Journal of Public Economics 127 (2015) 45–57

Contents lists available at ScienceDirect

Journal of Public Economics

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

Kantian optimization: A microfoundation for cooperation[☆]

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 13 April 2014

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

manner' means, as yet — that will comprise the details of this paper. Contrast this kind of thinking with the *autarkic* thinking postulated in Nash behavior — I change my action by myself, assuming that others in my community stand pat.

In Section 2, the economic environment for this inquiry is specified. Section 3 introduces two examples of Kantian optimization and proves that they produce Pareto efficient outcomes — they resolve different kinds of commons' tragedies that can afflict societies living in these economic environments. Section 4 takes up two possible objections to the approach, and argues more explicitly that Kantian optimization is not equivalent to altering agents' preferences. Section 5 presents a more general theory of Kantian optimization. Section 6 introduces altruism into agents' preferences, and studies whether Kantian optimization will continue to produce Pareto efficient outcomes. Section 7 contains a brief discussion of the existence of Kantian equilibria, and their dynamics. Section 8 concludes.[1]

2. The economic environment

There is a concave, differentiable production function G that produces a single output from a single input, called effort. Effort is supplied by individuals; it may differ in intensity or efficiency units, but effort, measured in efficiency units, can be aggregated across individuals. We assume, except in Section 6, that there are a finite number of individuals, n. If the sum of individual efforts is E^S then total production is $G(E^S)$. We denote the effort expended by an agent of type γ by E^γ. It is assumed that effort is unbounded above but bounded below by zero. Let the class of such production functions be denoted $\mathbf{G}$.

An individual of type γ has preferences represented by a utility function $u^\gamma(x, E)$ where x is consumption and E is effort. A person's utility depends only on her own consumption and effort, until Section 6 below.

An *allocation rule* is a mapping $X:\mathfrak{R}^n_+\times\mathbf{G}\rightarrow\mathfrak{R}^n_+$. If the vector of efforts is $E=(E^1,\ldots,E^\gamma,\ldots,E^n)$ then $X(E, G)$ is the allocation of output to individuals under the rule X. If we write $X=(X^1,\ldots,X^n)$ as a vector of real-valued functions, then $X^\gamma(E^1,\ldots,E^n, G)$ is the amount of output produced which agent γ receives. Thus, it is identically true that for any non-wasteful allocation rule, $\sum_\gamma X^\gamma\left(E^1,\ldots,E^n,G\right)\equiv G\left(E^S\right)$.

We will also at times write allocation rules in terms of the *shares* of output that they induce: that is an allocation rule X induces a vector of shares assigned to individuals, given by $X^\gamma(E, G)=\theta^\gamma(E, G)G(E^S)$. Of course, $\sum_\gamma\theta^\gamma\equiv 1$.

An *economic environment e* is specified by a profile of utility functions and a production function: $e=(u^1,\ldots,u^n, G)$ or $(\mathbf{u}, G, n)$ where $\mathbf{u}$ is the profile of utility functions. An *economy* is a pair (e, X). An economy induces a *game* among the population: for any vector of efforts, each can compute her utility. That is, define the payoff functions $\{V^\gamma\}$ by:

$$V^\gamma\left(E^1,\ldots,E^n\right)=u^\gamma(X^\gamma(E,G),E^\gamma),\text{where } E=\left(E^1,\ldots,E^n\right). \tag{2.1}$$

For example, consider the fishing economy described in Section 1. It is assumed that each fisher keeps his catch. Thus, statistically speaking, the amount of fish received by fisher γ will be proportional to the fraction of total labor, in efficiency units, that he expends. The allocation rule is given by:[2]

$$\theta^{\gamma,\Pr}\left(E^1,\ldots,E^n\right)=\frac{E^\gamma}{E^S}. \tag{2.2}$$

For obvious reasons, this is called the proportional (Pr) allocation rule. The game induced by the proportional allocation rule has payoff functions:

$$V^\gamma\left(E^1,\ldots,E^n\right)=u^\gamma\left(\frac{E^\gamma}{E^S}G\left(E^S\right),E^\gamma\right). \tag{2.3}$$

The 'tragedy of the commons' is the statement that if G is strictly concave, then the Nash equilibria of the game defined by Eq. (2.3) are Pareto inefficient: indeed all would be better off by suitable reductions in their effort from the Nash effort allocation.

Another important rule is the *equal-division allocation rule*, given by:

$$\theta^{\gamma,ED}\left(E^1,\ldots,E^n\right)=\frac{1}{n}, \tag{2.4}$$

and a third class of rules are the *Walrasian allocation rules,* given by:

$$\theta^{\gamma,Wa}\left(E^1,\ldots,E^n,G\right)=\frac{G'\left(E^S\right)}{G(E^S)}E^\gamma+\sigma^\gamma\left(1-\frac{G'\left(E^S\right)E^S}{G(E^S)}\right), \tag{2.5}$$

in which an agent receives output equal to her effort multiplied by the Walrasian wage plus her share (σ^γ) of profits. Note that the Walrasian shares *do* depend upon G, unlike the proportional and equal-division shares, and this illustrates why, in general, we allow θ to depend upon G as well as the effort vector.

Denote the class of economic environments $(\mathbf{u}, G, n)$ in which n is finite, $G\in\mathbf{G}$, and the u^γ are concave, differentiable functions, by $\mathfrak{E}$. Denote the sub-class of economic environments where G is linear by $\mathfrak{L}$.

3. Kantian equilibrium in non-altruistic economies

We may formalize the idea of Kantian optimization as follows. Let $\{V^\gamma\}$ be a set of payoff functions for a game played by types γ, where the strategy of each player is a non-negative effort E^γ, and E is the effort profile of the players. A *multiplicative Kantian equilibrium* is an effort profile E^* such that *nobody would prefer that everybody alter his effort by the same non-negative factor.* That is:

$$(\forall\gamma)(\forall r\geq 0)(V^\gamma(E*^\gamma)\geqslant V^\gamma(rE*^\gamma)) \tag{3.1}$$

In Roemer (1996, 2010), this concept was simply called 'Kantian equilibrium.'

The remarkable feature of multiplicative Kantian equilibrium is that it resolves the tragedy of the commons in the fishers' economy. It is proved in the two citations just mentioned that *if a strictly positive effort allocation is a multiplicative Kantian equilibrium in the game defined by* Eq. (2.3)*, then it is Pareto efficient in the economy* $e=(u, G)$. This is a stronger statement than saying the allocation is efficient in the game $\{V^\gamma\}$: for in the *game*, only certain types of allocation are permitted — ones in which fish are distributed in proportion to effort expended. But the *economy* defines any allocation as feasible, as long as $\sum_\gamma x^\gamma=G\left(E^S\right)$. So Kantian behavior, if adopted by individuals, resolves the tragedy of the commons.

The Kantian counterfactual (that each agent consider only deviations from an effort allocation if *all* deviate by a common factor) forces

[1] Ioriginally proposed a definition of Kantian equilibrium in Roemer (1996), and showed its relationship to the 'proportional solution,' of Roemer and Silvestre (1993). In Roemer (2010), I investigated a special case of Kantian equilibrium, that I now call *multiplicative* Kantian equilibrium. The present paper shows that there are many versions of Kantian optimization, and characterizes when they deliver efficient outcomes in the presence of the various kinds of externalities in which Nash equilibrium performs poorly. As well as extending the results of Roemer (2010) in a number of ways, this paper offers a clearer argument about the distinction between preferences and optimization protocols.

[2] For this rule, the shares θ^γ do not depend upon G.

each to consider the negative externality others would impose on her, if all deviated from a proposed effort vector in a particular way. Thus, universalizing the action she is considering (to re-scale her effort by a positive factor) forces her to internalize the externality in a certain way. It is interesting that she does not consider the effect of her action, if universalized, upon others, but only on herself. She evaluates the counterfactual from her own self-interested viewpoint. It is therefore somewhat surprising that Kantian thinking, so defined, generates Pareto efficient equilibria.[3]

A *proportional solution* in the fisher economy is defined as an allocation $(x, E) = (x^1,\ldots, x^n, E^1,\ldots, E^n)$ with two properties:

(i) for all γ, $x^\gamma = \frac{E^\gamma}{E^S} G\left(E^S\right)$, and
(ii) (x, E) is Pareto efficient.

The proportional solution was introduced in Roemer and Silvestre (1993), although the concept of (multiplicative) Kantian equilibrium came later. The proportional solutions of the fisher economy are exactly its positive multiplicative Kantian equilibria (see Theorem 1). In the small societies which Ostrom has studied, which are (in the formal sense) usually 'economies of fishers' where each individual 'keeps his catch,' she argues that internal regulation assigns 'fishing times' that often engender a Pareto efficient allocation. If this is so, these allocations are proportional solutions, and therefore (by the theorem just quoted) they are multiplicative Kantian equilibria in the game where participating fishers/hunters/miners propose labor times for accessing a commonly owned resource. This suggests that small societies discover their multiplicative Kantian equilibria. Ostrom (1990), however, does not provide any evidence for Kantian thinking among citizens of these societies: as mentioned earlier, she explains these good allocations as *Nash* equilibria of games with altered payoffs. Knowing the theory of multiplicative Kantian equilibrium, one is tempted to ask whether a 'Kantian optimization protocol' exists in these small societies, which leads to the discovery of the Pareto efficient equilibrium.

I now introduce a second Kantian protocol which leads to a notion of *additive Kantian equilibrium.*[4] An effort profile E is an additive Kantian equilibrium if and only if no individual would have all individuals add the same amount of effort (positive or negative) to everyone's present effort. That is:

$$(\forall\gamma)\left(\forall r \geq -\inf_{\tau} E^\tau\right)\left(V^\gamma(E) \geq V^\gamma(E+r)\right), \tag{3.2}$$

where $E + r$ is the effort profile in which the effort of type γ individuals is $E^\gamma + r$. The lower bound stipulated on r is necessary to avoid negative efforts. Additive Kantian equilibrium again postulates that each person 'internalizes' the effects of his contemplated change in effort, but now the variation is *additive* rather than multiplicative.

In the sequel, I will denote these two kinds of Kantian behavior as $K^\times$ and K^+.

We have:

Theorem 1

A. *Any strictly positive* $K^\times$ *equilibrium with respect to the proportional allocation rule is Pareto efficient on the domain. Any strictly positive* K^+ *equilibrium with respect to the equal-division allocation rule is Pareto efficient on the domain.*
B. *Conversely, any proportional allocation which is Pareto efficient is a* $K^\times$ *equilibrium and any equal-division allocation that is Pareto efficient is a* K^+ *equilibrium.*

Proof

1. Part *A*. Let $E = (E^1,\ldots, E^n)$ be a strictly positive equilibrium w.r.t. the proportional allocation rule θ^{Pr}. The first-order condition stating this fact is:

$$(\forall\gamma)\frac{d}{dr}\bigg|_{r=1} u^\gamma\left(\frac{rE^\gamma}{rE^S}G\left(rE^S\right), rE^\gamma\right) = 0, \tag{3.3}$$

which means:

$$(\forall\gamma)\left(u_1^\gamma \cdot \left(\frac{E^\gamma}{E^S}G'\left(E^S\right)E^S\right) + u_2^\gamma E^\gamma\right) = 0. \tag{3.4}$$

Since $E^\gamma > 0$, divide through Eq. (3.4) by E^γ, giving:

$$(\forall\gamma)\left(-\frac{u_2^\gamma}{u_1^\gamma} = G'\left(E^S\right)\right). \tag{3.5}$$

Eq. (3.5) states that the marginal rate of substitution between income and effort is, for every agent, equal to the marginal rate of transformation, which is exactly the condition for Pareto efficiency at an interior solution. This proves the first claim.
2. For the second claim, let E be a K^+ equilibrium w.r.t. the equal-division allocation rule θ^{ED} for any economy in $\mathfrak{E}$. Then:

$$(\forall\gamma)\left(\frac{d}{dr}\bigg|_{r=0} u\left(\frac{G\left(E^S + nr\right)}{n}, E^\gamma + r\right) = 0\right), \tag{3.6}$$

which expands to:

$$(\forall\gamma)\left(u_1^\gamma \cdot G'\left(E^S\right) + u_2^\gamma = 0\right) \tag{3.7}$$

(Strict positivity of E is here used so that the range of r includes a small neighborhood of zero.) Clearly Eq. (3.7) implies Eq. (3.5), and again the allocation is Pareto efficient.
3. Part *B*. This follows simply by reading the proof backwards. ■

Examine the proof of the first part of this proposition, and compare the reasoning that agents who are Kantian employ to Nash reasoning. When a fisher contemplates increasing his effort on the lake by 10%, she asks herself, "How would I like it if everyone increased his effort by 10%?" She is thereby forced to internalize the externality that others would impose upon her, were her action universalized, when G is strictly concave.

A similar story applies to the additive Kantian equilibrium with respect to the equal-division rule. The Nash equilibrium of the game induced by the equal-division rule is Pareto inefficient, as long as G is strictly concave — but in this case, agents apply too *little* effort at the Nash allocation. But with the K^+ optimization protocol, agents internalize the effect of their working too little. The equal-division allocation rule is often said to apply to hunting economies: unlike fishers, when tribes hunted for big game, it was common to divide the catch equally among all. Hunting economies, using the equal-division rule, will be plagued by the inefficiency of individuals shirking (taking a nap behind a bush while others carry on), but their problem can be resolved if all use the additive Kantian protocol. Some of the early Israeli kibbutzim used the equal-division rule: regardless of efforts expended, the product was divided equally among households (or perhaps in proportion to family size). An additive Kantian optimization protocol would therefore have generated Pareto efficient allocations.

Theorem 1 states that *each* method of Kantian optimization (multiplicative or additive) engenders Pareto efficient results in the games induced by *particular* allocation rules (proportional, equal division). Although generally Kantian optimization forces agents to internalize

[3] I am grateful to a referee for making this observation.
[4] This variation of Kantian equilibrium was proposed to me by J. Silvestre in 2004.

externalities associated with strictly concave production functions in these economic environments, the optimization protocols do not *completely* resolve the inefficiencies associated with these externalities except when the allocation rule is the right one.

We emphasize that, in Kantian optimization, agents evaluate deviations from their own viewpoints, as in Nash optimization. They do not put themselves in the shoes of others, as they do in Rawls's original position, or in Harsanyi's (1977) thought experiment in which agents employ *empathy*. In this sense, Kantian behavior requires *less of a displacement from self* than 'veil-of-ignorance' thought experiments require. Agents require *no empathy* to conduct Kantian optimization: what changes from Nash behavior is the supposition about the counterfactual.

It remains to ask, when we discover an example of a society that appears to implement one of these allocation rules in a Pareto efficient manner, whether Kantian thinking among its members plays a role in maintaining its stability. This is an empirical question. Just as a Nash equilibrium is self-enforcing, so a Kantian allocation will be self-enforcing if the players in the game employ Kantian optimization.

We close this section with another example of how Kantian optimization can overcome inefficiencies — this time, with respect to income taxation. Suppose G is linear: $G(x) = ax$, some $a > 0$. Suppose each worker is paid his marginal product per unit effort (which is a). The *affine tax rule for tax rate t* is given by the allocation:

$$X^\gamma\left(E^1,\dots,E^n\right) = (1-t)aE^\gamma + ta\frac{E^S}{n}. \tag{3.8}$$

We know that the Nash equilibrium in the game induced by this allocation rule is Pareto inefficient for any $t > 0$: this is the familiar deadweight loss of taxation. But we have:

Theorem 2. *On the domain of economies $\mathfrak{E}$, the strictly positive K^+ equilibria with respect to the affine tax rules are Pareto efficient, for any $t \in [0, 1]$.*

Proof. The vector of efforts E comprises a strictly positive K^+ in such an economy exactly when:

$$(\forall\gamma)\left(\frac{d}{dr}\Big|_{r=0} u^\gamma\left((1-t)a\left(E^\gamma+r\right)+t\frac{a\left(E^S+nr\right)}{n}, E^\gamma+r\right) = 0\right), \tag{3.9}$$

which expands to:

$$u_1^\gamma\cdot(1-t+t)a + u_2^\gamma = 0,$$

which says that $-\frac{u_2^\gamma}{u_1^\gamma} = a$, the condition for Pareto efficiency. ■

What is the intuition? In Nash equilibrium, when the agent chooses his effort supply, he assumes there is negligible impact on the lumpsum demogrant he will receive from the tax. But if an agent uses the additive optimization protocol, he only reduces his effort by a quantum if he would prefer that all others reduce their effort by the same quantum. The effect on the demogrant will then be significant. Thus, the additive optimization protocol makes the agent internalize the externality of his choice of labor supply, should others behave like him — in this case, the positive externality that taxes are distributed to all in a lumpsum fashion. The fact that the internalization is *exactly right*, in the sense of inducing Pareto efficiency, is not a priori obvious. And the theorem does not hold if G is strictly concave.

4. Two possible objections

Readers may find the conceptualization of Kantian optimization to be too complex. Would it not be more faithful to Kant to say that a Kantian expends that effort level that he would like all others to expend as well? Why introduce the complexity that Kantian optimization means 'at an effort allocation, each believes that he can deviate in a particular way, only if he would prefer all others deviate *in similar fashion?*' The answer is this: the simpler version is equivalent to the more complex version exactly when all agents are identical (have the same preferences). The more complex version is, I maintain, the proper generalization of the simpler version when agents are heterogeneous.

Brekke et al. (2003), for example, present a model of moral motivation, in which all agents are identical. They write, "To find the morally ideal effort e_i^*, the individual asks herself, 'Which action would maximize social welfare, given that everyone acted like me?'"

We have the following easy proposition, in our economic environment.

Proposition 1. *Let X be any anonymous allocation rule.*[5] *Suppose all utility functions are identical. Then:*

A. *If each chooses the effort level that she would most like all others to choose as well, then the allocation is Pareto efficient.*
B. *The effort level that all (universally) choose in part A is both a K^+ and a $K^\times$ equilibrium of the game with identical players.*

Proof of A. If X is any anonymous rule, then it immediately follows that, for any effort level, and any i, $X^i(E, E, \dots, E) = \frac{G(nE)}{n}$. In part *A*, each agent *i*, solves the problem:

$$\max_E u\left(\frac{G(nE)}{n}, E\right);$$

the first-order necessary condition for an interior solution is

$$u_1G'(E) + u_2 = 0, \tag{4.1}$$

where the derivatives of u are evaluated at $\left(\frac{G(nE)}{n}, E\right)$. Thus, the solution is indeed Pareto efficient. Denote the solution of this problem by E^*.

Proof of Part B. To check that the vector $(E^*, E^*,\dots, E^*)$ is a multiplicative Kantian equilibrium, we examine the definition:

$$\frac{d}{dr}\Big|_{r=1} u\left(\frac{G(nrE^*)}{n}, rE^*\right) = u_1\cdot G'(E^*)E^* + u_2\cdot E^* = 0, \tag{4.2}$$

where the second equality follows from Eq. (4.1). Hence, $(E^*, E^*,\dots, E^*)$ is a $K^\times$ equilibrium.

The proof that $(E^*, E^*,\dots, E^*)$ is a K^+ equilibrium is equally straightforward. ■

The proposition proves that Kantian equilibrium in the way it is defined in the present article, is a generalization of the 'simpler' version of Kantian equilibrium proposed by Brekke et al. (2003). Unfortunately, the simpler version does not work when agents are heterogeneous — that is, the simpler kind of Kantian equilibrium is generally not Pareto efficient with heterogeneous agents. This is unsurprising. What is perhaps surprising is that the relatively natural change – from thinking about expending *identical efforts* to making *similar deviations* at a vector of efforts – is sufficient to generate socially desirable outcomes (in the sense of Pareto efficiency), at least in the cases discussed in Section 3.[6]

The second objection that some have raised is against the distinction I have drawn in Section 1 between optimization protocols and preferences. They ask, 'Cannot the Kantian protocol be shown really to be a

[5] An anonymous allocation rule is one such that, if the effort levels are permuted, then the output assignments are likewise permuted.

[6] Ostrom and Gardner (1993) argue that commons' problems are more easily solved when the individuals involved are 'symmetric' (i.e., identical). But they also argue that, even heterogeneous agents, can solve commons' problems. When the individuals have identical preferences, the simpler Kantian protocol of Brekke et al. (2003) leads to efficiency, and that is an easier one to learn than Kantian optimization protocols needed for groups of heterogeneous individuals.

kind of preference, and Kantian equilibria transform into Nash equilibria of the game with these new preferences?' I now argue that this is not, in general, so.

The most general kind of preferences would be defined over the entire allocation, $(x^1,\ldots, x^n, E^1, E^2,\ldots, E^n)$ where (x^i, E^i) is the effort-consumption vector of agent i. The question can then be posed as follows:

> Given an arbitrary economic environment $(\mathbf{u}, G, n)$ of the kind defined in Section 2, are there preferences, represented by utility functions $v^i : \mathfrak{R}^{2n}_+ \to \mathfrak{R}$, where the argument of v^i is an allocation $(x^1,\ldots, x^n, E^1, E^2,\ldots, E^n)$, such that, for any allocation rule X, the *Kantian* equilibria of the game induced by X on $(\mathbf{u}, G)$ are the *Nash* equilibria of game induced by X on $(\mathbf{v}, G)$?

The next proposition shows that this may be partially accomplished in a very special case, that of quasi-linear utility functions u^i.

Proposition 2. *Let $(\mathbf{u}, G, n)$ be an economic environment where for all i, $u^i(x, E) = x - h^i(E)$. Define $v^i\left(x^1,\ldots,x^n,E^1,\ldots,E^n\right) = \sum_j x^j - h^i\left(E^i\right)$. Let X be any allocation rule such that the $K^\times$ equilibria of the economy $(\mathbf{u}, G, n, X)$ are Pareto efficient. Then these $K^\times$ equilibrium allocations are Nash equilibria of the game induced by $\{\{v^i\}, X\}$, where the strategies of the agents are their efforts.*

Proof

1. The game induced by $\{\{v^i\},X\}$ in the economy has payoff functions V^γ defined by

$$V^\gamma\left(E^1,\ldots,E^n\right) = v^\gamma\left(X^1(E),\ldots,X^n(E),E^1,\ldots,E^n\right) = \sum X^j(E) - h^\gamma(E^\gamma) = G\left(E^S\right) - h^\gamma(E^\gamma) \tag{4.3}$$

where $E = (E^1,\ldots, E^n)$. Hence the first-order conditions defining Nash equilibrium are:

$$(\forall\gamma)\quad 0 = \frac{d}{dE^\gamma} V^\gamma\left(E^1,\ldots,E^n\right) = G'\left(E^S\right) - (h^\gamma)'(E^\gamma). \tag{4.4}$$

But Eq. (4.4) says that $(E^1,\ldots, E^n)$ is the vector of effort levels uniquely associated with all Pareto efficient allocations of the economy. Thus, the (strictly positive) Nash equilibria of this game comprise exactly the Pareto efficient allocations of the economy $(\mathbf{u}, G, n)$.

2. Since, by hypothesis, the $K^\times$ equilibria of $(\mathbf{u}, G, n)$ are Pareto efficient, it follows that they are Nash equilibria of the game $\{\{v^i\}, X\}$. ■

Proposition 2 remains true if we substitute 'K^+' for '$K^\times$'. However, it is not true that the Nash equilibria of the game $\{\{v^i\}, X\}$ contain the Kantian equilibria of the game induced by $(\mathbf{u}, G, n, X)$ if the latter equilibria are not Pareto efficient. For example, let X be the equal-division allocation rule, $X^\gamma(E) = \frac{G(E^S)}{n}$. Then, even with quasi-linear preferences, the $K^\times$ equilibria are not efficient, for the condition defining $K^\times$ equilibrium is:

$$\begin{aligned}(\forall\gamma)\quad 0 &= \left.\frac{d}{dr}\right|_{r=1}\left(\frac{G\left(rE^S\right)}{n} - h^\gamma(rE^\gamma)\right) = \frac{G'\left(E^S\right)}{n}E^S - E^\gamma(h^\gamma)'(E^\gamma) \\ &\Leftrightarrow G'\left(E^S\right)\frac{E^S}{nE^\gamma} = (h^\gamma)'(E^\gamma)\end{aligned} \tag{4.5}$$

which does not define a Pareto efficient allocation except in the singular case that all the effort levels are identical. However, the game $\{\{v^i\}, X\}$ remains exactly the same for any allocation rule X, since $\sum_\gamma X^\gamma \equiv G$, and so in this case the $K^\times$ equilibria of the game $(\mathbf{u}, G, n)$ are *not* Nash equilibria of the game $\{\{v^i\}, X\}$.

Even in the case that Proposition 2 examines, we can ask: Is it more reasonable to believe that communities, with quasi-linear preferences, which achieve Pareto efficient outcomes, are using the utility functions v^γ in which they do not care at all about their own consumption, but only community consumption, than to believe they are optimizing self-interested utility functions u^γ, but with the Kantian optimization protocol?

I have not proved that the question posed prior to the statement of Proposition 2 cannot be answered affirmatively, but I conjecture it cannot be for any natural transformation of the utility functions u into utility functions v. — even for the simple case of economic environments with quasi-linear preferences, let alone other preferences. Hence, I believe that the Kantian optimization protocol cannot be viewed as equivalent to Nash equilibria with agents' having exotic preferences.

5. Other versions of Kantian equilibrium

We can define a general 'Kantian variation' which includes as special cases additive and multiplicative Kantian equilibrium. We say a function $\varphi : \mathfrak{R}^2_+ \to \mathfrak{R}^2_+$ is a *Kantian variation* if:

$$\forall x \quad \varphi(x, 1) = x,$$

and if, for any $x \neq 0$, the function $\varphi(x)$ maps onto the non-negative real line. Denote by $\varphi[E(\cdot),r]$ the effort profile $\widetilde{E}$ defined by $\widetilde{E}^\gamma = \varphi(E^\gamma, r)$.

Then an effort profile $E(\cdot)$ is a φ — Kantian equilibrium of the game $\{V^\gamma\}$ if and only if:

$$(\forall\gamma)(V^\gamma(\varphi[E(\cdot),r])\text{is maximized at } r = 1). \tag{5.1}$$

If we let $\varphi(x, r) = rx$, this definition reduces to multiplicative Kantian equilibrium; if we let $\varphi(x, r) = x + r - 1$, it reduces to additive Kantian equilibrium.

Let $\varphi(x, r)$ be any Kantian variation that is concave in r, and let the payoff functions generated by some allocation rule, $\{V^\gamma\}$, be concave. Then a positive effort schedule E is a φ — Kantian equilibrium if and only if:

$$\forall\gamma \quad \left.\frac{d}{dr}\right|_{r=1} V^\gamma(\varphi[E(\cdot),r]) = 0. \tag{5.2}$$

Eq. (5.2) follows immediately from the definition (Eq. (5.1)), since $V^\gamma(\varphi[E(\cdot), r])$ is a concave function of r, and hence its maximum, if it is interior, is achieved where its derivative with respect to r is zero. Note that both the additive and multiplicative Kantian variations are concave (indeed, linear) functions of r.

The next theorem states that there is a unidimensional continuum of allocation rules, with the proportional and equal-division rules as its two extreme points, each of which can be efficiently implemented on $\mathfrak{E}$ using a particular Kantian variation. Define the allocation rules:

$$(\forall\beta\in\mathfrak{R}_+)(\forall\gamma = 1,\ldots,n)\left(X^\gamma_\beta\left(E^1,\ldots,E^n\right) = \frac{E^\gamma+\beta}{E^S+n\beta}G\left(E^S\right)\right) \tag{5.3}$$

and the Kantian variations:

$$\varphi_\beta(x, r) = rx + (r-1)\beta, \quad 0 \leq \beta \leq \infty. \tag{5.4}$$

Note that for $\beta = 0, X_\beta$ is the proportional rule and φ_β is the multiplicative Kantian variation, and as $\beta \to \infty, X_\beta$ approaches the equal-division rule and φ_β approaches the additive Kantian variation (this last fact is perhaps not quite obvious). Thus we identify X_∞ as the equal-division allocation rule. We will call a Kantian equilibrium associated with the variation φ_β, a K^β equilibrium.

First, fix β and an effort vector $E \in \mathfrak{R}^n_+$. Define $r^j_i = \frac{E^i+\beta}{E^j+\beta}$. Now consider the set of vectors in $\mathfrak{R}^n_+$ of the form $(\varphi_\beta(x, r^j_1), \varphi_\beta(x, r^j_2), \ldots, \varphi_\beta(x, r^j_n))$ where x varies over the real numbers, but restricted to an interval that keeps the defined vectors non-negative. This is a ray in $\mathfrak{R}^n_+$ which I denote by $M^j_\beta(E)$. We have:

Lemma. *Fix a vector $E \in \mathfrak{R}^n_{+\,+}$ and a non-negative number β. Then the ray $M^j_\beta(E)$ does not depend on j.*

Proof. Let $v = (\varphi_\beta(x, r^j_1), \varphi_\beta(x, r^j_2), \ldots, \varphi_\beta(x, r^j_n))$ be an arbitrary vector in $M^j_\beta(E)$. We wish to show that, for any $k \neq j$, $\nu \in M^k_\beta(E)$. This is accomplished if we can produce a number $\hat{x}$ such that $v = \left(\varphi_\beta\left(\hat{x}, r^k_1\right), \ldots, \varphi_\beta\left(\hat{x}, r^k_n\right)\right)$. Check that $\hat{x} = \frac{E^k+\beta}{E^j+\beta}x + \beta\left(\frac{E^k-E^j}{E^j+\beta}\right)$ works. ■

As a consequence of the lemma, we may drop the superscript 'j' and refer to the ray just defined as $M_\beta(E)$.

Theorem 3. [7]*For $0 \leq \beta \leq \infty$:*

A. *If E is a strictly positive K^β equilibrium w.r.t. the allocation rule θ^β at any economy in $\mathfrak{E}$*, then the induced allocation is Pareto efficient.
B. *X_0 is the only allocation rule for which the $K^\times$ equilibrium is Pareto efficient on the domain $\mathfrak{G}$.*
C. *For any $\beta > 0$, the only allocation rules that are efficiently implementable on $\mathfrak{G}$* are of the form $X^j(E^1, \ldots, E^n, G) = X^j_\beta(E^1, \ldots, E^n) + k^j(E^1, \ldots, E^n)$ *where $k^j : \mathfrak{R}^n_+ \to \mathfrak{R}_+$ are any functions satisfying:*
   (i) $\sum_j k^j(E) \equiv 0$
   (ii) $(\forall j, E)(X^j_\beta(E) + k^j(E) \geq 0)$ and
   (iii) $(\forall j, E)(k^j$ is constant on the ray $M_\beta(E))$. *That is, on $M_\beta(E)$*

$$\nabla k^j \cdot (E + \beta) \equiv 0,$$

where $E + \beta = (E^1 + \beta, \ldots, E^n + \beta)$ and ∇f denotes the gradient of the function f.

D. *For any $\beta \in [0, \infty]$, and*

$$(\forall E \in \mathfrak{R}^n_{++})(\forall j = 1, \ldots, n)\left(X^j_\beta(E) = \lambda(E)X^j_0(E) + (1-\lambda(E))^j_\infty(E)\right),$$

where $\lambda(E) = \frac{E^S}{E^S+n\beta}$.

Proof. See Appendix A.

The theorem states first that for all $\beta \geq 0$, the pair (X_β, φ_β) is an *efficient Kantian pair*: i.e., that the allocation rule X_β is efficiently implementable in K^β equilibrium on the domain $\mathfrak{E}$ Part C states that the only other allocation rules that are K^β implementable are ones which add numbers to the X_β rule that are constant on certain rays in $\mathfrak{R}^n_+$. Part B states that in the unique case when $\beta = 0$, these constants must be zero. Part D states that the allocation rules X_β are 'convex combinations' of the proportional rule X_0 and the equal-division rule X_∞. The quote marks in this sentence are meant to alert the reader to the fact that the weights in the convex combination depend on the equilibrium effort vector, but not on the component j.

Unfortunately, part C makes Theorem 1 difficult to state. One may ask, is it necessary? That is, do there in fact exist allocation rules satisfying conditions $C(i)$–$C(iii)$ of the theorem where the functions k^j are not identically zero? The following example shows that there are.

Example 4. We consider K^+ equilibrium (i.e., $\beta = \infty$) where $n = 2$. In this case

$$\theta^j_\infty\left(E^1, E^2\right) = \frac{1}{2},$$

that is, the equal-division allocation rule. Now consider:

$$\widetilde{\theta}^1(E) = \begin{cases} \dfrac{1}{2} + \dfrac{G\left(E^1 + E^2\right)}{2G(E^1 + E^2)}, & \text{if } E^1 \geq E^2 \\[2ex] \dfrac{1}{2} - \dfrac{G\left(E^1 + E^2\right)}{2G(E^1, E^2)}, & \text{if } E^1 < E^2 \end{cases}. \tag{5.5}$$

The $\widetilde{\theta}$ rule satisfies conditions $C(i)$–$C(iii)$.

Example 5. We now provide an example of a similar kind for any $\beta > 0$. Let $n = 2$. Fix E. The ray $M_\beta(E)$ has a smallest element: it is a vector with at least one component equal to zero. (This vector is dominated, component-wise, by all other vectors in the ray.) Denote this vector by $M_\beta(E)^{\min}$, and the sum of its components by $M^S_\beta(E)^{\min}$. Define the allocation rules:

$$\widetilde{\theta}^1(E) = \begin{cases} \theta^1_\beta(E) - \dfrac{G\left(M^S_\beta(E)^{\min}\right)}{2G(E^S)}, & \text{if } E^1 \geq E^2 \\[2ex] \theta^1_\beta(E) + \dfrac{G\left(M^S_\beta(E)^{\min}\right)}{2G(E^S)}, & \text{if } E^1 < E^2 \end{cases}.$$

Since $M^S_\beta(E)^{\min} < E^S$, we have $\widetilde{\theta}^i_\beta(E) \in [0, 1]$. Moreover the function $G(M^S_\beta(E)^{\min})$ is constant on the ray $M_\beta(E)$. Hence the allocation rule satisfies conditions $C(i)$–$C(iii)$ of the theorem.

From the history-of-thought vantage point, the case $\beta = 0$ is the classical 'socialist' economy: that is, it is an economy where output is distributed in proportion to labor expended *and efficiently so*. The rule X_∞ is the classical 'communist' economy: output is distributed 'according to need' (here, needs are identical across persons), *and efficiently so*. Indeed, the allocation rules X_β associated with $\beta \in (0, \infty)$ are convex combinations of these two classical rules, in the sense that part D states. The fact that the allocation rules that can be efficiently implemented with various kinds of Kantian optimization define a unidimensional continuum between these two classical concepts of cooperative society provides further support for viewing the Kantian optimization protocols as models of cooperative behavior.

I conjecture that there are no other allocation rules, than the ones described in Theorem 3, which can be efficiently implemented with respect to any Kantian variation on the domain $\mathfrak{E}$.

As we have noted, history displays examples of both the proportional and equal-division allocation rules. The former have been discussed in relation to Ostrom's work on fisher economies. And anthropologists conjecture that many hunting societies employed the equal-division rule. (Whether they found Pareto efficient equal-division allocations is another matter.) Although Theorem 3 suggests that we look for societies that implemented some of the other allocation rules in the β continuum, the Kantian variations involved for $\beta \notin \{0, \infty\}$ may be too arcane for human societies, lacking the simplicity of the additive and multiplicative rules.

[7] Theorem 3 of Roemer (2010) stated something similar to part B of the present theorem, but the proof offered there is incorrect. Consider the present theorem to constitute a corrigendum.

There is an analogous, but negative result to Theorem 3 for *Nash* equilibrium:

Theorem 4.

A. *There is no allocation rule that is efficiently implementable in Nash equilibrium on the domain* $\mathfrak{E}$.
B. *On continuum economies, Walrasian rules (with no taxation) are efficiently Nash implementable.*[8]

Proof. Appendix A.

The reason that the Walrasian allocation rules, as defined in the previous footnote, are not efficiently implementable in Nash equilibrium on *finite* economies is that an individual's Nash optimization behavior at the Walrasian allocation rule must take account of her effect on $G'(E^S)$ and on her share of profits as she deviates her effort. That is, in finite economies, Nash-optimizers are not price takers. It is essentially only in the continuum economy that the agent rationally ignores such effects, and hence, Nash behavior induces efficiency.

To conclude this section, I provide a geometric interpretation of the various Kantian equilibria defined in Theorem 3. Let $n = 2$. In Fig. 1, the allocation under consideration is $\left(\hat{E}^1, \hat{E}^2\right)$. Under the multiplicative Kantian protocol, both agents consider whether they would prefer an allocation on the ray labeled $K^{\times}$. Under the additive Kantian protocol, they both consider whether they would prefer an allocation on the 45^0 ray through $\left(\hat{E}^1, \hat{E}^2\right)$. Any of the Kantian variations listed in Theorem 3 will generate a common ray – a typical one is the dashed ray labeled K^{β} – which passes through $\left(\hat{E}^1, \hat{E}^2\right)$ and lies between the K^+ and $K^{\times}$ rays. On the other hand, under the Nash protocol, agent 1 asks whether he would prefer an effort vector on the dashed line N^1, and agent 2 asks whether she would prefer an effort vector on the dashed line N^2. Thus, the key distinction is that in Kantian reasoning, agents ask whether they would prefer an alternative in a *common set* of counterfactual effort vectors, whereas in Nash reasoning, agents consider *different sets* of counterfactuals. I am proposing that the consideration by each player of a social deviation to a common set is the mathematical characterization of cooperative behavior.

6. Economies with other-regarding preferences (ORPs)

It is appropriate to begin this section with a thought of the political philosopher, Cohen (2009), who offers a definition of 'socialism' as a society in which earnings of individuals at first accord with a conception of equality of opportunity that has developed in the last thirty years in political philosophy (see Rawls, 1971; Dworkin, 1981; Arneson, 1989; Cohen, 1989), but in which inequality in those earnings is then reduced because of the necessity to maintain 'community,' an ethos in which '…people care about, and where necessary, care for one another, and, too, care that they care about one another.' Community, Cohen argues, may induce a society to reduce material inequalities (for example, through taxation) that would otherwise be acceptable according to 'socialist' equality of opportunity. But, Cohen writes:

> …the principal problem that faces the socialist ideal is that we do not know how to design the machinery that would make it run. Our problem is not, primarily, human selfishness, but our lack of a suitable organizational technology: our problem is a problem of design. It may be an insoluble design problem, and it is a design problem that is undoubtedly exacerbated by our selfish propensities, but a design problem, so I think, is what we've got.
>
> [Cohen (2009, p.57)]

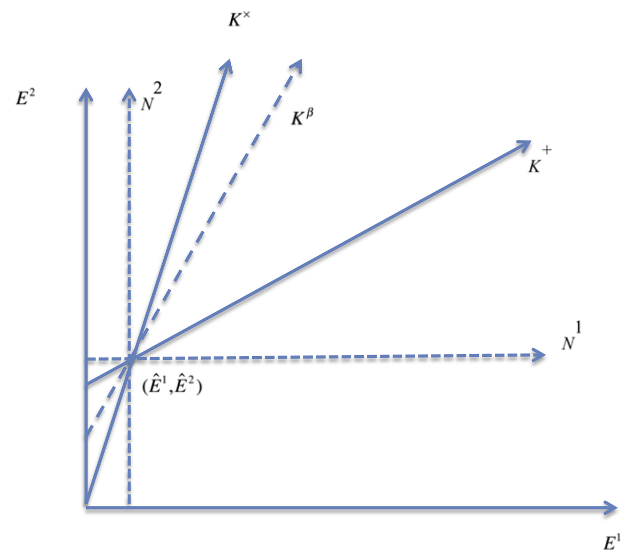

Fig. 1. Illustration of Kantian counterfactual rays for $n = 2$.

An economist reading these words thinks of the first theorem of welfare economics. A Walrasian equilibrium is Pareto efficient in an economy with complete markets, private goods, and the absence of externalities. But under Cohen's communitarian ethos, people care about the welfare of others – which induces massive consumption externalities – and so the competitive equilibrium will not, in general, be efficient. What economic mechanism can deliver efficiency under these conditions,[9,10]?

We proceed, now, to study Kantian equilibrium where agents have *all-encompassing utility functions* consisting of a person utility function, of the kind we have been working with thus far, plus a social welfare function, which responds positively to the utility of other agents in the society. Such economies are synonymously referred to as ones with a *social ethos*, or with other-regarding preferences (ORPs).

Let $S : \mathfrak{R}^n_+ \to \mathfrak{R}$ be a *social welfare function* for a society of n individuals. S is strictly increasing in its arguments, differentiable and concave. The *all-encompassing utility function* of an individual γ is given by:

$$U(x, E) = u^{\gamma}\left(x^{\gamma}, E^{\gamma}\right) + \alpha S(u(x, E)) \tag{6.1}$$

where (x, E) is a feasible allocation for the economy, and

$$u(x, E) = \left(u^1\left(x^1, E^1\right), \ldots, u^N\left(x^N, E^N\right)\right)$$

is the profile of utility functions evaluated at the allocation. The non-negative constant α measures the *degree of social ethos*. For some results, we allow α to vary with the type (thus, α^{γ}). We denote the economic environment now as $(\mathbf{u}, G, n, \alpha)$. The case $\alpha = 0$ reduces to the economy with self-regarding preferences, and the case $\alpha = \infty$ is one in which every type is fully altruistic, caring only about social welfare.

The choice to model other-regarding preferences as represented by the addition of a social welfare function to a personal utility function

[8] Walrasian allocation rules are defined in Eq. (2.5).

[9] In war-time Britain, many spoke of 'doing their bit' for the war effort — voluntary additional sacrifice for the sake of the common good. But, if I want to contribute to the common struggle, how *much* extra should I do? The price mechanism does not coordinate 'doing their bit' well.

[10] A recent contribution which is relevant to this inquiry is that of Dufwenberg et al. (2011), which studies the veracity of the first and second welfare theorems in the presence of other-regarding preferences — what I here call social ethos. From the viewpoint of the evolution of economic thought, it is significant that their article is the result of combining three independent papers by subsets of the five authors: in other words, the problem of addressing seriously the efficiency consequences of the existence of other-regarding preferences is certainly in the air at present.

is classical. There are various other ways in which one might model 'social ethos,' some motivated by the literature in experimental economics. More generally, instead of thinking of all-encompassing preferences as embodying an altruistic element, we might think of them as embodying a sense of justice. In this case, an individual would not necessarily be concerned with the welfarist formulation of a social welfare function as in Eq. (6.1), but rather with some theory of just distribution that might be non-welfarist. The extensive literature in non-welfarist theories of justice could be brought to bear (see Roemer, 1998; Fleurbaey, 2008).

A remark is in order. Up until now, the theory of Kantian optimization has been entirely ordinal: the results are independent of the choice of utility function to represent individuals' preferences. But Eq. (6.1) is a cardinal representation: the evaluation of social welfare will not be independent of the representations of personal preferences. We may, however, make the analysis ordinal as follows. Suppose we begin with a specification of the utility-function profile as $\mathbf{u}$. For each personal preference order (that is, the one that the personal utility function represents), we may choose the representation v^γ defined by:

$$v^\gamma(x^\gamma, E^\gamma) = x^* \text{ where } u^\gamma(x^*, 0) = u^\gamma(x^\gamma, E^\gamma).$$

The utility function v^γ represents γ's personal preference order, and it is invariant with respect to choice of u^γ. If we use the utility functions $\{v^\gamma\}$ then the ORPs given by Eq. (6.1) are purely ordinal, in the sense that their values are independent of the choice of the u^γ. Of course, we cannot interpret the statement "$v^i(x^i, E^i) = v^j(x^j, E^j)$" as meaning that i and j have the same level of welfare, unless we postulate that the welfare of all agents is the same at any vector $(x^*,0)$. This might not be a bad way of implementing interpersonal comparisons. (We are all equally well off consuming the same amount of the good, and not working.)

To maintain the ordinal flavor of Kantian equilibrium, I will work with the utility functions v^γ in this section. There is no difference in the results, but using the $\{v^\gamma\}$ emphasizes the ordinal nature of the theory.

6.1. Efficiency results

We begin by characterizing interior Pareto efficient allocations in economies where individuals have all-encompassing utility functions as in Eq. (6.1). At an allocation (x^*, E^*), we write $\mathrm{v}^\gamma(x^*\gamma, E^*\gamma) = v[^*, \gamma]$, and for the two partial derivatives of v at the allocation, $v_1[^*, \gamma]$ and $v_2[^*, \gamma]$.

Theorem 5. *Let all-encompassing preferences be given by Eq. (6.1). Then an interior allocation (x^*, E^*) is Pareto efficient for the economy $(\mathbf{u}, G, n, \alpha)$ if and only if:*

(a)

$$\textit{for all}\ \ \gamma, \quad -\frac{v_2[*,\gamma)}{v_1[*,\gamma]} = G'\left(E^{*S}\right) \tag{6.2a}$$

and

(b)

$$\alpha \le \left(\max_\gamma\left(v_1^\gamma S_\gamma \sum_k \left(v_1^k\right)^{-1} - \sum_k S_k\right)\right)^{-1}, \tag{6.2b}$$

where all functions are evaluated at the allocation.

Proof. Appendix A.

I offer some remarks about and corollaries to Theorem 5.

1. Note the separate roles played by the conditions (a) and (b) of Theorem 5. Condition (a) assures allocative efficiency in the economy with $\alpha = 0$ — it says that for all types, $\mathrm{MRS}^\gamma = \mathrm{MRT}$. Condition (b) is entirely responsible for the efficiency requirement induced by social ethos: it concerns only distribution, not production, which is to say the function G does not appear in (b).
   Indeed, it is obvious that any allocation which is Pareto efficient in the α-economy (for any α) must be efficient in the economy with $\alpha = 0$. For suppose not. Then the allocation in question is Pareto-dominated by some allocation in the 0-economy. But immediately, that allocation must dominate the original one in the α-economy, as it causes the social welfare function to increase (as well as the private parts v^γ of all-encompassing utility). It is therefore not surprising that the characterization of Theorem 5 says that 'the allocation is efficient in the 0-economy (part (a)) and satisfies a condition which becomes increasingly restrictive as α becomes larger (part (b)).'
2. Define $PE(\mathbf{u}, G, n, \alpha)$, which we will here-to-for abbreviate as $PE(\alpha)$ as the set of interior Pareto efficient allocations for the α-economy $(\mathbf{u}, G, n, \alpha)$. It follows from condition (b) of Theorem 3 that the Pareto sets are nested, that is:

   $$\alpha > \alpha' \Rightarrow PE(\alpha) \subset PE(\alpha').$$

   Hence, denoting the fully altruistic economy by $\alpha = \infty$, we have:

   $$PE(\infty) = \cap_{\alpha \ge 0} PE(\alpha).$$

   $PE(\infty)$ will generally be a unique allocation — the allocation that maximizes social welfare.
3. Let $\alpha \to \infty$; then condition (b) of Theorem 3 reduces to:

   $$\text{for all } \gamma, \quad v_1^\gamma S_\gamma \sum_k \left(v_1^k\right)^{-1} = \sum_k S_k. \tag{6.3}$$

   We have:

Corollary 1. *An interior allocation is efficient in the fully altruistic economy (i.e., maximizes social welfare) if and only if condition (a) of Theorem 5 holds and for some K (b′) $(\forall\ \gamma)(v_1^\gamma S_\gamma = K)$.*

Proof. Let

$$K = \frac{\sum_k S_k}{\sum_k \left(v_1^k\right)^{-1}}. \tag{6.4}$$

Then (b′) follows immediately from Eq. (6.3). Conversely, if (b′) holds, then we immediately verify Eq. (6.3). ■

4. Consider the quasi-linear economy in which:

   $$v^\gamma(x, E) = x - h^\gamma(E). \tag{6.5}$$

   Then $v_1^\gamma = 1$. Suppose that S is a symmetric function. Then corollary 1 implies that the partial derivatives S_j are all equal and so *in the quasi-linear economy, the only Pareto efficient interior allocation as $\alpha \to \infty$ is the equal-utility allocation for which condition* (a) $(\mathrm{MRS}^\gamma = \mathrm{MRT})$ *holds.*

6.2. Kantian equilibrium

Fix an allocation rule X. The first remark is *there may be no Pareto efficient allocations in $(\mathbf{u}, G, n, \alpha)$ that can be implemented with the rule X*. Think, for instance, of the proportional rule X^{Pr}. Suppose α is very large — say, infinity. Then the unique Pareto efficient allocation in $(\mathbf{u}, G, n, \alpha)$ is the one that maximizes social welfare. But this allocation may not (in general, it *will* not) be a proportional allocation. Consider the quasi-linear example of Eq. (6.5). Assuming S is symmetric, the unique maximizer of the social welfare function (and therefore the unique Pareto efficient point in this economy with ORPs) is the one which maximizes the surplus (this determines the effort vector) *and* distributes output to equalize utilities. This allocation will only, by coincidence, be a proportional allocation. It therefore follows that we cannot expect the Kantian equilibrium of economies $(\mathbf{u}, G, n, \alpha)$ to be Pareto efficient (always, with respect to the ORPs).

Denote the set of K^{β} equilibria for the economy $(\mathbf{u}, G, n, \alpha)$ when the allocation rule is X by $\mathbf{K}^{\beta}(\alpha, X)$. We have:

Theorem 6. For all $\alpha \geq 0$ and $\beta \geq 0$, and for all allocation rules X, $\mathbf{K}^{\beta}(\alpha, X) = \mathbf{K}^{\beta}(0, X)$.

Proof

1. To avoid perhaps confusing notation, we will prove this for $X = X^{\mathrm{Pr}}$ the proportional rule and $\beta = 0$ — that is, multiplicative Kantian equilibrium. An allocation (x, E) is a $K^{\times}$ equilibrium for the proportional rule in the economy $(\mathbf{u}, G, n, \alpha)$ iff:

$$(\forall\gamma)\quad \left.\frac{d}{dr}\right|_{r=1}\left(v^{\gamma}\left(\frac{E^{\gamma}}{E^{S}}, rE^{\gamma}\right) + \alpha S\left(v^{1}\left(\frac{E^{1}}{E^{S}}, rE^{1}\right), \ldots, v^{n}\left(\frac{E^{n}}{E^{S}}, rE^{n}\right)\right)\right) = 0. \tag{6.6}$$

Denote $\left.\frac{d}{dr}\right|_{r=1} v^{\gamma}\left(\frac{E^{\gamma}}{E^{S}}, rE^{\gamma}\right) \equiv D_r v^{\gamma}$. Then Eq. (6.6) can be written:

$$(\forall\gamma)\quad D_r v^{\gamma} + \alpha\sum_{k} S_k\left(D_r v^{k}\right) = 0 \tag{6.7}$$

from which it follows that for all γ, $D_r v^{\gamma} = K$, a constant. Substituting this constant into Eq. (6.7), we have:

$$K + \alpha K\sum_{k} S_j = 0.$$

Since $\sum_{k} S_j > 0$, it immediately follows that $K = 0$. But this says that $D_r v^{\gamma} = 0$ for all γ, which is exactly the condition that the allocation is a $K^{\times}$ equilibrium in the economy $(\mathbf{u}, G, n, 0)$, proving the claim. ■

Theorem 6 says that *the Kantian equilibria for an economy with positive social ethos, with respect to an allocation rule, are identical to the Kantian equilibrium for the associated economy with purely self-regarding preferences.* Indeed, the theorem is more general than stated: it is easy to check that different agents can have different values of the altruistic parameter α and the proof goes through.

Because, as was noted above, there in general will not exist Pareto efficient allocations that can be implemented by a given rule X, in the economy with $\alpha > 0$, we should look at *second-best allocations.*

Definition. Let $PE^{X}(\mathbf{u}, G, n, \alpha)$ be the allocations that are implementable with the rule X in the economy (u, G, α) and are not Pareto-dominated by any X-implementable allocation. Without confusion, we abbreviate the notation to $PE^{X}(\alpha)$.

For example, there are reasons that fishing economies use the proportional allocation rule — because it implements the simple rule 'each fisher keeps his catch.' Likewise, the equal-division allocation rule may be a good rule in hunting societies. So one should ask: What are the best allocations that can be found, *given* that a society is using a particular rule X? The second-best allocations $PE^{X}(\alpha)$ comprise good candidates, from the efficiency viewpoint.

We now state the main theorem of this section:

Theorem 7. *Given an allocation rule X and a Kantian variation β such that in the self-interested economies $(\mathbf{u}, G, n, 0)$, the β — Kantian allocations are Pareto efficient. Then*:

$$PE^{X}(\alpha) \subset \mathbf{K}^{\beta}(\alpha, X).$$

In other words, any X-implementable allocation which is second-best efficient for the α economy is a Kantian equilibrium in the economy with ORPs. The converse, however, is not true: that is there may be Kantian equilibria which are not in $PE^{X}(\mathbf{u}, G, \alpha)$.

Proof

1. We will show this for the multiplication Kantian equilibrium and the proportional allocation rule, X^{Pr} — again, to preserve simplicity of notation.
   We first observe that if an allocation is in $PE^{X}(\alpha)$, it must be Pareto efficient in the 0-economy. To show this, we need only show that $MRS^{\gamma} = MRT$ for all agents γ, where the marginal rate of substitution is computed with the personal utilities functions v^{γ}. Suppose, to the contrary, that for some γ, $MRS^{\gamma} < G'(E^{S})$ at the allocation $X(E^{1}, \ldots, E^{n})$. Now consider a small positive increase ε in E^{γ}, holding all other efforts fixed, and look at the new allocation determined by the proportional rule X^{Pr}. The *personal* utility of γ will increase if:

$$\frac{d}{dE^{\gamma}} v^{\gamma}\left(\frac{E^{\gamma}}{E^{S}} G\left(E^{S}\right), E^{\gamma}\right) > 0 \tag{6.8}$$

   This derivative evaluates to:

$$v_1^{\gamma}\left(\frac{E^{\gamma}}{E^{S}} G'\left(E^{S}\right) + \frac{E^{S} - E^{\gamma}}{\left(E^{S}\right)^2} G\left(E^{S}\right)\right) + v_2^{\gamma} \overset{?}{>} 0$$

   Now use the fact that $v_1^{\gamma}G' + v_2^{\gamma} > 0$ (that is, $MRS^{\gamma} < G'$) and the desired inequality will follow if:

$$-v_2^{\gamma}\frac{E^{\gamma}}{E^{S}} + v_1^{\gamma}\frac{E^{S} - E^{\gamma}}{\left(E^{S}\right)^2} G\left(E^{S}\right) \overset{?}{>} -v_2^{\gamma}$$

   which can be rewritten as:

$$-\left(\frac{v_2^{\gamma}}{v_1^{\gamma}}\right)\left(\frac{E^{\gamma} - E^{S}}{E^{S}}\right) + \frac{E^{S} - E^{\gamma}}{\left(E^{S}\right)^2} G\left(E^{S}\right) \overset{?}{>} 0.$$

   Dividing by the positive number $(E^{S} - E^{\gamma})/E^{S}$, this inequality reduces to:

$$\frac{G\left(E^{S}\right)}{E^{S}} \overset{?}{>} MRS^{\gamma}$$

   which is true, because the concavity of G implies that $\frac{G\left(E^{S}\right)}{E^{S}} > G'\left(E^{S}\right)$. This proves Eq. (6.8).
2. Now this small increase in γ's effort also *increases* the utilities of all other agents to the first-order, because they care about person γ via S; it also *decreases* their utility, but only to the second order, because there is a second order decrease in G' and hence in the consumptions of the others. Net, the all-encompassing utilities of all the other agents increase. However, because the consumption of the other agents decrease, γ's all-encompassing utility *decreases*, because he cares about the others, but this decrease is only to the second order. Therefore, to the first-order, all all-encompassing utilities increase, and this contradicts the assumption that the original allocation was in $PE^{X}(\alpha)$.
3. It follows that it must be that $MRS^{\gamma} = MRT$ for all agents γ and hence the allocation is 0-Pareto efficient. We can state this fact as: $PE^{X}(\alpha) \subset PE(0)$.
4. Denote by $X[\mathbf{u}, G]$ the set of allocations that can be implemented by the allocation rule X in the economy $(\mathbf{u}, G, n)$. Theorem 1 states that $\mathbf{K}^{\times}(0, X^{\mathrm{Pr}}) = X^{\mathrm{Pr}}[\mathbf{u}, G] \cap PE(0)$. (Part A of that theorem states one direction of this set equality, and part B the other.) By virtue of Theorem 6, we have $\mathbf{K}^{\times}(\alpha, X^{\mathrm{Pr}}) = X^{\mathrm{Pr}}[\mathbf{u}, G] \cap PE(0)$. By virtue of step 3 of this proof, we therefore have $PE^{X}(\alpha) = X[\mathbf{u}, G] \cap PE^{X}(\alpha) \subset X[\mathbf{u}, G] \cap PE(0) = \mathbf{K}(\alpha, X)$, proving the theorem. ■

Why does the converse to Theorem 7 not hold? Suppose there are several allocations in $\mathbf{K}^{\times}(0, X) = \mathbf{K}^{\times}(\alpha, X)$. Generically, they will be strictly ranked by the social welfare function *S*. So if $\alpha = \infty$, only one of them will be a member of $PE(\infty)$. The same argument applies for large finite α because the set $PE(\alpha)$ shrinks to a singleton as α becomes large. Of course, if $\mathbf{K}^{\times}(0, X)$ is a singleton, then the converse does hold.

6.3. An example of Kantian implementation of an allocation in PE(∞)

There are, however, examples where an allocation which is Pareto efficient in the fully altruistic ORP economy can be Kantian implemented, by suitable choice of the Kantian variation.

Consider the family of quasi-linear economies, where, for some fixed $\rho > 1$:

$$u^{\gamma}(x, E) = x - \frac{E^{\rho}}{\rho \gamma}. \tag{6.9}$$

For these economies we can always choose a value β so that the K^{β} equilibrium w.r.t. the allocation rule X_{β} is efficient for economies with *any* value of α: that is to say, the (K^{β}, X_{β}) allocation maximizes social welfare (and so is in $PE(\infty)$).

Theorem 8. *Let* $u^{\gamma}(x,E) = x - \frac{E^{\rho}}{\rho\gamma}$, *some* $\rho > 1$. *Let G be any concave production function. Define* E^S *by the equation* $E^S = \bar{\gamma}_{\rho} G'(E^S)^{1/(\rho-1)}$ *where* $\bar{\gamma}_{\rho} \equiv \sum_{\gamma} \gamma^{1/(\rho-1)}$. *Then for this economy:*

(a) *An allocation is* PE(0) iff $E^{\gamma} = \gamma^{1/(\rho-1)} G'(E^S)^{1/(\rho-1)}$.

(b) *Define* $\beta(\rho) = \rho \frac{G(E^S)}{G'(E^S)} - E^S$. *The* $K^{\beta(\rho)}$ *allocation w.r.t. the allocation rule* $X_{\beta(\rho)}$ *is in* $PE(\infty)$.

Proof. Appendix A.

6.4. Taxation in private-ownership economies

The K^{β} equilibria for the allocation rules X_{β} are not implementable with markets in any obvious way. This is most easily seen by noting that the proportional rule and the equal-division rules are not so implementable. According the second theorem of welfare economics, there is some division of shares in the firm which operates the technology *G* which would implement these rules in Walrasian equilibrium in continuum economies, but to compute those shares, one would have to know the preferences of the agents. The advantage of the Kantian approach is that the Kantian allocations are decentralizable in the sense that agents need only know the production function *G*, total effort E^S, and their own preferences, to compute the deviation they would like (everybody) to make.

Nevertheless, one would like Kantian optimization to be useful in market economies as well. For the linear economies, we have a hopeful result — namely, Theorem 2. Before stating it, let us define the allocation rules associated with linear taxation. Define the affine tax allocation rule $X_{[t]}$ for *linear* economies with production function *G(x)* = *ax* by:

$$X_{[t]}^{\gamma}\left(E^1, \ldots, E^n\right) = (1-t)aE^{\gamma} + t\frac{aE^S}{n} \tag{6.10}$$

Theorem 9

A *For any* $t \in [0,1]$, *the* K^{+} *equilibria for the linear tax rule* $X_{[t]}$ *is Pareto efficient on* $\mathfrak{L}$.

B. *The only allocation rules which are efficiently implementable in* K^{+} *on* $\mathfrak{L}$ *are of the form* $X^{\gamma}(E^1, \ldots, E^n) = X_{[t]}^{\gamma}(E^1, \ldots, E^n) + k^{\gamma}(E^1, \ldots, E^n)$ *for some* $t \in [0,1]$ *where:*

(i) *for all* $E \in \mathfrak{R}_{+}^{n}$ $\sum k^{\gamma}(E) = 0$,

(ii) *for all* (j, E) $X^{\gamma}(E) \geq 0$, and

(iii) *for all* γ, *for all* $E \in \mathfrak{R}_{+}^{n}$, $\nabla k^{\gamma}(E) \cdot E = 0$.

Proof. Part A is simply Theorem 2; part B is proved in Appendix A.

By virtue of Part *A* of the above theorem, in a society with other-regarding preferences and linear production, citizens could choose a high tax rate to redistribute income substantially, without sacrificing allocative efficiency, thereby addressing the positive externality due to their concern for others. Part *B* of the theorem is analogous to part *C* of Theorem 3.

As in Theorem 3, one is entitled to ask whether there are examples of allocation rules where the functions k^j are not identically zero. There are, as the next example shows.

Example 5. Let $n = 2$, and consider the allocation rule:

$$\theta^1(E) = \begin{cases} (1-t)\dfrac{E^1}{E^S} + \dfrac{t}{2} + \dfrac{t^2\left(E^1 - E^2\right)}{2E^S}, & \text{if } E^1 \geq E^2 \\[2ex] (1-t)\dfrac{E^1}{E^S} + \dfrac{t}{2} + \dfrac{t^2\left(E^2 - E^1\right)}{2E^S}, & \text{if } E^1 \geq E^2 \end{cases}, \tag{6.8}$$

for $t \in (0,1)$. It is easy to verify that these rules satisfy conditions *B(i)–(iii)* of Theorem 8, and these rules are clearly not linear tax rules.

We are not interested in linear economies as such, because they are so special. Theorem 8 is presented because it motivates us to ask how linear taxation performs in concave economies. Let us postulate that a linear taxation allocation rule is applied to a person's income, which is equal to his effort times the Walrasian wage plus an equal-per-capita share of the firm's profits. One may compute that the effort allocation $E(\cdot)$ is a K^{+} equilibrium for the *t*-linear tax rule only if:

$$(\forall \gamma) \quad u_1^{\gamma} \cdot \left((1-t)\left(E^{\gamma} - E^S\right)G''\left(E^S\right) + G'\left(E^S\right)\right) + u_2^{\gamma} = 0, \tag{6.11}$$

and so the marginal rate of substitution of type γ is:

$$-\frac{u_2^{\gamma}}{u_1^{\gamma}} = G'\left(E^S\right) + (1-t)\left(E^{\gamma} - E^S\right)G''\left(E^S\right). \tag{6.12}$$

What is noteworthy is that the wedge between the MRS and the MRT, which is $(1-t)(E^{\gamma} - E^S)G''(E^S)$, goes to zero as *t* approaches one. This must be the case, since the allocation at $t = 1$ is the equal-division allocation, which we know is in $PE(0)$ on convex economies.

Compare Eq. (6.12) with Nash–Walras equilibrium in the same private-ownership economy with taxation, which is given by:

$$-\frac{u_2^{\gamma}}{u_1^{\gamma}} = (1-t)G'\left(E^S\right) \tag{6.11}$$

Here, the wedge between the MRS and the MRT is $tG'(E^S)$ which becomes equal to the whole MRT as *t* goes to one. If there is positive social ethos, citizens might well wish to redistribute market incomes via taxation. Under Nash optimization, *it becomes increasingly costly to do so (as taxes increase), while with* K^{+} *optimization,* Eq. *(6.12) suggests it becomes decreasingly costly to do so, in terms of deadweight loss.*

7. Existence and dynamics

The existence of *proportional solutions*, which are the $K^{\times}$ equilibria of convex economies $(\mathbf{u}, G, n)$ was proved in Roemer and Silvestre (1993).

Here, we provide conditions under which K^β equilibria exist, with respect to the allocation rules described in Theorem 3.

Theorem 10. *Let* $(\mathbf{u}, G, n)$ *be a finite economy where the component functions of* $\mathbf{u}$ *are strictly concave.*

A. *If for all,* γ, $\frac{\partial^2 u^\gamma}{\partial x \partial E} \leq 0$ *then a strictly positive* K^+ *equilibrium w.r.t. the equal-division allocation rule* X_∞ *exists.*
B. *Let* $0 \leq \beta < \infty$. *If for all* γ, u^γ *is quasi-linear, then a strictly positive* K^β *equilibrium w.r.t. the allocation rule* X_β *exists.*

Proof. Appendix A.

The premises of this theorem can surely be weakened.[11]

We turn briefly to dynamics. There will not be robust dynamics for Kantian equilibrium, as there are not for Nash equilibrium. There is, however, a simple dynamic mechanism that will, in well-behaved cases, converge to a Kantian equilibrium from any initial effort vector. The mechanism is based on the mapping Θ defined in the proof of Theorem 9. Informally, the dynamics are as follows. Beginning at an arbitrary vector of effort levels, each agent adds to his own effort the amount r that he would like *all* agents to add to their efforts. This produces a new effort vector, and the process is then iterated. This is the Kantian analog of iterating the best-response function to arrive at a Nash equilibrium.

We illustrate it here for the case of a profile of quasi-linear utility functions and the equal-division allocation rule. Thus, let $u^j(x, y) = x - c^j(y)$, for $j = 1,\ldots,n$, where c^j is a strictly convex function. For any vector $E_0 \in \mathfrak{R}^n_{+\,+}$, define $r^j(E_0)$ as the unique solution of:

$$\arg\max_r \left(\frac{G(E_0 + nr)}{n} - c^j \left(E_0^j + r \right) \right) \tag{7.1}$$

Define $\Theta^j(E_0) = E_0^j + r^j(E_0)$. The mapping $\Theta = (\Theta^1,\ldots, \Theta^n)$maps $\mathfrak{R}^n_+ \rightarrow \mathfrak{R}^n_+$ and is analogous to the best-reply correspondence in Nash equilibrium. A fixed point of Θ is a K^+ equilibrium for the equal-division allocation rule, since at a fixed point E^*, $r^j(E^*) = 0$ for all j. Since the example is special, the next result is proved only for the case $n = 2$, although it is true for finite n. The next proposition shows that if we iterate the mapping Θ indefinitely from any initial starting vector E_0 it converges to (the unique) K^+ equilibrium for the equal-division allocation rule.

Theorem 11. *For* $n = 2$, *there exists a unique fixed point of the mapping* Θ, *which is a* K^+ *equilibrium for the equal-division allocation rule with quasi-linear preferences. The dynamic process defined by iterating the application of* Θ *from any initial effort vector converges to the* K^+ *equilibrium.*

Proof. Appendix A.

The point Theorem 11 makes is that Kantian equilibrium is like Nash equilibrium in that we can define a 'best-reply' function, which in well-behaved cases will converge, if iterated to the Kantian equilibrium.

8. Discussion

My analysis has been positive rather than normative. I have argued that *if* agents optimize in the Kantian way, then certain allocation rules will produce Pareto efficient allocations, while Nash optimization will not. While the *analysis* is positive, Kantian optimization, if people follow it, is motivated by a moral attitude or social norm: each must think that he should take an action if and only if he would advocate that all others take a similar action. Optimization protocols differ from preferences: thus, optimizing according to the Kantian protocol implies nothing about whether one's preferences are other-regarding or self-interested — rather, it has to do with cooperation. You and I may cooperate, to our mutual benefit, whether or not we care about each other. Is it plausible to think that there are (or could be) societies where individuals do (or would) optimize in the Kantian manner?

Certainly parents try to teach Kantian behavior to their children, at least in some contexts. "Don't throw that candy wrapper on the ground: How would you feel if everyone did so?" The golden rule ("Do unto others as you would have them do unto you") is a special case of Kantian ethics. And wishful thinking ["if I do *X*, then all those who are similarly situated to me will do *X*"], although a predictive claim, rather than an ethical one, will also induce Kantian equilibrium — if all think that way. This may explain why people vote in large elections, make charitable contributions, and recycle their garbage. Kantian equilibria may be much more common than those of us trained in Nash equilibrium are bound to think.

Consider the relationship between the theoretical concept of Nash equilibrium and the empirical evidence that agents play the Nash equilibrium in certain social situations that can be modeled as games. We do not claim that agents are consciously computing the Nash equilibrium of the game: rather, we believe there is some process by which players *discover* the Nash equilibrium, and once it is discovered, it is stable, given autarkic reasoning. We now know there are many experimental situations in which players in a game do not play (what we think is) the Nash equilibrium. Conventionally, this 'deviant' behavior has been rationalized by proposing that players have different payoff functions from the ones that the experimenter is trying to induce in them, or that they are adopting behavior that is Nash in repeated games generated by iterating the one-shot game under consideration. Another possibility, however, is that players in these games are playing some kind of Kantian equilibrium. In Roemer (2010), I showed that if, in the prisoners' dilemma game, agents play mixed strategies on the two pure strategies of {Cooperate, Defect}, then all multiplicative Kantian equilibria entail both players' cooperating with probability at least one-half (i.e., no matter how great is the payoff to defecting against a cooperator). It can also be shown that, in a stochastic dictator game, where the dictator is chosen randomly at stage 1 and allocates the pie between herself and the other player in stage 2, the unique $K^\times$ equilibrium is that each player gives one-half the pie to the other player, if he is chosen.

The non-experimental (i.e., real-world) counterpart, as I have said in the introduction, may be the games that the societies that Elinor Ostrom has studied are playing. If these games can be modeled as 'fisher' economies, with common ownership of a resource whose use displays congestion externalities, and if, as Ostrom contends, these societies figure out how to engender efficient allocations of labor applied to the common resource, then they are discovering the multiplicative Kantian equilibrium of the game. Perhaps Kantian reasoning helps to maintain the equilibrium: optimizing behavior may be cooperative and not autarkic. Ostrom explains the maintenance of the efficient labor allocation by invoking the community's use of sanctions and punishments, but that may not be the entire story: it may be that many fishers are thinking in the Kantian manner, and that punishments and monitoring are needed only to control a minority who are Nash-optimizers. I am proposing that an ethic may have evolved, in these societies, in which the fisher says to himself, "I would like to increase my fishing time by 5% a week, but I should do so only if all others could similarly increase their fishing times, and that I would not like. " Armed only with the theory of Nash equilibrium, one naturally thinks that these Pareto efficient solutions to the tragedy of the commons require punishments to keep *everyone* in line.

As I noted earlier, Kantian ethics, and therefore the behavior they induce, require *less* selflessness than another kind of ethic: putting oneself in the shoes of others. Consider charity. "I should give to the unfortunate, because I could have been that unfortunate soul — indeed, there but for the grace of God go I. " The Kantian ethic says, in contrast: "I will give to the unfortunate an amount which I would like all others who are similarly situated to me to give." Assuming that there is a social

[11] As with Nash equilibria, there is no guarantee that Kantian equilibria are unique.

ethos (that is, $\alpha > 0$) this kind of reasoning may induce substantial charity — or, in the political case, fiscal redistribution. Cooperation is the active behavior rather than empathy.

To the extent that human societies have prospered by exploiting the ability of individuals of members of our species to cooperate with each other, it is perhaps likely that Kantian reasoning is a cultural adaptation, selected by evolution (the classic reference is Boyd and Richerson, 1985). Because we have shown that Kantian behavior can resolve, in many cases, the inefficiency of autarkic behavior, cultures which discover it, and attempt to induce that behavior in their members, will thrive relative to others. Group selection may produce Kantian optimization as a meme. Imagine, for example, a time when there were many societies of fishers. Suppose that in a small number of these societies, a clever priest or shaman proposed that fishers optimize using the multiplicative Kantian protocol. These societies, given the proportional allocation rule, will achieve Pareto efficient allocations. If utility tracks fitness, these societies will prosper while those using the Nash protocol will not. The meme of Kantian optimization could spread.[12]

One can rightfully ask whether it is utopian to suppose that the allocation rules studied here can be used in large economies.[13] Even if the allocation rules of Theorem 3 are not employed, one may ask what happens if agents in a private-ownership economy with markets optimize by choosing their effort supplies in the Kantian manner. I have done some simulations of the affine tax allocation rules where the market allocation is Walrasian, and the production function is strictly concave.[14] We do not get full Pareto efficiency, but the results are better when agents optimize in the additive Kantian way than when they are Nash optimizers.

One of the motivations I gave for studying Kantian optimization was to resolve the inefficiencies in economies with a social ethos, due to the consumption externalities that they entail. One might think that, if a society is altruistic in the sense of possessing a social ethos, then it is more likely that its members would behave in a cooperative fashion. The behavior upon which I have focused in this article is optimizing behavior. I have not argued, however, that there is a link between a community's possessing a social ethos and its members' learning and employing Kantian optimization, although I suspect that there is. I leave the reader with this question.

Appendix A. Proofs of theorems

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jpubeco.2014.03.011.

References

Alger, I., Weibull, J.W., 2014. Homo moralis: preference evolution under incomplete information and assortative matching. Econometrica (in press).
Arneson, R., 1989. Equality and equality of opportunity for welfare. Philos. Stud. 56, 77–93.
Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press, Princeton.
Boyd, R., Richerson, P., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.
Brekke, K.A., Kverndokk, S., Nyborg, K., 2003. An economic model of moral motivation. J. Public Econ. 87, 1967–1983.
Cohen, G.A., 1989. On the currency of egalitarian justice. Ethics 99, 906–944.
Cohen, G.A., 2009. Why not Socialism? Princeton University Press, Princeton.
Curry, P., Roemer, J., 2012. Evolutionary stability of Kantian optimization. Rev. Public Econ. (Hacienda Publica Espanola) 200-(I/2012), 131–148.
Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F., Sobel, J., 2011. Other regarding preferences and general equilibrium. Rev. Econ. Stud. 78, 613–639.
Dworkin, R., 1981. What is equality? Part 2: equality of resources. Philos. Public Affairs 10, 283–345.
Fleurbaey, M., 2008. Fairness, Responsibility, and Welfare. Oxford University Press.
Harsanyi, J., 1977. Rational Behavior and Bargaining Equilibrium in Games and Social Situations. Cambridge University Press.
Henrich, N., Henrich, J., 2007. Why Humans Cooperate: A Cultural and Evolutionary Explanation. Oxford University Press, Oxford.
Olson, M., 1965. The Logic of Collective Action. Harvard University Press, Cambridge MA.
Ostrom, E., 1990. Governing the Commons: the Evolution of Institutions for Collective Action. Cambridge University Press, Cambridge.
Ostrom, E., Gardner, R., 1993. Coping with asymmetries in the commons: self-governing irrigation systems can work. J. Econ. Perspect. 7, 93–112.
Rabin, M., 2006. The Experimental Study of Social Preferences. Soc. Res. 73, 405–428.
Rawls, J., 1971. A Theory of Justice. Harvard University Press, Cambridge, MA.
Roemer, J., 1996. Theories of Distributive Justice. Harvard University Press, Cambridge, MA.
Roemer, J., 1998. Equality of Opportunity. Harvard University Press, Cambridge, MA.
Roemer, J., 2010. Kantian equilibrium. Scand. J. Econ. 112, 1–24.
Roemer, J., Silvestre, J., 1993. The proportional solution for economies with both private and public ownership. J. Econ. Theory 59, 426–444.
Scheve, K., Stasavage, D., 2012. Democracy, war and wealth: lessons from two centuries of inheritance taxation. Am. Polit. Sci. Rev. 106, 81–102.
Tomasello, M., 2009. Why We Cooperate. MIT Press, Cambridge, MA.

[12] For some preliminary evolutionary analysis of Kantian behavior, see Curry and Roemer (2012).

[13] An interesting recent example is the behavior of the small island nation of Mauritius with regard to global warming, which will affect it severely, through rising sea levels. Mauritius has undertaken serious steps to reduce its carbon footprint, although this will have negligible effect on its own situation (namely, the sea level). It is behaving as a Kantian optimizer, taking the action it would like all other nations to take. Kantian optimization, in this case, is an attempt to set a moral example. See the Maurice Ile Durable website (http://www.gov.mu/portal/sites/mid/index.html). We can think of many other examples where individuals have attempted to induce cooperative behavior in others by their moral example.

[14] Available from the author.