Games and Economic Behavior 90 (2015) 1–16

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Direct implementation with minimally honest individuals

Juan Ortner [1]

Boston University, Department of Economics, 270 Bay State Road, Boston, MA 02215, United States

ARTICLE INFO

Article history:
Received 18 September 2012
Available online 16 January 2015

JEL classification:
C72
C73
D71
D78

Keywords:
Implementation
Mechanism design
White lie aversion
Direct mechanisms

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.

1. Introduction

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

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

The implementation literature addresses the issue of multiple equilibria by seeking more complicated mechanisms with richer message spaces. In other words, the literature focuses on mechanisms that require players to make additional announcements besides the information that is directly relevant to the environment. The most notable example of this augmentation of the message space is Maskin's (1999) integer game. Despite the success that the theory has had in characterizing implementable social choice functions, the complex message spaces and game forms required to achieve full

E-mail address: jortner@bu.edu.

[1] I am indebted to Faruk Gul and Satoru Takahashi for their advice and guidance, and to Avidit Acharya, Ben Brooks, Edoardo Grillo, Navin Kartik and Stephen Morris for useful comments. I also thank the editor and an anonymous referee for very thoughtful suggestions. All remaining errors are my own.

http://dx.doi.org/10.1016/j.geb.2015.01.002
0899-8256/

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.[2]

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

The first refinement I consider is *fault tolerant Nash equilibrium*, introduced by Eliaz (2002). The idea behind fault tolerant Nash equilibrium is that players may not know whether all of their opponents are rational. Suppose that players believe that there are at most k irrational agents in the population, but that they know neither the identity of the irrational players, nor how irrational players behave. A *k-fault tolerant Nash equilibrium* (k-FTNE) is a strategy profile that is robust to the presence of k irrational agents: under a k-FTNE each agent has an incentive to play her equilibrium action, regardless of the identity and actions of irrational players, as long as $n-k-1$ of her opponents adhere to equilibrium behavior.

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

The direct mechanism I use is a majoritarian aggregation rule. The planner asks each agent to announce the state of nature. If more than half the population announces the same state θ, the mechanism's outcome is $f(\theta)$. In any other case, the outcome is some fixed alternative a^*. Importantly, a^* need not be a particularly bad outcome, and it may even be a Pareto optimal alternative. Besides its simplicity, from a practical point of view this mechanism has the attractive feature of being anonymous (i.e., the outcome is unchanged if agents are permuted) and completely independent of the preferences of the agents. The strategic game that this mechanism induces may have multiple equilibria. However, if there are at least five players and they are all white lie averse, then both fault tolerance and stochastic stability yield the same unique prediction: all agents make truthful announcements and the planner is able to implement the desired alternative.[3]

Other papers have studied implementation under the refinements that I consider. Eliaz (2002) studies complete information implementation in k-FTNE. Adapting Maskin's (1999) canonical mechanism, he shows that any social choice correspondence satisfying *k-monotonicity* and no veto power can be implemented in k-FTNE. In contrast, the current paper shows that any social choice function can be implemented in k-FTNE with a simple direct mechanism when players have a small preference for honesty. Sandholm (2007) studies implementation in stochastically stable equilibrium in an environment with externalities in which a planner wants the agents to choose an utilitarian action profile. Sandholm (2007) shows that the planner can achieve this objective by introducing a simple tax scheme under which each agent pays for the externalities she creates. In contrast to Sandholm (2007), the current paper studies a general implementation problem in which transfers among individuals may not be possible.[4] Neither Eliaz (2002) nor Sandholm (2007) study implementation with white lie averse agents.

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

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

The idea of studying implementation when agents have a minimal degree of honesty is due to Matsushima (2008a), who considers the problem of implementing a social choice function in a complete information setup with three white lie averse agents. He shows that, in this environment, any social choice function can be exactly implemented in iteratively undominated strategies with a mechanism similar to the one in Abreu and Matsushima (1992). Matsushima (2008b) considers implementation in Bayesian environments when agents have an intrinsic preference for being honest and shows that any

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

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

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