Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Kantian optimization: A microfoundation for cooperative behavior[☆]

John E. Roemer [*]

Department of Political Science, Yale University, United States
Department of Economics, Yale University, United States
Cowles Foundation, Yale University, United States

ARTICLE INFO

Article history:
Received 18 January 2013
Received in revised form 19 January 2014
Accepted 27 March 2014
Available online xxxx

JEL classification:
D60
D62
D64
C70
H30

Keywords:
Kantian equilibrium
Tragedy of the commons
Other-regarding preferences
Cooperation
Implementation

ABSTRACT

Although evidence accrues in biology, anthropology and experimental economics that *homo sapiens* is a cooperative species, the reigning assumption in economic theory is that individuals optimize in an autarkic manner (as in Nash and Walrasian equilibrium). I here postulate a cooperative kind of optimizing behavior, called Kantian. It is shown that in simple economic models, when there are negative externalities (such as congestion effects from use of a commonly owned resource) or positive externalities (such as a social ethos reflected in individuals' preferences), Kantian equilibria dominate the Nash–Walras equilibria in terms of efficiency. While economists schooled in Nash equilibrium may view the Kantian behavior as utopian, there is some – perhaps much – evidence that it exists. If cultures evolve through group selection, the hypothesis that Kantian behavior is more prevalent than we may think is supported by the efficiency results here demonstrated.

1. Introduction

Recent work in contemporary social science and evolutionary biology emphasizes that *homo sapiens* is a cooperative species. In evolutionary biology, scientists are interested in explaining how cooperation and 'altruism' may have developed among humans through natural selection. In economics, there is now a long series of experiments whose results are often explained by the hypothesis that individuals are to some degree altruistic. A recent summary of the state-of-the-art in experimental economics, anthropology, and evolutionary biology is provided by Bowles and Gintis (2011). Rabin (2006) provides a summary of the evidence for altruism from experimental economics. An anthropological view is provided in Henrich and Henrich (2007). Tomasello (2009) describes experiments that indicate that the urge to cooperate in human babies is inborn, while it does not exist in chimpanzees. Alger and Weibull (in press) model the evolution of altruism, and provide a useful bibliography.

Altruism may induce behavior that appears to be cooperative, but altruism and cooperation have different motivations. Altruism, at least when it is intentional in humans, is motivated by a desire to improve the welfare of others, while cooperation may be motivated (only) by the desire to help oneself. (For example, workers in a firm cooperate, but each may do so because she realizes that cooperative behavior advances her own welfare.) There is an important line of research, conducted by Ostrom (1990) and her collaborators, arguing that, in many small societies, people figure out how to cooperate to avoid, or solve, the 'tragedy of the commons.' That tragedy may be summarized as follows. Imagine a lake which is owned in common by a group of fishers, who each possess preferences over fish and leisure, and perhaps differential skill (or sizes of boats) in (or for) fishing. The lake produces fish with decreasing returns with respect to the fishing labor expended upon it. In the game in which each fisher proposes as her strategy a fishing time, it is well known that the Nash equilibrium is Pareto inefficient: there are congestion externalities, and all would be better off were they able to design a decrease, of a certain kind, in everyone's fishing. Ostrom studied many such societies, and maintained that many or most of them learn to regulate 'fishing,' without privatizing the 'lake.' Somehow,

[☆] My main debt is to Joaquim Silvestre, with whom I have discussed these ideas for many years. I also thank the following, for their suggestions and help: Roger Howe, François Maniquet, Juan Moreno-Ternero, Joseph Ostroy, Larry Samuelson, Rajiv Sethi, Joel Sobel, and referees for this journal.

[*] Department of Political Science, Yale University, United States.
E-mail address: john.roemer@yale.edu.

http://dx.doi.org/10.1016/j.jpubeco.2014.03.011
0047-2727/© 2014 Elsevier B.V. All rights reserved.

the inefficient Nash equilibrium is avoided. This example is not one in which fishers care about other fishers (necessarily), but it is one in which cooperation is organized to deal with a negative externality of autarkic behavior.

The ethos that motivates cooperation is called *solidarity*. Merriam-Webster's dictionary defines solidarity as 'unity (as a group or class) that produces or is based on community of interests or objectives.' There is no mention of altruism: we do not cooperate because we care about *others*, but because we recognize we are *all in the same boat*, and cooperation will advance each individual interest. Of course, *if* altruism exists, it may also motivate cooperation, but I wish to emphasize that cooperation does not require altruism.

Ostrom's observations pertain to small societies. In large economies, we observe the evolution of the welfare state, supported by considerable degrees of taxation of market earnings. It is conventionally argued that the successful welfare states had their genesis in solidarity: they provided insurance which was in everyone's self-interest. It was easier to organize welfare states where citizens were ethnically and linguistically homogeneous, because the 'unity' which Merriam-Webster refers to was more evident in this case. We do not need to invoke altruism among the citizens of Nordic societies to explain the welfare state: in other words, their *homogeneity* was the source of their recognition of common interests, but it need not have induced altruism to generate the welfare state.

There is, however, also an argument that welfare states expand after wars as a reward to returning soldiers; see Scheve and Stasavage (2012). Perhaps altruism develops in a population as a result of their participation in a cooperative venture: we identify more with others when we succeed in cooperating, and that identification may lead to altruism. Or we feel soldiers deserve a reward for having fought the war. Redistributive taxation appears to be at least to some degree a polity's reaction to the material deprivation of a section of society, which many view as undeserved, and desire to redress. To the extent that welfare states provide insurance which it is rational for self-interested agents to desire, it is a manifestation of cooperation; to the extent that citizens support the welfare state to redress unjust inequality, it is a manifestation of altruism, or at least of a sense of justice. Regardless of the motive, as is well known, redistributive taxation induces, to some degree, allocative inefficiency. I will argue that this is due in large part to non-cooperative behavior of individual workers when they face the tax regime. Each worker is computing his optimal labor supply in the Nash fashion: that is, assuming that all others are holding their labor supplies fixed.

Among economists, there have been a number of strategies to explain behavior that is not easily explained as the Nash equilibrium of the game that agents appear to be playing. Ostrom explains the avoidance of the tragedy of the commons among 'fishing communities' by the imposition of punishment of those who deviate from the cooperative behavior: in other words, the payoffs of the game are changed so that it becomes a Nash equilibrium for each fisher to cooperate. This is also the argument that Olson (1965) employs to explain cooperation: unions, for example, get workers to cooperate by offering side payments (carrots) to those who participate, and punishments (sticks) for those who deviate. In experimental economics, when individuals often do not play what appears to be the Nash equilibrium of a game (dictator and ultimatum games, for example), there are a number of moves. Perhaps individuals are using rules of thumb that are associated with strategies that are equilibria in repeated games, even though the game in the laboratory is not repeated. Or perhaps players have other-regarding preferences: they are to some degree altruistic. Or perhaps they have a sense of morality, which can be viewed as a kind of preference — a player feels better when, in the dictator game, she gives something to the opponent. Or, in the ultimatum game, the proposer offers a substantial amount to the opponent because she believes the opponent does not have classical preferences — that is, Opponent will reject an 'unfair' offer. Outcomes are then explained as Nash equilibria of games whose players have non-classical (i.e., non-self-interested) preferences.

Here, I introduce another approach. I propose that we can explain cooperation by observing that players may be *optimizing* in a non-classical (that is, non-Nash) manner. This leads to a class of equilibrium concepts that I call Kantian equilibria. Briefly, with Kantian optimization, agents ask themselves, at a particular set of actions/strategies in a game: If I were to deviate from my stipulated action, *and all others were to deviate in like manner from their stipulated actions*, would I prefer the consequences of the new action profile? I denote this kind of thinking *Kantian* because an individual only deviates in a particular way, at an action profile, if he would prefer the situation in which his action were *universalized* — that is to say, he'd prefer the action profile where all make the kind of deviation he is contemplating. Each agent evaluates *not* the profile that would result if *only he* deviated, but rather the profile of actions that would result if *all* deviated in similar fashion. Kant's categorical imperative says: take those and only those actions that are universalizable, meaning that the world would be better (according to one's own preferences) were one's behavior universalized. It is important that the new action profile be evaluated with one's own preferences, which need not be altruistic.

There is a distinction, then, between the approach of behavioral economics, which has by and large focused on amending *preferences* from self-interested ones to altruistic or other-regarding ones, or ones in which players possess a sense of justice, to the approach I describe, which amends *optimizing behavior,* but does not (necessarily) fiddle with preferences. Of course, one could be even more revisionist, and amend *both* optimizing behavior and preferences, leading to the four-fold taxonomy of modeling approaches summarized in Table 1.

The purpose of the present inquiry is to study whether the inefficiency of Nash equilibrium can be overcome with Kantian optimization — both cases in the bottom row of Table 1. I hope to clarify, in what follows, my claim that varying *preferences* as a modeling technique differs from the strategy of varying *optimizing protocols*. The first strategy alters the column of the matrix in Table 1 in which the researcher works, while the second alters the row.

Let me comment further on the distinction between Nash and Kantian behavior. It is noteworthy that economists have devoted very little thought to modeling cooperation. We have a notion of cooperative games, but that theory represents cooperation in an extremely reduced form. Cooperative behavior is not modeled, but is simply represented by defining values of coalitions. How do coalitions come to realize these values? The theory is silent on the matter. If an imputation is in the core of a cooperative game, it is, a fortiori, Pareto efficient: typically, one is concerned with whether cooperative games contain non-empty cores, but the behavior which leads to an imputation in the core is typically not studied. A major exception to this claim is the theorem that non-cooperative, autarkic optimizing behavior, in a perfectly competitive market economy, induces an equilibrium that lies in the core of an associated game. But this is an exception to my claim, not the rule. In contrast, the Shapley value of a convex cooperative game is in the core: but I do not think anyone derives the Shapely value as the outcome of optimizing behavior of individuals.

I wish to propose that Kantian optimization can be viewed as a model of cooperation. As a Kantian optimizer, I hold a norm that says: "If I want to deviate from a contemplated action profile (of my community's members), then I may do so only if I would have all others deviate 'in like manner.'" I have not spelled out what the phrase 'in like

Table 1
Taxonomy of possible models.

Optimization	Preferences	
	Self-interested	Other-regarding
Nash	Classical	Behavioral economics
Kantian	This paper, Sections 3 and 5	This paper, Section 6