t) \in S_a, \forall t \geqslant t'\}) = 1$.

Proof. See the Appendix.

The intuition for the stability part of the result is that if all agents agree on an outcome a and also announce the true ϕ (so that the message profile is in D_1), the test agent does not want to change the strategy and announce the test outcome by N1 (modified monotonicity), and any other change by any agent would only lead to an outcome in D_3^j, which the agent who changed would not like by N2.

The convergence result starts by showing that the message profile will go with positive probability to D_4 (where the integer game is played) if the initial state is not one where the social choice rule is implemented. For example, if all agents agree on an outcome and announce a false ϕ, the test agent would like to change her announcement to the test outcome by N1, and after that change, any player can announce the true preferences (which puts the message profile in D_4) and obtain her favorite outcome by announcing a high enough integer. For similar reasons, if the initial state is in D_2 or D_3^j, any player can announce something that puts the message profile in D_4 and obtain her favorite outcome.

Once the message profile is in D_4, announcing the true ϕ and some $a \in F(\phi)$ is a best response if a high enough integer is also announced.[6] If all agents announce it simultaneously the message profile will be in D_1, and the stability argument guarantees that the message profile becomes fixed at that point.

Notice that a similar argument would work for modulo games. This is important because one could reasonably argue that a practical problem

[6] If there are only 2 dissidents it may not be a best response for them to tell the truth, but then there is positive probability that some other player also becomes a dissident, which is a best response, and at that point it is a best response for all agents to announce ϕ and $a \in F(\phi)$.

with the canonical mechanism is that in real life the designer would have trouble with an infinite strategy space (time constraints could preclude describing arbitrarily high integers). But modulo games are not subject to that criticism and have the same dynamic properties.

We have assumed that the state of the world (the preference profile) remains unchanged while people learn. It is interesting to consider what would happen if the state of the world changed with positive probability while the agents were learning. In that case, a result analogous to Proposition 1 could be obtained by making the learning process operate on strategies as functions of the state into the messages. While the dynamics for the mechanism would still converge and be stable, this alternative assumption of a changing state of the world would make an important difference for a couple of reasons. First of all, the increase in dimensionality of the space in which the dynamics move would probably make convergence slower. But a state that changes is also important because it makes apparent the difference between implementation with complete and incomplete information. In order to use a strategy where the message sent varies with the state of the world (which is likely to be the best response eventually), the agent needs to *know* the state. When the state does not change over time, the player only needs to know the payoffs of the different message profiles (which she can learn by trial and error),[7] so the distinction between implementation with complete and incomplete information becomes blurred.

Another important issue is that the planner may be able to use the information that the agents are not fully rational in the design of the mechanism. We have already done this in part, since we have modified Repullo's [22] mechanism to make the socially desirable equilibria stable. But the planner may also be interested in accelerating the convergence to equilibrium. Addressing this issue properly would require a more formal treatment of the speed of adjustment. This, in turn, would require more specific assumptions about the dynamics, and it is likely to be more dependent on the particular environment than the questions of convergence and stability. Nevertheless, we now make some conjectures about how the planner may be able to modify the mechanism to accelerate convergence to equilibrium using the agents' bounded rationality. Notice, however, that while these changes may accelerate convergence, the outcomes on the way to equilibrium may be quite bad for some agents. To properly evaluate this tradeoff, it would be necessary to postulate preference rankings for the planner, something that is typically avoided in the implementation literature.

[7] We owe this observation to Tilman Börgers.

If the probability of a change in strategy were related to the difference in payoffs, the planner could accelerate convergence by making payoff differences between some outcomes of the mechanism as large as possible. For example, if there is more than one possible *test outcome*, $a(\phi, \phi')$ should be chosen to give the maximum utility possible for the *test agent* from the outcome of the consensus announcement so that she deviates soon (and sends play from D_1 to D_2). If possible, the *test outcome* should also be such that agents other than the test agent are punished (maybe by having them pay a penalty). This punishment would give them in turn more incentives to deviate from a strategy profile in D_2 and move play to D_4. Another thing that may delay convergence is the fact that in D_4 announcing the true state ϕ (and a high integer) so as to obtain $b_i(\phi)$ is a best response, but so may be announcing some other ϕ' (for example because $b_i(\phi) = b_i(\phi')$). If that is the case, and if there are outcomes $c_i(\phi)$ for all i, ϕ with the property that $v_i(c_i(\phi), \phi) > v_i(c_i(\phi'), \phi)$ for all ϕ, ϕ' (that is, every i prefers the c_i outcome corresponding to the "true" state of Nature), and $v_i(c_i(\phi), \phi) > v_i(c_j(\phi), \phi)$ (that is, every i prefers "her own" outcome to somebody elses'), then one could amend the mechanism to use $c_{i(m)}(m^2_{i(m)})$ in D_4 and the convergence and stability results would be maintained, but convergence may be faster.

4. UNDOMINATED AND VIRTUAL IMPLEMENTATION

4.1. *Implementation in Iteratively Undominated Strategies*

So far, we have only considered implementation in Nash equilibrium. What about other equilibrium concepts? Since the seminal work of Moore and Repullo [19], there has been considerable interest in implementation with equilibrium concepts that are more refined than Nash equilibrium.[8] The main advantage of these mechanisms is that the conditions for implementation are weaker. In particular, monotonicity is no longer required. This is important since in economic environments implementing a single-valued social choice rule and requiring monotonicity is equivalent to truthful implementation in dominant strategies (see Dasgupta, Hammond, and Maskin [8]).[9]

[8] Although the iterative elimination of weakly dominated strategies is not a refinement of the notion of Nash equilibrium (there are many games for which the iterative deletion of dominated strategies leaves a set that is larger than the set of Nash equilibria) for the game forms that we study the set of iteratively undominated strategies is a strict subset of the set of Nash equilibria.

[9] This is true only with the domain of preferences is very large. In economies with a unique Walrasian/Lindahl equilibrium the Walrasian/Lindahl correspondence is Nash-implementable (see Corchón [7], p. 68).

By comparison, implementation in undominated strategies requires basically no restrictions. Abreu and Matsushima [2] show that "any social choice function is exactly implementable in iteratively weakly undominated strategies," and Sjöström [25] "in economic environments any social choice rule can be implemented in undominated Nash equilibria." An additional advantage of some of these mechanisms (notably those of Abreu and Matsushima [2] and Sjöström [25]) is that "integer games" or "modulo games" are not used.

The purpose of this section is to show that these advances should be viewed with some suspicion if we believe that equilibrium is the outcome of a learning process, since the adaptive dynamic process leads to undesired outcomes even asymptotically.

To focus the discussion we will concentrate on the mechanism proposed by Abreu and Matsushima (henceforth AM) [2], but the results can be extended to other mechanisms that have been proposed in the literature.

We will begin by introducing some notation and describing the mechanism.

The first thing to notice is that AM [2] only consider single-valued social choice rules. Another important assumption is that there is a private good that can be used to levy (small) fines. Thus the utility function will be v_i: $A \times R \times \Phi_i \to R$. We will use (as AM [2] do) for simplicity the quasi linear utility function $v_i(a, T, \phi_i) = u_i(a, \phi_i) + T_i$. Since the fines that the mechanism in AM [2] imposes are arbitrarily small, quasi-linearity is used without loss of generality. Besides the outcome function $g(M)$ the mechanism specifies a *transfer rule*, $T = (T_i)_{i \in N}$: $M \to R^n$. The message space in AM [2] is,

$$M_i = \Phi_i \times \Phi_{i+1} \times \Gamma \times \cdots \times \Gamma = M_i^{-1} \times M_i^0 \times M_i^1 \times \cdots \times M_i^K,$$

where K may have to be quite large to make the fines very small. For expositional simplicity we will allow the fines to be large in which case it is enough to have $K = 1$. The arguments also go through (but notation and proofs are more cumbersome) when we have K large and small fines. Let then, in our case

$$\begin{aligned} M_i &= \Phi_i \times \Phi_{i+1} \times \Gamma = M_i^{-1} \times M_i^0 \times M_i^1, \\ M &= M_1 \times M_2 \cdots \times M_n, \\ M^h &= M_1^h \times M_2^h \cdots \times M_n^h; \end{aligned}$$

and let m_i, m, and m^h be generic elements of M_i, M and M^h.

By the lemma in AM [1] we have that there exists a function f_i: $\Phi_i \to A$, such that for every $\phi_i \in \Phi_i$,

$$u_i(f_i(\phi_i), \phi_i) > u_i(f_i(\phi_i'), \phi_i) \qquad \text{for all} \quad \phi_i' \in \Phi_i/\{\phi_i\}.$$

For any message profile m, the outcome function is,

$$g(m)=\frac{e(m^0, m^1)}{n}\sum_{i\in I} f_i(m_i^{-1})+(1-e(m^0, m^1))\,\rho(m^1),$$

where we define ρ: $M^1 \to A$ by

$$\rho(m^1)=\begin{cases} F(\phi) & \text{if } \quad m_i^1=\phi \text{ for at least } (n-1) \text{ agents} \\ b & \text{otherwise, where } b \text{ is an arbitrary element of } A \end{cases}$$

and if we let ε be a small positive number to be specified later, and $\tilde{m}_0=(m_n^0, m_1^0, ..., m_{n-1}^0)$, we define e: $M^0\times M^1\to R$ by

$$e(m^0, m^1)=\begin{cases} \varepsilon & \text{if } \quad m_i^1\neq\tilde{m}_0 \text{ for some } i\in I \\ 0 & \text{otherwise} \end{cases}$$

The outcome function g is a lottery with the following characteristics. With a probability determined by the function $e(\cdot)$ (which is nonzero when some agent's oneth announcement differs from $\tilde{m}^0$) the favorite outcome of agent i, given her m_i^{-1} announcement, is selected with probability $1/n$. With probability $1-e(\cdot)$ an outcome given by the function $\rho(m^0, m^1)$ is chosen. This function says that if all but one of the m_i^1 announcements coincide on ϕ, then $F(\phi)$ is implemented, otherwise an arbitrary outcome b is implemented.

To finish the description of the mechanism the penalty function has to be specified. Let γ, ξ, η be small positive numbers to be specified later. Three possible penalties are specified for each player i.

1. γ if his zeroth announcement differs from player $(i+1)$'s minusoneth announcement

2. ξ if his oneth announcement differs from $\tilde{m}^0$.

3. η if his oneth announcement is the only one to differ from the other players' oneth announcements.

We will now give names to the fines

$$\tau_i(m_{i+1}^{-1}, m_i^0)=\begin{cases} -\gamma & \text{if } \quad m_{i+1}^{-1}\neq m_i^0 \\ 0 & \text{otherwise} \end{cases}$$

$$d_i(m^0, m^1)=\begin{cases} -\xi & \text{if } \quad m_i^1\neq\tilde{m}_0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_i(m^1)=\begin{cases} -\eta & \text{if for some } \phi,\ m_i^1\neq\phi, \text{ but } m_j^1=\phi \text{ for all } j\in I/\{i\}, \\ 0 & \text{otherwise} \end{cases}$$

The total fine is thus $T_i(m) = \tau(m_{i+1}^{-1}, m_i^0) + d_i(m^0, m^1) + \mu_i(m^1)$.

To finish with the description of the implementation game we need to define the constants ε, η, ξ and γ. To do this, define first,

$$E_i(\phi_i) = \max_{m^{-1} \in M^{-1}, m^1 \in M^1} \left| \frac{1}{n} \sum_{j \in I} u_i(f_j(m_j^{-1}), \phi_i) - u_i(\rho(m^1), \phi_i) \right|$$

$$D_i(\phi_i) = \max_{m^1 \in M^1, \bar{m}_i^1 \in M_i^1} \{u_i(\rho(m^1), \phi_i) - u_i(\rho(m_{-i}^1, \bar{m}_i^1), \phi_i)\}$$

Fix ε (small) and choose η, ξ and γ to satisfy

AM1

$$\eta > \varepsilon \max_{i, \phi} E_i(\phi_i)$$

$$\xi > \max_{i, \phi} D_i(\phi_i) + \eta$$

$$\gamma > \varepsilon \max_{i, \phi} E_i(\phi_i) + \xi$$

With these three inequalities AM [2] show the following lemmas,[10]

LEMMA 1. *Let any m_i, and $\bar{m}_i = (\phi_i, m_i^0, m_i^1)$. Under* AM1, *for all m_{-i},*

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) \geqslant v_i(g(m), T(m), \phi_i)$$

LEMMA 2. *Under* AM1, *for all m with $m_i^{-1} = \phi_i$ and all i, if we let $\bar{m}_i = (\phi_i, \phi_{i+1}, m_i^1)$,*

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), T(m), \phi_i)$$

LEMMA 3. *Under* AM1, *for all m with $m_i^{-1} = \phi_i$ and $m_i^0 = \phi_{i+1}$ if we let $\bar{m}_i^1 = \phi$ then,*

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), T(m), \phi_i)$$

Lemma 1 says that announcing the true preference index at m_i^{-1} and keeping the rest of the strategy constant is weakly dominant. Lemma 2 says that if the true preference index is announced at m_i^{-1} then announcing the true preference profile at m_i^0 and keeping the rest of the strategy constant, is strictly dominant. Lemma 3 says that if the true preference index is

[10] Notice that in our case ξ and γ need not be arbitrarily small. If K were larger than 1, the second inequality would be $\xi > (1/K)\, D_i(\phi_i) + \eta$ and ξ and γ could be very small, if K were sufficiently large.

announced at m_i^{-1} and m_i^0, then announcing the true preference profile at m_i^1 is strictly dominant.

We now show that if the lemmas are true, the dynamics will go with positive probability to a state where the social choice rule is implemented.

Proposition 2. *Let the true preference profile be ϕ. Given dynamics that satisfy properties* (Y0), (Y1), (Y2), *if Lemmas* 1, 2 *and* 3 *are satisfied,*

$$P(\text{for some } t, m(t) \in S_{F(\phi)}) > 0.$$

Proof. See the Appendix.

The intuition of the result is simple. Lemma 1 shows that in a best response one has to tell the truth at m_i^{-1} from any initial position. Lemma 2 shows that in a best response one must tell the truth at m_i^0 once the truth is announced at m_i^{-1} and then telling the truth at level m_i^1 is part of a best response by Lemma 3. Given this, assumptions (Y2) and (Y1) guarantees that these strategy switches take place.

This shows that the mechanism of AM [2] will lead to implementation of the social choice rule. Unfortunately, it is also possible to diverge from the equilibrium in which the social choice rule is implemented.

Proposition 3. *Let the true preference profile be ϕ. Given dynamics that satisfy properties* (Y0), (Y1), (Y3), (Y4), *if $m(t) \in S_{F(\phi)}$, then P(for some $t' \geqslant t, m(t') \in S_{F(\tilde{\phi})}) > 0$ for any $\tilde{\phi}$.*

Proof. See the Appendix.

The intuition for this proposition is that starting from a message profile where the true preferences are announced at all levels, switching to announcing a false preference index at m_i^{-1} does not hurt agent i (it's a best-response to the current strategy profile). But if i changes the announcement of m_i^{-1}, then for agent $i-1$ switching to a false preference index (but consistent with m_i^{-1}) at m_{i-1}^0 is improving. And given the previous steps, switching to the new m^0 at level m_j^1 is improving for all agents.

Notice that a difference between Proposition 1 (and 2) and 3 is that the latter uses assumption (Y4), while the former uses (Y2). From the proof of Proposition 3 one can see that the first change away from implementing the social choice rule (announcing a false preference index at m_i^{-1}) is a best response. After that, agent $i-1$ changes m_{i-1}^0 to best respond to the new m_i^{-1}, but then the probability $e(m^0, m^1) = \varepsilon > 0$, which makes optimal announcing the true preferences at level m_i^{-1}. Agent i, however, does not modify its announcement of m_i^{-1} on the way to the new equilibrium, so her changes are *improving*, but not *best responses*. Assumption (Y4) guarantees that this can happen, and therefore that other (not socially desirable) equilibria are reached. In the absence of (Y2) the first deviation by i and then the deviation by $i-1$ are possible, but from then on it is not clear how far from the desired equilibrium the process can go, without further assumptions.

Notice that even with (Y2) the $i-1$ agent has to pay a γ fine as a result of the deviation by i. In principle the γ could be very small, but then the ε would also be very small (see AM1). In that case, the use of (Y4) would be more acceptable, because the *improving* strategies that are *not best responses* used on the way to the new equilibrium differ from the best responses by an amount that is of the order of magnitude of ε. This choice between a large penalty that has to be paid with high probability and a higher likelihood of ending up in the "wrong" outcome will appear again in subsection 4.2.

An implication of Proposition 3 is that while (our version of) the canonical mechanism is robust in the presence of agents who use improving strategies, the mechanism in AM [2] is not. We have concentrated on sufficient conditions for divergence from the desired equilibrium because our purpose was to highlight the relatively higher robustness of the Nash mechanism, but it is not hard to think of sufficient conditions to guarantee that the process converges and is stable at the "right" equilibrium. Suppose we have an initial condition where $m_i^1 \neq \tilde{m}_0$ for some $i \in I$, that agents only change strategies if there is a strict improvement, and that they always choose a best response to the past message profile with probability one. Then, the dynamics would converge and be stable at the socially desired outcome.[11]

4.2. *Virtual Implementation*

The idea behind virtual implementation is that to obtain implementability results under weaker sufficient conditions on the domain of preferences one can relax the notion of implementation (instead of strengthening the equilibrium concept). After all, the planner may well be satisfied as long as the social choice rule is implemented with a high probability. AM [1] show that if the planner only requires that the social choice rule is implemented with arbitrarily high probability, basically any social choice rule can be implemented, even with such a simple solution concept as iterative strictly undominated strategies.

This result would appear very congenial with the spirit of this paper. Since the solution concept is iterative strictly undominated strategies, both convergence and stability would be expected not only under the dynamics of this paper, but in a variety of evolutionary and learning models (see

[11] Cabrales and Ponti [5] also find that a positive or a negative result on the stability of a mechanism that implements in iterative undominated strategies depends on the "degree of best responsiveness" of the dynamics and the initial conditions.

Nachbar [21], Samuelson and Zhang [23] or Cabrales and Sobel [6]). There is a problem, however, if the planner wants to implement a social choice rule which is ε-close to the original social choice rule. In that case some of the dominated strategies which have to be eliminated for the process to converge are only ε-strictly dominated. In fact, we will show that if the agents can switch between strategies whose utilities are ε-close, then the same instability problem of the mechanism of the previous subsection is reproduced here. This assumption is not unreasonable given the idea behind virtual implementation that the planner does not care too much if the social choice rule is not implemented, as long as the function that is actually implemented is ε-close to the original social choice rule.

Following AM [1], we say that social choice rules x and y are ε-close if for all preference profiles, x and y map to lotteries that are ε-close. A social choice rule x is *virtually implementable* in iterative strictly undominated strategies if for all $\varepsilon > 0$, there exists a social choice rule y which is ε-close to x and which is exactly implementable in iterative strictly undominated strategies.

To make the presentation a little simpler, we will not use the same mechanism that AM [1] use but a modification based on AM [2]. As before, we use the quasi linear utility function $v_i(a, T, \phi_i) = u_i(a, \phi_i) + T_i$. Besides the outcome function $g(M)$ the mechanism specifies a *transfer rule*, $T = (T_i)_{i \in N}$: $M \to R^n$. The message space will again be,

$$M_i = \Phi_i \times \Phi_{i+1} \times \Gamma = M_i^{-1} \times M_i^0 \times M_i^1$$

Let $m_i = (m_i^{-1}, m_i^0, m_i^1)$, and $m^1 = (m_1^1, ..., m_n^1)$. The only change in the mechanism is that for any message profile m, the outcome function is now,

$$g(m) = \frac{\varepsilon}{n} \sum_{i \in I} f_i(m_i^{-1}) + (1 - \varepsilon)\, \rho(m^1),$$

where we define $\rho: M^1 \to A$ as in the previous subsection and ε is a small positive number as in the definition of *virtual implementation*. The penalty functions are also as specified in the previous subsection.

Note that with the modification made in the mechanism, Lemma 1 is now true with a strict inequality.

LEMMA 4. *Under* AM1. *Let any m_i, and $\bar{m}_i = (\phi_i, m_i^0, m_i^1)$, then for all m_{-i},*

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) > v_i(g(m), t(m), \phi_i)$$

Proof. Trivial from the proof of Lemma 1, and the definition of the mechanism. ▮

Lemma 4, plus Lemmas 2 and 3 imply that the implementation solution concept is the iterative deletion of strictly undominated strategies. Note also that the function exactly implemented now is ε-close to F. Since ε can be made arbitrarily small, this mechanism virtually implements F. Let's denote the social choice rule that is actually implemented for each value of ε, F_ε.

Proposition 4. *Let the true preference profile be ϕ. Given dynamics that satisfy properties* (Y0), (Y1), (Y2), *and* (Y3), *for all $m(0)$ there exists t' such that $P($for all $t \geqslant t', m(t) \in S_{F_\varepsilon(\phi)}) = 1$.*

Proof. A straightforward modification of the proof of Proposition 2 shows that with Probability 1 there exists t' such that $s(t') \in S_{F_\varepsilon(\phi)}$ and the message $m_i = (\phi_i, \phi_{i+1}, \phi)$ is sent by all players. Lemmas 4, 2 and 3 show that for all $\bar{m}_i \neq m_i$,

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), T(m), \phi_i) > 0$$

so by Assumption Y3 $P($for all $t \geqslant t', s(t) \in S_{F_\varepsilon(\phi)}) = 1$. ▮

The mechanism proposed guarantees convergence and stability to a message profile that implements the social choice rule with arbitrarily high probability, under assumptions (Y0), (Y1), (Y2) and (Y3).

The problem arises if assumption (Y2) is replaced by (Y5), with $U = \max_{i, \phi, \phi'} u_i(f(\phi'), \phi)$.[12] We can then show,

Proposition 5. *Let the true preference profile be ϕ. For all $\varepsilon \geqslant 0$, given dynamics that satisfy properties* (Y1), (Y3), (Y5) *if $m(t) \in S_{F(\phi)}$, then P (for some $t' \geqslant t, m(t') \in S_{F(\tilde{\phi})}) > 0$ for any $\tilde{\phi}$.*

Proof. If $s(t) \in S_{F(\phi)}$, then if agent n changes m_n^{-1} to some $\phi'_n \neq \phi_n$, her payoff does not change by more than εU by the definition of the mechanism. Thus, Y1 and Y5, and the definition of ε-improvement guarantees that this happens with positive probability. The rest of the proof retraces the steps of Proposition 3 closely. ▮

This result implies that the agents have to care about the outcomes of the implementation process orders of magnitude more than the planner to avoid the instability of the mechanism. While this may be justified under certain circumstances, it is by no means clear that it will always be so.

[12] U is well defined because Φ is a finite set.

Notice also that in the Nash mechanism we proposed earlier the equilibria are strict, so as long as ε is small, the equilibria are stable even if assumption Y5 holds.

This note of caution about virtual implementation is different from the one in Glazer and Rosenthal [10]. They think unlikely that agents will do many rounds of deletion of strictly dominated strategies, especially in circumstances where some alternatives to the social choice outcome are focal (like when an alternative outcome Pareto-dominates the social choice outcome).[13] We, on the other hand, are worried that even if agents reach the social choice outcome, the near indifference between that outcome and some alternatives will destabilize it.

5. CONCLUSIONS

The main message of this paper is that thinking explicitly about the equilibrating process in the implementation problem can be a fruitful experience. We hope that these results encourage more work into the implementation problem using dynamic tools. An important question that should be answered is how sensitive are our conclusions to the dynamics postulated. Also, we have not examined the question about the speed of adjustment; reaching the socially desirable outcome may be irrelevant if it takes a very long time (and the outcomes achieved on the way are undesirable). These considerations suggest the need for additional theoretical work, but no real progress can be made unless more empirical and experimental investigation is done in this field.

6. APPENDIX

Proof of Proposition 1. The proof will proceed through two lemmas. First we will show that a message profile which does not implement any social choice function outcome can lead to all profiles whose outcomes are outcomes of the social choice rule, and then we will show that a message in S_a, where everybody announces the true preference profile cannot exit that set.

LEMMA 5. *Let the true preference profile be* ϕ *and let* $m(t') \notin \bigcup_{a \in F(\phi)} S_a$. *Then, for all* $a \in F(\phi)$, $P($*for some* $t > t'$, $m(t) \in S_a) > 0$.

Proof. The proof will proceed by dividing the possible initial states into a series of subsets.

[13] The experimental evidence in Sefton and Yavaş [24] supports this view.

Claim 1. For a given ϕ, if $m(t') \notin S_a$, for any $a \in F(\phi)$ and $m(t') \in D_1$, then $P(\text{for some } t > t',\ m(t) \in D_4) > 0$.

Since $m(t') \in D_1$ and $m(t') \notin S_a$ all agents must be announcing a preference profile $\phi' \neq \phi$. By assumption (N1) and the definition of the mechanism, it is a best response for agent $i(\phi, \phi')$ to announce $(\phi, a'(\phi, \phi'))$. Then with positive probability, by assumptions (Y1) and (Y2), agent $i(\phi, \phi')$ will have a chance to update and will choose to announce ϕ. After agent $i(\phi, \phi')$ changes her announcement, any agent $i \neq i(\phi, \phi')$ announcing state ϕ will move the message profile to a state in D_4. If at the same time she announces a high enough integer so that $i = i(m)$, then it will be a best response to do so. Therefore this will happen with positive probability by (Y2).

Claim 2. Let $m(t') \in D_2$. Then $P(\text{for some } t > t',\ m(t) \in D_4) > 0$.

If $m(t')$ is in D_2, and the consensus is a, ϕ', there is some ϕ'' such that the dissident is $i(\phi', \phi'')$. Any agent j other than $i(\phi', \phi'')$ can move the message profile to D_4 by announcing the true preference profile ϕ and a (if $\phi = \phi''$) or $a(\phi', \phi'')$, (if $\phi'' \neq \phi$). In either case there will be three different messages, so the message profile will be in D_4. If j also chooses an integer high enough, she can obtain $b_j(\phi)$, which is a best response to the current strategy. Assumptions (Y1) and (Y2) guarantee that this happens with positive probability.

Claim 3. Let $m(t') \in D_3^j$. Then $P(\text{for some } t > t',\ m(t) \in D_4) > 0$.

If $m(t')$ is in D_3^j, $m_i(t') = (a', \phi', n_i)$ for all $i \neq j$. Any agent other than j can move the message profile to D_4 by announcing a different outcome than a', and by choosing an integer high enough, and the true preference profile ϕ (which may or may not be equal to ϕ'), she can obtain $b_j(\phi)$, which is a best response to $m(t')$. Assumptions (Y1) and (Y2) guarantee that this happens with positive probability.

Claim 4. Let the true preference profile be ϕ and $a \in F(\phi)$. Let $m(t) \in D_4$. Then $P(\text{for some } t' > t,\ m(t') = (a, \phi, n_i)) > 0$.

If $m(t) \in D_4$, we can study two cases. In the first case no agent can move the message profile outside of D_4 by changing the announcement to $(a, \phi, \cdot)$. In that case it is a best response for all agents to choose (a, ϕ, n_i) if n_i is sufficiently high. Assumptions (Y1) and (Y2) guarantee that this happens with positive probability.

In the second case some agent can move the message profile outside of D_4 by changing the announcement to $(a, \phi, \cdot)$. This happens if all but two agents are announcing $(a, \phi, \cdot)$. In that case it is a best response for one of the agents j who announce $(a, \phi, \cdot)$ to change to (a', ϕ, n_j), for $a' \neq a$ as long as n_j is large enough. It is also a best response for the rest of the agents to maintain their announcements about outcome and preference profile as long as they announce a high enough integer. Assumptions (Y1) and (Y2) guarantee that this change by j to a' and the other agents not changing happens with positive probability. But once this has occurred no single agent can move the message profile outside of D_4 by changing the announcement to $(a, \phi, \cdot)$ so we are in the previous case.

LEMMA 6. *Let the true preference profile be* ϕ *and let* $m(t) \in S_a$ *for* $a \in F(\phi)$ *and* $m_i^1(t) = \phi$, *for all i. Then* $m(t') \in S_a$ *for all* $t' > t$.

Proof. If $m(t) \in S_a$, all message profiles are in D_1 and the outcome is a. The only replacements that can change something will lead to a profile in D_2 or D_3^j. Since $m_i^2(t) = \phi$ for all i, assumptions (N1) and (N2) guarantee that these replacements do not mean an improvement for any agent, since a test agent announcing a test outcome for another profile ϕ' will obtain $v_i(a(\phi, \phi'), \phi) < v_i(a', \phi)$ by (N2) and any other deviating announcement (a', ϕ') obtains $v_i(a_i'(a, \phi), \phi) < v_i(a, \phi)$ by (N3). Since deviating messages produce strict losses, assumption (Y3) guarantees that they will not be sent. ▌

Lemma 5 establishes part (a) of Proposition 1. With the addition of Lemma 6 we have that from any message profile there is a lower bound $\varepsilon > 0$ on the probability of reaching $\bigcup_{a \in F(\phi)} S_a$ and staying there forever in a number of steps smaller than some fixed and finite k. So the probability of not reaching $\bigcup_{a \in F(\phi)} S_a$ in kn steps is bounded above by $(1-\varepsilon)^{kn}$. Since $\lim_{n \to \infty} (1-\varepsilon)^{kn} = 0$, part (b) follows. ▌

Proof of Proposition 2. Let an arbitrary $m(0) \notin S_{F(\phi)}$. Then with positive probability the players will change their messages so that $m_i^{-1}(t^{-1}) = \phi_i$ for all i and some $t^{-1} > 0$. That is, the minusoneth announcement of all players will be their true preferences. This happens because by Lemma 1 announcing the agent's own type truthfully in the minusoneth position is weakly dominant so assumptions Y1 and Y2 guarantee this will happen with positive probability. Similarly Lemma 2 and assumption Y1 and Y2 guarantee that with positive probability there is a $t^0 > t^{-1}$ such that $m_i^{-1}(t^0) = \phi_i$, $m^0(t^0) = \phi_{i+1}$ for all i and Lemma 3 and assumption Y1 and Y2 guarantee that there is a time period, $t^1 > t^0$ such that $m_i^{-1}(t^1) = \phi_i$; $m_i^0(t^1) = \phi_{i+1}$ and $m_i^1(t^1) = \phi$ for all i. ▌

Proof of Proposition 3. If $m(t) \in S_{F(\phi)}$, then if agent n changes m_n^{-1} to some $\phi_n' \neq \phi_n$, her payoff does not change by the definition of the mechanism. Y1 and Y4 guarantee that this happens with positive probability. Let $\tilde{\phi}$ be such that $\tilde{\phi}_n = \phi_n'$ and $\tilde{\phi}_i = \phi_i$ for all $i \neq n$. Through a series of claims we show that with positive probability the population message profile goes to $S_{F(\tilde{\phi})}$ that is, $F(\tilde{\phi})$ is implemented.

Claim 1. If $m^{-1}=\tilde{\phi}$, $m_i^0=\phi_{i+1}$ for all $i\in I$ and $m_i^1=\phi$ for all $i\in I$, then

$$v_{n-1}(g(\bar{m}_{n-1}, m_{-(n-1)}), T(\bar{m}_{n-1}, m_{-(n-1)}), \phi_{n-1}) \\ -v_{n-1}(g(m), T(m), \phi_{n-1})>0$$

where $\bar{m}_{n-1}=(\tilde{\phi}_{n-1}, \tilde{\phi}_n, \tilde{\phi})$

$$v_{n-1}(g(\bar{m}_{n-1}, m_{-(n-1)}), T(\bar{m}_{n-1}, m_{-(n-1)}), \phi_{n-1}) \\ -v_{n-1}(g(m), T(m), \phi_{n-1}) \\ =-\eta+\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,F(\phi)-(-\gamma+F(\phi))>0$$

where the equality follows from the definition of the mechanism and the inequality follows from AM1.

Claim 2. If $m^{-1}=\tilde{\phi}$, $m_i^0=\tilde{\phi}_{i+1}$ for all $i\in I$ and $m_i^1\in\{\tilde{\phi}, \phi\}$ (with at least $m_{n-1}^1=\tilde{\phi}$) and $m_i^1=\phi$ for all $i\in I$, then

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i)-v_i(g(m), T(m), \phi_i)>0$$

where $\bar{m}_i=(\tilde{\phi}_i, \tilde{\phi}_i, \tilde{\phi})$

If $m_i^1=\tilde{\phi}$ only for $i=n-1$, then for $i\neq n-1$,

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i)-v_i(g(m), T(m), \phi_i) \\ =\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,b-\left(-\xi+\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,F(\phi)\right)>0$$

If $m_i^1=\tilde{\phi}$ for more than 1 but less than $n-2$ individuals, then for i with $m_i^1=\phi$,

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i)-v_i(g(m), T(m), \phi_i) \\ =\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,b-\left(-\xi+\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,b\right)>0$$

If $m_i^1=\tilde{\phi}$ for $n-2$ individuals, then for i with $m_i^1=\phi$

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i)-v_i(g(m), T(m), \phi_i) \\ =\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,F(\tilde{\phi})-\left(-\xi+\frac{\varepsilon}{n}\sum_{i\in I} f(\tilde{\phi}_i)+(1-\varepsilon)\,b\right)>0$$

If $m_i^1 = \tilde{\phi}$ for $n-1$ individuals, then for i with $m_i^1 = \phi$

$$v_i(g(\bar{m}_i, m_{-i}), T(\bar{m}_i, m_{-i}), \phi_i) - v_i(g(m), T(m), \phi_i)$$

$$= F(\tilde{\phi}) - \left(-\xi - \eta + \frac{\varepsilon}{n} \sum_{i \in I} f(\tilde{\phi}_i) + (1-\varepsilon)\, F(\phi) \right) > 0$$

where the equalities follow from the definition of the mechanism and the inequalities follow from AM1.

The claims show that $S_{F(\tilde{\phi})}$ is attained with positive probability because they show a series of changes in the messages, all of which are improving. Thus assumptions Y1 and Y4 guarantee that the sequence will take place with positive probability.

We have shown that there is positive probability of a transition between $S_{F(\phi)}$ and $S_{F(\tilde{\phi})}$ where $\tilde{\phi}$ differs from ϕ only in $\phi'_n \neq \phi_n$. But if the $m(t) \in S_{F(\tilde{\phi})}$, it is costless for individual $n-1$ to change $m_{n-1}^{-1} = \phi'_{n-1} \neq \phi_{n-1}$. By applying analogs of Claims 1 through 4 we can then show that with positive probability there is a time t' such that $m(t') \in S_{F(\ddot{\phi})}$, where $\ddot{\phi} = (\phi_1, ..., \phi'_{n-1}, \phi'_n)$. If we iterate this argument, the result follows. ▌

REFERENCES

1. D. Abreu and H. Matsushima, Virtual implementation in iteratively undominated strategies: complete information, *Econometrica* **60** (1992), 993–1008.
2. D. Abreu and H. Matsushima, Exact implementation, *J. Econ. Theory* **64** (1994), 1–19.
3. K. Binmore and L. Samuelson, "Evolutionary Drift and Equilibrium Selection," Working Paper 26, Institute for Advanced Studies, Vienna, 1996.
4. G. W. Brown, Iterative solutions of games by fictitious play, *in* "Activity Analysis of Production and Allocation," Wiley, New York, 1951.
5. A. Cabrales and G. Ponti, "Implementation, Elimination of Weakly Dominated Strategies and Evolutionary Dynamics," Working Paper 221, Universitat Pompeu Fabra, 1997.
6. A. Cabrales and J. Sobel, On the limit points of discrete selection dynamics, *J. Econ. Theory* **57** (1992), 407–420.
7. L. C. Corchón, "The Theory of Implementation of Socially Optimal Decisions in Economics," McMillan Press, London, 1996.
8. P. Dasgupta, P. Hammond, and E. Maskin, The implementation of social choice rules: Some general results on incentive compatibility, *Rev. Econ. Stud.* **46** (1979), 185–216.
9. I. Gilboa and A. Matsui, Social stability and equilibrium, *Econometrica* **59** (1991), 859–867.
10. J. Glazer and R. W. Rosenthal, A note on Abreu-Matsushima mechanisms, *Econometrica* **60** (1992), 1435–1438.
11. T. Groves and J. Ledyard, Optimal allocation of public goods: a solution to the free rider problem, *Econometrica* **45** (1977), 783–809.
12. S. Hurkens, Learning by forgetful players, *Games Econ. Behav.* **11** (1995), 304–329.
13. M. O. Jackson, Implementation in undominated strategies: A look at bounded mechanisms, *Rev. Econ. Stud.* **59** (1992), 757–775.

14. M. O. Jackson, T. R. Palfrey, and S. Srivastava, Undominated Nash implementation in bounded mechanisms, *Games Econ. Behav.* **6** (1994), 474–501.
15. J. S. Jordan, Instability in the implementation of Walrasian allocations, *J. Econ. Theory* **39** (1986), 301–328.
16. Y. G. Kim and J. Sobel, An evolutionary approach to pre-play communication, *Econometrica* **63** (1995), 1181–1193.
17. E. Maskin, "Nash Implementation and Welfare Optimality," mimeo, Massachusetts Institute of Technology, 1977.
18. J. Moore, Implementation in environments with complete information, *in* "Advances in Economic Theory: Sixth World Congress" (J. J. Laffont, Ed.), Econometric Society Monograph, Cambridge University Press, Cambridge, 1992.
19. J. Moore and R. Repullo, Subgame perfect implementation, *Econometrica* **58** (1988), 1083–1099.
20. T. Muench and M. Walker, Are Groves-Ledyard equilibria attainable?, *Rev. Econ. Stud.* **50** (1984), 393–396.
21. J. Nachbar, Evolutionary selection dynamics in games: Convergence and limit properties, *Int. J. Game Theory* **19** (1990), 59–89.
22. R. Repullo, A simple proof of Maskin's theorem on Nash implementation, *Soc. Choice Welfare* **4** (1987), 39–41.
23. L. Samuelson and J. Zhang, Evolutionary stability in asymmetric games, *J. Econ. Theory* **57** (1992), 363–392.
24. M. Sefton and A. Yavaş, Abreu-Matsushima mechanisms: Experimental evidence, *Games Econ. Behav.* **16** (1996), 280–302.
25. T. Sjöström, Implementation in undominated Nash equilibria without integer games, *Games Econ. Behav.* **6** (1994), 502–511.
26. P. D. Taylor and L. B. Jonker, Evolutionary stable strategies and game dynamics, *Math. Biosci.* **40** (1978), 145–156.
27. P. de Trenqualye, Stability of the Groves and Ledyard mechanism, *J. Econ. Theory* **46** (1988), 164–171.
28. P. de Trenqualye, Stable implementation of Lindahl allocations, *Econ. Lett.* **29** (1989), 291–294.
29. F. Vega-Redondo, Implementation of Lindahl equilibrium: An integration of the static and dynamic approaches, *Math. Soc. Sci.* **18** (1989), 211–228.
30. M. Walker, A simple auctioneerless mechanism with Walrasian properties, *J. Econ. Theory* **32** (1984), 111–127.