Arpil 2, 2015

"How we cooperate … perhaps"

by

John E. Roemer*

Yale University

Abstract. Evolutionary psychologists argue that *homo sapiens*' ability to cooperate is a selected adaptation, unique to our species among the great apes. Economic theory, for the main, attempts to explain cooperative behavior as the *non-cooperative* equilibrium of a complex game with many stages. The innovation of behavioral economics is to include exotic arguments in preferences (for example, a sense of fairness) but the analytical structure is still Nash (non-cooperative) equilibrium. I argue that both these approaches are unsatisfying. Instead, I propose that cooperators possess classical (non-exotic) preferences, but optimize in a cooperative (Kantian) way, and that doing so is not irrational. I distinguish between cooperative behavior and altruism, argue that altruism is unnecessary for cooperation, and indeed may not induce outcomes that are different from those that occur with cooperation, absent altruism. My approach provides microfoundations for cooperative behavior that are precisely analogous to the microfoundation that Nash equilibrium provides for non-cooperative behavior.

* Departments of political science, economics, Cowles Foundation. Acknowledgments to be added. john.roemer@yale.edu

1. A cooperative species

It is frequently said that *homo sapiens* is a cooperative species. It is clearly not unique in this regard: ants and bees cooperate, and perhaps other mammalian species do as well. But Michael Tomasello (2014) argues, I think persuasively, that the only cooperative species among the five great apes (chimpanzees, bonobos, gorillas, orangutans, and humans) are the humans. Tomasello believes that the tendency to cooperate with other humans is inborn. He offers a number of examples of our features and behavior that are unique to humans among the five great apes, indicating that the tendency to cooperate must have evolved very early. Here are three: (1) among the great apes, humans are the only beings with sclera (the whites of the eyes); (2) only humans point and pantomime; (3) only humans have language. The conjecture is based on the fact that it is the sclera of the eye that enables you to see what I am looking at. If I am looking at an animal that would make a good meal, and if you and I cooperate in hunting, it is useful for me that you can see the animal I am looking at, because then we can catch and consume it together. Were you and I only competitors it would not be useful for me that you see the object of my gaze, as we would then fight over who gets the animal. Thus, one would expect the mutation of sclera to be selected in a cooperative species, but not selected in a competitive one. Miming and pointing probably first emerged in hunting as well, and were useful for members of a species who cooperated in hunting. Chimpanzees, who do not cooperate in hunting, do not mime or point[1]. Miming and pointing are the predecessors of language. Complex organs like the eye and the language organ must have evolved incrementally as the result of selection of many random mutations. Tomasello argues that language would not be useful, and would not evolve in a species that did not already have cooperative behavior. If you and I are only competitors, why should you believe anything I tell you? I am only out for myself, and must be trying to mislead you, because cooperation is not something in our toolkit. So language, were primitive forms of it to emerge in a non-cooperative species, would die out for lack of use.

Tomasello's main work consists of experiments in which he compares human infants to chimpanzees, who are set with a task in which cooperation would be useful. The general

[1] Tomasello disagrees with some who argue that chimpanzees do cooperate in hunting smaller monkeys.

outcome of these experiments is that human infants cooperate immediately, while chimpanzees do not. Often, Tomasello's cooperative project involves working together to acquire some food, which then must be shared. If chimpanzees initially cooperate in acquiring the food, they find they cannot share it peacefully, but fight over it, and hence they do not cooperate the next time the project is proposed to them, for they know that the end would be a fight, which is not worth the value of the food acquired. Human infants, however, succeed immediately and repeatedly in cooperating in both the productive and consumptive phase of the project.

There are, of course, a huge number of examples of human cooperation, involving projects infinitely more complex than hunting or acquiring a piece of food that is difficult to get. Humans have evolved complex societies, in which people live together, cheek by jowl, in huge cities, and do so relatively peacefully. We organize complex projects, including states and taxation, the provision of public goods, large firms, and so on, which are only sustained because most of those who participate do so cooperatively – that is, they participate not because of the fear of penalties if they fail to do so, but because they understand the value of contributing to the cooperative venture. (This may seem vague at this point, but will become more precise below.) We often explain these human achievements by the intelligence that we uniquely possess. But intelligence does not suffice as an explanation. The tendency to cooperate, whether inborn or learned, is surely necessary. If we are persuaded by Tomasello, then the tendency to cooperate is inborn and was necessary for the development of the huge and complex cooperative projects that humans undertake.

It is even possible that large brains that differentiate humans from the other great apes evolved as a result of the cooperative tendency. Why? Because large brains are useful for complex projects – initially, complex projects that would further the fitness of the members of the species. From an evolutionary viewpoint, it might well not be efficient to spend the resources on producing a large brain, were it not necessary from complex projects. Such projects will not be feasible without cooperation: by definition, complexity, here, means that the project is too difficult to be carried out by an individual, and requires coordinated effort. If humans did not already have a tendency to cooperate, then a mutation that enlarged the brain would not be selected, as it would not be useful. So not only language, but intelligence generally, may be the evolutionary product of a prior selection of the cooperative 'gene.'

Readers may object: cooperation, they might say, is fairly rare among humans, who are mainly characterized by competitive behavior. Indeed, what seems to be the case is that cooperation evolves in small groups – families, tribes – but that these groups are often at war with one another. Stone-age New Guinea, which was observable up until around the middle of the twentieth century, was home to thousands of tribes (with thousands of languages) who fought each other; but within each tribe, cooperation flourished. (One very important aspect of intra-tribal cooperation among young men was participating in warfare against other tribes. See Bowles and Gintis (2013).) Indeed, up until the middle of the twentieth century (at least), human society has been characterized by increasingly complex states, in which cooperative behavior is pervasive internally, who do not trust, and fight wars, with each other. So the human tendency to cooperate is, so it appears, not unlimited[2].

This paper's purpose is to review some of the main results on the microfoundations of cooperation that I have published earlier in Roemer (1996, 2009, 2014). As time goes by, one learns to simplify one's analysis, perhaps to argue for the ideas more adroitly, and this motivates the present paper. In particular, here, I spend half of the exposition discussing *simple Kantian equilibrium*, which, as the name says, is simpler than the kinds of Kantian equilibria I discovered earlier, and are discussed later in this paper.

My approach to cooperation is most similar to the idea of *strong reciprocity,* first formulated in Gintis (2000), and discussed in several other publications since then (e.g., Gintis et al (2008)). In addition, a serious attempt to conceptualize cooperation as reciprocity was published by Kolm (2005). What I hope to have added to their story are microfoundations – a proposal for exactly how cooperators may be thinking -- and also to propose a sharper distinction between cooperation and altruism. What their work adds to mine is a more thorough discussion of the evolution of cooperation, as well as a deeper understanding of the experimental evidence for the theory.

2. Economic theory and cooperation

[2] There is a recent spate of books on human cooperation: see Bowles and Gintis (2011), Tomasello (2009), Henrich and Henrich(2007), as well as Tomasello (2014).

If the secret of the success of *homo sapiens* is its ability to cooperate, one would think that economic theory would try to understand how cooperation takes place. But the great achievements of economic theory have been to model competitive behavior. These achievements are encapsulated in two models: the general-equilibrium model of a competitive economy, and the Nash equilibrium of a game. Both models were fully formed by the 1950s. The first formal statements of the models were Léon Walras's simultaneous equation model of a competitive economy (1874) and Auguste Cournot's model of duopoly (1838), respectively.

Both models are populated by agents who possess preference orders that they attempt to maximize. What's key is that each agent treats his environment as inert, or 'parametric.' I call this *autarchic optimization.* Assuming that his environment is inert and unchanging, the individual chooses that action from his feasible choices that maximizes his own preference order. An *equilibrium* of the model consists of a set of actions by each individual, where each action is optimal for the individual given the actions of all others, such that the actions are jointly consistent, in the sense of being feasible and individually optimal conditional on the play of all others.

In the general equilibrium model, the actions of consumer-workers are labor supplies, consumption orders and investment supplies, and the actions of firms are labor and investment demands. What is special about the model is the concept of a price vector, which puts a value on every commodity and every kind of labor. Remarkably, no individual need ever know anything about the *actions* of other individuals (consumers or firms): she need only study the price vector, and choose her optimal actions with respect to it. The price vector defines what actions are feasible for her, her preference order is defined on these actions, and so there is no need to see the actions of other individuals. A price vector is said to be an equilibrium if, when each agent optimizes autarchically against it, the 'sum' of all actions is consistent: that is, the aggregate demand for every commodity (including kinds of labor and investment goods) equals its aggregate supply. (All markets clear.) Indeed, there is no *social* activity at all in the model: each individual jousts with the price vector, as it were, and, willy-nilly, all actions mesh. (Marx might have chuckled at this formal incarnation of the commodity fetishism of the market economy.)

A *game* is formally defined as a situation in which each player's preferences are defined over the actions of all the players. John Nash proposed that an equilibrium be defined as a set of actions, one for each player, such that, treating the actions of others as given, each individual's action maximizes his own preference order (or payoff function). If such a vector of actions is proposed, then each player does the best that he can for himself, treating the other players as inert, and the result is consistent, in the sense that the proposed vector of actions is feasible and stable. Nash proved that in a large class of games, such an equilibrium always exists (there may, however, be many of them in a game). John von Neumann, who with Oscar Morgenstern had several years earlier published *A Theory of Games and Economic Behavior,* was non-plussed by Nash's idea. The von Neumann-Morgenstern concept of equilibrium was based on considering what various coalitions of players could achieve by cooperating with each other, and requiring an equilibrium to be a set of actions which, in various senses, was immune to challenge from any coalition. Perhaps von Neumann disliked the non-cooperative (that is, autarchic) nature of Nash's concept: we do not know the basis of his disregard for Nash equilibrium.

Cooperative game theory, which was initiated by von Neumann and Morgenstern in their book, has largely faded away in today's economics curriculum. I must, however, point out that their theory did not explain cooperation, or propose *how individuals might achieve it*: it treated cooperation as a black box. As Mas-Colell (1987, p. 659) writes:

> The typical starting point [of cooperative game theory] is the hypothesis that, in principle, any subgroup of economic agents (or perhaps some distinguished subgroups) has a clear picture of the possibilities of joint action and that its members can communicate freely before the formal play starts. Obviously, what is left out of cooperative theory is very substantial.

In the theory it is simply assumed that each coalition of players can avail itself of certain payoffs, should its members cooperate among themselves, but it is not explained how they achieve the cooperation that produces these payoffs. *Given* these data, players in the global economy then compete with each other, with their coalitional payoffs as backstops. What is lacking in the theory, to make Mas-Colell's point explicit, is a theory of how agents in a coalition achieve the cooperative payoffs. To be precise, the non-cooperative equilibrium of

Nash explains how non-cooperative agents reach a stable point, but 'cooperative' game theory does not explain how members of coalitions reach a cooperative solution within their coalitions.

In other words, what we need is a micro-foundation for cooperative behavior, in the sense that Nash equilibrium and Walrasian equilibrium provide micro-foundations for non-cooperative behavior. My goal in this paper is to propose one way of micro-founding cooperative behavior.

Now modern economists do not ignore the existence of cooperation, although it must be said that, because the main tools of the trade explain non-cooperative behavior, it is unsurprising that cooperation has received fairly short shrift until recent years. Economists, by and large, attempt to explain instances of cooperation using their non-cooperative tools: that is, they attempt to explain cooperation as the Nash *non-cooperative* equilibrium of a (quite complex) game. I will discuss this below, where I will suggest that there is something Ptolemaic about the approach. We observe that human behavior is sometimes (often?) cooperative. But we have only a theory of non-cooperation. So we will attempt to explain cooperation as, really, a non-cooperative outcome of a complex game. Not only is this intellectually unsatisfying, but, I will argue, it is incredible that we could maintain large examples of cooperation where everyone is, in fact, thinking like an autarchic optimizer, treating his environment as inert.

3. Cooperation, solidarity, and altruism

Altruism is often modeled by proposing that an individual's utility function contains as arguments the utilities achieved by other individuals, and that the individual's utility responds positively to an increase in the utilities of others. The classical assumption is that an individual is self-interested, in the sense that her preferences are defined only over goods that she consumes. (These may include public goods, which can be simultaneously consumed by many others.) Cooperation and altruism are very often confounded. In fact, I maintain they are independent concepts, and it is wise to keep them conceptually distinct[3].

[3] A typical example is the fine treatise by Bowles and Gintis (2011), entitled *A cooperative species: Human reciprocity and its evolution.* The first two chapters (after the introduction) are entitled 'The evolution of altruism in humans' and 'Social preferences.' In contrast social preferences and altruism play almost no role in the theory of cooperation I present here.

For members of a group to cooperate means that they 'work together, act in conjunction with one another, for an end or purpose (Oxford English Dictionary).' There is no supposition that they care about each other. Cooperation may be the only means of satisfying *one's own self-interested preferences*. You and I build a house together so that we may each live in it. We cooperate not because of interest in the other's welfare, but because cooperative production is the only way of providing *any* domicile. The same thing is true of the early hunters I described above: without cooperation, neither of us could capture that deer, which, when caught by our joint effort, will feed both of us. In particular, I cooperate with you because the deer will feed *me*. It is not necessary that I ascribe any value to the fact it will feed you, too.

Solidarity is defined as 'a union of purpose, sympathies, or interests among the members of a group (American Heritage Dictionary).' H.G. Wells is quoted there as saying , "A downtrodden class … will never be able to make an effective protest until it achieves solidarity." Solidarity, so construed, is not the cooperative action that the individuals take, but rather a characterization of their objective situation: namely, that they are all in the same boat. I take 'a union of interests' to mean we are all in the same situation and have common preferences. It does not mean we are altruistic towards each other. Granted, one might interpret 'a union of …sympathies..' to mean altruism, but I choose to focus rather on ' a union of purpose or interests.' The Wells quote clearly indicates the distinction between the joint action and the state of solidarity.

Of course, people may become increasingly sophisticated with respect to their ability to understand that they have a union of interests with other people. The old left-wing expression "we all hang together or we all hang separately" urges everyone to see that he does, indeed, have similar interests to others, and hence it may be logical to act cooperatively (to hang together). Notice the quoted expression does not appeal to our altruism, but to our self-interest, and to our solidaristic state.

My claim is that the ability to cooperate for reasons of self-interest is less demanding than the prescription to care about others. I believe that it is easier to explain the many examples of human cooperation from an assumption that people learn that cooperation can further their own interests, than to explain those examples by altruism. For this reason, I

separate in this paper the discussion of cooperation among self-interested individuals from cooperation among altruistic ones; the latter topic will be addressed in section 10 below.

4. Simple Kantian optimization and simple games

We can say that a set of individuals enjoy a union of interests when its members have the same preferences and face a common environment. This does not exhaust the class of cases when people have a union of interests, but it is an instance of such a case. Let us take the simplest example of a *symmetric game*, with two persons, who have the same preferences. This means that their payoff functions are identical up to a permutation of the individuals. Let the strategy space of each individual be some interval of positive real numbers *I*, and define the payoff functions V^i of the two persons as:

$$V^1(E^1,E^2) = V(E^1,E^2) = V^2(E^2,E^1), \text{ some function } V,$$

where $E^i \in I$. The actions are to be thought of as 'efforts;' hence, the notation. The prisoners' dilemma, though defined on a discrete strategy set of two actions, can be written in this form, as can all the other familiar symmetric two-person games (chicken, battle of the sexes, etc.). These games on a small set of strategies (usually two) induce games where the strategy space is an interval by considering mixed strategies.

My approach to modeling cooperation is to *alter the optimizing protocol* from the autarchic protocol. In a symmetric game, suppose each player asks himself: what action would I most like *all* of us to take? I call this a *simple Kantian optimizing protocol*, as the individual is applying the categorical imperative of Kant: take that action one would desire to have universalized. This means each player, in the above game, chooses $E \in I$ to maximize $V(E,E)$. By definition, the two players in the symmetric game will agree upon the solution: call it E^* .

Definition 1. In a game $\{V^i\}$ where all players have the same strategy space *I*, a *simple Kantian equilibrium* (SKE) is strategy $E^* \in I$ such that :

$$(\forall E \in I)(\forall i)(V^i(E^*,...,E^*) \geq V^i(E,...,E)) \ .$$

A SKE is the strategy that each would prefer that all players play, conditional upon their all playing the same strategy. If the game is not symmetric, typically a SKE will not exist. But

it will exist in symmetric games, since all players will be maximizing the same function $V(E,E)$[4] .

We now observe:

Proposition 1 *In the* 2×2 *symmetric game, if the function V is concave, then the simple Kantian equilibrium is Pareto efficient.*

Proof:

Suppose not, and there is a pair of actions (E^1,E^2) such that:

$$V(E^1,E^2)\geq V(E^*,E^*) \text{ and } V(E^2,E^1)\geq V(E^*,E^*)$$

with at least one strict inequality. Adding these inequalities gives:

$$\frac{1}{2}V(E^1,E^2)+\frac{1}{2}V(E^2,E^1)>V(E^*,E^*) \ .$$

By concavity, $V(\bar{E},\bar{E})\geq \frac{1}{2}V(E^1,E^2)+\frac{1}{2}V(E^2,E^1)>V(E^*,E^*)$, where $\bar{E}=\frac{E^1+E^2}{2}$. This contradicts that fact that E^* maximized the function $V(E,E)$, which proves the claim. ■

It is a little trickier to define a symmetric game with $n\geq 2$ players. To keep the argument very simple, let's avoid that complexity and treat a special case. I'll consider symmetric *n*-person games where the payoff function of player *i* is $V(E^i,E^{N\setminus i})$, where $E^{N\setminus i}=\sum_{j\neq i}E^j$. An example of such a game is one with congestion externalities: for example, fishers on a lake, where the fishing of others decreases the productivity of the lake. Many situations where the action of each contributes in the aggregate to a public bad can be modeled in this way; these are often called tragedies of the commons. We have:

---

4 I owe the formulation of simple Kantian equilibrium to Brekke et al (2003) who write "To find the morally ideal effort the individual asks herself, 'Which action would maximize social welfare, given that everyone acted like me?'" My definition is, however, different, because my decision maker does not think in terms of maximizing social welfare. Although this may appear to be a distinction without a difference in an economy where everyone has the same preferences, the difference becomes crucially important when we move to an environment with heterogeneous agents.

Proposition 2. *If V is concave, then the simple Kantian equilibrium of the n-person symmetric game is Pareto efficient.*

Proof:

1. Let E^* maximize $V(E^*,(n-1)E^*)$. $(E^*,...,E^*)$ is the simple Kantian equilibrium. Suppose it is Pareto dominated by a vector of actions $(E^1,...,E^n)$. Then:

$$(\forall i)(V(E^i,\sum_{j\neq i}E^j)\geq V(E^*,(n-1)E^*))$$

with a strict inequality for at least one *i*. By adding these *n* inequalities we have:

$$\sum_{i=1}^{n}\frac{1}{n}V(E^i,\sum_{j\neq i}E^j)>V(E^*,(n-1)E^*)\ .$$

By concavity of *V*, $\sum_{i=1}^{n}\frac{1}{n}V(E^i,\sum_{j\neq i}E^j)\leq V(\bar{E},(n-1)\bar{E})$ where $\bar{E}=\sum_{i=1}^{n}E^i/n$, and therefore

$$V(\bar{E},(n-1)\bar{E})>V(E^*,(n-1)E^*),$$

contradicting the definition of E^* . This proves the claim. ■

Definition Let $\{V^i \mid i=1,...,n\}$ be the payoff functions of an *n* person game, where the strategy space for each player is a real interval. The game is *monotone increasing* (*strictly monotone increasing*) if each player's payoff function is increasing (strictly increasing) in the strategies of the other players. The game is *monotone decreasing* (*str. monotone decreasing*) if each player's payoff function is decreasing (str. decreasing) in the strategies of the other players. A game is *monotone* (*strictly monotone*) if it is either monotone increasing or decreasing (strictly monotone increasing or decreasing).

Proposition 3. *Let a* 2×2 *symmetric game be strictly monotone. Then any SKE of the game is Pareto efficient.*

Proof:

Suppose the game is strictly monotone increasing. Let (E^*,E^*) be a SKE and suppose it is Pareto-dominated by (E^1,E^2) , so:

$$V(E^1,E^2)\geq V(E^*,E^*) \text{ and } V(E^2,E^1)\geq V(E^*,E^*)$$

with at least one inequality strict. Obviously $E^1\neq E^2$. For this would contradict the fact that (E^*,E^*) is a SKE. Suppose $E^1<E^2$. Then:

$$V(E^2,E^2) > V(E^2,E^1) \geq V(E^*,E^*) \ ,$$

where the first inequality follows by the strict monotone increasing property of the game, invoked for the second player. But this inequality contradicts the premise that (E^*,E^*) is SKE.

Essentially the same argument works if the game is strictly monotone decreasing.■

Clearly, in the classical prisoners' dilemma (PD), where there are only two strategies, it is obvious that the simple Kantian equilibrium is that both players play 'cooperate.' If we move to mixed strategies, then the equilibrium depends on the payoff matrix, which is, in general form[5]:

	Cooperate	**Defect**
Cooperate	(0,0)	(-c,1)
Defect	(1,-c)	(-b,-b)

where $0 < b < c$. The payoff function of the row player is $V^{PD}(p,q) = -p(1-q)c + (1-p)q - b(1-p)(1-q)$, where p (q) is the probability that Row (Column) plays Cooperate. The game is symmetric (thus, the payoff function of the column player is $V^{PD}(q,p)$). The common strategy space of the two players is $I = [0,1]$. Recall that in the mixed-strategy game, Pareto efficiency is defined in terms of expected utility (i.e., *ex ante* efficiency).

The function V^{PD} is only concave on its domain I^2 in the singular case that $c - b = 1$, in which case it is actually linear. However, the game is strictly monotone increasing: just note that

$$\frac{\partial V^{PD}(p,q)}{\partial q} = pc + (1-p)(1+b) > 0 \ .$$

---

[5] Since the payoffs are von Neumann- Morgenstern utilities, we are free to pick one payoff to be 0 and one to be 1 for each player. Thus, the PD game in mixed strategies is a two-parameter game – here, (b,c) .

It follows immediately from Proposition 3 that the SKE of the mixed-strategy PD game is Pareto efficient.

Proposition 4

a. The SKE of the PD game is Pareto efficient.

b. If $1 \le c \le 1+b$ *, the SKE of the PD game is* $(p^*, p^*) = (1,1)$.

c. If $c < 1$ *the SKE of the PD game is* $p^* = \dfrac{2b+1-c}{2(1+b-c)}$ *and* $0 < p^* < 1$.

d. If $1+b < c$ *, the SKE of the PD game is* $p^* = 1$.

Proof:

Part *a* follows from Proposition 3 since the PD game is str. monotone increasing.

The function $V(p,p)$ is concave if and only if $c-b \le 1$. In this case the first-order condition $\dfrac{d}{dp}V^{PD}(p,p) = 0$ gives the SKE. If $1 \le c$ the solution is a corner one, at $p^* = 1$ (part *b*). If $c < 1$, the solution is interior, and given by part *c*. If $c-b > 1$, the function $V^{PD}(p,p)$ is convex, and hence the SKE occurs either at $p = 0$ or 1 . The value is higher at $p = 1$, giving part *d*. ■

It is interesting that in the case of part *c*, although the simple Kantian equilibrium is Pareto efficient, it entails less than full cooperation. The intuition here is that the payoff to defecting against a cooperator (which is one) is high, and so it is optimal for both players not to cooperate fully. This shows that cooperation, in the Kantian sense, does not always deliver what we might intuitively consider to be 'ideal' cooperative behavior.

I will consider next the 'battle of the sexes,' also a symmetric game, whose payoff matrix is given by

	Dance	**Box**
Box	(0,0)	(*a*,1)
Dance	(1,*a*)	(*b*,*b*)

where 'she' is the row player and 'he' is the column player. Note that I have written the strategies in a non-traditional way (changing the order of their listing for the two players): this must be done to reveal the symmetric nature of the game. Thus, we let *p* be the probability that she plays 'Dance' and *q* the probability that he plays 'Box.' The game has two parameters, (a,b) where $0 < b < a < 1$. The payoff function for the row player is $V^{BS}(p,q) = bpq + p(1-q) + aq(1-p)$ and the column player's payoff is $V^{BS}(q,p)$. A simple Kantian equilibrium in pure strategies must be either (Dance, Box) or (Box, Dance). She prefers the first option, while he prefers the second: so there is no SKE in pure strategies in this game, which is one reason it is an interesting game.

The reader can check that the BS game in mixed strategies is not a monotone game. Nor is the game concave, for any parameter values, so neither propositions 1 or 3 are of much use. We have:

Proposition 5.

a.The SKE of the 2×2 *mixed-strategy BS game is* $(p^*, p^*) = \dfrac{1+a}{2(1+a-b)}$ *, and* $0 < p^* < 1$.

b. There are BS games in which the SKE is not Pareto efficient.

c. The Nash equilibrium of the mixed-strategy BS game is $\hat{p} = \hat{q} = \dfrac{1}{1+a-b}$. *It is strictly Pareto dominated by the SKE.*

d. $p^* < \hat{p}$.

Proof:

Compute that $V^{BS}(p,p) = (b-(1+a))p^2 + p(a+1)$, which is a strictly concave function of *p*. Hence the FOC gives us the SKE, which is $p^* = \dfrac{1+a}{2(1+a-b)}$. It is easy to compute that p^* is interior in [0,1]. Compute that $V^{BS}(p^*,p^*) = \dfrac{(a+1)^2}{4(a+1-b)}$. Let $a = 0.75, b = .01, p = 0, q = 0.6$. Then

$$V^{BS}(p^*,p^*) = 0.4400, V^{BS}(p,q) = 0.45, V^{BS}(q,p) = 0.6 ,$$

and so (p^*,p^*) is Pareto-dominated by (p,q) .

The Nash equilibrium of the mixed-strategy BS game is computed from the first-order conditions for Nash equilibrium. The payoff to each player at the equilibrium is $\frac{a}{1+a-b}$. This is strictly Pareto dominated by the SKE because the inequality

$$\frac{a}{a+1-b} < \frac{(a+1)^2}{4(a+1-b)}$$

is equivalent to $0 < (a-1)^2$, which is true, because $a < 1$. ■

In other words, simple Kantian optimization does not always deliver Pareto efficiency in the BS game, although the SKE always dominates the Nash equilibrium of the game. From part *d*, we have that in the SKE, both 'she' and 'he' offer to attend their favorite event with lower probability than in the Nash equilibrium (NE): in other words, they *compromise more* in SKE than in NE.

More generally, we must have that, in any symmetric game, the SKE Pareto dominates the symmetric NE, as long as the two equilibria are not the same, because the symmetric NE is of the form (p,p), and SKE maximizes the payoff of the players on the diagonal of strategy space I^2.

Nevertheless, the *BS* game is one that is harder to crack , from the cooperative viewpoint, than the *PD* game, because in the latter the SKE is always (ex ante) efficient.

Finally, we consider 'chicken,' also known as 'Hawk-Dove' game, which we take as the names of the strategies. The payoff matrix is given by:

	Dove	**Hawk**
Dove	(*c*,*c*)	(*b*,1)
Hawk	(1,*b*)	(0,0)

where $1 > c > b > 0$. The payoff function is $V^{HD}(p,q) = cpq + bp(1-q) + q(1-p)$, where *p* (*q*) is the probability that the row (column) player plays Dove. We immediately verify that *HD* is a strictly monotone increasing game, and so the SKE is Pareto efficient. The SKE is given by:

$$p^* = \begin{cases} 1, \text{ if } c \geq \dfrac{1+b}{2} \\ \dfrac{1+b}{2(1+b-c)}, \text{ if } c < \dfrac{1+b}{2}. \end{cases}$$

Thus, peace reigns if *c* is sufficiently large; otherwise, there is a positive probability that peace reigns although it is not assured. There are three Nash equilibria to HD: $(1,0),(0,1),$ and $(\frac{b}{1+b-c},\frac{b}{1+b-c})$. The SKE Pareto dominates the symmetric Nash equilibrium.

Besides the 2×2 games, three other simple games about which much has been written are the dictator ,ultimatum, and trust games. I will assume classical preferences: a player's von Neumann Morgenstern utility is some strictly concave increasing function of the monetary prize, $u(x)$, normalized so that $u(0)=0$ and $u(1)=1$. The second player's vNM utility function is v , similarly normalized. In the *stochastic dictator* game, Nature chooses one of two players to be the dictator, who then assigns a division of a dollar between herself and the other player. Thus, assuming each player is chosen to be the dictator with probability one-half, the expected utility of first player, if she keeps *x* and the second player decides to keep *y*, is $\frac{1}{2}(u(x)+u(1-y))$. In a simple Kantian equilibrium, the first player chooses x to maximize $\frac{1}{2}(u(x)+u(1-x))$, the solution to which is $x=\frac{1}{2}$. Clearly, the second player also chooses $x=\frac{1}{2}$. Strict concavity is necessary to generate this result.

In the stochastic ultimatum game, a player's strategy consists of an ordered pair (x,z) , where *x* is what he will give to the other player, should he be chosen to be the decision maker, and *z* is the minimum that he will accept, should the other player be chosen the decision maker. The game has three stages: first, Nature chooses the ultimator; second, the ultimator presents an offer; third, the other player either accepts or rejects. The unique subgame perfect Nash equilibrium is $(x,z)=(1,0)$.

It is not obvious how to model cooperation in the ultimatum game. This is the first time we have encountered a game where the strategy is multi-dimensional. It seems to me a Kantian should think as follows. If I were to offer *x*, if chosen to be the ultimator, this must be the amount I would also like the other person to offer, if he were the ultimator, and hence I

must accept any amount from him that is at least $1-x$. Therefore, $z \le 1-x$. Consequently, the simple Kantian solution solves the program:

$$\max \frac{1}{2}u(x)+\frac{1}{2}u(z)$$
$$\text{subj. to}$$
$$z \le 1-x$$

The unique solution, if *u* is strictly concave, is $(x,z)=(\frac{1}{2},\frac{1}{2})$.

Arguably, the simple Kantian equilibria, in these two games, is closer to what is often observed in experiments than the Nash equilibrium. Moreover, we have established this result without recourse to including a sense of fairness in the utility function. Granted, in the ultimatum game, players who reject offers of less than 0.25 may say they do so because the offer was unfair. My claim is that those offers are considered unfair *because these are not the offers a person should make* if he recognizes the arbitrariness of being chosen the ultimator. Thus, one uses the Kantian protocol because the situation strongly suggests that 'we are all in the same boat' -- Nature is just flipping a coin to choose the ultimator . In more conventional language, it is a social norm to play Kantian in situations of solidarity, and deviators are punished by norm followers. The same explanation applies in the dictator game, even though no retaliation is possible against a stingy dictator. The arbitrariness of Nature's choice induces, in players, use of the Kantian protocol.

Finally, I discuss the 'trust game.' There are two players, who draw lots to determine who moves first. Each player is endowed with *M* units of value. Player One chooses an amount, *x*, to give to Player Two. Player Two, however, receives *ax* units of value, where $a>1$ is a constant known to both. Then Player Two returns some amount, *y*, to Player One and the game is over. It is played only once.

Conventionally, this game is modeled as a stage game, with three stages: first, Nature chooses the order of players; second, the first player moves; third, the second player moves. The unique subgame perfect Nash equilibrium is $x=y=0$ if the players have self-interested preferences.

Suppose a player's von Neumann-Morgenstern utility function for money lotteries is *u*. Before the game begins, her expected utility is $\frac{1}{2}u(M-x+y)+\frac{1}{2}u(M+ax-y)$. She

chooses a strategy (x,y) that she would like both players to choose, which is the one that maximizes her expected utility:

$$\max \frac{1}{2}u(M-x+y)+\frac{1}{2}u(M+ax-y)$$

s.t.

$$0 \le x \le M$$

$$0 \le y \le M+ax$$

If the agent is risk averse (u is strictly concave), the unique solution to this program is

$$x = M, \quad y = \frac{(1+a)M}{2} .$$

Thus, the Kantian optimizer does not break the game up into stages. She recognizes that, before the game begins, both players are 'in the same boat,' and calculates the strategy (x,y) that she would like each to play. Total wealth is maximized when $x = M$ (regardless of what the second player does). At the simple Kantian equilibrium, the total wealth is split equally between the two players: the solution engenders ex post efficiency and equity (in an obvious sense). The game need not even be symmetric – players will converge on this equilibrium regardless of their risk preferences, so long as they are both risk averse.

Cox, Ostrom et al (2009) perform the trust game with students, and report the results. It appears from Figure 4.1 of their paper that out of 34 games played by different players, three played the simple Kantian equilibrium. (Cox, Ostrom et al (2009) do not call it that: I am imposing my interpretation on the results.) In 11 out of 34 games, the first player played $x = M$: that is, he played his part of the SKE. In only three of these cases, however, did the second player respond with the value of y associated with the SKE. However, in 9 out of these 11 cases, the second player returned at least M to the first player. When the second player returns exactly M, she is, of course, keeping the entire surplus generated from cooperation, rather than sharing it with the first player, but she leaves the first player whole. In four out of 34 games, the Nash equilibrium was played. In six out of 34 games, the first player contributed a positive amount to the second, and the second responded Nash, by returning zero to the first. The authors conduct interviews with the participants after the conclusion of the game, and discover, unsurprisingly, that playing $x = M$ is associated with having trust in others.

Very little interpretive gloss on the results is provided in Cox, Ostrom et al (2009); however, Walker and Ostrom (2009) do provide an interesting gloss on the results of the earlier paper. The authors discuss the results of experiments with three games: the trust game of Cox, Ostrom et al (2009), a public-goods game, and a common-pool-resource game. They write they each of these games are instances of 'social dilemmas:'

> Social dilemmas characterize settings where a divergence exists between expected outcomes from individuals pursuing strategies based on narrow self-interests versus groups pursuing strategies based on the interests of the group as a whole… individuals make decisions based on individual gains rather than group gains or losses; and environments that do not create incentives for internalizing group gains or losses into individuals' decision calculus.

From my point of view, these authors are confounding cooperation with altruism. As I showed above, the fully cooperative solution is attained by a Kantian optimizer who has no concern for others: caring about group gains is irrelevant. Saying that the problem in social dilemmas is based upon 'a divergence between …narrow self-interest versus …strategies based on the interests of the group as a whole' is, from my viewpoint, a gratuitous interpretation of the thought process. Playing the strategy that one would like everyone to play is, for me, motivated entirely by self-interest, not by a concern for the welfare of the group as a whole. It entails a recognition that cooperation can make *me* better off (incidentally, it makes all of us better off). But that parenthetical fact is not or *need not be* the motivation for my playing 'cooperatively.' The fact that these games are played only once by a team shows that building a reputation was not an issue.

My interpretation of the Cox, Ostrom (2006) results for the trust-game experiment is that about one-third of the players chosen to be first movers were playing (their part) of the simple Kantian equilibrium, because they had trust in their opponents/partners. About 27% of their partners responded by playing (their part) of the Kantian equilibrium. Another 54% of the second players in these matches shared the gains induced by the first players' transfers, but did not share as much as the simple Kantian equilibrium prescribes; none of the second players in these matches played the Nash solution in the subgame that they faced (i.e., returning nothing to the first player). A smaller fraction of players appear to be using

autarchic optimization. I cannot reject the hypothesis that a significant number of individuals are Kantian optimizers. I see no reason to suppose that group welfare motivated anyone.

5. Simple Kantian optimization and economic games

Somewhat more interesting than simple matrix games are economic games with production. The pre-eminent example is the game that illustrates the tragedy of the commons. There is a lake upon which a community of fishers live. In the symmetric case, each fisher has the same preferences over fish caught and labor expended, represented by the concave utility function $u(x,E)$, where x is fish consumed and E is effort or labor expended. The lake produces fish according to a concave production function G, where $G(E^S)$ is total fish caught when total fishing labor expended is $E^S = \sum_i E^i$. We assume that E is measured in efficiency units of labor, so that, randomness aside, the amount of fish caught by fisher i is given by $x^i = \frac{E^i}{E^S} G(E^S)$. This defines a game, where the payoff function of fisher i is:

$$V^i(E^i,E^S) = u(\frac{E^i}{E^S} G(E^S),E^i) \ . \qquad (5.1)$$

It is well-known that if G is strictly concave, the Nash equilibrium(a) of this game is (are) Pareto inefficient: all fishers could improve their welfare by cutting back a bit on their fishing time.

It is obvious that this game is strictly monotone decreasing, because the average yield $\frac{G(E^S)}{E^S}$ is strictly decreasing, since G is strictly concave. Therefore, if any j increases his fishing time, i's yield falls, holding constant her own effort. It therefore follows by Proposition 2 that the SKE of the game is Pareto efficient in the game. (Proposition 2, to be precise, continues to hold for symmetric games whose payoff functions are of the form $V^i(E^i,E^S) = V^{\Pr}(E^i,E^S) = u(\frac{E^i}{E^S} G(E^S),E^i)$ as is the case here.) So, simple Kantian optimization resolves the tragedy of the commons.

However, this resolution is thus far incomplete, in the following sense: saying that an allocation of efforts is Pareto efficient *in the game* is not the same as saying it's Pareto efficient *in the economy*. The only allocations that are admitted in the game defined by $V^{\Pr}$

are the proportional allocations, where fish are divided in proportion to labor expended. But there may be non-proportional allocations, feasible in the economy, that Pareto dominate the SKE of $V^{\Pr}$. Could this happen?

Pareto efficiency in the fishing *economy* is defined by the equations that state that each fisher's marginal rate of substitution between fish and labor equals the marginal rate of transformation between fish and labor: that is, an interior feasible allocation $\{(x^i, E^i)_{i=1,\ldots,n}\}$ is Pareto efficient in the economy if and only if:

$$(\forall i)(G'(E^S) = -\frac{u_2(x_i, E_i)}{u_1(x_i, E_i)}) \qquad (5.2)$$

where u_i is the i^{th} partial derivative of u and G' is the derivative of G. Let us compute the SKE of the fishing economy by examining the first-order condition for a simple Kantian optimizer:

$$\frac{d}{dE^i} u(\frac{1}{n} G(nE^i), E^i) = 0 \ ,$$

which expands to :.

$$u_1 \cdot \left(\frac{G'(nE^i)n}{n} \right) + u_2 = 0, \text{ or } -\frac{u_2}{u_1} = G'(E^S) \ . \qquad (5.3)$$

Therefore, indeed:

Proposition 6 *The SKE of the game* $V^{\Pr}$ *is Pareto efficient in the economy.*

In this sense, the SKE resolves the commons' tragedy completely. An allocation in this production economy which is *proportional* and *Pareto efficient* is called a *proportional solution.* A more general definition of study of proportional solutions in economies with many goods and kinds of labor was provided by Roemer and Silvestre (1993), and was shown to exist under very general conditions on preferences and production.

Fishing societies typically use an allocation rule 'divide the fish caught by the fishers in proportion to the efficiency units of labor expended,' because this rule is equivalent (randomness aside) to 'each fisher keeps his catch,' a rule that is easy to implement.

Since simple Kantian optimization seems to provide such a clear resolution to the commons

problem of over-fishing, one wonders whether communities whose livelihood depended on fishing discovered it – that is, because their communities would have been better off in the SKE than in the NE, and one might surmise that cultural evolution would select tribes that taught their members to optimize in the Kantian rather than Nash way. We will return to this question below.

Tribes which survived by hunting big game typically, in ancient times, used another allocation rule: equal division. A group of hunters fan out into the bush, and the catch is divided equally. In the symmetric situation where all have the same preferences, the game is defined by:

$$V^{ED}(E^i, E^S) = u(\frac{G(E^S)}{n}, E^i) \ . \quad (5.4)$$

We immediately see this a strictly monotone increasing game, and so it follows by Proposition 3 that the SKE is Pareto efficient in the game. We again should check whether the SKE is efficient in the economy; the first order condition defining the SKE is

$$\frac{d}{dE^i} u(\frac{G(nE^i)}{n}, E^i) = 0, \text{ or}$$

$$u_1 G'(nE^i) + u_2 = 0 \text{ or } -\frac{u_2}{u_1} = G'(E^S),$$

and so:

Proposition 7 *The SKE of the game* V^{ED} *is Pareto efficient in the economy.*

In contrast, the Nash equilibrium of the game V^{ED} is always Pareto inefficient and is Pareto dominated by the SKE, as long as *G* is strictly concave. In NE, each hunter hunts *too little*. It is in his autarchic interest to take a nap under a tree and let the others continue hunting. If everyone does this, production suffers sufficiently to render all hunters worse off than at the SKE.

One can, of course, ask the same evolutionary question with regard to hunting societies: might those that flourished have done so by discovering the Kantian optimization protocol?

There are many other examples of symmetric games derived from economies in which the SKE is Pareto efficient in the economy. Consider the generic public-good game, in which each person has the common utility function $u(E^i, Y) = v(Y) - h(E^i)$ and $Y = G(E^S)$, where *G* and *v* are concave and *h* is convex. The FOC for a SKE is:

$$\frac{d}{dE}(v(G(nE))-h(E))=0=v'(G(nE))G'(nE)n-h'(E)\ ,$$

or
$$\frac{1}{n}h'(E)=v'(G(nE))G'(nE)\ .\ (5.5)$$

An allocation in the public-good economy is Pareto efficient if:

$$(\exists a^i\geq 0)(\sum a_i=1)(a_i h'(E^i)=v'(G(E^S))G'(E^S)\ .\ \ (5.6)$$

Hence the SKE is Pareto efficient: simply let $a^i=1/n$ for every *i*. In like manner, the SKE in a symmetric public-bad economy is Pareto efficient. The Nash equilibria in public-good and public-bad economies are not Pareto efficient – the so-called free-rider problem is just another name for the tragedy of the commons.

6. Elinor Ostrom and the tragedy of the commons

E. Ostrom is justly famous for studying hundreds of communities that face commons problems, and articulating a view about how many of them – the successful ones – solve their commons' tragedies (for instance, Ostrom (1990)). Her general position is that success is achieved through regulation combined with punishments levied against those who break the rules – e.g., they fish more than their entitled time. It is true that, in all of these communities, one observes the existence of punishments and there are deviators who disobey the rules. (Often these deviators are new members to the community.) My conjecture is that Ostrom's explanation – that the good equilibrium is established through the existence of a structure of punishment – is only part of the story, and perhaps a very small part. Perhaps the main explanation of successful resolution of commons' tragedies is that most participants are applying the simple Kantian protocol: they are consciously acting in they way they believe all others should act.

Let us think about how to model Ostrom's proposal as the Nash equilibrium of a game. It is a game with stages. Consider a fishing community where all have the same self-interested preferences (we are still in the symmetric case). In the first stage, individuals choose how long to fish: it has previously been announced what the SKE is; everyone understands, let us suppose, what the Kantian-cooperative action is, which will achieve a Pareto efficient solution. However, some people cheat in the first stage. In the second stage, some fishers among those who cooperated – chosen by some rule – are assigned to punish

those who cheated in stage one.[6] In stage two, if any of the chosen punishers fail to carry out the punishment, *they* are punished in stage three by other punishers. This description fits well with the definition of a social norm: a norm is a prescription concerning how to behave, such that those who fail to behave are punished by others. (See Elster (2009).)

The problem, however, is that punishing others is costly: it reduces the utility of the punisher, because he incurs some danger or cost in carrying out the punishment. One can easily see that the stage game just constructed will not support cooperative behavior as a Nash equilibrium unless it has an infinite number of stages. For suppose the game had only two stages: in stage two, the appointed punishers would not carry out the punishments, since doing so would reduce their utility, and there is no stage three in which someone would punish them for shirking. The same argument holds for any finite number of stages. In the last stage, the appointed punishers would not carry out their punishment of those who failed to punish in the previous stage. The 'good' equilibrium unravels: the only Nash equilibrium of the game is that everybody cheats and nobody punishes.

One might respond: well, fishers are not completely self-interested, they have a sense of fairness, an argument in their utility function that causes them to punish others. Let us leave this objection aside for the moment: the demonstration above shows that fishers *with classical self-interested preferences* cannot, in Nash equilibrium, be induced to play the good solution unless the game has an infinite number of stages.

I contend that such an explanation is Ptolemaic: the infinite-stage game is like an epicycle that must be invoked to explain planetary motion. I find it more parsimonious to propose that most fishers are saying to themselves, 'I'll fish the amount of time that I'd like all others to fish.' Or, perhaps in a more regulated environment, 'I'll fish the assigned time because I understand that if we all fish our assigned time, the result will be better for me than if we were all to optimize autarchically.' The counterfactual the individual constructs in the thought experiment is not the autarchic one (that only he deviates from the cooperative action) but rather that *all* deviate. The constraint is we all choose the same action.

---

[6] In lobster fishing communities in Maine, the first time a lobsterman put out too many nets, other fishers left messages on his buoys. If he cheated again, a committee visited him to warn him. If a cheated a third time, his nets were destroyed.

Nevertheless, the argument is a self-interested one: the individual is not calculating the welfare of others – or at least *he need not.* Cooperation does not presume altruism. It does, perhaps, assume that fishers possess a sense of fairness, which is embodied in the Kantian optimization protocol.

In reality, there are usually some fishers who cheat, and they are indeed punished – even though the real 'game' does not have an infinite number of stages. My explanation for why punishers carry out punishments is that they are optimizing in the Kantian way: they punish the cheaters because they would like all others assigned to punish to do their jobs as well. So the enforcement of social norms – since in reality these games do not have an infinite number of stages – is itself due , I contend, to Kantian optimization.

Let us now consider a second category of explanation for why punishers punish cheaters. It is, many would say, because their preferences are defined over more complex arguments than their own consumption and labor: fairness is an argument of their preferences. A cheater offends one's sense of fairness, and the welfare loss I thereby experience by your cheating can be at least partially attenuated by my punishing you. A more positive version of this argument occurs in other contexts: that people derive a 'warm glow' from taking the cooperative action, or being altruistic. (See Andreoni (1990).)

I do not deny that a sense of fairness is real, and that warm glows exist. What I say is that conceptualizing a sense of fairness or a warm glow as an argument of preferences is unparsimonious. *Why* is my sense of fairness offended by a cheater? I think it is *because the cheater is not doing what we all should do.* He is thereby taking advantage of me. So the offense to my sense of fairness is only descriptive, it is not fundamental: what's fundamental is that the cheater is deviating from the Kantian protocol. Likewise, why do I get a warm glow from taking the cooperative action? The warm glow is not the *explanation*, the *cause* of my behavior, it is an unintended side effect that follows from my doing the fair thing – that is, optimizing according to the Kantian protocol[7].

What I am proposing is that optimizing behavior in games involves specifying two 'parameters': one's preferences and one's optimizing protocol. Classically, in economic

[7] When I help my child with a task, and he succeeds, I enjoy a warm glow. But getting the glow was not the motivation for my helping him: I wanted him to accomplish a task. The warm glow is what Elster (1983) calls a 'state that is essentially a by-product.'

theory, the second parameter is not recognized as one, because it's assumed that a *unique* optimizing protocol characterizes rationality – namely, the autarchic protocol that defines Nash (and Walrasian) equilibrium. The difference between classical and behavioral economists is that the latter *expand the domain of preferences* to include senses of fairness, desires for equality, altruism, and so on. But both classical and behavioral economists maintain the Nash optimizing protocol. In contrast, I propose *not* to expand the domain of preferences – let's be parsimonious and keep preferences classical, at least for now – but alter the optimizing protocol from Nash's to Kant's.

Rationality, in a decision problem, seems quite well-defined. But in a game, when everyone's welfare depends on everyone's actions, it is not so clear what rationality entails. We, schooled in Nash equilibrium, have come to accept the conclusion that 'rational behavior often entails collectively sub-optimal results.' I am uncomfortable with that statement. If cooperation could render everyone better off than in the Nash equilibrium, would it not be rational to cooperate? The response, of course, is that someone (myself?) could be *even better off* if all others cooperate and I cheat. So a self-interested, rational agent should cheat in this situation, as long as this is the last stage in the game, etc. But if evolution has endowed us with an ability to optimize in the Kantian manner (and I am implying this is what the work of Tomasello and others is indicating), is it sensible or useful to call that behavior *irrational?*

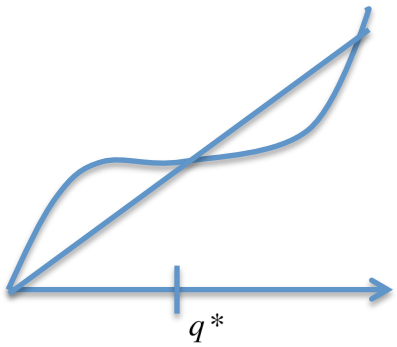
I do not deny that pure Kantian optimizers – those who universally apply the Kantian optimization protocol, independent of the behavior of others – are fairly rare. I believe many people are *conditional* Kantian optimizers: they apply the Kantian protocol in situations where they trust that many others will do so as well. (This will be formalized below.) Thus, there are two necessary conditions for the existence of Kantian behavior: solidarity, in the sense defined in section 1, and trust, which is a belief that others, too, will optimize in the Kantian way. We see many situations in which solidarity (in the sense of unity of interests exists) but trust is lacking, and Kantian optimization does not occur. We also see many situations where both solidarity and trust exist and it does occur.

Consider recycling trash. Recycling entails a small personal cost. The marginal benefit that I produce by recycling, in terms of a clean environment, is trivial. There is, usually, no punishment for failing to recycle. Yet many -- in some countries, most -- people

recycle. This is not a Nash equilibrium but it is a simple Kantian equilibrium. I invoke again the caveat against explaining such behavior by the warm glow recyclers achieve. Solidarity is obvious in this example – many of us have essentially the same preferences over the public good achieved by recycling and the disutility of our own effort -- and trust has been built by observing that, indeed, many others are recycling.

In World War II Britain, a simple Kantian equilibrium was to 'do one's bit,' some extra voluntary contribution to the war effort[8]. Both solidarity and trust existed in this instance. Evidently the sense of solidarity was palpable, as it doubtless is in many such situations. The *sense of solidarity* is very close to trust: for it means that we all understand we are in a situation of common interests, from which it may be a small step to trusting others will reason in the Kantian manner.

Those who employ the simple Kantian protocol even if they do not trust that others will are called *saints*. They hope to start a movement; sometimes they do. It is probably true that virtually all successful examples of achieving Kantian equilibrium involve some saints, who get the ball rolling. Consider this description of the process. There is a community of individuals each of whom is a conditional Kantian optimizer. There are two behaviors to take: the simple Kantian action or the autarchically rational action (Nash). An individual *i* is characterized by the fraction q^i of Kantians he must observe in order for him to take the Kantian action. q^i is *i*'s *threshold*. Let the distribution function of q^i be *Q*: that is fraction $Q(q)$ have a threshold of $q^i \leq q$. Suppose the distribution function *Q* has the graph in figure 1a:



[8] Foyle's War, an excellent BBC series, invokes 'doing one's bit' in almost every episode. The theme of the series is that most people did their bit; the detective Foyle's job was to chase down those who took advantage of the war by doing well for themselves at the expense of others. The series describes a culture of cooperation that many say characterized Britain during the war years, and that has largely faded away, exemplified, among other things, by the growing conservatism of the Labour Party.

Figure 1a

Then we will observe an equilibrium where exactly fraction q^* take the Kantian action. For suppose that exactly $q < q^*$ are cooperating. Then fraction $Q(q) > q$ wish to cooperate, and so the fraction of cooperators increases. Similarly, if $q > q^*$ the fraction of cooperators decreases. The only stable equilibrium is q^*. On the other hand, in figure 1b, the stable equilibria are $q^* \in \{0,1\}$.

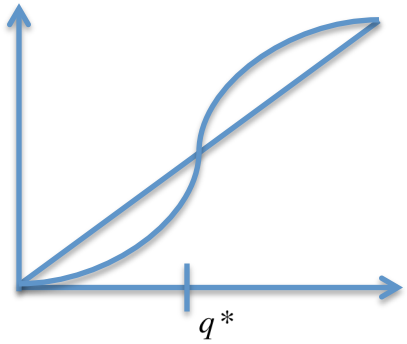


Figure 1b

A saint is an individual whose threshold is zero. If there is a non-trivial number of saints, then $Q(0) > 0$. Unconditional Nash players have a threshold of one. Figure 1c illustrates a situation with both non-trivial fraction of saints and a non-trivial fraction of Nash players. A moment's thought will convince the reader that, when saints exist, there is at least one stable equilibrium with a positive fraction of cooperators, as in figure 1c at q^*.

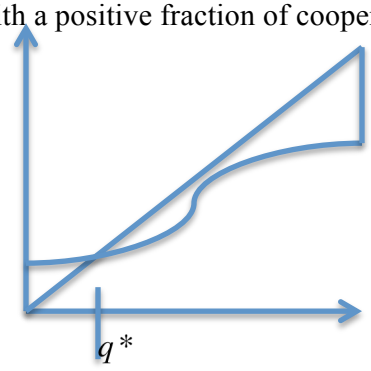


Figure 1c

Absent the existence of saints, we might be in the situation of figure 1b, and universal Nash play would be a stable point, even though everybody is a conditional Kantian optimizer.

7. Non-symmetric economic games

I now turn to non-symmetric games. First, I briefly define two generalizations of simple Kantian equilibrium in a general context. Let a game with *n* players be specified by their payoff functions V^i where the strategy space of all players is a common non-negative interval *I* (which may be infinite) and so the domain of each payoff function is I^n .

Definition 3 A strategy profile $E = (E^1,...,E^n)$ is a *multiplicative Kantian equilibrium* if *no* player would advocate multiplying *all* strategies by *some* common non-negative factor. That is:

$$(\forall i)(\arg\max_{r\geq 0} V^i(rE^1,...,rE^n) = 1)$$

I will call *E* a $K^{\times}$ *equilibrium.*

Here, the Kantian fisher, who is contemplating expanding his fishing time, thinks as follows. "I would like to increase my fishing time by 10%; but I should do so only if I would prefer that all fishers increase their fishing time by 10%." What 'taking the same action' means is now a more complex move than in simple Kantian equilibrium. It follows that an equilibrium, subject to this optimization protocol, is a strategy vector as defined above. Mathematically, the counterfactual the player imagines is not that *only she* deviates (as in Nash) but that deviations are restricted to the ray in $\Re^n_+$ through the current strategy profile.

We have:

Proposition 8 *Let* $\{V^i\}$ *be any monotonic game. Let* $E^* = (E^{1*},...,E^{n*})$ *be a strictly positive multiplicative Kantian equilibrium. Then* E^* *is Pareto efficient in the game.*

Proof: See Roemer (2014).

Let us apply Proposition 8 to a heterogeneous fishing economy, where the preferences of the *n* players are now $u^1,...,u^n$ and the production function is *G*. Fish are allocated in proportion to labor expended. The game is defined , as before, by:

$$V^{\Pr,i}(E^i,E^S) = u^i(\frac{E^i}{E^S}G(E^S),E^i) \ ,$$

which is equation (5.1), altered by indexing each payoff function with *i*. It is clear the game is *monotone decreasing*, as long as *G* is concave. It follows from Proposition 8 that any $K^{\times}$ is Pareto efficient in the game. But as before, this does not mean it is Pareto efficient in the *economy*, which requires (at a positive effort vector) that the marginal rate of substitution between labor and fish is equal, for each fisher, to the marginal rate of transformation.

Applying the definition, we have that E^* is a $K^{\times}$ equilibrium when:

$$\left.\frac{d}{dr}\right|_{r=1} u^i\left(\frac{rE^{*i}}{rE^{*S}}G(rE^{*S}), r\frac{rE^{*i}}{rE^{*S}}\right) = 0 \qquad (7.1)$$

which expands to:

$$u_1^i \cdot \left(\frac{E^{*i}}{E^{*S}}G'(E^{*S})E^{*S}\right) + u_2^i \cdot E^{*i} = 0$$

$$G'(E^{*S}) = -\frac{u_2^i}{u_1^i},$$

where the last step uses the positivity of E^{*i} . Hence:

Proposition 9. *Any positive $K^{\times}$ equilibrium of the proportional game is Pareto efficient in the economy.*

The hunting game, with the equal-division payoff functions V^{ED} , requires a somewhat different approach.

Definition 4 Let $\{V^i\}$ be any game where the players' strategy spaces are a common non-negative interval *I*. An *additive Kantian equilibrium* is a strategy profile E^* such that *no* player would prefer to add to *all* strategies *any* (feasible) number. That is:

$$(\forall i)(\forall r \geq -\min E^{*i})(\arg\max_r V^i(E^{*1}+r,...,E^{*n}+r) = 0) \ .$$

We call such an equilibrium a K^+ *equilibrium.*

We have:

Proposition 10. *Let $\{V^i\}$ be any monotone game. Let $E^* = (E^{1*},...,E^{n*})$ be any additive Kantian equilibrium. Then E^* is Pareto efficient in the game.*

Proof: Roemer (2014).

In the hunting game where players have heterogeneous preferences, the payoff functions are:

$$V^{ED,i}(E^1,E^S) = u^i\left(\frac{G(E^S)}{n}, E^i\right) \ .$$

Clearly, this game is *monotone increasing*, and so it follows that any K^+ equilibrium is Pareto efficient in the game. To see that such an allocation is also Pareto efficient in the economy, apply the definition:

$$\left.\frac{d}{dr}\right|_{r=0} u^i\left(\frac{G(E^S+nr)}{n}, E^i+r\right)=0$$

$$u_1^i \cdot \frac{1}{n} G'(E^S)n + u_2^i = 0$$

$$G'(E^S) = -\frac{u_2^i}{u_1^i}.$$

Proposition 11. *Any K^+ equilibrium of the hunting game is Pareto efficient in the economy.*

The next proposition states that these two equilibrium concepts are in fact *generalizations* of SKE:

Proposition 12. *Suppose $\{V^i\}$ is a symmetric game. Then its simple Kantian equilibria, multiplicative Kantian equilibria, and additive Kantian equilibria coincide.*

Proof: Easy to check.

When discussing simple Kantian equilibrium, I asked the reader to imagine that it is possible for a fisher or a hunter or a potential recycler to ask himself, "What is the action I would like everyone to take?" and to follow the prescription himself. Is it credible that individuals could optimize in the way required by the two more complex protocols introduced in this section, which are necessary to achieve Pareto efficient outcomes in situations of heterogeneous preferences?

With heterogeneity, we lack solidarity – at least in the sense that common preferences imply common interests. It may still be the case that a sense of solidarity exists when preferences are heterogeneous, but it will be more difficult to achieve. Let us think of how heterogeneity may occur in a fishing society. We might suppose that all fishers have the same preferences over fish and leisure, but that they possess different sized boats, capable of trolling the waters at different speeds, so some catch more fish per unit time than others. Suppose their common preferences over fish and labor are represented by $u(x,L)$. In terms of efficiency labor, where $E^i = \alpha^i L^i$ and α^i is speed at which one's canoe travels, we have:

$$u^i(x,E) = u(x, \frac{L}{\alpha^i}),$$

and so preferences become heterogeneous in the arguments that are necessary to define the proportional solution. Nevertheless, these fishers could understand that they have a common interest in not over-exploiting the lake. Invoking the multiplicative Kantian counterfactual – multiplying everyone's labor by a constant as representing the fair alternative – requires the implicit moral supposition that a fisher is *entitled* to the advantage bestowed by owning a faster boat or having more fishing skill. It is not obvious why fishing communities would possess this morality.

Similarly, for the hunting economy to achieve Pareto efficient outcomes when preferences or skills are heterogeneous, given its equal-division sharing rule, hunters must learn to optimize in the additive way. "I should only take a two-hour nap under that tree if I would advocate that we all take a two-hour nap.[9]" It does not seem to me there is any ethically interesting relationship between the nature of the allocation rule (proportional versus equal division) and the conception of fairness (multiplicative versus additive) that must be applied in the Kantian counterfactual needed to achieve successful cooperation with that rule. My view is that *if* fishing and hunting societies indeed did learn to use these optimization protocols, thereby solving their commons' tragedies, this happened through random cultural evolution. (See Boyd and Richerson (1985).) A clever priest (Archimedes?) in some fishing tribe deduced the merits of multiplicative Kantian optimization, he taught it to the tribe, and they thrived. Another clever priest (Pythagoras?) in a hunting tribe figured out the beauty of additive Kantian optimization. Knowing the virtually infinite complexity that biological evolution has created in living things, is it such a stretch to suppose that our ancestors could have discovered these very useful *ways of thinking* through selective and cultural adaption? (I need not here rehearse the controversy over group selection.)

Let us pause momentarily to reflect, once more, on the distinction between altruism and cooperation in the application of these more complex Kantian protocols to common-pool resource games. In the fishers' game, the Kantian optimizer rejects expanding her fishing time by 5%, at an allocation, if she would not like everyone to expand his fishing time by 5%.

---

[9] To be precise, if hunters have different efficiencies in hunting, because the additive Kantian protocol requires each to imagine changing all hunting times by a constant, more efficient hunters would receive shorter naps in the counterfactual.

Why might she not like that alternative? Because the *negative externality* that *others* would impose upon *her* by the 5% increase in total fishing time, due to the decreasing-returns nature of the technology, makes such an expansion not worthwhile *to her*. I emphasize that the Kantian counterfactual does not induce her to worry about the negative externality *she* imposes upon *others*[10]. Her decision not to expand her fishing time, in this situation, is motivated by considerations of her own welfare, not the welfare of others. There is, it is true, a conception of fairness embodied in her use of the Kantian optimization protocol, and to the extent that fairness implies a concern for the welfare of others, then she does care about others. But fairness need not be so motivated: it can be motivated, instead, by the symmetry of the problem. Indeed, the human brain loves symmetry; and the evolutionary path through which we learned to optimize in the Kantian manner may have been through a prior partiality to symmetry that had evolved in us.

It is worth pointing out that the information required to calculate the Kantian best-response at a particular profile of actions is the same as is required for an autarchic optimizer. One needs to know the current strategy profile, one's own preferences, and, in the economic examples, the production function. I showed earlier (Roemer (2014)) that the simple dynamics of iterative best responses converges (in well-behaved cases) to the Kantian equilibrium. At a strategy profile $(E^1,...,E^n)$, each player calculates the optimal multiplicative constant r^i (in the case of the fishing economy) he would like to apply to the whole vector, and he plays as his next action r^iE^i; thus, at the next stage the strategy profile is $(r^1E^1,...,r^nE^n)$ because, out of equilibrium, different players will have different optimal re-scalings, r^i. This dynamic procedure converges to the multiplicative Kantian equilibrium (as I said, with well-behaved preferences). Thus, from a mathematical viewpoint, Kantian equilibrium seems very similar to Nash equilibrium, where best-response dynamics converge to the Nash equilibrium in well-behaved cases: if we believe people have learned to think in the Nash manner, we should extend that belief to their having the cognitive capacity to think Kantian.

To say that Nash or Kantian equilibrium characterizes a stable point for the fishers does not, of course, require them to be mathematicians (although perhaps the visionary priest

[10] I owe this observation to an anonymous referee of my article Roemer (2014).

was). To calculate whether or not to change one's behavior, a fisher using the Nash protocol has to compare his marginal rate of substitution to the *average yield* of the lake. That is a fisher decides to increase his fishing time if:

$$\frac{d}{dE^i}u(\frac{E^i}{E^S}G(E^S),E^i) \approx (\text{for large } n)\, u_1 \frac{G(E^S)}{E^S} + u_2 > 0$$

$$\Leftrightarrow MRS < \frac{G(E^S)}{E^S}.$$

On the other hand, the rule for the multiplicative Kantian optimizer is to increase his fishing time as long as his marginal rate of substitution is less than the *marginal* productivity of the lake. (This is why the latter protocol gives a stable point that is Pareto efficient.) It does not seem that one rule of thumb is more complex than the other.

For hunting communities, Nash equilibrium *if n is large* leads to a total failure of hunting:

$$\frac{d}{dE^i}u(\frac{G(E^S)}{n},E^i) \approx \ (\text{for large } n)\, u_2 < 0 \Rightarrow \text{always take a nap!}\ .$$

If the hunting band is small then the rule of thumb is:

$$\frac{d}{dE^i}u(\frac{G(E^S)}{n},E^i) = u_1 \frac{G'(E^S)}{n} + u_2 > 0 \Leftrightarrow MRS < \frac{G'(E^S)}{n}\ .$$

The rule of thumb for an additive Kantian optimizer is to increase his hunting time if and only if $MRS < G'(E^S)$. Mathematically, these two rules of thumb are of similar complexity.

Consider the public-good game with heterogeneous preferences: $u^i(Y,E^i)$ is the utility function of player *i,* where Y is the value of the public good and E^i is the contribution of agent *i*, with $Y = G(E^S)$. The game is defined by :

$$V^i(E^i,E^S) = u^i(G(E^S),E^i)\ ,$$

and the $K^\times$ equilibrium is defined by:

$$\left.\frac{d}{dr}\right|_{r=1} u^i(G(rE^S),rE^i) = u_1^i \cdot G'(E^S)E^S + u_2^i E^i = 0 \Leftrightarrow MRS^i \cdot \frac{E^i}{E^S} = MRT\ .$$

Thus, the $K^\times$ equilibrium of the public-good game is characterized by a condition that is a special case of the Samuelson condition for Pareto efficiency in a public-good economy (see (5.6)). On the other hand, the K^+ equilibrium of the public-good economy is characterized by:

$$\left.\frac{d}{dr}\right|_{r=0} u^i(G(E^S+nr),E^i+r) = 0 = u_1^i G'(E^S)n + u_2^i \Leftrightarrow MRS^i \frac{1}{n} = MRT\ ,$$

which is, again, a special case of the Samuelson condition. So both additive and multiplicative Kantian optimization protocols deliver Pareto efficiency in the public-good economy, although they lead to different allocations (except in the case of symmetry).

Walker and Ostrom (2009) report on experiments that they performed using a common-pool resource game, defined as follows. Each of 8 players receives an endowment of M tokens. There are two possible investments: a common-pool resource (CPR), for which total monetary returns will equal $F(x^S) = (.01)(23x^S - \frac{1}{4}(x^S)^2)$, where $x^S = \sum x^i$ is the sum of investments in this resource, and a 'treasury bill' which yields a return of $.05x$ to an investor who invests x. The monetary returns from the CPR are divided in proportion to investments. Thus, the pay off function of individual i is:

$$\frac{x^i}{x^S}F(x^S) + .05(M - x^i) \ .$$

Because the CPR exhibits diminishing marginal returns, this is a CPR problem, where the Nash equilibrium is Pareto inefficient. If M is sufficiently large, then we can ignore M in the objective function, and simply view the term $-.05x^i$ as a disutility from investment. Thus, this is simply a fishing economy where an agent's utility equals consumption minus a linear term in effort. To compute the SKE, the individual maximizes:

$$\frac{1}{n}F(nx) + .5(M - x) \ .$$

The solution is $x^S = 36$, and so with $n = 8$, each individual should invest 4.5 tokens – it is unfortunate that this investment was not feasible in the Walker-Ostrom game, since investments had to be integers. In the symmetric Nash equilibrium, $x = 8$, which is feasible as long as $M \geq 8$. Many variations on this game were performed, and it is beyond my scope to discuss them here. What interests me is that the authors focus on cooperation which they conceptualize as the maximization of total income. Granted, this is achieved in the SKE. However, it is not achieved because each individual is *thinking* to maximize total income, in the Kantian proposal, but rather because she is thinking about doing what she'd like everyone to do. I believe this distinction is important, because the motivations in the two explanations are different. Walker and Ostrom are implicitly saying that the reason cooperation often fails is that participants are not thinking of the collective good. (See my earlier quotation from their article in section 5.) This would lead to the prescription that a

social engineer try to induce cooperation in a group by teaching them to think of the collective good. Far less demanding, I say, and more natural for *homo sapiens,* is to think in terms of behaving fairly, à la Kant.

An interesting field experiment is conducted by Bandiera et al (2006). The authors describe the experiment, as follows, which was implemented on a British farm. There are approximately 40 field workers on a given day who will work in a given field. The farm has a fixed wage bill, *w*, for the day, for this field. It allocates the wage bill to workers in proportion to the amount of fruit they pick. The foreman also announces each day what he thinks the productivity of the workers will be. To model this situation, let's assume a worker's utility equals his income minus a disutility that increases with the speed at which he picks. Then his payoff is:

$$w\frac{x^i}{x^S} - h(x^i)$$

where x^i is the amount he picks and $x^S = \sum x^i$, and we assume *h* is a convex function. The SKE of this game is: $x^i = 0$ for all *i,* and the wage bill is divided equally among the workers. Clearly, the game has not been described accurately, as the owner would not pay the workers if they do not produce. Presumably the foreman's announcement of expected productivity acts as a guide. In a second field, owned by the same farmer, workers were paid a fixed amount per kilo picked. The authors state their results:

> First, individuals cooperate more, namely their productivity is significantly lower, as their exposure to the relative incentive scheme increases [i.e., in the first field]. This effect is significantly larger for the cohort of early worker arrivals, namely individuals who started working at the beginning of the peak season when the scheme was first introduced. Second, individuals cooperate more when they work with co-workers who have been exposed to the scheme for longer and hence are more familiar with the norm.

It is interesting that the authors write in the introduction of their paper:

> It is important to stress that cooperation can arise either because of altruism or collusion; workers might cooperate either because they truly care about colleagues' payoffs, or because they have established an implicit collusive agreement enforced by credible

> threats of punishment. In this paper we focus on how cooperation evolves with time, regardless of its underlying motives.

Here, the authors are saying that *either* altruism *or* Nash equilibrium with punishments explains what is occurring in the first field: yet they give no evidence of either altruism or the existence of punishments or ostracism. (They do explicitly refer to the Nash equilibrium of the game, in which neither altruism nor punishments exist, and claim that the productivity of workers in the first field is quite far below that value. They do not, however, explicitly write down a utility function for workers.) It seems to me more likely that workers have learned Kantian thinking, and are attempting to play the Kantian equilibrium.

In private-goods economies of the form $(u^1,..,u^n,G)$ we saw that each allocation rule (proportional or equal-division) was associated with a different Kantian optimization protocol. It turns out that, in these economies, one can construct a continuum of allocation rules, running from 'proportional division' at one pole to 'equal division' at the other, each one with an associated 'Kantian optimization protocol,' with the property that using the right protocol with a given allocation rule always produces Pareto efficient outcomes as equilibria. (See Roemer (2014).) I do not review this material here, because it is more technical, and because it seems to me that discovering these new protocols would be more unlikely than discovering either the simple additive or multiplicative ones. One cannot, however, find a way of implementing *any* allocation rule efficiently with some Kantian protocol: the allocation rules which can be so implemented span, as I said, a one-parameter family 'between' the proportional and equal-division rules. (It is only a little misleading to say that these rules are 'convex combinations' of proportional and equal division.) Because division in proportion to effort expended and equal division are the two ubiquitous conceptions of fairness that arise in a multitude of contexts, there is at least poetry in the fact a family with these two rules as its extremes comprises those that can be efficiently implemented by Kantian optimization.

8. Can Kantian equilibrium be rationalized as Nash equilibrium?

As I've argued, the distinguishing move that behavioral economists make is to include non-traditional arguments in the preferences of individuals: these may endow people

with altruism, or a sense of fairness, or a sense of dignity or duty. One might protest that the view that optimizers are using a Kantian protocol cannot be distinguished, empirically or in a logical sense, from the view that people's preferences contain these exotic arguments. In this section, I address the question whether Kantian equilibrium can be distinguished from a logical viewpoint from Nash equilibrium in which players are endowed with exotic preferences.

Let me formulate this question in a precise way. Let us consider the set of two-person economies $(u^1,u^2,G,X^{\mathrm{Pr}})$ where u^1 and u^2 are any concave self-interested utility functions over consumption and efficiency units of labor, G is any concave production function transforming total efficiency units of labor into the consumption good, and X^{Pr} is the proportional allocation rule, in which output is divided in proportion to efficiency units of labor expended. Denote the domain of such economies by Ω. We know that multiplicative Kantian equilibrium chooses exactly the Pareto efficient, proportional allocations on this set of economies. (If an economy possesses more than one such allocation, they are all multiplicative Kantian equilibrium.) Let us ask: Is it possible to define a rule for transforming the utility functions u^1 and u^2 in any such economy into utility functions with more arguments v^1,v^2 such that the *Nash* equilibrium (or equilibria) on the extended environment $(v^1,v^2,G,X^{\mathrm{Pr}})$ is (or are) the Kantian equilibria of $(u^1,u^2,G,X^{\mathrm{Pr}})$? Were this possible, one could claim that Kantian equilibria could always be rationalized as Nash equilibria where preferences are more complex than one might have thought.

Let me give an example of economies for which this can be done: they are the quasi-linear economies. Let $u^1(x,E)=x-h^1(E)$ and $u^2(x,E)=x-h^2(E)$, where h^i are convex functions. The Pareto efficient allocations in any economy with these preferences are those that maximize $G(E^1+E^2)-h^1(E^1)-h^2(E^2)$ -- that is, for which

$$G'(E^S)=h^{1\prime}(E^1)=h^{2\prime}(E^2)\ . \qquad (8.1)$$

Hence the multiplicative Kantian equilibria of such an economy are characterized by (8.1) and (8.2):

$$x^1=\frac{E^1}{E^S}G(E^S),\quad x^2=\frac{E^2}{E^S}G(E^S)\ . \qquad (8.2)$$

Now consider the 'extended' economy where we define the new preferences:

$$v^1(x^1,x^2,E^1,E^2)=u^1(x^1,E^1)+u^2(x^2,E^2)=v^2(x^1,x^2,E^1,E^2)\ . \qquad (8.3)$$

Let us calculate the Nash equilibria of the game defined by $(v^1, v^2, G, X^{\Pr})$. The payoff functions for the two players are both:

$$V(E^1, E^2) = \frac{E^1}{E^S} G(E^S) + \frac{E^2}{E^S} G(E^S) - h^1(E^1) - h^2(E^2). \qquad (8.3)$$

Because V is a concave function, the Nash equilibrium is characterized by the two first-order conditions, which we now calculate:

$$\begin{aligned} \frac{\partial}{\partial E^1} V(E^1, E^2) &= G'(E^S) - h^{1\prime}(E^1) = 0 \\ \frac{\partial}{\partial E^2} V(E^1, E^2) &= G'(E^S) - h^{2\prime}(E^2) = 0, \end{aligned} \qquad (8.4)$$

which is the same as (8.1). Because, in addition, the allocation must be proportional, we conclude that the *Nash* equilibria of this game comprise precisely the multiplicative Kantian equilibria of the original game with classical preferences. Therefore, if preferences are quasi-linear, we cannot distinguish, simply from observing outcomes, whether the participants are optimizing using the Kantian protocol with classical, self-interested preferences, or the Nash protocol with the extended *altruistic* preferences, in which each player is maximizing the sum of utilities of all players.

If such a transformation of every game induced by the economies in Ω could be constructed, we would have to say that Kantian optimization and Nash optimization are observationally equivalent. What the next proposition states is that such is not the case.

Consider a pair of transformations of utility functions $V^i(u^1, u^2) : \Re_+^4 \to \Re$, $i = 1,2$, where each 'extended' utility function V^i is defined on the argument (x^1, x^2, E^1, E^2). We can then speak of a game with extended preferences $(V^1, V^2, G, X^{\Pr})$ where the payoff function of player *i* is:

$$\mathbf{V}^i(E^1, E^2) = V^i(u^1(\frac{E^1}{E^S} G(E^S), E^1), u^2(\frac{E^2}{E^S} G(E^S), E^2)) \ . \quad (8.5)$$

Proposition 13 *There do not exist transformations V^1, V^2 on ordered pairs of concave utility functions such that, on Ω , the Nash equilibria of the extended games $(V^1(u^1, u^2), V^2(u^1, u^2), G, X^{\Pr})$ coincide with the multiplicative Kantian equilibria of the classical games $(u^1, u^2, G, X^{\Pr})$.*

Proof:

1. Assume, to the contrary, that such transformations V^1 and V^2 do exist, with payoff functions for the associcated games as defined in (8.5). . The strategies for the players continue to be their effort levels. The FOCs for a Nash equilibrium of the extended games are:

$$\frac{\partial \mathbf{V}^1(E^1,E^2)}{\partial E^1} = \frac{\partial \mathbf{V}^2(E^1,E^2)}{\partial E^2} = 0 \ .$$

Writing out the derivative[11] for $\mathbf{V}^1$, we have:

$$\left(V_1^1 u_1^1 \frac{E^1}{E^S} + V_2^1 u_1^2 \frac{E^2}{E^S}\right) G'(E^S) + \frac{G(E^S)}{E^S}\frac{E^2}{E^S}(V_1^1 u_1^1 - V_2^1 u_1^2) + V_1^1 u_2^1 = 0 \ . \quad (8.6)$$

Dividing by u_1^1 and using the fact that $-G'(E^S) = \frac{u_2^1}{u_1^1}$, since by hypothesis the Nash equilibrium is the multiplicative Kantian equilibrium, and is hence Pareto efficient, we have:

$$\left(V_1^1 \frac{E^1}{E^S} + V_2^1 \frac{u_1^2}{u_1^1}\frac{E^2}{E^S}\right) G'(E^S) + \frac{G(E^S)}{E^S}\frac{E^2}{E^S}(V_1^1 - V_2^1 \frac{u_1^2}{u_1^1}) - V_1^1 G'(E^S) = 0$$

$$G'(E^S)\left(V_1^1(\frac{E^1}{E^S} - 1) + V_2^1 \frac{u_1^2}{u_1^1}\frac{E^2}{E^S}\right) + \frac{G(E^S)E^2}{(E^S)^2}\left(V_1^1 - V_2^1 \frac{u_1^2}{u_1^1}\right) = 0$$

$$G'(E^S)\frac{E^2}{E^S}(-V_1^1 + V_2^1 \frac{u_1^2}{u_1^1}) + \frac{G(E^S)E^2}{(E^S)^2}\left(V_1^1 - V_2^1 \frac{u_1^2}{u_1^1}\right) = 0 \quad (8.7)$$

from which it follows that either $V_1^1 = V_2^1 \frac{u_1^2}{u_1^1}$ or $G' = \frac{G(E^S)}{E^S}$. But the second possibility is false for any strictly concave *G*. Therefore we must have:

$$V_1^1 = V_2^1 \frac{u_1^2}{u_1^1} \quad (8.8)$$

at the proportional solution on the whole domain Ω . In like manner, we have:

$$V_2^2 = V_1^2 \frac{u_1^1}{u_1^2} \quad (8.9)$$

on the whole domain. Therefore, $\frac{V_2^2}{V_1^2} = \frac{V_2^1}{V_1^1}$ at all proportional solutions on the whole domain.

2. Now for any two numbers *a,b*, we can generate a proportional solution, by appropriate choice of $\{u^1, u^2, G\}$, so that the utilities of the two players are *a* and *b* at the

[11] V_j^1 is the derivative of V^1 with respect to the utility of the *j*th player.

solution. Thus, $\frac{V_2^2}{V_1^2} = \frac{V_2^1}{V_1^1}$ is an *identity* on $\Re^2$. It therefore follows that the marginal rate of substitution of V^1, when viewed as a function on $\Re^2$, is identical to the marginal rate of substitution of V^2 on the whole plane – that is, V^1 and V^2 possess the same map of indifference curves. Therefore V^1 is just an ordinal transform of V^2, so without loss of generality, we may take $V^1 = V^2$, since only the ordinal properties of V^1 and V^2 matter – both Nash and Kantian equilibrium are ordinal concepts[12].

3. Now choose a pair of utility functions (u^1,u^2) and two points $((x^1,E^1),(\hat{x}^1,\hat{E}^1))$ such that such that the following hold:

* u^1 and u^2 possess a pair of identical indifference curves, I^1 and I^2,
* $(x^1,E^1),(\hat{x}^1,\hat{E}^1) \in I^1$ and therefore $(x^1,E^1),(\hat{x}^1,\hat{E}^1) \in I^2$
* $\frac{u_1^2(\hat{x}^1,\hat{E}^1)}{u_1^2(x^1,E^1)} \neq \frac{u_1^1(\hat{x}^1,\hat{E}^1)}{u_1^1(x^1,E^1)}$.

This can surely be done: notice the third condition is one on the ratios of first derivatives of the two points for the two different utility functions, and although the marginal rates of substitution are identical at these two points for the two utility functions this says nothing about the ratios of the first derivatives.

Since $MRS_{u^1}(x^1,E^1) = MRS_{u^2}(x^1,E^1)$, we can choose *G* such that $2x^1 = G(2E^1)$ and $G'(2E^1) = MRS(x^1,E^1)$; it follows that $((x^1,E^1),(x^1,E^1))$ is a proportional solution for the economy (u^1,u^2,G), and so it must be case by the above that:

$$V_1^1 = V_2^1 \frac{u_1^2}{u_1^1} \qquad (8.10)$$

where the function V^1 is evaluated at (a,b) where $a = u^1(x^1,E^1)$ and $b = u^2(x^1,E^1)$.

4. But by the same argument,

$$V_1^1 = V_2^1 \frac{u_1^2}{u_1^1} \qquad (8.11)$$

[12] Note that, in our example with quasi-linear preferences, indeed $V^1 = V^2 = u^1 + u^2$.

when the functions are evaluated at $(\hat{x}^1,\hat{E}^1),(\hat{x}^1,\hat{E}^1)$, since we can produce a production function $\hat{G}$ such that $(\hat{x}^1,\hat{E}^1),(\hat{x}^1,\hat{E}^1)$ is a proportional solution of the economy $(u^1,u^2,\hat{G})$. Notice the utilities *a,b* are the same at this point as they are at $((x^1,E^1),(x^1,E^1))$, and so it therefore follows from (8.10) and (8.11) that:

$$\frac{u_1^2(\hat{x}^1,\hat{E}^1)}{u_1^1(\hat{x}^1,\hat{E}^1)}=\frac{u_1^2(x^1,E^1)}{u_1^1(x^1,E^1)} \ .$$

But this contradicts the choice of the two points , because $\frac{u_1^2(\hat{x}^1,\hat{E}^1)}{u_1^1(\hat{x}^1,\hat{E}^1)}\neq\frac{u_1^2(x^1,E^1)}{u_1^1(x^1,E^1)}$, which proves the claim: that is, there are no functions V^1,V^2 such that, on the whole domain Ω, the Nash equilibria of the game induced by V^1,V^2 are the Kantian equilibria of the original game. ■

Proposition 13 tells us that on the entire domain of economies Ω there is no way of transforming preferences by transformations $V^1(u^1,u^2),V^2(u^1,u^2)$ such that the Nash equilibria of these 'extended' games are the Kantian equilibria of the original games. We have, however, shown that if the preferences of both players are quasi-linear, it is possible to find such extended preferences: namely $V^1(u^1,u^2)=V^2(u^1,u^2)=u^1+u^2$. One can ask: Are there more examples like this? I do not have the complete answer; but the next proposition shows this cannot be done if both preferences are Cobb-Douglas. My conjecture is that the only case in which such extensions can be made is the quasi-linear one.

Define the domain $\Omega^{(u^1,u^2)}$ as the set of economies where the two agents' preferences are *fixed* and represented by utility functions u^1 and u^2 , and *G* can be any concave production function. Thus the economies are of the form $(u^1,u^2,G,X^{\mathrm{Pr}})$ where the proportional rule and the preferences are fixed and *G* varies. Note that we may have different utility functions to represent the same preferences. But Nash and Kantian equilibrium use only (ordinal) preferences. Thus, for example, the domains $\Omega^{(u^1,u^2)}$ and $\Omega^{(qu^1,ru^2)}$ where $q,r>0$ are identical.

Proposition 13a *Let* u^1,u^2 *represent different Cobb-Douglas preferences over consumption and effort. On the domain* $\Omega^{(u^1,u^2)}$ *there are no extended preferences* $V^1(u^1,u^2),V^2(u^1,u^2)$

such that for all G, the Nash equilibria of the game $(V^1(u^1,u^2),V^2(u^1,u^2),G,X^{\Pr})$ *coincide with the multiplicative Kantian equilibria of the game* $(u^1,u^2,G,X^{\Pr})$.

Proof:

1. Let $u^1(x,E)=x^{\alpha}(1-E)^{1-\alpha},u^1(x,E)=x^{\beta}(1-E)^{1-\beta}$, $\alpha\neq\beta$. The condition for an allocation (x^1,E^1,x^2,E^2) to be an interior multiplicative Kantian equilibrium on the domain $\Omega^{(u^1,u^2)}$ is precisely the conjunction of:

$$\frac{1-\alpha}{\alpha}\frac{x^1}{1-E^1}=\frac{1-\beta}{\beta}\frac{x^2}{1-E^2} \qquad (8.12a)$$

$$\frac{x^1}{E^1}=\frac{x^2}{E^2} \qquad (8.12b) \ .$$

$$\frac{1-\alpha}{\alpha}\frac{E^1}{1-E^1}\leq 1 \qquad (8.12c)$$

To see this, note that (8.12a) states that the marginal rates of substitution of the two agents are equal, (8.12b) states that allocation is proportional, and (8.12c) states that $\frac{1-\alpha}{\alpha}\frac{x^1}{1-E^1}<\frac{x^1}{E^1}$. The last condition implies that there exists a concave *G* such that $G'(E^S)=MRS$ and the average product, $G(E^S)/E^S=\frac{x^1+x^2}{E^1+E^2}$ is greater than or equal to the marginal product (which equals the common marginal rate of substitution), which is exactly the condition for there existing a concave production function that renders this allocation Pareto efficient.

2. Let (x^1,E^1,x^2,E^2) and $(\hat{x}^1,\hat{E}^1,\hat{x}^2,\hat{E}^2)$ be two interior allocations satisfying (8.12abc) and such that $E^1\neq\hat{E}^1$, so they are two different proportional solutions for the domain $\Omega^{(u^1,u^2)}$. Define Q^1,Q^2 so that:

$$Q^iu^i(\hat{x}^i,\hat{E}^i)=u^i(x^i,E^i)\equiv a^i \ . \qquad (8.13)$$

3. Suppose, to the contrary of what we aim to prove, there do exist transformations V^1,V^2 for which the Nash equilibria on economies $(V^1(u^1,u^2),V^2(u^1,u^2),G,X^{\Pr})$ are the Kantian equilibria on the domain $\Omega^{(u^1,u^2)}$. The first step of the proof of Proposition 13 continues to apply here, so we must have, from (8.8):

$$\frac{V_1^1(a^1,a^2)}{V_2^1(a^1,a^2)}=\frac{u_1^2(x^2,E^2)}{u_1^1(x^1,E^1)} \text{ and } \frac{V_1^1(a^1,a^2)}{V_2^1(a^1,a^2)}=\frac{Q^2u_1^2(\hat{x}^2,\hat{E}^2)}{Q^1u_1^1(\hat{x}^1,\hat{E}^1)}$$

and so: $$\frac{u_1^2(x^2,E^2)}{u_1^1(x^1,E^1)} = \frac{Q^2 u_1^2(\hat{x}^2,\hat{E}^2)}{Q^1 u_1^1(\hat{x}^1,\hat{E}^1)}. \qquad (8.14)$$ [13]

But (8.14) says that $\frac{\beta a^2}{x^2} / \frac{\alpha a^1}{x^1} = \frac{\beta a^2}{\hat{x}^2} / \frac{\alpha a^1}{\hat{x}^1}$ or

$$\frac{x^2}{\hat{x}^2} = \frac{x^1}{\hat{x}^1} \ . \qquad (8.15)$$

However, it is false that if two allocations satisfy (8.12abc) they necessarily satisfy (8.15).

For (8.12ab) imply that:

$$\frac{1-\alpha}{\alpha}\frac{E^1}{1-E^1} = \frac{1-\beta}{\alpha}\frac{E^2}{1-E^2} \text{ and } \frac{1-\alpha}{\alpha}\frac{\hat{E}^1}{1-\hat{E}^1} = \frac{1-\beta}{\alpha}\frac{\hat{E}^2}{1-\hat{E}^2}$$

which together give us:

$$\left(\frac{E^1}{\hat{E}^1}\right)\left(\frac{1-\hat{E}^1}{1-E^1}\right) = \left(\frac{E^2}{\hat{E}^2}\right)\left(\frac{1-\hat{E}^2}{1-E^2}\right), \qquad (8.16)$$

while (8.15) and (8.12b) give us $\frac{E^1}{\hat{E}^1} = \frac{E^2}{\hat{E}^2}$. Together with (8.16) this implies $E^1 = \hat{E}^1$, the contradiction that establishes the proposition. ■

It is key that preferences be different in Proposition 13a.

Proposition 13b *Consider the domain of production economies* (u,u,G,X^{Pr}) *for any concave u. Let the extended preferences for each player be* $V(u(x^1,E^1),u(x^2,E^2)) = u(x^1,E^1)+u(x^2,E^2)$. *The symmetric Nash equilibrium of the extended game* $(V(u,u),V(u,u),G,X^{\text{Pr}})$ *is the multiplicative Kantian equilibrium of the game* (u,u,G,X^{Pr}).

Proof:

1. The payoff function of the first player in the extended game is:

$\mathbf{V}(E^1,E^2) = u(\frac{E^1}{E^S}G(E^S),E^1)+u(\frac{E^2}{E^S}G(E^S),E^2)$. The FOC for Nash equilibrium is:

$$\frac{\partial \mathbf{V}}{\partial E^1} = u_1(x^1,E^1)\cdot\left(\frac{E^1}{E^S}G'(E^S)+G(E^S)\frac{E^2}{(E^S)^2}\right)+u_2(x^1,E^1)+u_1(x^2,E^2)\cdot\left(E^2\frac{E^S G'(E^S)-G(E^S)}{(E^S)^2}\right)=0$$

[13] This step requires us to understand that the functions V^i be applied to whatever representations are chosen to represent the given preferences.

Notice at a symmetric allocation (i.e., where $E^1 = E^2$ and $x^1 = x^2$) the terms containing G annihilate each other, and we are left with:

$$u_1 G'(E^S) + u_2 = 0$$

at the allocation, which is the condition that the MRS for player 1 equals the MRT. The same analysis holds for player 2. Hence the allocation is Pareto efficient, and hence it is the multiplicative Kantian equilibrium. ■

Notice that Proposition 13b requires that Nash players find the symmetric Nash equilibrium, and that they have the same preferences. The same 'extended preference' is used as in the example above with quasi-linear preferences (namely, taking the sum of the utility functions), but the quasi-linear result is stronger, as it does not require the players have the same preferences.

I conclude (from Propositions 13 and 13a) that Kantian optimization and Nash optimization with extended preferences are distinct protocols. These propositions suggest that one might be able to distinguish experimentally between the behavioral-economics explanation of cooperative behavior (which is that players are playing the Nash equilibrium with extended preferences) and that players are playing Kantian equilibrium with classical preferences. However, Proposition 13b tells us we cannot expect to discover this distinction if experiments endow players with the same preferences.

[There is more to be said on this topic, which will appear in a subsequent version.]

9. Production and taxation

Paying taxes is, in many countries, sustained by Kantian thinking. Many studies have established that the fines and punishments, combined with the probability of being detected to have cheated if one does, are insufficient to explain the degree of tax compliance. (citations) One can again argue whether compliance is explained by Kantian thinking ('I pay my taxes because it's the action I'd like everyone to take') or by exotic preferences ('I feel like a lousy citizen if I cheat , or I have a duty to pay'). Once more, my view is that, indeed, a person may feel like a lousy citizen, or feel she is abrogating her duty, should she cheat, but that such feelings are not the *explanation* of compliance – they are the byproducts of feeling one is shirking if one fails to take the action one believes all should take. Another explanation for why people may cheat is that they have little faith that the state will make

good use of tax revenues. It is difficult, however, to evaluate this justification, as it may be one created out of cognitive dissonance to make it easier to justify to oneself one's cheating.

Even if one agrees that the decision to pay taxes honestly is a result of Kantian thinking, this is insufficient to render taxation Pareto efficient; this is because individuals do not (by and large) extend their Kantian thinking to their labor-supply decisions. Consider the following simple production economy with taxes. Individuals have preferences over income and labor in efficiency units represented by utility functions $u^i(x,E)$. Production is linear: total output is $G(E^S) = a\sum E^i$, some $a > 0$. There is a linear income tax given by (t,q) where *t* is the tax rate and q is the lumpsum demogrant returned to citizens. The firm is competitive and pays a wage of a per unit labor to worker *i*. Thus the worker chooses his labor supply to maximize $u^i((1-t)aE+q,E)$ according to the FOC:

$$u_1^i \cdot (1-t)a + u_2^i = 0 \quad , \qquad (9.1)$$

where I have assumed, as is usual, that the number of workers is so large that an individual can rationally ignore the effect of his labor on q. By virtue of (9.1), the marginal rate of substitution for worker *i* is equal to $(1-t)a$, while efficiency requires that it be equal to the marginal rate of transformation, which is a. The equilibrium so generated is indeed a Nash equilibrium of the game where workers' strategies are their labor supplies. Note, the larger the tax rate is, the larger is the 'wedge' between the marginal rates of substitution and the marginal rates of transformation.

Now let us examine the additive Kantian equilibrium of this game, using the same preferences. This is a profile of labor supplies $(E^1,E^2,...)$ such that nobody would advocate that everybody change her labor supply by any additive constant. Under this protocol, workers cannot rationally ignore the effect of the deviation on the demogrant *q*. The FOC defining such an equilibrium is:

$$(\forall i) \quad \frac{d}{dr}u^i((1-t)a(r+E^i)+\frac{ta}{n}\sum(E^j+r),E^i+r)=0$$

$$u_1^i \cdot ((1-t)a+\frac{ta}{n}n)+u_2^i=0$$

$$a=-\frac{u_2^i}{u_1^i}$$

and so the additive Kantian equilibrium is Pareto efficient. Achieving efficiency requires that people extend their Kantian thinking from the decision to comply with taxation to the labor-supply decision. Were people to think in this way, then, at least with linear production functions, the decision concerning redistribution could be entirely separated from the question of efficiency, for efficiency occurs for any tax rate *t.* Thus is resolved the well-known equity-efficiency trade-off.

This result does not extend to non-linear production. Let's examine additive Kantian equilibrium with taxation when the production function is a strictly concave function, *G*. With concave production, profits are non-zero and must be allocated: I will assume equal division of profits for the example. The competitive wage for an efficiency unit of labor is $w = G'(E^S)$, the marginal product. The FOC characterizing K^+ equilibrium is:

$$\frac{d}{dr}u^i((1-t)w(E^i+r)+\frac{t}{n}G(E^S+nr)+\frac{G(E^S+nr)-w(E^S+r)}{n},E^i+r)=0. \quad (9.2)$$

If workers are contemplating changing everyone's labor supply, they must understand that the wage will change as well, so we must substitute $G'(E^S+nr)$ for the wage, giving us:

$$\frac{d}{dr}u^i((1-t)G'(E^S+nr)(E^i+r)+\frac{t}{n}G(E^S+nr)+$$
$$(1-t)\frac{G(E^S+nr)-G'(E^S+nr)(E^S+nr)}{n},E^i+r)=0$$

Expanding this FOC gives:

$$MRS_i=-\frac{u_2^i}{u_1^i}=G'(E^S)+(1-t)G''(E^S)(nE^i-E^S)\ . \quad (9.3)$$

It follows that there are four cases in which the K^+ equilibrium of the tax problem is Pareto efficient:

(i) when $t=0$,

(ii) when *G* is linear (so $G''(E^S)=0$) ,

(iii) when the economy is symmetric (so $nE^i=E^S$),

and (iv) when $t=1$.

Indeed, we knew (iv) already, because in this case the economy becomes the equal-division hunting economy, for which we have shown that K^+ is efficient. And we knew (iii), since

we have shown that in symmetric economies, both K^+ and $K^\times$ equilibrium reduce to simple Kantian equilibrium, and we know the simple Kantian equilibrium in efficient in production economies. More generally, we can surmise that if we are close to one of these four situations, then K^+ equilibrium does quite well in terms of efficiency.

On the other hand, the Nash equilibrium in the labor-supply game is given by the FOC:

$$\frac{d}{dE^i}u^i((1-t)wE^i + \frac{tG(E^S)}{n} + (1-t)\frac{G(E^S)-wE^S}{n}, E^i) = 0 \quad .$$

If each agent (rationally) ignores the effect of changing his labor supply on the wage, the demogrant, and profits, this reduces to:

$$MRS_i = -\frac{u_2^i}{u_1^i} = (1-t)w = (1-t)G'(E^S) \ . \qquad (9.4)$$

The Nash equilibrium becomes decreasingly efficient, so to speak, as the tax rate increases, while the K^+ equilibrium becomes increasingly efficient. Nash equilibrium is only efficient when the economy is laissez-faire. We knew this, as it is the first theorem of welfare economics.

[To attempt to make the comparison more precise, we can study an example with quasi-linear preferences, for in such economies, there is a simple measure of the inefficiency of an allocation – namely, how far the surplus it generates is from the maximum surplus (Pareto efficiency). Consider this example:

$$u^i(x,E) = x - \frac{\alpha_i}{2}E^2, \quad G(E^S) = aE^S - \frac{b}{2}(E^S)^2 \ .$$

Let E^* be the maximum possible total effort and suppose that $\frac{a}{b} > E^*$, so G is monotone increasing on its domain. An appendix carries out a comparison between the Nash and K^+ equilibria of the game (perhaps omit this).]

10. Altruism

Until now, I have argued that cooperation is conceptually distinct from altruism. I've also said that behavioral economists sometimes suppose that altruism is reflected in a decision maker's preferences, and this is used to explain observations that are not easily

explained as Nash equilibria of games with self-interested preferences. My strategy has been to argue that often cooperation among self-interested agents suffices to explain these non-classical outcomes. In this section, I introduce a concern for others' welfare into the preferences of individuals, and study the conjunction of altruism and cooperation.

We continue to work with the production economies $(u^1,u^2,...,u^n,G,X)$, where X is some allocation rule, except we now view the functions u^i as *personal* utility functions, and contrast these to a person's *all-encompassing* utility function, which we define as:

$$U^i(x^1,E^1,...,x^n,E^n) = u^i(x^i,E^i) + \alpha^i S(u(x,E)) \qquad (10.1)$$

where S is a Bergson-Samuelson social welfare function, and

$$u(x,E) = (u^1(x^1,E^1),...,u^n(x^n,E^n)) \ .$$

$S:\Re^n \to \Re$ is assumed to be monotone increasing in its n arguments, concave, and symmetric. We assume that $\alpha^i \geq 0$, and call this parameter i's *degree of altruism.* We take the statement $'\alpha^i = \infty'$ to mean that i's all-encompassing utility function is just the social-welfare function.

Assume that $\alpha^i = \alpha$, for some α, and all i. Denote by $PE(\alpha)$ the set of Pareto efficient interior allocations of the economic environment $(u^1,...,u^n,G,S,\alpha)$. Of course, the introduction of altruism engenders consumption externalities. We can expect that, as α increases, the set of Pareto efficient allocations shrinks, and this is made precise by the following:

Proposition 14

A. An allocation $\{(x^i,E^i)_{i=1,...,n}\} \in PE(\alpha)$ if and only if:

A1. *For all i,* $-\dfrac{u_2^i(x^i,E^i)}{u_1^i(x^i,E^i)} = G'(E^S)$

A2. $\alpha \leq \left(\max_i \left(u_1^i \cdot S_i \sum_{j=1}^n (u_1^j)^{-1} \right) - \sum_{j=1}^n S_j \right)^{-1}$

where $S_j = \dfrac{\partial S(a^1,...,a^n)}{\partial a^j}$ *, and all functions are evaluated at the stated allocation.*

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

C. $PE(\infty) = \bigcap_{\alpha \geq 0} PE(\alpha)$.

Part A is established in Roemer (2014, Theorem 5). Note that the first condition, A1, in part A simply states that marginal rates of substitution (of the personal utility functions) equal the marginal rate of transformation for all persons. Hence condition A1 simply says that the allocation must be in $PE(0)$. The second condition, A2, does not involve *G*: it is this condition that takes care of efficiency in the presence of the consumption externalities. Note that as α increases, condition A2 becomes more demanding, and this establishes Part B – the sets $PE(\alpha)$ are nested. Part C follows immediately from Part B.

In particular, $PE(\infty)$ is simply the set of allocations that maximize social welfare. Typically this will consist of a single point.

We now observe that typically, when an allocation rule *X* is given – for specificity, take it to be the proportional rule $X^{\Pr}$ -- there will be *no* Pareto efficient point in $PE(\alpha)$ that satisfies *X*, for large enough α. For a typical economy, there is one proportional allocation that is Pareto efficient in the self-interested economy – that is, is a member of $PE(0)$. Now increase α -- eventually, $PE(\alpha)$ shrinks to a point, and in all likelihood, that point will not be the allocation that is proportional and efficient in the 0-economy.

To solidify this observation, consider an economy with quasi-linear preferences: $u^i(x,E) = x - h^i(E)$. *S* can be any social-welfare function with the properties delineated earlier. Because $u_1^i \equiv 1$, condition A2 becomes:

$$\alpha \leq \left(\max_i (nS_i - \sum_{j=1}^{n} S_j) \right)^{-1} . \qquad (10.2)$$

For (10.2) to be satisfied at $\alpha = \infty$ it is necessary that for all *i*, $nS_i = \sum_{j=1}^{n} S_j$ and so all the first partial derivatives of *S* are equal. Therefore all the utilities are equal, since *S* is symmetric. But we know that $PE(0)$ consists of all allocations of a given output $G(E^{S^*})$ where E^{S^*} maximizes the surplus in the 0-economy, and the efforts of all individuals are fixed. And we know that $PE(\infty) \subset PE(0)$. It follows that $PE(\infty)$ consists of that unique distribution of the product $G(E^{S^*})$ that equalizes the utilities of all agents. There is no reason that this allocation should distribute output in proportion to labor or in any other pre-arranged way. Therefore, it is almost certainly the case that there are no allocations in $PE(\infty)$ that satisfy any particular allocation rule (other than the rule that equalizes utilities).

It is therefore too much to ask that any equilibrium for an economy $(u^1,...,u^n,G,S,X,\alpha)$ -- be it Kantian or any other kind – be Pareto efficient for the all-encompassing preferences. The most we can hope for is that a Kantian equilibrium be *second-best efficient*.

Definition. Let $\mathbf{X}(u^1,...,u^n,G)$ be the set of allocations that can be implemented by the allocation rule X in the economy $(u^1,...,u^n,G)$. We say an allocation (x,E) is *second-best efficient in* $(u^1,...,u^n,G,S,X,\alpha)$ if $(x,E)\in\mathbf{X}(u^1,...,u^n,G)$ and there is no allocation in $\mathbf{X}(u^1,...,u^n,G)$ that Pareto-dominates it, according to the utility functions $U^1,...,U^n$. Denote the set of such allocations by $PE^X(\alpha)$.

The definition of Kantian equilibrium for an economy $(u^1,...,u^n,G,S,X,\alpha)$ is the same as before. For specificity, let us write it out for an economy using the proportional rule, X^{Pr}. Let us call such an (interior) allocation a $K^{\times}(\alpha,X^{\mathrm{Pr}})$ equilibrium, understanding that the other parameters $(u^1,...,u^n,G,S)$ are all fixed. It must be the case that no player would advocate that all players re-scale their efforts by any non-negative multiple. Given the concavity of the problem, and the extended preferences defined by (10.1), the first-order condition characterizes such allocations:

$$(\forall i)\quad \left.\frac{d}{dr}\right|_{r=1}[u^i(\frac{E^i}{E^S}G(rE^S),rE^i)+\alpha^i S(u^1(\frac{E^1}{E^S}G(rE^S),...,u^n(\frac{E^n}{E^S}G(rE^S))]=0\ . \quad (10.3)$$

Denote by $K^{\times}(\alpha,X^{\mathrm{Pr}})$ the multiplicative Kantian equilibria for the an economy with the extended preferences (10.3) where $\alpha=(\alpha^1,...,\alpha^n)$ and by $K^{\times}(0,X^{\mathrm{Pr}})$ the multiplicative Kantian equilibrium of the economy with self-interested preferences ($\alpha^i=0$ for all *i*).
We now have the somewhat surprising fact:

Proposition 15 *If for all i* $\alpha^i\geq 0$, $K^{\times}(\alpha,X^{\mathrm{Pr}})=K^{\times}(0,X^{\mathrm{Pr}})$.

That is, the Kantian equilibria of an economy are the same, regardless of the degrees of altruism α^i! Kantian optimization is completely insensitive to altruism, as modeled here.

This result holds also for K^+ equilibria, and for any allocation rule – I use $K^{\times}$ and X^{Pr} to keep the notation relatively simple.

Proof:

1. Denote $\left.\frac{d}{dr}\right|_{r=1} u^i(\frac{E^i}{E^S}G(rE^S),rE^i) \equiv D_1u^i$. Then we can write (10.3) as:

$$(\forall i) \quad D_1u^i + \alpha^i \sum_{j=1}^{n} S_j \cdot D_1u^j = 0 \ . \qquad (10.4)$$

If $\alpha^i = 0$ then $D_1u^i = 0$. We now suppose that $\alpha^i > 0$.

2. It follows that there is a constant *k* such that

$$(\forall i \text{ s.t. } \alpha^i > 0) \quad -\alpha^i k = D_1u^i \ ,$$

namely, $k = \sum_j S_j \cdot D_1u^j$.

Substituting these equations into step 1, we have:

$$-\alpha^i k + \alpha^i \sum_{j \text{ s.t. } \alpha^j>0} S_j(-\alpha^j k) = 0 \ ,$$

or $$k\alpha^i(1 + \sum_{j \text{ s.t. } \alpha^j>0} \alpha^j S_j) \ .$$

Since $S_j \geq 0$, this implies that $k = 0$.

3. But this means that $D_1u^i = 0$ for all *i* such that $\alpha^i > 0$, and therefore $D_1u^i = 0$ for all *i,* by step 1. This is exactly the condition that the allocation is in $K^{\times}(0,X^{\text{Pr}})$. ■

We now return to the idea of second-best efficiency. As I have argued, the most that could be asked for is that if an allocation is second-best efficient for an economy, then it is a Kantian equilibrium for that economy. We have:

Proposition 16 *Consider the class of economies* $(u^1,...,u^n,G,S,X,\alpha)$ *where* α *can take on any non-negative value. Suppose the K – equilibria in economy* $\alpha = 0$ *are Pareto efficient. Then* $PE^X(\alpha) \subset K(\alpha,X)$ *for any* α .

For example, let $X = X^{\text{Pr}}$ and $K = K^{\times}$. Then the premise is true: multiplicative Kantian equilibria of economies with the proportional allocation rule are Pareto efficient in the economy without altruism. The proposition states that any second-best efficient allocation which is proportional, for any α , is also a multiplicative Kantian equilibrium in the α economy. The converse $(K(\alpha,X) \subset PE^X(\alpha))$ is generally false, because of Proposition 15: it is not true that any proportional allocation that is efficient in the 0-economy

(that is, a member of $K^{\times}(0,X^{\Pr})$) is Pareto-efficient in the economy with large α . Recall the above example with quasi-linear economies when $\alpha = \infty$.

The good news from Proposition 16 is that if there is a second-best efficient allocation for an economy with altruism that uses a particular allocation rule *X*, then it can always be implemented as a *K* equilibrium of the economy where players ignore their altruism – which makes their calculation simpler. If the set $K(0,X)$ is a singleton, this makes things very easy. Consider, for example, an economy of fishers who are to some degree altruistic towards each other. They employ the proportional allocation rule (each keeps his catch), and suppose that the $K^{\times}(0,X^{\Pr})$ is a singleton. They need only implement this allocation, disregarding their altruism. If there is a second-best efficient allocation under the proportional rule, *given* their altruism, this is it. If there is not, nothing can be done, except, perhaps, to change their allocation rule.

An important consequence of Proposition 15 is that *it will be difficult to deduce that members of a community who are cooperating are or are not altruistic towards each other*, because the Kantian equilibria are the same in both cases. Looking only at the Kantian equilibria, economies with altruism are *observationally equivalent* to ones without altruism. We may, of course, look for other kinds of evidence of altruism, but we cannot infer altruism from observing the nature of the cooperative equilibrium. Occam's razor would seem to dictate that altruism not be assumed. The caveat is that we have here modeled altruism in a particular way, and the analog of Proposition 15 might fail to be true with other models of altruistic preferences.

11. Extensions to more complex economies

In the production economies studied in sections 7-9, the production input is unidimensional: total efficiency units of labor. In this section, I discuss whether Kantian optimization will still engender Pareto efficiency if the argument of the production function is multi-dimensional: that is, production is defined on a vector of effort levels, $G(E^1,...,E^m)$. Let $\frac{\partial G}{\partial E^j} = G_j$,

Definition. A production function *G* is *homothetic* if for any vector *E* and for all positive α , and for all components (i,j) , there is a constant k_{ij} such that $G_i(\alpha E) = k_{ij} G_j(\alpha E)$. (Indeed, $k_{ij} = G_i(E)/G_j(E)$.) In other words, on any ray through the origin, the slopes of the iso-level curves are constant.

A production function *G* is *C-homothetic* if for any vector *E* and for all α for which $\alpha + E \geq 0$, and for all components (i,j) , there is a constant k_{ij} such that $G_i(\alpha + E) = k_{ij} G_j(\alpha + E)$, where $\alpha + E = (\alpha + E^1, ..., \alpha + E^n)$. (Indeed, $k_{ij} = G_i(E)/G_j(E)$.) In other words, on any 45^0 line in the plane (with $n = 2$) , the slopes of the intersected iso-level curves are constant.

I call the second condition *C-homotheticity* because it was discussed in Chipman (1965), who gives as an example of such a production function $G(E_1, E_2) = -e^{-E_1} - e^{-E_2} + 2$.[14]

Lemma Let *n* = 2. *Let G be twice differentiable. Let* **H** *be the Hessian matrix of G.*

a. G is homothetic iff for all E

$$(E^1, E^2)\mathbf{H}\begin{pmatrix} G_2 \\ -G_1 \end{pmatrix} = 0 \ . \qquad (11.1)$$

b. *G is C-homothetic if and only if for all E*

$$(1,1)\mathbf{H}\begin{pmatrix} G_2 \\ -G_1 \end{pmatrix} = 0 \ . \qquad (11.2)$$

Proof:

In case *a* we have $\frac{d}{d\alpha}(G_2(\alpha E) - kG_1(\alpha E)) = 0$. This reduces to (11.1). A similar argument produces (11.2), expanding the equation $\frac{d}{d\alpha}(G_2(\alpha + E) - kG_1(\alpha + E)) = 0$. ■

We now assume that in an economy with *n* producers, each produces a kind of labor that may be idiosyncratic, and so the production function is $G(E^1, ..., E^n)$. The competitive firm will pay a wage of $w^i = G_i(E^1, ..., E^n)$ for each unit of i' s labor. We assume, as before,

[14] I thank J. Silvestre for this citation.

that preferences are $u^i(x^i,E^i)$. The generalization of a proportional allocation now allocates output in proportion to the *value* of one's labor:

Definition. The *proportional allocation rule* in an economy with multi-dimensional labor is:

$$X^i(E)=\frac{G_i(E)E^i}{\sum G_j(E)E^j}G(E),\ \text{for } E=(E^1,...,E^n)\ .$$

(See Roemer and Silvestre (1993).)

The definitions of $K^{\times}$ and K^{+} equilibrium remain unchanged.

Proposition 17

A. *Let* $G:\Re_+^n\to\Re_+$ *be homothetic and twice differentiable. Consider the game, whose strategies are effort supplies, where the allocation rule is proportional. If* E^* *is a positive* $K^{\times}$ *equilibrium of this game, then it is Pareto efficient.*

B. *Let G be C-homothetic and twice differentiable. Consider the allocation rule* $Y^i(E)=\frac{G_i(E)}{\sum_j G_j(E)}G(E)$. *If* E^* *is a positive* K^{+} *equilibrium of this game, then it is Pareto efficient.*

Proof: We do the calculation for $n=2.$

Part A. At a $K^{\times}$ equilibrium, we have:

$$\frac{d}{dr}|_{r=1}\ u^1\left(\frac{G_1(rE)E^i}{G_1(rE)E^1+G_2(rE)E^2}G(rE),rE^1\right)=0.$$

Differentiating and expanding gives:

$$u_1^1\left(G_1(E)E^1+\frac{G(E)}{(\nabla G\cdot E)^2}(E^1,E^2)\mathbf{H}\begin{pmatrix}G_2(E)\\-G_1(E)\end{pmatrix}\right)+u_2^1E^1=0\ ,$$

where ∇G is the gradient of *G* evaluated at *E*. By the lemma and (11.1), and the fact that $E^1>0$, this reduces to $u_1^1G_1(E)+u_2^1=0,$ or 1's marginal rate of substitution is equal to his marginal rate of transformation. The same argument applies to player 2. Hence the allocation is Pareto efficient, as this is the condition for interior Pareto efficiency.

Part B. At a K^{+} equilibrium, we have:

$$\frac{d}{dr}|_{r=0}\ u^1(\frac{G_1(r+E)}{G_1(r+E)+G_2(r+E)}G(r+E),r+E^1)=0.$$

Differentiating, this reduces to:

$$u_1^1\left(G_1(E)+\frac{G(E)}{(G_1(E)+G_2(E))^2}(1,1)\mathbf{H}\begin{pmatrix}G_2\\-G_1\end{pmatrix}\right)+u_2^1=0\ .$$

Since G is C-homothetic, this reduces to $G_1(E)=-\frac{u_2^1}{u_1^1}$, using (11.2). Again, we have Pareto efficiency. ■

Clearly, the allocation rule Y in part B of Proposition 17 is a generalization of the equal-division rule. It would, however, be strange to call this multi-dimensional analog 'equal-division.'

From a formal viewpoint, suppose one agent supplies labor, and the other capital. The production function is $G(L,K)$. In this interpretation, we must consider $u^2(x,K)$ to be a utility function over consumption and capital supplied. If G is constant-returns-to-scale and homothetic, then the proportional solution is just the Walrasian equilibrium: each of the two agents receives the value of his input, priced at their marginal-productivity values. If G is strictly concave, the entire output is divided, efficiently, between labor and capital, where each receives the value of his marginal product.

Let us look at the symmetric case, where all the u's are the same and the function G is symmetric. The simple Kantian equilibrium is characterized by:

$$\frac{d}{dE}u(\frac{G(E,...,E)}{n},E)=0\ ,$$

since at such a vector of efforts, the marginal products cancel out. The solution is:

$$u_1\left(\frac{\nabla G\cdot\mathbf{1}}{n}\right)+u_2=0\ ,$$

which reduces to $u_1G_1+u_2=0$ -- hence, the allocation is Pareto efficient. We do not require homotheticity of G.

The generalization in the heterogeneous case, however, does not extend beyond these homothetic production functions. The lesson may be that it is difficult to organize cooperation when production is complex. Furthermore, even in the homothetic cases, the link between fairness and the Kantian counterfactuals is harder to see. If we supply different

kinds of effort, does 'multiplying all efforts by a constant' represent a fair change? I find that hard to argue. I think the appropriate inference is that the Kantian micro-foundation for cooperation does not extend in any clear way to these more complex situations.

12. Existence of *K* equilibria

Most of the propositions state that *if* an allocation is a Kantian equilibrium, then it has certain properties. In the simple games of section 2, and more generally in the symmetric production economies, it is immediately clear from the calculations that SKE exist. It is not obvious, however, that $K^{\times}$ and K^{+} equilibria exist in the economies in Ω . Roemer (2014) shows that under weak conditions, they do.

13. Evolutionary considerations

Suppose there is a large population, some of whom are Kant and some Nash optimizers. There is random pairing of individuals at each date, who play the prisoners' dilemma. A player, however, cannot tell ex ante if she is paired with a Kantian or Nash optimizer. Nash players always play *D*. Suppose Kantian players play *C* with probability *p*, knowing that they have a positive probability of encountering a Nash player. Is there a population frequency *q* of Kantian players that is stable – that is, such that if Kantians choose *p* optimally, knowing *q*, the expected utilities of the two kinds of player are the same, and so we assume they have the same fitness? It turns out the answer is no: Nash optimizers will drive the Kantians to extinction – unless the Kantians choose $p = 1$, in which case the society looks as if it consists only of Nash players.

However, suppose the Kantian players punish Nash players who defect against them. In this formulation, Kantian players always play *C*; if they meet a Nash player who defects on them, they punish the Nash player: they impose on him a penalty of one. Imposing this punishment costs the Kantian player $\delta > 0$. Therefore, if the frequency of Kantian players is *q*, the Nash player either faces the standard PD game, should he meet a Nash optimizer, or if he meets a Kantian, he faces the payoffs $(0,0)$ for the plays *C* and *D*. Now it is the Nash players' turn to randomize. If all Nash players play *C* with probability *p*, their expected payoff is:

$V^N = q \cdot 0 + (1-q)(p^2 \cdot 0 + p(1-p)(1-c) - b(1-p)^2) = (1-q)(1-p)(p(1-c) - b(1-p))$.
(13.1)

The Nash players choose *p* to maximize V^N .

We consider case *b* of Proposition 4, where in a Kantian world, the optimal mixed strategy is to cooperate with probability one. In (13.1), the sign of p^2 is negative, because $1-c+b>0$, and so we examine the FOC for maximizing V^N with respect to *p*, which solves to $\hat{p} = \frac{1-c+2b}{2(1-c+b)}$. But, since $c>1$ in this case, $\hat{p}>1$: therefore, the solution for the Nash player is $p^* = 1$, to cooperate fully. Therefore both Nash and Kantian players receive an expected payoff of one; the size of δ is immaterial, since punishment is meted out with probability zero. Thus, any population frequency *q* is stable.

It follows that if the Kantian protocol includes a prescription to punish those do not cooperate, as I have argued that it does, then the economy will look as if everyone is Kantian: there is full cooperation.

14. Historical examples (to be written)

15. Conclusion

I have been at pains to argue that one can understand cooperation without asserting that individuals are altruistic or possess social preferences. This view is at odds with most of the recent literature on cooperation and reciprocity. I maintain that the main action, in distinguishing cooperative from non-cooperative behavior, lies in the different optimization protocols that decision makers employ, not in their having different preferences for different occasions. I have tried to argue this distinction is not semantic. I characterize behavioral economics as attempting to explain non-classical behavior as Nash equilibria of games with players who possess exotic preferences – whose arguments include things like a sense of duty or fairness or a love of equality or a warm glow. My counter-proposal is that the sense of fairness is important, but it is not parsimoniously modeled as an argument of preferences: rather it induces people to optimize in a Kantian way. Nor do I deny that humans enjoy warm glows from behaving cooperatively, but that does not entail that the *reason* they do so

is to generate the warm glow. The warm glow is an unintended by-product, in the words of Elster.

My view, following Michael Tomasello, is that humans are able to optimize in the Kantian way, by contemplating the universalization of their actions, because they evolved to understand *joint intentionality*. As Tomasello (2014, 33) writes:

> Early humans' new form of collaborative activity was unique among primates because it was structured by joint goals and joint attention into a kind of second-personal *joint intentionality* of the moment, a 'we' intentionality with a particular other, within which each participant had an individual role and an individual perspective. Early humans' new form of cooperative communication – the natural gestures of pointing and pantomiming – enabled them to coordinate their roles and perspectives on external situations with a collaborative partner toward various kinds of joint objectives.

The autarchic reasoning that is postulated in Nash equilibrium is just not the way we naturally think – at least in some situations.

Nor is 'magical thinking' necessary. I need not believe, weirdly, that my taking an action will cause others to take it. It suffices that there be a situation of solidarity – that we are all in the same boat – and that we trust one another. To the extent that solidarity is reduced by differentiation or that trust is reduced by non-cooperative behavior of a significant fraction, decision-makers may switch their optimization protocol from Kant to Nash. The understanding of switching protocols – in either direction – is the domain of psychology. (See the work of David Rand.)

Proposition 15 can be interpreted as saying that economies where people cooperate are (under certain conditions) observationally equivalent to ones where they cooperate *and are altruistic towards each other.* Another way of saying this is that teaching people to be altruistic will not necessarily enable them to expand their degree of cooperation. I believe this is a hopeful result for our species, because it is, I think, much easier for people to employ their *conception of fairness* in dealing with others than to *extend their altruism* to others. Many have argued that altruism is not easily extended beyond one's kin and close friends. But we have many examples where a sense of fairness induces cooperation among millions.

References

Andreoni, J. 1990. “Impure altruism and donations to public goods: A theory of warm-glow giving,” *Econ. Jour.*100, 464-477

Bandiera, O., I. Barankay, and I. Rasul, 2006. “The evolution of cooperative norms: Evidence from a natural field experiment,”*Advances in economic analysis & policy* 6; article 4

Bowles, S. and H. Gintis, 2011. *A cooperative species: Human reciprocity and its evolution,* Princeton University Press

Boyd, R. and P. Richerson, 1985. *Culture and the evolutionary process*, University of Chicago Press

Brekke, K.A., S. Kverndokk, K. Nyborg, 2003. “An economic model of moral motivation,” *J. Public Econ.* 87, 1967-1983

Chipman, J.S. 1965. "A Survey of the Theory of International Trade: Part 2, The Neo-Classical Theory,” *Econometrica*, 33, 685-760.

Cournot, A. 1838. *Recherches sur les principes mathématiques de la théorie de richesse*, Paris

Cox, J., E. Ostrom, J. Walker, A.J. Castillo, E. Coleman, R. Holahan, M. Schoon and B. Steed, 2009. “Trust in private and common property experiments,” *Southern Econ. Jour.* 75, 957-975

Elster, J. 1983. *Sour grapes*, Cambridge University Press

Elster, J. 2009. “Norms” in P. Hedström and P. Berman, *The Oxford Handbook of Analytical Sociology* Oxford: Oxford University Press

Gintis. H. 2000. “Strong reciprocity and human sociality,” *J. Theoretical Biology* 206, 169-179

Gintis, H., J. Henrich, S. Bowles, R. Boyd and E. Fehr, 2008. “Social reciprocity and the roots of human morality,” *Social justice research*, DOI: 10.1007/s11211-008-0067-y

Henrich, N. and J. Henrich, 2007. *Why humans cooperate:A cultural and evolutionary explanation*, Oxford University Press

Mas-Colell, A. 1987. “Cooperative equilibrium,” in J. Eatwell, M. Milgate and P. Newman, *The Palgrave Dictionary of Economics*, vol. 1, London: Macmillan

Kolm, S.-C. 2006. “Reciprocity: Its scope, rationales, and consequences,” in S.-C. Kolm and J. M. Ythier, *Handbook of the economics of giving, altruism and reciprocity,* North-Holland Press

Ostrom, E. 1990. *Governing the commons*: *The evolution of institutions for collective action,* Cambridge University Press

Roemer, J. and J. Silvestre, 1993. “The proportional solution for economies with both private and public ownership,” *J. Econ. Theory* 59, 426-444

Roemer, J. 2010 . “Kantian equilibrium,” *Scandinavian J. Econ.*112, 1-24

Roemer, J. In press. “Kantian optimization: A microfoundation for cooperation,” *J. Public Economics*

Tomasello, M. 2009. *Why we cooperate,* Cambridge, MA: MIT Press

Tomasello, M. 2014. *A natural history of human thinking,* Cambridge, MA: Harvard University Press

Walker, J. and E. Ostrom, 2009. “Trust and reciprocity as foundations of cooperation,” in K.S. Cook, M. Levi, and R. Hardin, *Whom can we trust?*, New York: Russell Sage Foundation

Walras, L. 1874. *Eléments d’économie pure ou théorie de la richesse sociale*