Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 163 (2016) 557–603

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Mechanism design and intentions

Felix Bierbrauer [a,b], Nick Netzer [c,*]

[a] *Max Planck Institute for Research on Collective Goods, Bonn, Germany*
[b] *University of Cologne, Center for Macroeconomic Research, Albertus-Magnus-Platz, 50923 Köln, Germany*
[c] *University of Zurich, Department of Economics, Bluemlisalpstrasse 10, CH-8006 Zurich, Switzerland*

Received 24 November 2014; final version received 26 February 2016; accepted 27 February 2016

Available online 2 March 2016

Abstract

We introduce intention-based social preferences into mechanism design. We explore information structures that differ with respect to what is commonly known about the weight that agents attach to reciprocal kindness. When the designer has no information on reciprocity types, implementability of an incentive-compatible social choice function is guaranteed if it satisfies an additional insurance property. By contrast, precise information on reciprocity types may imply that all efficient social choice functions are implementable. We show how these results extend to a two-dimensional mechanism design setting where the agents have private information about their material payoff types and their reciprocity types. We also provide a systematic account of the welfare implications of intentionality.

JEL classification: C70; C72; D02; D03; D82; D86

Keywords: Mechanism design; Psychological games; Social preferences; Reciprocity

1. Introduction

Agents with intention-based social preferences are willing to give up own material payoffs in order to either reward behavior by others that they attribute to good intentions, or to punish behavior that they attribute to bad intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004).

* Corresponding author.
E-mail addresses: bierbrauer@wiso.uni-koeln.de (F. Bierbrauer), nick.netzer@econ.uzh.ch (N. Netzer).

http://dx.doi.org/10.1016/j.jet.2016.02.013

The behavioral relevance of such preferences is well established (e.g. Andreoni et al., 2002; Falk et al., 2003; Falk et al., 2008). In this paper, we explore their implications for the theory of mechanism design. Specifically, we provide answers to the following questions:

There is a rich literature on mechanism design that has proceeded under the assumption that agents are selfish. To what extent are these mechanisms robust to the possibility that the participants may be motivated by intention-based social preferences?

How do intention-based social preferences affect the set of implementable social choice functions relative to a benchmark with selfish agents? In particular, does intentionality make it easier or more difficult to implement good outcomes?

Suppose that the designer seeks not only good material outcomes but also good attitudes among the participants of the mechanism. Is there a trade-off between these objectives? Do we have to sacrifice efficiency if we want kindness among the agents, or are sensations of kindness helpful for the implementation of efficient outcomes?

For clarity of exposition, our analysis is based on one particular model of intention-based social preferences. Specifically, we adapt the model by Rabin (1993) to games of incomplete information and work with the solution concept of a Bayes–Nash fairness equilibrium, in the context of an otherwise conventional independent private values model of mechanism design.

We approach the questions above in three different ways. We first characterize social choice functions that are *strongly implementable*. Our notion of strong implementability is attractive from the perspective of a mechanism designer who acknowledges the possibility that the agents may be motivated by intention-based social preferences, but who wishes to remain agnostic about the intensity of these preferences. A strongly implementable social choice function is implementable irrespective of the mixture between selfish and reciprocal individuals among the participants of a mechanism. We then consider social choice functions that are *weakly implementable*, i.e., which are implementable if the mechanism designer has precise information on the strength of intention-based social preferences. This concept is of interest for two different reasons. First, it allows for a clear exposition of the conceptual issues that arise due to the procedural nature of intention-based preferences. For instance, we show that the revelation principle does not hold. We also discuss alternative notions of welfare and the modelling of participation constraints in a model with intentions. Second, looking first at weakly implementable social choice functions sets the stage for our analysis of the *two-dimensional mechanism design* problem that emerges if the agents have private information both about their material payoffs and about the weight that kindness sensations have in their utility function.

Strongly implementable social choice functions Our first main result (Theorem 1) states that a social choice function is strongly implementable if it is implementable in a model with selfish agents and, moreover, is such that the agents cannot affect each other's payoff by unilateral deviations from truth-telling. We refer to the latter property as the *insurance property*, since it implies that the expected payoff of agent i does not depend on the type of agent j, i.e., each agent is insured against the randomness of the other agent's type. The insurance property shuts down the transmission channel for reciprocal behavior. If agent i cannot influence the payoff of agent j, then j has no reason to interpret i's behavior as kind or as unkind. Agent j thus neither has a reason nor an opportunity to reward or punish agent i, so she focusses on her own expected