SPRING 2019
NEW YORK UNIVERSITY
SCHOOL OF LAW

"A theory of cooperation in games with an application to market socialism"
AND
"Cooperation, altruism and economic theory"

John Roemer
Yale University

February 12, 2019
Vanderbilt Hall – 208
Time: 4:00 – 5:50 p.m.
Week 4

SCHEDULE FOR 2019 NYU TAX POLICY COLLOQUIUM

(All sessions meet from 4:00-5:50 pm in Vanderbilt 208, NYU Law School)

1. Tuesday, January 22 – Stefanie Stantcheva, Harvard Economics Department.
2. Tuesday, January 29 – Rebecca Kysar, Fordham Law School.
3. Tuesday, February 5 – David Kamin, NYU Law School.
4. Tuesday, February 12 – John Roemer, Yale University Economics and Political Science Departments.
5. Tuesday, February 19 – Susan Morse, University of Texas at Austin Law School.
6. Tuesday, February 26 – Ruud de Mooij, International Monetary Fund.
7. Tuesday, March 5 – Richard Reinhold, NYU School of Law.
8. Tuesday, March 12 – Tatiana Homonoff, NYU Wagner School.
9. Tuesday, March 26 – Jeffery Hoopes, UNC Kenan-Flagler Business School.
10. Tuesday, April 2 – Omri Marian, University of California at Irvine School of Law.
11. Tuesday, April 9 – Steven Bank, UCLA Law School.
12. Tuesday, April 16 – Dayanand Manoli, University of Texas at Austin Department of Economics.
13. Tuesday, April 23 – Sara Sternberg Greene, Duke Law School.
14. Tuesday, April 30 – Wei Cui, University of British Columbia Law School.

June 1, 2018. To appear in a symposium on this topic in the *Review of Social Economy,* edited by Roberto Veneziani

"A theory of cooperation in games with an application to market socialism"

by

John E. Roemer[**]
Yale University
John.roemer@yale.edu

Abstract. Economic theory has focused almost exclusively on how humans compete with each other in their economic activity, culminating in general equilibrium (Walras-Arrow-Debreu) and game theory (Cournot-Nash). Cooperation in economic activity is, however, important, and is virtually ignored. Because our models influence our view of the world, this theoretical lacuna biases economists' interpretation of economic behavior. Here, I propose models that provide micro-foundations for how cooperation is decentralized by economic agents. It is incorrect, in particular, to view competition as decentralized and cooperation as organized only by central diktat. My approach is not to alter preferences, which is the strategy behavioral economists have adopted to model cooperation, but rather to alter the way that agents optimize. Whereas Nash optimizers view other players in the game as part of the environment (parameters), Kantian optimizers view them as part of action. When formalized, this approach resolves the two major failures of Nash optimization from a welfare viewpoint -- the Pareto inefficiency of equilibria in common-pool resource problems (the tragedy of the commons) and the inefficiency of equilibria in public-good games (the free rider problem). An application to market socialism shows that the problems of efficiency and distribution can be completely separated: the dead-weight loss of taxation disappears.

Key words: Kantian equilibrium, cooperation, tragedy of the commons, free rider problem, market socialism

JEL codes: D70, D50, D60, D70

[**] I am grateful to Roberto Veneziani for organizing the symposium that led to this issue, and to the authors of the contributions herein for stimulating my thinking on Kantian cooperation.

1. Man, the cooperative great ape

It has become commonplace to observe that, among the five species of great ape, *homo sapiens* is by far the most cooperative. Fascinating experiments with infant humans and chimpanzees, conducted by Michael Tomasello and others, give credence to the claim that a cooperative protocol is wired into to the human brain, and not to the chimpanzee brain. Tomasello's work, summarized in two recent books with similar titles (2014, 2016), grounds the explanation of humans' ability to cooperate with each other in their capacity to engage in *joint intentionality,* which is based upon a common knowledge of purpose, and trust.

There are fascinating evolutionary indications of early cooperative behavior among humans. I mention two: pointing and miming, and the sclera of the eye. Pointing and miming are pre-linguistic forms of communicating, probably having evolved due to their usefulness in cooperative pursuit of prey. If you and I were only competitors, I would have no interest in indicating the appearance of an animal that we, together, could catch and share. Similarly, the sclera (whites of the eyes) allow you to see what I am gazing it: if we cooperate in hunting, it helps me that you can see the animal I have spotted, for then we can trap it together and share it. Other great apes do not point and mime, nor do they possess sclera.

Biologists have also argued that language would likely not have evolved in a non-cooperative species (Dunbar[2009]). If we were simply competitive, why should you believe what I would tell you? Language, if it began to appear in a non-cooperative species, would die out for lack of utility. The problem of cheap talk would be severe. In addition, language is useful for coordinating complex activities – that is, ones that require cooperation. It would not have been worth Nature's investment in a linguistic organ, were the species not already capable of cooperation, so the argument goes.

Cooperation must be distinguished from altruism. Altruism comes in three varieties: biological, instrumental, and psychological. Biological altruism is a hard-wired tendency to sacrifice for others of one's species, which sometimes evolved through standard natural selection, as with bees and termites. Some people speak of

instrumental altruism, which is acting to improve the welfare of another, in expectation of a reciprocation at some time in the future. It is questionable whether this should be called altruism at all, rather than non-myopic self-interest. Psychological altruism is caring about the welfare of others: it is a kind of preference. It is intentional, but not motivated by self-interest, as instrumental altruism is. Psychological altruism is what economists usually mean by the term.

Cooperation is not the same as psychological altruism. I may cooperate with you in building a house because doing so is the only way I can provide myself with decent shelter. It is of no particular importance to me that the house will also shelter you. Cooperation is, I believe, a more generalized tendency in humans than altruism. One typically feels altruism towards kin and close friends, but is willing to cooperate with a much wider circle. With the goal of improving human society, I think it is much safer to exploit our cooperative tendencies more fully, than our altruistic ones.

The examples I gave above of cooperation are quite primitive. Humans have, of course, engaged in much more protracted and complex examples of cooperation than hunting. We live in large cities, cheek by jowl, with a trivial amount of violence. We live in large states, encompassing millions or hundreds of millions, in peace. Early human society (in its hunter-gatherer phase) was characterized by peace in small groups, up to perhaps several hundred, but by war between groups. Our great achievement has been to extend the ambit of peaceful coexistence and cooperation to groups of hundreds of millions, groups *between* which war continues to exist. In this sense, cooperation has expanded immeasurably since early days.

Within large states, of an advanced nature, a large fraction of the economic product is pooled, via taxation, and re-allocated according to democratic decisions. We have huge firms, in which cooperation is largely decentralized. Trade unions show the extent of cooperation in firms that is decentralized and tacit when, in labor struggles, they instruct their members to ‘work to rule.’ In other words, it is wrong to view cooperation as primarily organized centrally; it’s a false dichotomy to say that competition is decentralized and cooperation must be centrally planned. By far most instances of human cooperation are decentralized as well.

From this perspective, it is quite astonishing that economic theory has hardly anything to say about cooperation. Our two greatest contributions to understanding economic activity – the theory of competitive equilibrium, and game theory with its concomitant concept of Nash equilibrium -- are theories of how agents compete with each other. Behavior of agents in these theories is autarchic: I decide upon the best strategy for me under the assumption that others are inert. Indeed, in Walrasian general equilibrium, a person need not even observe the *actions* that others are taking: she need only observe prices, and optimize as an individual, taking prices as given. Nothing like Tomasello's joint intentionality exists in these theories: rather, other individuals are treated as parameters in an agent's optimization problem.

It would, however, be a mistake to say that economic theory has ignored cooperation. Informally, lip service is paid to the cooperative tendency of economic actors: it is commonplace to observe that contracts would not function in a cut-throat competitive society. There must be trust and convention to grease the wheels of competition. Nevertheless, this recognition is almost always in the form of the *gloss* economists put on their models, not in the *guts* of the models.

There is, however, one standard theory of cooperation, where cooperative behavior is enforced as the Nash equilibrium of a game with many stages. There are typically many Nash equilibria in such games. The 'cooperative' one is often identified as a Pareto efficient equilibrium, where the cooperative behavior is enforced by punishing players at stage t+1 who failed to play cooperatively at stage t. Since punishing others is costly to the punisher, those assigned to carry out punishment of deviants must themselves be threatened with punishment at stage $t+2$, should they fail to punish. Only if such games have an infinite or indefinite number of stages can this behavior constitute a Nash equilibrium. For if it were known that the game had only three stages, then no person in stage 3 will punish deviators from stage 2, because there is no stage 4 in which they would punished for shirking. So in stage 2, agents will fail to play cooperatively. By backward induction, the 'good' equilibrium unravels. (See Kandori [1992].)

What's interesting about this explanation of cooperation is that it forces cooperation into the template of *non-cooperative* Nash equilibrium. I will maintain that

this is an unappealing solution, and too complex as well. It is a Ptolemaic attempt to use non-cooperative theory to explain something fundamentally different.

Let me give a simple example, the prisoners' dilemma, with two players and two strategies, *C*(ooperate) and *D*(efect). In fact, the strategy profile (*D,D*) is something stronger than a Nash equilibrium: it's a dominant strategy equilibrium. If the game is played with an indefinite number of stages, then the behavior where both players cooperate at each stage can be sustained as a Nash equilibrium, if punishments are applied to defectors. I propose, alternatively, that in a symmetric game like this one, each player may ask himself "What's the strategy I'd like both of us to play?" This player is not considering the welfare of the other player: she is asking whether for *her own welfare* a strategy profile (*C,C*) is better than the profile (*D,D*). The answer is yes, and if both players behave according to this Kantian protocol ('take the action I'd like everyone to take'), then the Pareto efficient solution is achieved in the one-shot game.

What is needed for people to think like this? I believe it is being in a *solidaristic situation*. Solidarity is defined as 'a union of purpose, interests, or sympathies among the members of a group (American Heritage Dictionary).' Solidarity, so defined, is not the action we take together, or the feeling I have towards others, it is a *state of the world* that might induce unison action. Solidarity may promote joint action, in the presence of trust: if I take the action that I'd like all of us to take, I can trust others will take it as well. To be precise, as we will see, this behavior has good consequences when the game is symmetric (to be defined below). Symmetry is the mathematical form of 'a union of purpose or interests.' Thus Tomasello's joint intentionality, for me, is what comes about when there is a union of a solidaristic state and trust.

Trust, however, must be built up from past experience. I therefore do not claim that it is rational in a truly one-shot game to ask the Kantian question. Nash equilibrium is the rational solution of the truly one-shot game. But in real life, we are very often in situations where trust is warranted, either because of past personal experience with potential partners, or because of social conventions, of culture. In these situations, trust exists, and the Kantian question is a natural one to ask.

Now you might respond that, if the game is *really* embedded in a multi-stage game of life, then the reason that I take the action I'd like all of us to take is for fear that

if, instead, I played 'Defect' (say, in the prisoners' dilemma) I will be punished in the future, or I will fail to find partners to play with me. Indeed, I think some people do think this way. But many people, I propose, do not. They have embedded the morality that playing the action I'd like all to play is 'the right thing to do' and a person should do the right thing. This behavior is not motivated by fear of punishment, but by morality. The morality, however, is not appropriately modeled as an object of preferences, but by a *manner of optimizing*. This may seem like a pedantic distinction, but I will argue that it is not.

Indeed, we now come to the second way that contemporary economics explains cooperation, and that is under the rubric of *behavioral economics*. Behavioral economics has many facets: here I am only concerned with its approach to explaining cooperation. I claim that the general strategy adopted by behavioral economists to explain cooperation is to insert *exotic arguments* into preferences – like a sense of fairness, a desire for equality, a care for the welfare of others, experiencing a warm glow – and then to derive the 'cooperative' solution as a *Nash* equilibrium of this new game. Thus, for example, a player in the prisoners' dilemma plays *C* because it would be unfair to take advantage of an opponent playing *C* by playing *D*. In this formulation both (*C,C*) and (*D,D*) would be Nash equilibria, if I incur a psychic cost for playing *D* against your *C*. Or suppose we simply say that the player gets a 'warm glow' from playing *C* (see Andreoni [1990]). Then the unique Nash equilibrium, if the warm glow is sufficiently large, will be (*C,C*).

Indeed, Andreoni's 'warm glow' merits further comment. I think it's true that many people get a warm glow from playing the Kantian action, from doing the right thing. But the warm glow is an unintended side effect, to use Elster's (1981) terminology, not the motivation for the action. I teach my daughter the quadratic formula. She gets it: I enjoy a warm glow. But I did not teach her the formula *in order to generate* the warm glow, which came along as a result that I did not intend. Andreoni has reversed cause and effect. The same criticism applies to explanations of charity. The Kantian explanation is that I give what I'd like everyone in my situation to give, rather than my giving because it makes me feel good – which is not to deny that I do feel good when I do the right thing.

The Kandori explanation of cooperation as a Nash equilibrium in a multi-stage game with punishments is what Elster (1989) calls a *social norm.* To be precise, it is part of Elster's characterization of social norms that those who deviate be punished by others, and those who fail to punish deviators are themselves punished by others. Doubtless, many examples of cooperation are social norms: but not all are. It has often been observed by economists that normal preferences for risk will not explain the extent of tax compliance, given the probabilities of being caught for evading, and the subsequent (small) fines. In some countries, tax evaders' names are published in the newspaper, and there it may well be that compliance is a social norm. In many cities, large numbers of people recycle their trash. Often, nobody observes whether or not one recycles. There is no punishment, in these cases, for failing to recycle: but many recycle nevertheless. Assuming that recycling is somewhat costly, the Nash equilibrium – even if people value a clean environment – is not to recycle. (I should not recycle if the cost of recycling to me is greater than the marginal contribution my recycling makes to a clean environment.) Recycling, I think, is better explained as a Kantian equilibrium. Not everyone recycles because not everyone thinks Kantian.

People's trust in others may come with thresholds. I will recycle if I see or read that fraction q of my community recycles. There is a distribution function of the thresholds q in the community. In figure 1 such a distribution function is graphed; there is a stable equilibrium where fraction q^* recycle. (There are also unstable equilibria where fraction 0 or fraction 1 recycle.) I have called these people conditional Kantians; Elster (2017) calls them quasi-moral (if $1 > q > 0$) and reserves the label 'Kantian' for those for whom $q = 0$. Nash players have $q = 1$: they always play Nash, no matter how many others are playing Kantian.

[place figure 1 here]

There are, I think, three explanations of how workers cooperate when they go on strike, or why people join revolutionary movements or dangerous demonstrations and protests against the government. The first, promulgated by Olson (1965), is of the repeated-game-with-punishments variety. Workers who cross the picket line are beaten

up. Or, there is a carrot: joining the union comes with side payments. Olson's explanation is clearly cooperation-as-a-Nash-equilibrium-with-punishments. Recently, Barbera and Jackson (2017) explain these actions as occurring because participants enjoy an expressive value from the action: they value expressing their opposition to the regime or the boss. This is what I've called the behavioral-economics approach: putting exotic arguments into preferences. I (in press) model strikes as games where players' strategies are their probabilities of striking: in the case where all preferences are the same, the Kantian equilibrium is a probability that will maximize my expected income if everyone strikes with that probability. (With heterogeneous preferences, the story is more complicated.) Preferences are straight-forward economic preferences, with no expressive element; nor are there punishments. It is not unusual in this model for the Kantian equilibrium to be that each strikes with probability one. In reality, we do not often observe this, because not everyone is a Kantian. There may well be punishments, in reality, to deter those who would not strike when the strike is on. But it is wrong to infer that those punishments are the reason that *most* people strike. The punishments may be needed only to control a fairly small number of Nash optimizers. And if the workers are conditional Kantians, or *q*-Kantians, it is possible for a strike to unravel if even a small number of Nash players are not deterred from crossing the picket line[1].

I have thus far only discussed symmetric games – true solidaristic situations, .where all payoff functions are the same, up to a permutation of the strategies. I will now formalize what I have proposed, before going on to the more complicated problem of games that are not symmetric, where payoff functions are heterogeneous.

12. Simple Kantian equilibrium

Consider a game where all players choose strategies from an interval *I* of real numbers, and the payoff function of player *i* is

$$V^i(E^i, \mathbf{E}^{-i}) \tag{2.1}$$

[1] Unravelling will not occur at the q^* equilibrium in Figure 1, which is a stable equilibrium. But it will occur if the equilibrium is at $q = 1$.

where the vector of strategies is denoted $\mathbf{E} = (E^1,...,E^n)$ and $\mathbf{E}^{-i}$ is the vector $\mathbf{E}$ without its *i*th component.

Definition 2.1 A *simple Kantian equilibrium* is a vector $\mathbf{E}^* = (E^*,E^*,...,E^*)$ such that:

$$E^* = \arg\max_E V^i(E,E,...,E) \quad . \tag{2.2}$$

That is, among all vectors where everyone plays the same strategy[2], the strategy that everyone would choose is E^*. This is the formalization of the statement that each plays the strategy he'd like everyone to play.

Definition 2.2 A game is *strictly monotone increasing* if each player's payoff is strictly monotone increasing in the strategies of the *othe*r players. It is *strictly monotone decreasing* if it is strictly monotone decreasing in the strategies of the other players.

The standard example of a strictly monotone increasing game is when a person's *E* is her contribution to a public good. The more others contribute, the higher my welfare. The standard example of monotone decreasing game is the common-pool resource problem. We all fish on a common lake, and the more others fish, the less productive is the lake for me.

A symmetric game is one where all agents have the same payoff function, subject to a permutation of the strategy profile. For my purposes, we may consider symmetric games where each player's payoff is a function of her strategy and the sum of all strategies, that is:

$$\text{for all } i,\ V^i(E^i,\mathbf{E}^{-i}) = V^*(E^i,E^S)\ \text{, some } V^*\ , \tag{2.3}$$

where $E^S \equiv \sum_i E^i$. It is immediate to observe that if a game is symmetric, in the sense of satisfying (2.3), then a simple Kantian equiliabrium exists[3].

---

[2] The set of strategy profiles where all players play the same strategy is called an *isopraxis,* by J. Silvestre, in his contribution to this issue.

[3] Simple Kantian equilibria exist for a broader class of games than symmetric ones (see Roemer(in press, chapter 2).

The Nash equilibria of strictly monotone games are Pareto inefficient. The failure of efficiency of Nash equilibrium in monotone decreasing games is called *the tragedy of the commons,* while the failure in monotone increasing games is called the *free rider problem.* In the case of monotone increasing games, in Nash equilibrium, people contribute too little; in the case of monotone decreasing games, they fish too much. But we have:

Proposition 2.1 *The simple Kantian equilibrium of a strictly monotone game* (symmetric or not) *is Pareto efficient.*

This result is what I referred to earlier when I said that in symmetric game, if everyone plays the strategy he'd like everyone to play, the result is 'good.' Because this result is so central to the idea of Kantian optimization, I will prove it here.

Proof of Proposition 2.1:

Let the game $\mathbf{V} = (V^1,...,V^n)$ be strictly monotone decreasing. Let $(E^*,...,E^*)$ be a simple Kantian equilibrium. If it is not Pareto efficient, then there is a vector $\mathbf{E} = (E^1,...,E^n)$ such that:

$$(\forall i)(V^i(\mathbf{E}) \geq V^i(E^*,E^*,....,E^*)), \tag{2.4}$$

where the inequality is strict for at least one index *i*. Let *j* be an index such that $E^j = \min_i E^i$. Then $V^j(E^j,E^j,...,E^j) > V^j(\mathbf{E})$. This follows because we have reduced the efforts (*E*) of some players other than *j* and increased the efforts of no players, because E^j is minimal among the $\{E^i\}$. So the strict inequality just stated follows from the fact that the game is strictly monotone decreasing. But this last inequality implies that:

$$V^j(E^j,E^j,...,E^j) > V^j(E^*,E^*,...,E^*)\ , \tag{2.5}$$

by the application of (2.4). This contradicts the assumption that **E*** is a simple Kantian equilibrium, for agent *j* would prefer that everyone play E^j to E^*, which proves the proposition.

An analogous proof establishes the claim for monotone increasing games. ■

In this sense, simple Kantian equilibrium resolves the inefficiencies due to both negative and positive externalities that are characteristic of Nash equilibrium[4].

The general version of a 2×2 symmetric prisoners' dilemma is given by the payoff matrix in table 1:

	C	*D*
C	(0,0)	(-*c*,1)
D	(1,-*c*)	(-*b*,-*b*)

Table 1. The payoffs are (row player, column player).

where $0 < b < c$. (2.6)

Let's look at the more complicated version of the PD where each player plays a mixed strategy: with probability *p* the row player plays *C*, and with probability *q* the column player plays *C*. Then the payoff function of the row player is:

$$V^{row}(p,q) = -b(1-p)(1-q) - cp(1-q) + (1-p)q \ , \tag{2.7}$$

and by symmetry, the payoff function of the column player is:

$$V^{col}(p,q) = -b(1-p)(1-q) + p(1-q) - c(1-p)q \ . \tag{2.8}$$

Calculate that :

$$\frac{dV^{row}}{dq} = (b+1)(1-p) + pc > 0 \ \text{ for all } p, \tag{2.9}$$

so this is a strictly monotone increasing game. Hence the simple Kantian equilibrium of the game is Pareto efficient, by Proposition 2.1. This means there is no mixed-strategy pair that can give a higher expected utility to both players than the simple Kantian equilibrium. The exact form of the simple Kantian equilibrium of the game depends

[4] Indeed, one can prove that if the payoff functions are differentiable, any interior Nash equilibrium in a monotone game is Pareto inefficient (Roemer[in press, Proposition 3.3]).

upon the values of b and c. The Nash equilibrium of the game is always (0,0): both players defect for sure, and the outcome is inefficient.

Let me now make precise one of my criticisms of the behavioral economics. One can compute that if $c < 1$, then the simple Kantian equilibrium of the PD game is that both players cooperate with probability:

$$p^* = \frac{2b + 1 - c}{2(1 + b - c)}, \tag{2.10}$$

a probability strictly between zero and one. In other words, *full cooperation* ($p = 1$) is *not* the simple Kantian equilibrium! Of course, it remains true that (p^*, p^*) is Pareto efficient: in particular, it is better for both players (in ex ante utility) than the strategy profile (1,1). Now suppose you are a behavioral economist, and you want to insert an exotic argument into the preferences of the agents so that the *Nash* equilibrium of the game will be (p^*, p^*) ? How would you do it? Recall that the preferences you create must not work just for *this* PD game but for *all* PD games.

There turns out to be a way, and I believe only one sensible way, of doing this. Assign each player a new payoff function, which is the *sum* of the payoff functions of the two players in the original PD. That is, define:

$$Q(p,q) = V^{row}(p,q) + V^{col}(p,q) . \tag{2.11}$$

Now consider the game where both players have the payoff function Q – of course, the row player continues to play p and the column player q. It is not hard to check that the *Nash* equilibrium of this game is indeed:

$$\hat{p} = \hat{q} = p^* = \frac{2b + 1 - c}{2(1 + b - c)} . \tag{2.12}$$

So we *can* rationalize the simple Kantian equilibrium of the PD game as the Nash equilibrium of a game where each player is maximizing an 'exotic' preference order, the *total payoff* of the original game.

This might appear to support the behavioral economist's strategy. Here indeed – and this can be checked – the simple Kantian equilibrium of a symmetric game is always

a Nash equilibrium of an altered game where each player's payoff function is the *sum* of all players' payoff functions in the original game. There are, however, two problems with this move: first, is it credible to think that's what players are doing when they play the cooperative strategy in such a game – that they are attempting to maximize the total payoff? This is something that can be studied experimentally, and I conjecture it will not be borne out. But secondly, this trick (of transforming a Kantian equilibrium into a Nash equilibrium of a game with altered preferences) only works when the game is symmetric. A bit more on this later.

3. Simple production economies

We now consider a class of simple production economies in which we can study Kantian optimization in environments with negative and positive externalities. Suppose there is a production function that produces a desirable consumption good from the efforts of individuals, according to the production function $G(E^S)$, where G is strictly concave, increasing and differentiable, E^i is the labor expended by person i, measured in efficiency units, and $E^S = \sum E^i$. Player i has a utility function $u^i(x,E)$ where x is the consumption good produced. As usual u is concave, increasing in x, decreasing in E, and differentiable. For the moment, we allow the preferences of individuals to differ.

A. The fishing economy

In the fishing economy, we think of G as production of fish from a lake, which suffers from congestion externalities, the more labor is expended in fishing: hence, the strict concavity of G. Recall that an interior allocation is Pareto efficient exactly when :

$$\text{for all } i, \quad -\frac{u_2^i(x^i,E^i)}{u_1^i(x^i,E^i)} = G'(E^S) \ , \tag{3.1}$$

for this is the statement that the marginal rate of substitution between labor and fish for each fisher (MRS^i) equals the marginal rate of transformation of labor into fish (*MRT)*.

The *allocation rule* in this economy is given by:

$$x^i = \frac{E^i}{E^S} G(E^S) \ ; \tag{3.2}$$

that is, except for random noise, each fisher receives fish in proportion to the efficiency units of labor he expends. Another way of putting this is that 'each fisher keeps her catch.'

We can define a game: the payoff to fisher *i*, is given by the function:

$$V^i(E^1,...,E^n) = u^i(\frac{E^i}{E^S} G(E^S), E^i) \ . \tag{3.3}$$

It is well-known, that due to the strict concavity of *G*, the *Nash* equilibrium of this game is Pareto inefficient. This is the classical example of the tragedy of the commons. In Nash optimization, players do not take into account the negative externality they impose on other fishers by their own fishing. For each additional hour I fish, I lower the marginal productivity of everyone's labor on the lake. Indeed, the Nash equilibrium is the solution of these equations:

$$\text{for all } i, \quad -\frac{u_2^i}{u_1^i} = \frac{E^i}{E^S} G'(E^S) + (1 - \frac{E^i}{E^S}) \frac{G(E^S)}{E^S} \ . \tag{3.4}$$

In words, the marginal rate of substitution of a fisher is equal to a convex combination of the marginal rate of transformation and the average product. Only in the case of one fisher is this allocation Pareto efficient – as long as *G* is not linear.

To apply the concept of simple Kantian equilibrium, we will assume for the moment that $u^i = u$ for all *i*. What is the simple Kantian equilibrium of the game? It is given by solving:

$$\frac{d}{dE} u(\frac{1}{n} G(nE), E) = 0 \, , \tag{3.5}$$

for the solution of (3.5) is the amount of fishing time that each player would like all to expend. Taking the derivative, we have:

$$\begin{aligned} &\frac{d}{dE} u(\frac{1}{n} G(nE), E) = u_1 \frac{1}{n} G'(nE) n + u_2 = 0 \\ &\text{or} \quad G'(E^S) = -\frac{u_2}{u_1} . \end{aligned} \tag{3.6}$$

But this is the condition for Pareto efficiency! Hence the simple Kantian equilibrium is Pareto efficient.

Now, a caveat. You might have observed that the game defined in (3.3) is a monotone decreasing game, and hence concluded from Prop. 2.1 that the simple Kantian equilibrium (SKE) is Pareto efficient, and so the derivation (3.6) is redundant. But that inference is false. What Proposition 2.1 shows is that the SKE is Pareto efficient *among all allocations that can be achieved in the game, so defined :* that means among all allocations in which consumption of fish is proportional to labor expended. But (3.6) is a much stronger statement: it says that the SKE is Pareto efficient in the set of *all feasible allocations* for this economy. There is no way of allocating the total fish caught to the fishers, through any intricate system of redistribution, that can Pareto dominate the SKE of the game that models the fishing economy. In other words, we achieve full efficiency even though we restrict ourselves to allocations where each fisher keeps her catch. The demonstration in (3.6) is stronger than Proposition 2.1.

Again, I must make the comparison with the behavioral-economics approach. Because this is a symmetric game (when all utility functions are the same), it is indeed the case that the *Nash* equilibrium of the game where each player maximizes the sum of utilities of all players is the simple Kantian equilibrium given by (3.6). So we have, at this point, two explanations if a fishing community achieves the Pareto efficient allocation in which each keeps his catch: either each is fishing the amount of time he would all everyone to fish, or each is a complete altruist, desiring to maximize total utility, but optimizing in the manner of Nash.

B. Heterogeneous preferences

We now relax the assumption that all utility functions are the same, and assume the profile of utility functions is $\mathbf{u} = (u^1,...,u^n)$. One simple way to generate heterogeneous preferences is to say that everyone has the same preferences over fish and *labor time*, but since fishers have different skill levels, this induces heterogeneous preferences over fish and efficiency units of labor. Recall that the relevant labor in our models is the latter.

It's now the case that simple Kantian equilibria will generally not exist: the labor that I would most like everyone to expend is the different from the labor you would most

like everyone to expend. There is, however, a generalization of Kantian optimization that we can use with heterogeneous preferences.

Suppose at an allocation $\mathbf{E} = (E^1,...,E^n)$ I am contemplating increasing my fishing time by 5%. I ask myself: How would I like it if everyone increased her fishing time by 5%? The implicit moral commandment here is that I should only increase my fishing time by 5% if I'd be happy were everyone to do the same. It's a way of making me take into account the negative externality created by my contemplated action. But I don't ask myself how would my increased fishing would affect others, but rather *how, if they emulated my action, their increased fishing times would affect me.* Don't worry too much at this point about the psychological realism of this question. Instead, let's study the properties of an equilibrium when everyone thinks in this way.

Definition 3.1 An allocation of efforts $\mathbf{E} = (E^1,...,E^n)$ is a *multiplicative Kantian equilibrium* (a $K^\times$ equilibrium) if *nobody* would like to rescale *everybody's* fishing time by *any* non-negative scale factor. That is:

$$(\forall i)\left(\arg\max_{r\geq 0} u^i(\frac{rE^i}{rE^S}G(rE^S),rE^i) = 1 \right) \qquad (3.7)$$

I have stated the definition for the fishing economy, but in the attached footnote, I state it for a general game in normal form.[5]

We have the general result, using the definition in the footnote:

Proposition 3.1 *If* $\mathbf{E} = (E^1,...,E^n)$ *is a positive* $K^\times$ *equilibrium of a strictly monotone game* $\{V^i\}$ *, then it is Pareto efficient in the game.*

The proof of this proposition is very similar to the proof of proposition 2.1.

In particular, the fishing game of (3.7) is a strictly monotone decreasing game. So Proposition 3.1 implies that the multiplicative Kantian equilibrium, if it exists, is

---

[5] Let the payoff functions of the game be $\{V^i\}$ where the strategy space for each player is an interval of real numbers. Then $(E^1,...,E^n)$ is a $K^\times$ equilibrium of the game when $(\forall i)(\arg\max_{r\geq 0} V^i(r\mathbf{E}) = 1)$.

Pareto efficient in the game. As before, we must recall that being efficient 'in the game' means in the class of allocations where each keeps his catch. But we could ask for a stronger result: is the $K^{\times}$ equilibrium *efficient in the economy*?

We have:

Proposition 3.2 *Any strictly positive $K^{\times}$ equilibrium of the fishing game is Pareto efficient in the economy.*

Again, because this is a key result, let's prove it.

Proof:

1. By concavity of the utility functions and *G*, it is only necessary to show that :

$$(\forall i)\left(\frac{d}{dr}|_{r=1}\, u^i(\frac{E^i}{E^S}G(rE^S), rE^i = 0 \right) \Rightarrow \text{ the allocation is Pareto efficient.}$$

This suffices, because the antecedent to the implication sign is the first-order characterization of $K^{\times}$ equilibrium.

2. Compute that:

$$\frac{d}{dr}\Big|_{r=1} u^i(\frac{E^i}{E^S}G(rE^S), rE^i) = u_1^i \cdot \frac{E^i}{E^S}G'(E^S)E^S + u_2^i E^i = 0 \ . \tag{3.8}$$

Using the assumption that $E^i > 0$, this equation reduces to:

$$(\forall i)\left(-\frac{u_2^i}{u_1^i} = G'(E^S) \right) , \tag{3.9}$$

which is exactly the condition for Pareto efficiency at an interior solution. ■

As with the case of homogeneous preferences, we can therefore assert a stronger statement than Proposition 3.1: in the fishing economy, multiplicative Kantian optimization resolves the tragedy of the commons that afflicts Nash equilibrium. (Recall the Nash equilibrium of the fishing game is given by (3.4).)

Do such equilibria exist? These are allocations in which fish consumed is proportional to labor expended *and* the allocation is Pareto efficient. The idea of looking for such allocations is due to Joaquim Silvestre. In Roemer and Silvestre (1993), we

proved that in economies much more general than the simple production economies studied here, such allocations always exist. Silvestre's question was asked in the context of thinking about the socialist ideal: an allocation in which the total product is distributed in proportion to labor expended, *and* is Pareto efficient. Remarkably, perhaps, this question had not been raised earlier by mathematical socialist economists like Oscar Lange and Michio Morishima. It was not, however, until several years later that I observed that these allocations, which Silvestre and I called *proportional solutions*, had the property of being stable with respect to the kind of optimization that I now call Kantian (see Roemer [1996, Theorem 6.6]). Also, as an illustration of 'thinking slow,' the much simpler idea of *simple* Kantian equilibrium in symmetric games did not occur to me for another twenty years.

We can now repeat our earlier question: Is there a way of altering preferences in the fishing game so that the Nash equilibrium of the *altered* game is the Pareto efficient allocation in which each keeps his catch (i.e., the multiplicative Kantian equilibrium)? (There is often a unique such allocation; in any case, there is a finite number of such allocations.) The answer is there is no simple way of doing this: the idea of giving all players the desire to maximize total utility no longer works. For further discussion of the representation of Kantian equilibria as Nash equilibria of games with exotic preferencers, I refer the reader to Roemer (in press, chapter 6).

In other words, in games that are fairly complicated, the behavioral economist's strategy of altering preferences in order to achieve the good (cooperative) solution as the *Nash* equilibrium of a game does not work. Essentially, one has to know what the good solution is *a priori* and then one can jimmy preferences to give the right answer under Nash optimization. But this procedure cannot be regarded as decentralizing cooperation. In contrast, multiplicative Kantian optimization decentralizes cooperation, in the exact sense that Nash optimization decentralizes competition. As with Nash, there is no obvious solution to the dynamic problem of how one converges to a Nash (or Kantian) equilibrium. But, as with Nash, both kinds of equilibria are stable given the kind of optimization that players are using, once equilibrium is achieved.

I will have more to say about the realism of proposing that players in a game might optimize in the multiplicative Kantian fashion – but later. At this point, however, I still want to explicate the properties of this kind of thinking.

C. The hunting economy

Consider the game that according to anthropologists characterized early hunting societies. Unlike fishers, hunters divided the catch equally (more or less). A group of hunters goes out into the bush for a few days. They return with their catch, and divide it equally among the group.

Here the allocation rule is:

$$x^i = \frac{G(E^S)}{n} \ , \qquad (3.10)$$

which is the *equal-division rule.* The game defined by the equal-division rule is given by:

$$V^{EDi}(E^1,...,E^n) = u^i(\frac{G(E^S)}{n}, E^i) \ . \qquad (3.11)$$

This is a game with positive externalities: the total catch is a public good. The game is strictly monotone *increasing.* Again, the Nash equilibrium, which is given by the next set of equations, is Pareto inefficient as long as there is more than one hunter:

$$-\frac{u_2^i}{u_1^i} = \frac{G'(E^S)}{n} \ . \qquad (3.12)$$

Why is the Nash equilibrium inefficient? Because a hunter might like to take a nap for several hours under a bush, given what others are doing. He contributes less to the public good than efficiency requires. There is a free rider problem.

It turns out that multiplicative Kantian optimization does not resolve this inefficiency. But there is a kind of Kantian optimization that does. Suppose the hunter asks himself, "I'd like to take a two hour nap. But how would I feel if everyone took at two hour nap?" If the moral commandment is to take the nap only if the answer to this question is affirmative, then an equilibrium is defined as follows:

Definition 3.2 An allocation $\mathbf{E} = (E^1,...,E^n)$ is an *additive Kantian equilibrium* (K^+) if and only if:

$$(\forall i)\left(\arg\max_r u^i(\frac{G(E^S + nr)}{n}, E^i + r) = 0 \right). \qquad (3.13)$$

In other words, *nobody* would like to *translate* the effort vector by *any* number (positive or negative)[6].

I will now leave it to the reader to verify:

Proposition 3.3 *Any additive Kantian equilibrium of the hunting game is Pareto efficient in the economy.*

Thus, additive Kantian optimization resolves the under-provision of the public good in the hunting economy.

Kantian optimization always takes the form of each asking: "If we all take a similar symmetric action, what would I like that action to be?" In the case when the game is symmetric, the action is 'making the same contribution.' In the fishing game, the action is 'rescaling the current allocation by a common factor;' in the hunting game, it is 'translating the current allocation by a common factor.'

D. More general allocation rules in simple production economies

Let's define an allocation rule for an economic environment specified by the data $(u^1,...,u^n,G)$ as the share of output θ^i each individual receives as a function of the vector of effort contributions. We have studied two rules, the proportional and equal-division rules, whose share rules are:

$$\theta^{\Pr,i}(E^1,...,E^n) = \frac{E^i}{E^S}, \quad \theta^{ED,i}(E^1,...,E^n) = \frac{1}{n} \text{ for all } i. \qquad (3.14)$$

---

[6] For general games, $(E^1,...,E^n)$ is an additive Kantian equilibrium if:
$(\forall i)\left(0 = \arg\max_r V^i(E^1 + r,...,E^n + r) \right)$.

Indeed, these are surely the two classical notions of fair division of a jointly produced output. We have shown that each of these rules can be efficiently implemented -- more strongly, efficiently decentralized—by a specific kind of Kantian optimization.

We can also think more generally about kinds of Kantian symmetrical treatment. Define a *Kantian variation* as a function $\varphi(E,r)$ where E is a contribution and r is a number where the following holds:

$$\varphi(E,1)\equiv 1, \quad \varphi(E,r) \text{ is increasing in } r \ . \tag{3.15}$$

The *multiplicative Kantian variation* is given by:

$$\varphi^{\times}(E,r)=rE \tag{3.16}$$

and the *additive Kantian variation* is given by:

$$\varphi^{+}(E,r)=E+r-1 \ . \tag{3.17}$$

Now we say that the allocation rule θ is *efficiently implemented by the Kantian variation* φ on the set of economies $\{(u^1,...,u^n,G)\}$ if for all economies:

$$(\forall i)\left(1=\arg\max_{r} u^i(\theta^i(\varphi(\mathbf{E},r)G(\sum\varphi(E^i,r)),\varphi(E^i,r))\right) , \tag{3.18}$$

where $\varphi(\mathbf{E},r)\equiv(\varphi(E^1,r),...,\varphi(E^n,r))$.

Equations (3.18) simply state the generalization of multiplicative and additive Kantian equilibrium to other allocation rules, and other types of Kantian variation. They say that an allocation has the property that, for the given allocation rule θ, nobody would like to vary the entire effort vector according to any transformation permitted by the variation defined by φ.

Definition 3.2 The pair (θ,φ) is an *efficient Kantian pair* if the allocation rule θ is efficiently implemented as a K^{φ} equilibrium on all economies of the form $(u^1,...,u^n,G)$.

What we have thus far shown is that $(\theta^{\mathrm{Pr}},\varphi^{\times})$ and $(\theta^{ED},\varphi^{+})$ are efficient Kantian pairs. The question is: Are there any others?

The answer is that there is a unidimensional continuum of such pairs, of which the two we have studied are the endpoints. Consider allocation rule of the form:

$$\theta^{\beta,i}(E^1,...,E^n)=\frac{E^i+\beta}{E^S+n\beta} \tag{3.19}$$

where $0\le\beta<\infty$. Notice that θ^0 is the proportional rule, and that as β approaches infinity, θ^β approaches the equal-division rule. Because of this, let's define $\theta^\infty=\theta^{ED}=\frac{1}{n}$. We have:

Proposition 3.4 *For every rule θ^β, there is a Kantian variation that implements it efficiently on the domain of production economies we are considering. Furthermore, there are no other allocation rules that can be efficiently implemented on this domain with respect to any Kantian variation*[7].
(For proof, see Roemer (in press, Corollary 4.4).)

What do the rules θ^β look like? It is easy to show that they are 'convex combinations' of the proportional rule and equal division rule. Let's solve the following equation for λ:

$$\frac{E^i+\beta}{E^S+n\beta}=\lambda\frac{E^i}{E^S}+(1-\lambda)\frac{1}{n}. \tag{3.20}$$

The solution is $\lambda=\frac{E^S}{E^S+n\beta}$. The important fact is that the value of λ is independent of *i*. This means that the rule θ^β is simply this: it takes a share λ of the output $G(E^S)$ and distributes it to the participants in proportion to the efficiency units of labor expended, and it distributes the rest of the output equally to all. The share λ, however,

[7] There are two caveats. If $\beta<0$, the allocation rules can also be efficiently implemented; however, Kantian allocations may not exist. And for $\beta=0$ we must insist that the allocation be strictly positive. (The zero vector is a multiplicative Kantian equilibrium but it is not Pareto efficient.)

depends upon the equilibrium. We know that as β travels from zero to infinity, the share λ travels from one to zero: but that is all that we can say.

It's for this reason that I put 'convex combination' in quotations marks earlier. We cannot specify the share λ to be – say – one-half, and then quickly choose the right allocation rule to implement that share. To say this mathematically, the mapping from λ to β is complicated, as to know it we have to compute the equilibrium to know E^S. Only after knowing the equilibrium for each β can we find the one that implements a particular combination of proportional and equal division.

The upshot of this discussion is that the class of allocation rules that can be efficiently implemented by *any* kind of Kantian thinking is precisely the class of rules generated by the two classical egalitarian methods of distribution: in proportion to effort and equally. As these rules have a venerable history as rules of fair allocation, Proposition 3.4 re-enforces the view that Kantian optimization is the right way of thinking about morality and, may we say, cooperation.

I conclude this section with a remark about the behavioral feasibility of Kantian optimization. I think simple Kantian optimization is natural – there are many symmetric games that describe our social interactions, and in those games, I think many people ask the Kantian question: what's the action I'd like everyone to take? Multiplicative and additive Kantian optimization are, however, not natural. I think symmetry plays a large role in our conceptions of fairness, and the symmetrical treatment of all characteristic of Kantian optimization recommends it as a way of implementing fairness. But I do not claim societies have discovered these complex ways of optimizing. Rather, I see the approach as prescriptive. If the fairness involved in these kinds of Kantian optimization appeals to people, we may recommend –- to a fishing community, for example – the multiplicative Kantian equilibrium as a solution to their problem. Doing so, of course, would require the planner to know the preferences of the fishers, unless some process I do not yet understand could lead dynamically to the Kantian equilibrium of the game.

I have given examples where Kantian optimization resolves a *tragedy of the commons*, and a *free rider problem.* In fact, we have two general results: first, 'any' kind of Kantian optimization results in Pareto efficient equilibria in all strictly monotone games, and second, generically, the Nash equilibria in strictly monotone games are Pareto

inefficient. (See Roemer (in press, chapter 2.) Of course, the value of these statements depends upon the existence of the equilibria in question.

Thus, cooperation in the sense of Kantian reasoning resolves, generally, the two major failures of Nash optimization.

4. Market socialism

Thus far, I have described economies that do not trade on markets. Indeed, there is nothing that can be thought of as comprising trade in the fishing and hunting economies. We now ask what role Kantian optimization can play in market economies.

Indeed, I will propose that Kantian optimization can play an important role in a design for market socialism. The 'design problem' for socialism is stated by the philosopher G.A. Cohen as follows:

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

What Cohen means is this. Capitalism has a design consisting of private property rights in labor and productive assets, and free trade, mediated by prices that equilibrate supply and demand. The *ethos* that makes capitalism work is the maximization of self-regarding preferences in an autarchic manner: the individualist ethos is captured not only in the nature of preferences, but in the manner in which people optimize (à la Nash). Cohen also views the motivation for economic activity under capitalism as being 'greed and fear,' a point with which I do not completely agree.

Socialism, however, is supposed to be an economic system characterized by cooperation. The natural question becomes, how can one design an economic system based on cooperation to deliver good results? The first theorem of welfare economics for capitalist economies is the main formalized example of capitalism's good result. Can

we design a mechanism, that given a socialist ethos, would deliver good results? A good result for socialism should demand a higher standard than for capitalism: we desire to have not only efficiency but also a substantial degree of *income equality.*

Cohen, in the above quotation, says that it is not primarily (selfish) human nature that poses a problem for socialism, it is the lack of a solution to this design problem. Cohen, evidently, is optimistic about human nature – and there is indeed good reason to be so. Many people, in today's capitalist societies, do not plan their lives to maximize their wealth, but to do work that is useful for society. I agree with Cohen that human nature is not the primary obstacle to socialism, but the lack of a design that can *decentralize economic activity* to achieve good results, *given* that a socialist ethos exists in the population.

To state the problem slightly differently, recall the huge effect that Marx's theory of historical materialism had on engendering an international socialist movement. Marx offered – if not a design in the sense here being discussed – a *vision* of the feasibility and indeed historical necessity of socialism. We need not here debate the validity of that vision: what's salient is that, possessing this vision, millions of people organized to attempt to realize it. Now that vision has soured, due to the experiences of twentieth century socialism, constrained as they were by political authoritarianism and the fear of introducing markets. Having a *design* for how socialism could work, *given* a willing population, is of primary importance in rekindling the socialist vision.

I am less suspicious of markets than Cohen was, and so I will propose how incorporating *cooperative optimization* into a market economy can produce eqiuilibria which are *decentralized, efficient, and equitable.*

The model is considerably more general than the one I now exposit, for purposes of clarity. Assume there is one produced good, which is used both for investment and consumption. The good is produced from capital and labor, according to a production function $G(K,L)$, which is operated by a firm. The firm is owned by private citizens and the state. The share in the firm's profits of private citizen *i* is θ^i , for $i = 1,....,n,$ and the state's share is θ^0 : so $\sum_{i=0}^{n} \theta^i = 1$. Individuals, as well as owning shares of the firm, have endowments of efficiency units of labor in the amount ω^i . Citizen *i* has a

preference order over consumption of the good and labor expended, denoted in efficiency units, represented by a utility function $u^i(x,E)$. The state is endowed with an amount of capital K_0 -- this endowment was presumably produced in the previous period, not formalized in the model. Individuals own no capital, nor do they hold inventories of the consumption good. (The capital good and consumption good are, as I said, the same good, except for their vintage.)

So far, I have described exactly the set-up of an Arrow-Debreu economy, except for the state's partial ownership of the firm. There will be three prices in this economy, a price for output *p*, an interest rate for borrowing capital *r*, and a wage *w* per unit of efficiency labor. (Because the capital with which the state is endowed was produced in the past, its rental price will differ from the price of the good produced in the present period.) The firm behaves exactly as an Arrow-Debreu firm: it demands capital and labor and supplies the good in order to maximize profits. All incomes in the economy, except the state's revenues, will be taxed at an exogenous rate $t \in [0,1]$, and the revenues will returned to the citizenry as a demogrant: that is, divided equally among them. This is a flat, or affine, income tax.

How does the state behave? It supplies capital to the firm to maximize its income, which it uses to purchase the good produced by firm, to be used for next period's investment. In fact, the model is truncated: there is only one period, for purposes of simplicity. So the firm demands capital in the present period, which is uniquely supplied by the state, and supplies the state with the good produced at present, which the state purchases with its revenues (rents and capital income).

How do consumer-workers behave? Given a worker's labor supply E^i to the firm, he uses his revenues to purchase the good for consumption. Workers' revenues come from three sources: their after-tax wage and profit income, and the demogrant.

The only substantial way in which the model differs from the standard Arrow-Debreu model, besides the state's role, is in the determination of the labor supplies of workers. The vector of labor supplies $(E^1,...,E^n)$ must be an *additive Kantian equilibrium* of a game to be defined below.

Finally, a price vector (p,w,r) comprises a *Walras-Kant equilibrium at the tax rate t* if, given the optimizing behaviors described, all markets clear: that is, the supply of the good by the firm equals the demand of consumers for the good plus the demand by the state for the good (investment), the supply of capital by the state equals the firm's demand for capital, and the total efficiency units of labor supplied by workers equals the demand for labor by the firm.

We state the definition formally.

Definition 4.1 A *Walras-Kant market-socialist equilibrium at tax rate t,* consists of:

i. a price vector (p,w,r) ,

ii. labor and capital demands by the firm of *D* and *K* , respectively,

iii. labor supplies E^i by all workers $i = 1,...,n$ to the firm,

iv. for all private agents *i,* commodity demands x^i for the good, and a demand for the good by the state of x^0 ,

such that:

v. at given prices, (K,D) maximizes profits of the firm;

vi. the labor supply vector $\mathbf{E} = (E^1,...,E^n)$ constitutes an additive Kantian equilibrium at the given prices of the game $\mathbf{V}_+$ defined in equation (4.4) below;

vii. x^i maximizes the utility of agent *i*, given prices, her labor supply, and her income (that is, its purchase exhausts her budget);

viii. x^0 maximizes the state's utility u^0 subject to its budget constraint $px^0 \leq \theta^0\Pi + rK_0$, and

ix. all markets clear; that is, $D = E^S$, $x^S = G(K,D)$, and $K_0 = K$.

We must now be more precise in order to define the game of which the effort vector must be an additive Kantian equilibrium. First, define the income of the state, which will be:

$$I^0(p,r,w,E^S) = \theta^0\Pi(K_0,E^S) + rK_0 \ , \tag{4.1}$$

where the firm's profits are:

$$\Pi(K_0, E^S) = pG(K_0, E^S) - wE^S - rK_0 \ . \tag{4.2}$$

In words, (4.1) states that the state's income consists of its share of firm profits plus the interest on its investment. The income of worker *i* is given by:

$$I^i(E^1,...,E^n) = (1-t)wE^i + (1-t)\theta^i\Pi(K_0, E^S) + \frac{t}{n}(pG(K_0, E^S) - I^0(K_0, E^S)) \ . \tag{4.3}$$

The first term on the right-hand side of (4.3) is agent *i*'s after-tax wage income, the second term is her after-tax profit income, and the third term is the demogrant, which is the agent's per capita share of tax revenues, where taxes are levied on all incomes except the state's.

The payoff function for worker *i* is given by :

$$V_+^i(E^1, E^2, ..., E^n) = u^i\left(\frac{I^i(E^1,...,E^n)}{p}, E^i\right) \ . \tag{4.4}$$

That is, the worker's utility is generated by spending her entire income on the consumption good, and her labor. This completes the definition of equilibrium.

If we assume that the depreciation rate of capital is zero, then the state's capital, to be used in the 'next period,' will be $K_0 + \frac{I^0(K_0, E^S)}{p}$.

I emphasize that the only substantial ways in which this equilibrium differs from a private-ownership Arrow-Debreu equilibrium are that the state carries out all the investment, and the labor-supply decisions do not comprise a *Nash* equilibrium for workers, but an additive Kantian equilibrium. In determining their supplies of labor, workers do not ask whether supplying an extra day's labor generates an after-tax wage that compensates for the extra disutility of labor, *assuming all other workers' supplies of labor remain fixed,* but rather, whether the extra day's income compensates for the extra day's work, *assuming all others expend an extra day's labor as well,* which would increase substantially the value of the demogrant.

The consequence of this *single* change in optimizing behavior is the following:

Proposition 4.1 *Let* $(p,w,r,K_0,E^1,...,E^n,x^0,...,x^n)$ *be a Walras-Kant equilibrium at any tax rate* $t \in [0,1]$. *The equilibrium is Pareto efficient.*

One might call this proposition the 'first welfare theorem of market socialism.' It states that the equilibrium is Pareto efficient *at any income tax rate*. In particular, society, by democratic choice of the tax rate, can achieve any degree of income equality it desires – even perfect income equality, when $t = 1$ -- with no sacrifice in efficiency.

To be precise, I must explain what Pareto efficiency means here. Think of the state's having the utility function $u^0(x) = x$; the state simply cares about the amount of the investment good it acquires for the 'next period,' placed in quotes because the formal model has only one period. An allocation is Pareto efficient if there is no other feasible allocation that renders some agent better off, and none worse off, where the state is included as an agent. Now since there is no future in this model, we cannot assert that the state's investment is that which trades off efficiently present and future consumption of workers. Most economists believe that the Soviet Union and China invested too high a fraction of the national product – these states were (probably) not trading off future against present consumption of their citizens properly. That, too, can happen in the present model. But to deal with this problem, one would have to articulate a multi-period model, and I have elected here to keep things simple by truncating the future.

What is the trick that generates what appears to be an astounding result – that there is no trade-off between efficiency and equality in the model? The usual dead-weight loss of income taxation has been, apparently, entirely nullified by Kantian optimization in the labor supply decision. Without giving a full proof of Proposition 3.1, I can indicate how this occurs.

Let's ask how a worker sees his income as varying if he *and all other workers* increase their labor supply by a small amount. We must calculate:

$$\left.\frac{d}{d\rho}\right|_{\rho=0} I^i(E^1+\rho,...,E^n+\rho); \tag{4.5}$$

expanding this derivative shows it is equal to:

$$(1-t)w+(1-t)\theta^{i}\Pi_2(K_0,E^S)n+\frac{t}{n}(pG_2(K_0,E^S)n-\theta^{0}\Pi_2(K_0,E^S)n) \quad (4.6)$$

where G_2 and Π_2 are the derivatives of G and Π with respect to their second component. Now at equilibrium, we have:

$$pG_2 = w \text{ and } \Pi_2 = 0 \ ; \quad (4.7)$$

both statements are directly the consequence of the firm's having maximized profits. Therefore the mathematical expression in (4.6) reduces to:

$$(1-t)w+0+\frac{t}{n}wn+0=w \ . \quad (4.8)$$

That is, in contemplating the counterfactual stipulated in the definition of additive Kantian optimization, the worker computes that an extra's days work will supply her *not* with the after wage $(1-t)w$, but with the *gross* wage, *w*. This is because what the worker loses in her personal tax bite, she regains in the increased value of the demogrant. So the appropriate optimization condition is that the worker equate her gross wage to the marginal rate of substitution between labor and consumption. But this is the condition for Pareto efficiency!

In contrast, the worker who Nash-optimizes assumes that if she increases her labor supply by a small amount, all other workers stand pat, and this means the worker's optimality condition is equality of the after –tax wage $(1-t)w$ and the marginal rate of substitution between labor and consumption. Thus is generated the dead-weight loss of taxation in the standard Arrow-Debreu model.

Do these equilibria exist? Under standard conditions, they do:

Proposition 4.2 *If the production function G is concave and satisfies standard conditions, and each utility function is strictly concave and the demand for the commodity is normal, then a Walras-Kant equilibrium exists for all tax rates* $0 \le t < 1$.[8]

[8] The 'standard conditions' on *G* are the Inada conditions and homotheticity; a consumption good is *normal* if an increase in income generates an increase in the good's consumption. These conditions are sufficient for equilibrium; they may not be necessary.

Furthermore, the propositions remain true with many firms and many goods. What happens if private citizens can also invest in the firm(s)? To preserve the first welfare theorem, we would have to require that the vector of investments also be an additive Kantian equilibrium. This strikes me as being a less plausible assumption than the labor-supply assumption[9].

5. Discussion

The first model of market socialism is credited to Oscar Lange and Fred Taylor (1938). I have reviewed their model elsewhere (Roemer [1994]), and the discussion it generated -- importantly, the critiques made by Friedrich Hayek (1940). I shall not do so again here, except to say that the model's principal feature was state control of investment. Lange and Taylor also proposed a kind of tâtonnement process, in which the central planning board imitated convergence of prices to what they imagined markets did in reality. That part of their model has been discredited by later work of general-equilibrium theorists, who have shown that the tâtonnement process does not in general converge to the equilibrium, even when one exists.[10] There have been several market-socialist models where the state or other public institutions carry out investment; the most recent contribution is by Giacomo Corneo (2017). The focus of these models has been upon equalizing the distribution of capital income in society: the state spends its profit income either on investment or public goods and services or demogrants. But these models have had relatively little to say about labor income, which in advanced economies comprises at least 60% of national income. Informally, we can imagine that the state would invest substantially more in compensating children from disadvantaged households to build up their skills, and it may also engage in redistributive taxation. However, these facets have not been formalized by market-socialist theorists in novel ways.

One must also mention the theory of labor management, in which workers own firms, and maximize valued added per worker. The most rigorous presentation is due to

[9] The results exposited in this section, in a somewhat more complicated model, are given in Roemer (in press).

[10] See Sonnenschein (1972), Debreu (1974), Mantel (1974).

Jacques Drèze (1989), who shows that the equilibria in a labor-managed economy are isomorphic to Arrow-Debreu equilibria (without taxation), and are hence Pareto efficient.

The novelty of the present approach is that it embeds a formal model of worker cooperation into a market economy, something that has not been done earlier, probably because no simple model of cooperation existed. Of course, forming a worker-owned firm, or having the state manage investment, can be thought of as a form of social cooperation, and so my latest statement may seem too narrow. What I intend to say is that *cooperative behavior in decision making* has not been formally injected into market socialist models heretofore, and it should be, if we take seriously the idea that *socialist ethos* is something quite different from capitalist ethos, and deserves a formalization. To repeat my earlier point in this context, the socialist ethos in the model of market socialism presented here is not altruistic, but cooperative.

Finally, I wish to address the question that most readers will undoubtedly have about the proposed market-socialist model – and indeed, about the rather complicated forms of behavior stipulated in multiplicative and additive Kantian optimization more generally. It might be useful to say that Nash optimization was probably viewed with some trepidation as a characterization of human behavior when Nash introduced it as a graduate student in mathematics in 1950. Von Neumann apparently rejected it out of hand when Nash explained it to him. We still lack an adequate theory of how players converge to a Nash equilibrium; often, a sequence of iterated best responses does converge to an equilibrium, but the weakness of this mathematical fact as an *explanation* of convergence is that at each step in the process, the assumption that players make that other players' actions are fixed is belied[11]. Yet we do seem to have ample evidence that many stable situations appear to be describable as Nash equilibria of social interactions that can be described as games. And as I've said, even behavioral economists, who challenge the most classical interpretation of these equilibria do not depart from describing them as Nash equilibria of games, albeit with exotic preferences.

I think three pre-requisites are necessary for believing that a society of workers could actually optimize their labor-supply decisions in the manner postulated by additive

[11] In like manner, a sequence of iterated best responses also often converges to a Kantian equilibrium. See Roemer (in press, chapter 7).

Kantian optimization: desire, understanding and trust. Workers must *desire* to cooperate, they must see themselves as part of a solidaristic venture, in which cooperation could further their interests. They must *understand* that if they act in concert as the Kantian optimization process stipulates, the results will be good, in the sense that the issue of efficiency can be separated from that of income distribution. Citizens can decide on how much redistribution they believe is warranted, due to the unfairness of the distribution of endowments that will necessarily exist, without having to compromise because too much redistribution might be inefficient. Thirdly, each must *trust* that if he optimizes in the Kantian manner so will most others: he will not be taken advantage of by Nash players.

We have many examples of Kantian behavior in history, and desire, understanding and trust surely characterized those occasions. Although the labor-supply decision may seem more complex in the additive-Kantian protocol, in fact it is not cognitively more difficult than Nash optimization. Each person need only know his own preferences, as in Nash optimization; and instead of standing pat when his marginal rate of substitution is equal to the net wage, he stands pat when it is equal to the gross wage. Perhaps more realistically, if workers are represented by unions in a national confederation, the union representatives can implement the additive Kantian equilibrium.

Granted, education will be required. At present, in the United States, many people do not seem to understand that their taxes are used for *any* useful purpose, let alone returned to them in transfer payments, services and public goods. Europeans, being less state-phobic, may be more receptive to understanding the virtues of Kantian optimization. This survey, however, has not attempted to address how we get there from here: rather, its more modest aim has been to offer a design that can motivate our imagination about a possible future.

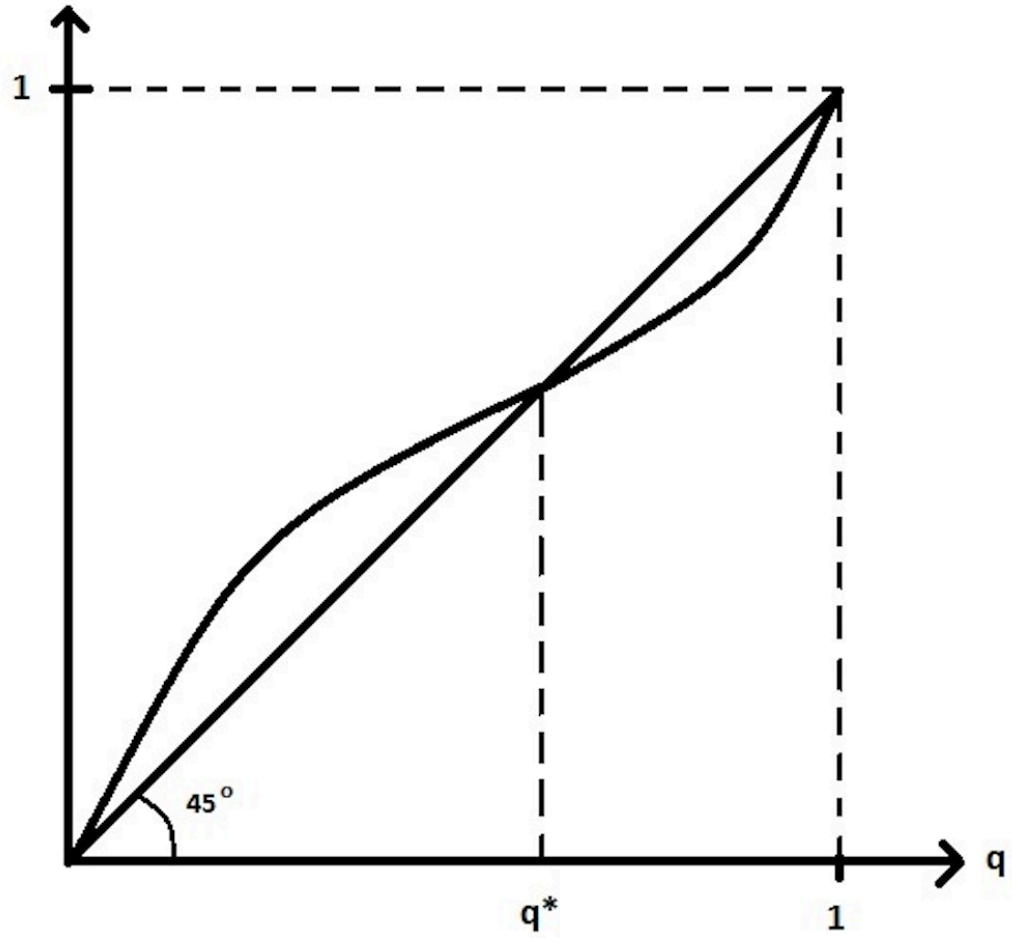

Figure 1 The curve is the distribution function of thresholds q. The stable equilibrium is that fraction q^* of the population recycles.

References

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

Barbera, S. and M. Jackson 2016. “A model of protests, revolution, and information,”

Cohen, G.A. 2009. *Why not socialism?* Princeton university Press

Corneo, G . 2018. *Is capitalism obsolete?* Harvard Univ. Press

Debreu, G. 1974. “Excess demand functions,” *J. Math. Econ.* 1, 15-21

Drèze,J. 1989. *Labour management, contracts and capital markets: A general equilibrium approach,* Oxford Univ. Press

Dunbar, R.I.M. 2009. “The social brain hypothesis and its implications for social evolution,” *Annals of human biology* 36, 562-572

Elster, J. 1981. “States that are essentially by-products,” *Social science information* 20, 431-473

Elster, J. 1989. “Social norms and economic theory,” *J. Econ. Perspectives* 3, 99-117

Elster, J. 2017. “On seeing and being seen,” *Social choice and welfare* 49, 721-734

Hayek, F.1940. “Socialist calculation: The competitive ‘solution’,” *Economica* 7, 125-149

Kandori, M. 1992. “Social norms and community enforcement,” *Review of economic studies* 59, 63-80

Lange, O. and F. Taylor 1938. *On the economic theory of socialism*, Univ. of Minnesota Press

Mantel, R. 1974. “On the characterization of aggregate excess-demand,” *J. Econ. Theory* 7, 348-353

Olson, M. 1965. *The logic of collective action* , Harvard University Press

Roemer, J. 1994. *A future for socialism,* Harvard Univ. Press

Roemer, J. 1996. *Theories of distributive justice,* Harvard Univ. Press

Roemer, J. In press. *How we cooperate: A Kantian explanation,* Yale University Press

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

Sonnenschein, H. 1972. “Market excess demand functions,” *Econometrica* 40, 549-563

Tomasello, M. 2014. *A natural history of human thinking*, Harvard UP

Tomasello, M. 2016. *A natural history of human morality,* Harvard UP

Word count: 6198; 11.14.17

Chapter 1. Cooperation, altruism and economic theory

1.1 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 (2014a, 2014b, 2016) argues that the only cooperative species among the five great apes (chimpanzees, bonobos, gorillas, orangutans, and humans) is our own[1]. 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. 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 mime; (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 prey I see, 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 to be selected in a competitive one[2]. 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[3] -- either with other chimpanzees, or with humans. Miming and pointing are the predecessors of language. Complex organs like the eye and language must have evolved incrementally as the result of selection from among many random mutations. Tomasello argues that language would not be useful, and therefore 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

---

1 Tomasello's view is extreme. Others, such as Frans de Waal (1996) and Philip Kitcher (2011) argue that limited cooperation exists in chimpanzees and other great apes.

2 See Kobayoshi and Kojima (2001).

3 Tomasello disagrees with some who argue that chimpanzees do cooperate in hunting smaller monkeys.

something in our toolkit. So language, were primitive forms of it to emerge in a non-cooperative species, would die out for lack of utility.

Tomasello conducts experiments in which he compares human infants to chimpanzees, who are set with a task in which cooperation would be useful. The general outcome of these experiments is that human infants (ten months or older) cooperate immediately, while chimpanzees do not. Often, the cooperative project that Tomasello designs in the lab 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 that might be acquired. Human infants, however, succeed immediately and repeatedly in cooperating in both the productive and consumptive phase of the project[4].

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 other social organizations, and complex social conventions, 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 be made more precise below.) We often explain these human achievements by the high 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 that tendency is inborn and was necessary for the development of the huge and complex cooperative projects that humans undertake.

---

[4] Formally, the game being played here is the game of 'chicken.' The issue is whether to share the captured food peacefully or to fight over it. In chapter 2, we show that the cooperative solution to the game of chicken is usually to share peacefully, but this depends upon the precise values of the payoffs.

Of course, Tomasello's claim (that humans are extremely cooperative great apes) does not fall if cooperation is learned through culture rather than transmitted genetically. In the former case, cooperation would be a meme, passed down in all successful human societies.

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 to produce a large brain, were it not necessary for 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, perhaps, 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.' See Dunbar (2009) for further elaboration of this hypothesis.

Readers, especially economists, 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) that 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 (2011), who attribute the participation of young men in warring parties against other tribes to their altruism towards co-tribals. I am skeptical that altruism is the key here, rather than cooperation.) Indeed, up until the present, human society has been characterized by increasingly complex states, within which cooperative behavior is pervasive, but between which there is lack of trust. Sharp competition between states (war) has been pervasive. So the human tendency to cooperate is, so it appears, not unlimited, but generally, as history has progressed the social units within which

cooperation is practiced have become increasingly large, now sometimes encompassing more than a billion humans.

1.2 Cooperation versus altruism

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 the individuals 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 an 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 all are in the same boat and understand that fact. 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 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, as the action *proceeds* from the solidaristic state.

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 venerable expression 'we all hang together or we will each hang separately' urges everyone to see that she does, indeed, have similar interests to others, and hence it may be logical to act cooperatively. Notice the quoted expression does not appeal to our altruism, but to our self-interest, and to the solidaristic state in which we find ourselves.

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 the discussion of cooperation among self-interested individuals from cooperation among altruistic ones; altruism will not be addressed until chapter 5 below.

Altruism and cooperation are frequently confounded in the literature. I do not mean the example I gave from Bowles and Gintis (2011), which explicitly views altruism as the characteristic that induces young men to undertake dangerous combat for their community. If they are right, this is a case of altruism's engendering cooperative action. I mean that writers often seem not to see a distinction between altruism and cooperation. The key point is that cooperation of an extensive kind can be undertaken because it is in the interest of *each,* not because each cares about others. I am skeptical that humans can, on a mass scale, have deep concern for others whom they have not even met, and so to base grand humanitarian projects on such a psychological propensity is risky. I do, however, believe that humans quite generally have common interests, and it is natural to pursue these cooperatively. (One can hardly avoiding thinking of the control of global greenhouse gas emissions as a leading such issue at present.) It seems the safer *general* strategy is to rely on the underlying motive of self-interest, active in cooperation, rather than on love for others, active in altruism.

The necessary conditions for cooperation are solidarity (in the sense of our all being in the same boat) and trust – trust that if I take the cooperative action, so will enough others to advance our common interest. Solidarity comes in different degrees – recall the familiar expression that first the tyrants come after the homosexuals and the Jews, then the gypsies… and eventually they come after *us*. The listener is being urged, here, to see that 'we are all in the same boat,' even if superficial differences among us may frustrate that understanding. Trust usually must be built by past experience of cooperation with the individuals concerned. Trust may be distributed in a somewhat continuous way in a population: some people are unconditional cooperators, who will cooperate regardless of the participation of others, some will cooperate when a certain

threshold is reached (say, 20% of others are cooperating), and some will never cooperate, even if all else are doing so. The common name we have for persons of the first kind is *saint.*

1.3 Cooperation and economic theory

Economic theory has focused not on our cooperative tendencies but on our competitive ones. Indeed, the two great theoretical contributions of micro-economics are both models of competition: the theory of competitive or Walrasian equilibrium, and game theory, with its associated stability concept, Nash equilibrium. It is clear that cooperation does not exist in the everyday meaning of the word in these theories. There is indeed nothing that can be thought of as social action. The kind of reasoning, or optimization, that individuals engage in in these theories is *autarkic:* other humans' actions are treated as parameters of the individual's problem, not as part of the action.

In general equilibrium theory, at least its most popular Walrasian version, individuals do not even observe what other people are *doing*: they simply observe the price vector and optimize against prices[5]. Prices summarize all the relevant information about what others are doing, and so it is superfluous for the individual to have specific information about others' actions. This indeed is usually championed as one of the beauties of the model – its ability to decentralize economic activity in the sense that each person need only know information about itself (preferences for humans, technologies for firms) and prices for Pareto efficiency to be achieved. To be precise, the 'achievement' of efficiency is an incomplete story, as it lacks dynamics: we only know that *if* an equilibrium is reached, it will be Pareto efficient, and the theory of dynamics remains incomplete. (The first theorem of welfare economics, which states that a competitive equilibrium is Pareto efficient, only holds under stringent and unrealistic conditions: economic problems that require cooperation, such as the financing of public goods and the regulation of public bads, are stipulated not to exist.) In the Nash equilibrium of a

[5] The Walrasian model is to be contrasted with the general-equilibrium model of Makowski and Ostroy (2001) who formalize the 19th century Austrian tradition in which equilibrium is produced by many bargaining games, where each attempts to extract as much surplus as she can from her opponents. Prices, for these authors, are what one sees after the 'dust of the competitive brawl clears,' and do not decentralize economic activity as with the Walrasian auctioneer. Their model cannot be accused of being asocial, although it is hyper-competitive.

game each player treats his competitors as inert: he imagines a counterfactual where he alone changes his strategy, while the others hold theirs fixed. A Nash equilibrium is a strategy profile such that each person's strategy is optimal (for himself) given the inertness of others' strategies. One can say that a Nash optimizer treats others as parameters of the environment, rather than persons like herself.

There is no doubt that general equilibrium and game theory are beautiful ideas; they are the culmination of what is probably the deepest thinking in the social sciences over the past several centuries. But they are not designed to deal with that aspect of behavior that is so distinctive of humans (among the great apes), our ability to cooperate with each other.

Economic theory does not entirely ignore cooperation, but attempts to fit it into the procrustean bed of the competitive model. Until behavioral economics came along, the main way of explaining cooperation – which here can be defined as the overcoming of the Pareto inefficient Nash equilibria that standardly occur in games – was to view cooperation as a *Nash* equilibrium of a complex game with many stages. (See Kandori (1992).) Think of a game like the prisoners' dilemma, where there is a cooperative strategy and a non-cooperative one. These strategies inherit their names from the fact that if both players play the cooperative strategy, each does better than if both play the non-cooperative one. In this well-known game, the unique Nash equilibrium is for both players to play the non-cooperative strategy. The complex stage game in which the one-shot prisoners' dilemma can be embedded stipulates that if a player fails to cooperate at stage *t,* then she is punished at stage $t+1$ by another player. However, punishment, being costly for the enforcer, is only carried out against non-cooperators in stage *t* if there is a stage $t+2$ in which those enforcers who fail to punish are themselves punished. The game must have an infinite number of stages, or at least an *unknown* number of stages, for this approach to support a cooperative equilibrium. For if it were known that the game had only three stages, say, then enforcers in the third stage would not punish the lazy enforcers who failed to punish in the second stage, because nobody would be around to punish *them* for failing to do so (there being no fourth stage). So those who are charged with punishing in the second stage will not do so (punishing being costly), and so a player can play the non-cooperative strategy in the first stage without fear of

punishment. Thus, with a known, finite number of stages, the good equilibrium (with cooperation) unravels.

But is this really the explanation of why people cooperate? Mancur Olson (1965) argued that it is. Workers join strikes only because they will be punished by other workers if they do not; they join unions not in recognition of their solidaristic situation, but because they are offered side-payments to do so.

Communities that suffer from the 'free rider problem' in the provision of public goods often do adopt punishment strategies to induce members to cooperate. Fishers must often control the total amount of fishing to preserve the fishery. Common-pool resources, like fisheries, are over-exploited, absent cooperation. Lobster fishermen in Maine apparently had a sequence of increasing punishments for those who deviated from the prescribed rules. If a fisherman put out too many lobster nets, the first step was to place a warning note on the buoys of the offending nets. If that didn't work, a committee went to visit him. If that didn't work, his nets were destroyed. Now consider the optimization problem of those who were appointed to do these acts of warning and punishment. If they failed in their duty, there must be another group who was charged with punishing them – or perhaps this would be accomplished simply by social ostracism. But is it credible that the whole system was maintained although *everyone* was in fact optimizing in the autarkic Nash way, carrying out his duty to punish only because of fear of punishment should he shirk in this duty? I am skeptical. It is perhaps more likely that there were may who were committed to implementing the cooperative solution, many who did not require the threat of punishment to take the cooperative action, at any stage of the game. The complex equilibrium in which cooperation is maintained by a elaborate chain of punishments is, I think, too fragile to explain the real thing. The explanation is Ptolemaic, an effort to fit an observed phenomenon into a theory that cannot explain it in a simple way.

Elster (2017) introduces useful distinctions. A *social norm* is a behavior that is enforced by punishment of those who deviate from it, and those who observe the deviation and fail to punish the deviator are themselves punished by others who observe this. A social norm is thus a Nash equilibrium of a game with stages, in which those who fail to cooperate are punished, and so on and on. A person obeys a social norm because

he is afraid of *being seen* if he fails to, and hence punished by the observer. In contrast, a *quasi-moral norm* is one that is motivated by wanting to do the right thing. But the 'right thing' is defined in large part by what others do. If I observe that most others are recycling their trash, and therefore I recycle, I am behaving according to a quasi-moral norm. In this case, I cooperate not because I am afraid of being seen should I fail to; rather, I cooperate because I *see others* taking the cooperative action. A *moral norm* is, in contrast, unconditional. I take the cooperative action regardless of what others are doing. The Kantian categorical and hypothetical imperatives are moral norms. The behavior of the lobster fisherman described above could be a social norm or a quasi-moral norm. It is unlikely that it constitutes a moral norm. Because I believe trust is a necessary condition, I view cooperation as a quasi-moral norm. For trust is established by observing that others are taking the cooperative action, or have taken similarly cooperative actions in the past.

The second place where we find cooperation addressed in neoclassical economic theory is in the theory of cooperative games. A cooperative game with a player set N is a function v mapping the subsets of N into the real numbers. Each subset $S \in 2^N$ is a coalition of players, and the number $v(S)$ is interpreted as the total utility (let us say) that S's members can achieve by cooperation among themselves. A solution to a cooperative game is way of assigning utility to the members of N that does not violate the constraint that total utility cannot exceed $v(N)$. For instance, the *core* is the set of 'imputations' or utility allocations such no coalition can do better for itself by internal cooperation. If $(x^1,...,x^n)$ is a utility imputation in the core, then the following inequality must hold:

$$(\forall S \in 2^N)(v(S) \leq \sum_{i \in S} x^i) \ . \qquad (1.1)$$

While cooperation is invoked to explain what coalitions can achieve on their own, the core itself is a competitive notion: the values $v(S)$ are backstops that determine the nature of competition among the player set as a whole. It is therefore somewhat of a misnomer to call this approach 'cooperative.' Indeed, 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.

Indeed!

Behavioral economists have challenged this unlikely rationalization of cooperative behavior as a Nash equilibrium of a complex game with punishments by altering the standard assumption of self-interested preferences. There are many versions, but they share in common the move of putting new and 'exotic' arguments into preferences – arguments like a concern with fairness (Fehr [1999] and Rabin [2003]),or giving gifts to one's opponent (Akerlof [1968]) , or of seeking a warm glow (Andreoni [1990]). Once preferences have been so altered then the cooperative outcome can be achieved as a *Nash* equilibrium of the new game. Punishments may indeed be inflicted by such players against others who fail to cooperate, but it is no longer necessarily costly for the enforcer to punish, because his sense of fairness has been offended, or a social norm has been broken that he values. Or he may even get a warm glow from punishing the deviator! I will discuss these approaches more below. My immediate reaction to them is that they are too easy – in the sense of being non-falsifiable. The invention of the concept of a preference order is extremely important, but one must exercise a certain discipline in using it. Just as econometricians are not free to mine the data, so theorists should not allow everything ('the kitchen sink') to be an argument of preferences. It is, of course, a personal judgment, to draw the line as I have suggested it be drawn.

If the undisciplined use of preferences were my only critique of behavioral economics, it might be minimized. A more formidable critique, I think, is that the trick of modifying preferences only works – in the sense of producing the 'good' or cooperative Nash equilibrium – when the problem is pretty simple. ('Simple' usually means a player has only a few strategies, and that the 'cooperative' strategy is obvious to everyone. This is true in most 2 x 2 matrix games. In laboratory games involving the voluntary contribution to a public good, and in ultimatum and dictator games, there are many strategies but it is nevertheless clear what the cooperative action is.) If we consider, however, the general problem of the tragedy of the commons in common-pool resource games, the cooperative strategy profile –in which each player plays her part of a

Pareto-efficient solution – is not obvious. Either some kind of *decentralization of cooperation* is needed, or cooperation must be organized by a central authority.

Just as the Walrasian equilibrium of a market economy is not obvious to anyone, and requires decentralization, so does cooperation with any degree of complexity. Although we have many examples of cooperation that are organized by a central authority, it is surely the case that the vast majority of cases of cooperation in human experience are not centrally organized. A normal person encounters hundreds of situations a year, in which cooperation would be profitable, but is not centrally organized. How, then, do people manage to cooperate in these cases?

I do not believe the strategy of behavioral economics supplies *micro-foundations for cooperation* of a general kind. And if cooperation is a major part of what makes us human, we should be looking for its general micro-foundations.

1.4 Simple Kantian optimization

This book will offer a partial solution to the problem of specifying micro-foundations for cooperation, which I call Kantian optimization, with its concomitant concept of Kantian equilibrium. The new move is, instead of altering preferences from classical, self-interested ones, to alter *how people optimize.* In the simplest case, consider a symmetric game. A two-person game is symmetric if the payoff matrix is symmetric, as in the prisoners' dilemma of figure 1.1.

	A	*B*
A	(1,1)	(-1,2)
B	(2,-1)	(0,0)

Figure 1.1 The payoff matrix of a prisoners' dilemma game. The first number in parentheses is the payoff to the row player, and the second number is the payoff to the column player.

A symmetric game is one in which players are identically situated: they are all in the same boat. In the game of figure 1.1, a Nash optimizer asks himself, " Given the strategy chosen by my opponent, what is the best strategy for me?" The answer, regardless of the opponent's choice, is that I should play *B* . *B* is a 'dominant strategy' in the language of game theory. But a Kantian optimizer – so I propose -- asks "What is the strategy I would like both of us to play?" Clearly the answer is *A* because I do better if we both play *A* than if we both play *B*. It is not relevant to me that *you* also do better when we bot play *A* – altruism is not my motivation. It is, however, important that I understand the symmetry of the game, and hence know that the answer to the proposed question is the same for both of us.

It is the symmetry of the situation that naturally suggests we ask the Kantian question. Tomasello argues that the ability to cooperate is founded in our ability to form 'joint intentionality.' My interpretation of this concept is that we each think 'what would I like each of us to do?', and *if* we trust each other, we understand that each of us is thinking in this way, and will behave in the way the answer instructs. I will elaborate on this in chapter 2.

Definition 1.1 In a symmetric game, the strategy that *each* would prefer *all* to play is a *simple Kantian equilibrium* (SKE)[6].

Invoking Kant is due to his categorical and hypothetical imperatives, stating one should take those actions one would like to see universalized[7]. I understand that it would be more precise to call this 'quasi-moral optimization,' because Kant's imperatives are unconditional, as mine is not. I opt, however, for the more imprecise 'Kantian' nomenclature because there is a history of using it in economics, as I review in section

---

[6] Symmetry of the game is clearly sufficient for the existence of a simple Kantian equilibrium. It is, however, not necessary. Consider a prisoners' dilemma, which is asymmetric (the off-diagonal payoffs are not symmetric across the two players). It remains the case that both players prefer (cooperate, cooperate) to (defect, defect). If the strategy space consists of only these two strategies, then (cooperate, cooperate) is an SKE. If, however, the game is one with mixed strategies, an SKE may not exist.

[7] "Act always in accordance with that maxim whose universality as a law you can at the same time will. (Kant, 2002)" It may be more textually accurate to justify the Kantian nomenclature by invoking Kant's hypothetical imperative. I use the term for its suggestive meaning, and do not wish to imply that there is a deeper, Kantian justification of my proposal.

2.7, and because it is aptly described by Kant's phrase, "Take those actions you would will be universalized," even if Kant meant this in an unconditional way.

The concept of Kantian equilibrium will be generalized beyond the case of symmetric games later, but it is useful to consider these games first, as they are the simplest games. Many laboratory experiments in economics involve symmetric games, and it is in symmetric games that Kantian optimization takes its simplest and most compelling form.

It is important to note that the Kantian optimizer asks what common strategy (played by all) would be *best for him*: he is not altruistic, in thinking about the payoffs of others. To calculate the strategy he would like everyone to play, he need only know his own preferences. But to invoke joint intentionality, he must also know that others are similarly situated – that is, that the game is symmetric. This implies that the common strategy that is best for him is also best for others, a fact that does not appeal to his perhaps non-existent altruism, but motivates his expectation that others will act in like manner. That expectation, however, must also be engendered by trust, or an experience of past cooperation.

What I emphasize is that cooperation, in this view, is achieved not by inserting a new argument into preferences, such as altruism or a warm glow, but by conceptualizing the optimizing process in a different way. These are different ways of modeling the problem – one involves altering preferences but keeping the Nash optimization protocol, and the other involves keeping preferences classical but altering the optimization protocol. Despite the conceptual distinction, it may be difficult to test which model better explains the reality of cooperation, a problem to which we will return.

A quite *different* question, to which I have no complete answer, is when, in a game, do players choose to invoke the Kantian protocol and when the Nash protocol. Often, I believe, this depends upon the degree of trust in the other players. Of course, trust is irrelevant for a Nash optimizer.

1.5 Some examples

I conclude this chapter with several examples of what I believe to be Kantian optimization in real life.

A. Recycling. In many cities, many or most people recycle their trash. There is no penalty for failing to do so. Often, others do not observe if one does not recycle. The cost of recycling may be non-trivial – certainly greater than the marginal benefit in terms of the public good of a clean environment one's participation engenders. Andreoni's (1990) view, that one cooperates in order to receive a 'warm glow,' is an example of explaining recycling by inserting an exotic argument into preferences. I think this puts the cart before the horse: one may indeed enjoy a warm glow, but that's *because* one has done the right thing – that is, taken the action one would like all to take. The warm glow is an unintended by-product of the action, not its cause. Suppose I help my child with her algebra homework: she masters the quadratic formula. I feel a warm glow. But seeking that glow was not my motivation: it was to teach her algebra, and the warm glow follows, unintendedly, as a consequence of success in that project. While recycling may be a quasi-moral norm, teaching my daughter algebra is probably due to altruism. In either case, I find the 'warm glow' no explanation at all.

B. 'Doing one's bit' in Britain in World War II. This was a popular expression for something voluntary and extra one did for the war effort. Is it best explained by seeking the respect or approval of others, or doing what one wished everyone to do? For some this, could be a social norm, punished, if avoided, by ostracism. For others, it was a quasi-moral norm, done because it was the right thing to do, as evidenced by what others were doing.

C. Soldiers protecting comrades in battle. This can be a Kantian equilibrium, but also could be induced by altruism. One becomes close to others in one's unit. In this case, the Kantian equilibrium is also an instance of the golden rule – "Do unto others what you would have them do unto you." Golden-rule optimization is a special case of simple Kantian equilibrium.

D. Voting. The voting paradox is not one from the Kantian viewpoint: I vote because I'd like everyone to vote, rather than not to vote, to contribute to the public good of democracy. A somewhat different form is that I vote because I would like everyone *similarly situated* to me (that is, sharing my politics) to vote.

E. Paying taxes. It has often been observed that the probability of being caught for tax evasion and the penalties assessed for doing so are far too small to explain the relatively small degree of tax evasion in most advanced countries. In most countries (though not all), tax cheaters are not publicly identified, so shame (an exotic argument in preferences) is not an issue. Elster (2017), however, points out that in Norway, everyone's tax payment is published on the internet, and this increases compliance. A caveat to the example is that the practice of withholding tax owed minimizes the possibility of evasion.
F. Tipping. A practice viewed by some as a paradox (Gambetta (2015)) is not one from the Kantian viewpoint: here, there is an altruistic element, but it is not the interesting part of the behavior. The thought process is that I tip what I would like each to tip. I understand what I think it's proper to tip by observing what the custom is – hence, the quasi-moral nature of the behavior.
G. Charity. The Nash equilibrium is often not to donate, even if I value the public good produced. There is a Kantian and a Rawlsian explanation of charity: the Kantian gives what he'd like all others (like him) to give. For the Rawlsian, charity is the random dictator game: behind the veil of ignorance, who will be the donor and who the recipient of charity? These two ways of looking at the problem generate different levels of charity (I may give much more in the so-called Rawlsian version). My conjecture is that the so-called Kantian thought process is more prevalent[8].

I have organized the book as follows. Part 1, comprising chapters 2 through 10, studies Kantian optimization in games. The main result is that in many cases, Kantian optimization solves the two major problems that afflict Nash equilibrium: the inefficiency of equilibrium in the presence of congestion externalites, known as the *tragedy of the commons*, and the inefficiency of equilibrium in the presence of public goods or positive externalities, known as *the free-rider problem*. In two important classes of games – those with positive and negative externalities -- Kantian equilibrium is Pareto efficient. Moreover, we will see that in such games, Nash equilibrium is *always* Pareto inefficient. So Kantian optimization 'solves' what must appear as the two greatest failures of Nash optimization, from the viewpoint of human welfare.

---

[8] Readers should not be distracted by the fact that Rawls called himself a Kantian. He was referring to his attempt to construct justice as a corollary to rationality, not to the specific use of the hypothetical imperative in daily decisions.

In Part 2, chapters 11 through 14, I apply Kantian optimization to market economies: that is, I embed cooperation in general-equilibrium models. I show how the problem of controlling global carbon emissions can be decentralized using a cap-and-trade regime, as a 'unanimity equilibrium;' how Kantian optimization in the labor-supply decision by workers in a 'market-socialist' economy produces Pareto efficient equilibria with any desired degree of income redistribution, which is to say that the equity-efficiency trade-off dissolves; how public goods can be produced efficiently in a market economy; and how an economy consisting of worker-owned firms can achieve efficient equilibria, again with many degrees of freedom in the distribution of income, using Kantian optimization. Chapter 15 offers some final reflections.