Word count: 35,338
This version: December 14, 2015

(tentative and awkward title)

How we (do and could) cooperate

by

John E. Roemer
Yale University
john.roemer@yale.edu

Words: 245

Table of Contents

Chapter 5 Altruism

5.1 Altruistic preferences and Pareto efficiency

5.2 Kantian equilibrium and efficiency

Chapter 6 Can one rationalize Kantian optimization as Nash optimization where players have extended preferences?

Chapter 7 Existence and dynamics of Kantian equilibrium

7.1 Existence of strictly positive $K^{\times}$ equilibria for production economies

7.2 Existence of K^{β} equilibria for $0 < \beta \leq \infty$

7.3 Is there an allocation rule that Nash equilibrium implements efficiently on the domain $\tilde{\mathbf{D}}$?

7.4 Dynamics

Chapter 8 Evolutionary considerations

8.1 The random dictator game

8.2 The prisoners' dilemma

8.3 Summary

Chapter 9 Alternative approaches to cooperation

9.1 The traditional approach

9.2 Strong reciprocity

9.3 Conditional cooperation

9.4 Rabin and the kindness function

9.5 *Homo moralis* (Alger and Weibull)

9.6 Surveys and interviews of players

9.7 The acequias of northern New Mexico (to be written)

Chapter 10 A generalization to more complex production economies

Final Remarks

Word count: 4207

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) argues, I think persuasively, that the only cooperative species among the five great apes (chimpanzees, bonobos, gorillas, orangutans, and humans) are the humans. Tomasello believes that the tendency to cooperate with other humans is inborn. He offers a number of examples of our features and behavior that are unique to humans among the five great apes, indicating that the tendency to cooperate must have evolved very early. Here are three such examples: (1) among the great apes, humans are the only beings with sclera (the whites of the eyes); (2) only humans point and pantomime; (3) only humans have language. The conjecture is based on the fact that it is the sclera of the eye that enables you to see what I am looking at. If I am looking at an animal that would make a good meal, and if you and I cooperate in hunting, it is useful for me that you can see the prey I am looking at, because then we can catch and consume it together. Were you and I only competitors it would not be useful for me that you see the object of my gaze, as we would then fight over who gets the animal. Thus, one would expect the mutation of sclera to be selected in a cooperative species, but not to be selected in a competitive one[1]. 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[2] -- 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 something in our toolkit. So language, were

---

[1] See Kobayoshi and Kojima (2001).

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

primitive forms of it to emerge in a non-cooperative species, would die out for lack of utility.

Tomasello's main work consists of experiments in which he compares human infants to chimpanzees, who are set with a task in which cooperation would be useful. The general outcome of these experiments is that human infants (ten months old 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 acquired. Human infants, however, succeed immediately and repeatedly in cooperating in both the productive and consumptive phase of the project[3].

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 become 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 the tendency to cooperate is inborn and was necessary for the development of the huge and complex cooperative projects that humans undertake.

---

[3] Formally, the game being played here is the game of chicken. The issue to 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 the uniquely cooperative great ape) 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.'

Readers may object: cooperation, they might say, is fairly rare among humans, who are mainly characterized by competitive behavior. Indeed, what seems to be the case is that cooperation evolves in small groups – families, tribes – but that these groups are often at war with one another. Stone-age New Guinea, which was observable up until around the middle of the twentieth century, was home to thousands of tribes (with thousands of languages) who fought each other; but within each tribe, cooperation flourished. (One very important aspect of intra-tribal cooperation among young men was participating in warfare against other tribes. See Bowles and Gintis (2011), who attribute the participation of young men in warring parties against other tribes as due to their altruism towards co-tribals. I am skeptical that altruism is the key here, rather than cooperation.) Indeed, up until the middle of the twentieth century (at least), human society has been characterized by increasingly complex states, in which cooperative behavior is pervasive internally, but between whom 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 over 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. I take 'a union of interests' to mean we are all in the same situation and have common preferences. It does not mean we are altruistic towards each other. Granted, one might interpret 'a union of …sympathies' to mean altruism, but I choose to focus rather on 'a union of purpose or interests.' The Wells quote clearly indicates the distinction between the joint action and the state of solidarity, as the action *proceeds* from the solidaristic situation.

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

My claim is that the ability to cooperate for reasons of self-interest is less demanding than the prescription to care about others. I believe that it is easier to explain the many examples of human cooperation from an assumption that people learn that cooperation can further their own interests, than to explain those examples by altruism. For this reason, I separate the discussion of cooperation among self-interested individuals from cooperation among altruistic ones; the latter topic will be addressed mainly in chapter 8.

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. 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 an emotion 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 greenhouse gas emissions as a leading such issue at present.) It seems a safer *general* strategy to rely on the underlying motive of self-interest, active in cooperation, 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… and finally they come after *us*. The listener is being urged, here, to see that 'we are all in the same boat,' even if differences among us may frustrate that understanding. Trust usually must be built by past experience of cooperation with the individuals concerned. Trust is usually 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.

(More on this in chapter 5.) The general 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 economics are both models of competition: the theory of competitive or Walrasian equilibrium, and game theory, with 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.*

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[4]. Prices summarize all the 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 has to know only information about itself (preferences for humans, technologies for firms) 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. (Of course, the first theorem of welfare economics 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 game each player treats his competitors as inert: he imagines a counterfactual where he alone changes his strategy, the others' holding theirs

---

[4] 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.

fixed. A Nash equilibrium is a strategy profile such under each person's strategy is optimal (for himself) given the inertness of others' strategies.

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 apparently unique to humans (among the great apes), our ability to cooperate with each other.

Of course, economic theory does not ignore cooperation, but it attempts to fit it into the procrustean bed of competition. 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. Players are induced to take the cooperative action for fear of being punished by others in the next stage if they fail to do so; and punishment, being costly for the enforcer, is only carried out against shirkers in the first stage if there is a third stage in which those enforcers who fail to punish are themselves punished. Clearly, 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 deviators 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), which means that people can deviate from cooperation in the first stage without fear of punishment. Thus, with a known, finite number of stages, the good equilibrium (with cooperation) unravels. For the sake of specificity, the reader can think of the repeated prisoners' dilemma as an example where playing the 'cooperative' strategy can be enforced with an infinite number of stages.

But is this really 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. 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 or punishment. If they failed in their duty, there must be another group who were 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? I think not. There must have many who were committed to implementing the cooperative solution, many who did not require the threat of punishment to behave properly, at any stage of the game.

The explanation of cooperation as a Nash equilibrium of a game with punishments seems Ptolemaic to me. It is an effort to fit an observed outcome into a theory that indeed cannot explain it in a simple way.

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 which 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 in the core, then the following inequality must hold:

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

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 responded to this unlikely rationalization of cooperative behavior as a Nash equilibrium of a complex game with punishments by fiddling with 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. 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 was a wonderful conceptualization, 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. This is, of course, a personal judgment that can be challenged.

If this were the 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' 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.) If we consider, however, the general problem of the tragedy of the

commons in common-pool resource games, the cooperative strategy – that is, the one that is part of a Pareto-efficient solution – is not obvious. Some kind of decentralization of cooperation is needed, just as the Walrasian equilibrium of a market economy is not obvious to anyone, and requires decentralization.

In other words, the moves of behavioral economics do not supply , as far as I know, *microfoundations 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 microfoundations.

1.4 Simple Kantian optimization

This book will offer a partial solution, which I call Kantian optimization, with its concomitant concept of Kantian equilibrium. The new move is not to fiddle with preferences but with *how people optimize.* In the simplest case, the game is symmetric. A two-person game is symmetric if the payoff matrix is symmetric. In a symmetric game, each player asks himself, "What is the action I would like all of us to take?" Suppose players have a common strategy set S, which is an interval of real numbers. Let the payoff function of player *i* be a function V^i of the profile of strategies of all players. Denoting strategies as $p,q \in S$ then each of the *n* players solves the problem:

$$\max_{p \in S} V^i(p,p,...,p) \ . \tag{1.2}$$

Each will have the same optimal solution to this problem, some p^*, because all the V^i coincide on the diagonal of S^n in a symmetric game.

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

Invoking Kant is due to his categorical imperative, stating one should take those actions one would like to see universalized. The concept of Kantian equilibrium will be generalized beyond the case of symmetric games below, but it is useful to consider these games first – for one, laboratory experiments in economics almost all involve symmetric

games, and secondly, it is in symmetric games that Kantian optimization takes its simplest form.

It is important to note that the Kantian optimizer asks what common strategy would be *best for him*: he is not altruistic, in thinking about the payoffs of others. He need only know his own preferences and that the others have the same set of strategies. Assuming he does know the game is symmetric, he also knows, however, that this strategy will be chosen by all others who ask the same question. And surely that question is motivated by our 'all being in the same boat,' here modeled as the game's being symmetric. I do not deny that a social norm is involved: but the norm is not an argument of preferences (we may, for instance, assume the payoff functions are those of the prisoners' dilemma), but rather is in *how* we optimize, which is to ask what is the *common strategy* we would (each) like played. I do not wish to motivate this kind of behavior as magical thinking: I am not proposing that players think 'if I decide to this , then everyone situated like me *will* also so decide.' Rather, it is symmetry of the situation that implies that we *should* all do the same thing, and therefore, of course, it should be the best 'same thing' that we can do – best for me (and, as it happens, for you, too).

Words: 4571

Chapter 2. Simple Kantian equilibrium

To keep the exposition simple, I will assume that the games in sections 2.1 and 2.2 have two players. Generalization to *n* players is straight-forward.

2.1 Monotonicity and Pareto efficiency

In a symmetric game with two players, the payoffs are $V(p,q)$ and $V(q,p)$ for players One and Two, respectively, for some common payoff function V, where p (q) is the strategy of player One (Two), assumed here to be chosen from a set of real numbers, S. The game is (*strictly) monotone decreasing* if each player's payoff is (strictly) monotone decreasing in the strategy of the other player(s); it is (*strictly) monotone increasing* if it is strictly monotone increasing in the strategy of the other player(s).

Definition 2.1 A game is (*strictly)monotone* if it is either (strictly) monotone increasing or (strictly) monotone decreasing.

Proposition 2.1 *In a symmetric strictly monotone game, the SKE is Pareto efficient.*
Proof:
Let the game be strictly monotone increasing. Let p^* be an SKE, and suppose it is Pareto-dominated by (p,q), so:

$$V(p,q) \geq V(p^*,p^*) \text{ and } V(q,p) \geq V(p^*,p^*)$$

with at least one inequality strict. Obviously $p \neq q$, for otherwise we contradict the assumption that p^* is a SKE. Suppose $p < q$. Then:

$$V(q,q) > V(q,p) \geq V(p^*,p^*),$$

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

An analogous argument works if the game is strictly monotone decreasing.

■

2.2 Two-person symmetric games

The prisoners' dilemma is given by the payoff matrix below. In the discrete version of the game, call 'Defect' strategy 0 and 'Cooperate' strategy 1. Then the game is strictly monotone increasing, and so the simple Kantian equilibrium, which is (Coop, Coop), is Pareto efficient (by Proposition 2.1). If we move to mixed strategies, where the strategy space is $S = [0,1]$ then the equilibrium depends on the payoff matrix, which is, in general form[1]:

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

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

The PD game is strictly monotone increasing: just note that

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

It follows immediately from Proposition 2.1 that the SKE of the mixed-strategy PD game is Pareto efficient.

Proposition 2.2

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

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

---

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

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

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

Proof:

Part *a* follows from Proposition 2.1 since the PD game is strictly monotone increasing. The function $V(p,p)$ is concave if and only if $c-b\le 1$. In this case the first-order condition $\frac{d}{dp}V^{PD}(p,p)=0$ gives the SKE. If $1\le c$ the solution is a corner one, at $p^*=1$ (part *b*). If $c<1$, the solution is interior, and given by part *c*. If $c-b>1$, the function $V^{PD}(p,p)$ is convex, and hence the SKE occurs either at $p=0$ or 1 . The value is higher at $p=1$, giving part *d*. ■

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

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

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

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

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

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

We finally consider the 'battle of the sexes.' For the game to be symmetric (i.e., for $V^{Row}(p,q) = V^{Col}(q,p)$) we must write the payoff matrix unconventionally, as follows:

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

That is, the 'first' strategy for the Row player ("He") is the event he prefers, and the first strategy for the Column player ("She") is the event she prefers. The game has two parameters, (a,b) where $0 < b < a < 1$. The payoff function for the row player is $V^{BS}(p,q) = bpq + p(1-q) + aq(1-p)$ and the column player's payoff is $V^{BS}(q,p)$. The simple Kantian equilibrium in pure strategies is (Dance, Box). It is not Pareto efficient, being dominated by both (Dance, Dance) and (Box, Box).

The reader can check that the BS game in mixed strategies is not a monotone game. We have:

Proposition 2.3

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

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

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

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

Proof:

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

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

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

The Nash equilibrium of the mixed-strategy BS game is computed from the first-order conditions for Nash equilibrium. Write $V^{BS}(p,q)=p(bq+1-q-aq)+aq$. Therefore, Row's best response to q is:

$$p=\begin{cases} 1, \text{ if } bq+1-q-aq>0 \\ 0, \text{ if } bq+1-q-aq<0 \\ [0,1], \text{ if } bq+1-q-aq=0 \end{cases}$$

It follows that (1,1) is not a Nash equilibrium, because if $q=1$, the best response of Row is 0. Likewise (0,0) is not a Nash equilibrium, because if $q=0$ the best response of Row is 1. The only Nash equilibrium occurs in the third case, when $p=q=\dfrac{1}{1+a-b}$.

■

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

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

For Nash equilibrium, it does not matter in which order we write the strategies. But for Kantian equilibrium it does, because Kantian optimization requires a conception of which strategies are the 'same' for the two players. In the above formulation, of the battle of the sexes, we identified the first strategy for the two players as the event that he or she preferred. If we write the payoff matrix in its traditional form, then the payoff matrix is:

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

The game in this form is not symmetric. We cannot suppose that a simple Kantian equilibrium exists, and in fact one does not exist. His payoff function is now $\hat{V}(p,q) = pq + bp(1-q) + (1-p)(1-q)a$, and $\hat{V}(p,p)$ is maximized at $p = 1$. Her payoff function is maximized at $q = 0$, and so a SKE , indeed, does not exist.

2.3 Some simple non-symmetric games

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

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

It is not obvious how to model cooperation in the ultimatum game. This is the first time we have encountered a game where the strategy is multi-dimensional. It seems to me a Kantian should think as follows. If I were chosen the ultimator, and were to propose to keep x, this must be the amount I would also like the other person to keep, were she chosen to be the ultimator, and hence I must accept any amount from her that is at least $1-x$. Therefore, $z \leq 1-x$. Consequently, the simple Kantian solution solves the program:

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

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

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

These games demonstrate what is a general feature of Kantian optimization in stage games. *The notion of subgame perfection does not apply.* Fairness enters not as an argument of preferences, but as the realization that either player could have been chosen by Nature to be the first. Thus a Kantian optimizer in these games ask: How would I like each of us to play if each of us could be chosen to be the first or second player?

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

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

Suppose a player's von Neumann-Morgenstern utility function for money lotteries is u. Before the game begins, her expected utility is $\frac{1}{2}u(M-x+y)+\frac{1}{2}u(M+ax-y)$. She chooses a strategy (x,y) that she would like both players to choose, which is the one that maximizes her expected utility:

$$\max \frac{1}{2}u(M-x+y)+\frac{1}{2}u(M+ax-y)$$
$$\text{s.t.}$$
$$0 \le x \le M$$
$$0 \le y \le M+ax$$

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

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

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

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

second, and the second responded Nash, by returning zero to the first. The authors conduct interviews with the participants after the conclusion of the game, and discover, unsurprisingly, that playing $x = M$ is associated with having trust in others.

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

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

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

My interpretation of the Cox, Ostrom (2009) results for the trust-game experiment is that about one-third of the players chosen to be first movers were playing (their part) of the simple Kantian equilibrium, because they had trust in their opponents/partners. About 27% of their partners responded by playing (their part) of the Kantian equilibrium.

Another 54% of the second players in these matches shared the gains induced by the first players' transfers, but did not share as much as the simple Kantian equilibrium prescribes; none of the second players in these matches played the Nash solution in the subgame that they faced (i.e., returning nothing to the first player). A smaller fraction of players appear to be using autarkic optimization. I cannot reject the hypothesis that a significant number of individuals are Kantian optimizers. I see no reason to suppose that group welfare motivated anyone.

2.4 Some examples of simple Kantian equilibria

A. Recycling. In many cities, many or most people recycle. There is no penalty for failing to do so. 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 one's participation produces. Most of the behavioral-economic explanations listed earlier do not explain this: Andreoni's is an exception. Perhaps one recycles in order to get a 'warm glow.' I think this puts the cart before the horse: one may indeed get a warm glow, but that's *because* one has done the right thing. 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.

B. 'Doing one's bit' in Britain in World War 2. 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?

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.

D. Voting. The voting paradox is not one from the Kantian viewpoint: I vote because I'd like all others to vote, as a contribution to the public good of democracy.

E. Paying taxes. It has often been observed that the probability of being caught for evasion and the penalties are far too small to explain the relatively small degree of tax

evasion in some countries. In most countries (though not all), tax cheaters are not publicly identified.

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.

G. Charity. The Nash equilibrium is not to donate. 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? 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.

H. Do unto others as you would have them do unto you. Strictly speaking, however, the golden rule and reciprocity more generally require repetition, while many of the practices described above do not (*C* is an exception).

2.5 Economies with production

We introduce simple production economies. There are *n* producers, each with a concave utility function u^i defined over consumption (*x*) and effort (*E*). Effort is measured in efficiency units (if s is a person's skill level and he exerts *E* units of efficiency effort, then his labor time is E/s). Production is defined by a concave function mapping total units of efficiency labor into total output. Defining $E^S = \sum E^i$, then total output at the effort vector $\mathbf{E} = (E^1,...,E^n)$ is $G(E^S)$.

Suppose we consider a fishing economy: fishers fish on a lake, and there are decreasing returns to scale in labor expended fishing, due to congestion effects. Each fisher keeps his catch, so the *proportional allocation rule* is

$x^i = X^{\Pr,i}(E^1,...,E^n) = \frac{E^i}{E^S} G(E^S)$: that is, except for noise, the fish caught by a fisher will be proportional to the labor in efficiency units he expends. Traditionally, this allocation

rule has been used in fishing communities. Given preferences, technology, and the allocation rule, a game is defined, where the payoff for fisher *i* at an effort allocation is:

$$V^i(E^1,E^2,...,E^n) = u^i(\frac{E^i}{E^S}G(E^S),E^i) \tag{2.1}$$

In this chapter, we assume homogeneous preferences, and so $u^i = u$ for all *i.*

It is well-known that, if *G* is strictly concave, then the Nash equilibrium of this game is Pareto inefficient. Fishers fish too much: each does not take into account the fact that his labor contributes to a public bad, the reduction of the productivity of the lake. The Nash equilibrium of the game $\{V^i\}$ is given by:

$$(\forall i) \quad -\frac{u_2[i]}{u_1[i]} = \frac{E^i}{E^S}G'(E^S) + (1-\frac{E^i}{E^S})\frac{G(E^S)}{E^S} \ , \tag{2.2}$$

where $u_j[i]$ is the j^{th} derivative of *u* evaluated at the consumption bundle of individual *i.* This says that the marginal rate of substitution for each player is equal to a convex combination of the marginal product ($G'(E^S)$) and the average product. But the condition for Pareto efficiency at an interior solution is:

$$(\forall i)\ MRS^i \equiv -\frac{u_2[i]}{u_1[i]} = MRT = G'(E^S) \ . \tag{2.3}$$

Only in the case where *G* is linear (and so the average and marginal products are equal) does (2.2) reduce to (2.3). In general the MRS^i is greater than the MRT (because the marginal product is less than the average product for strictly concave *G*) and each fisher could benefit from a reduction in her effort. This example is the simplest form of the 'tragedy of the commons' (Hardin [1968]).

Now let's compute the SKE for this game. Each fisher solves the problem:

$$\max_E u(\frac{E}{nE}G(nE),E) \ ; \tag{2.4}$$

the first-order condition is:

$$u_1[i]G'(nE) + u_2[i] = 0 \text{ or } -\frac{u_2[i]}{u_1[i]} = G'(nE) \ , \tag{2.5}$$

and the SKE is Pareto efficient. Kantian reasoning overcomes the commons' tragedy.

Indeed the argument is more general. Let an allocation rule be specified by $(X^1,...,X^n)$ where $X^i:\Re_+^n\to\Re_+$ with the identity $\sum_j X^i(E^1,...,E^n)=G(E^S)$ for all effort vectors $(E^1,...,E^n)$. Suppose the rule is *symmetric*, meaning that:

$$(\forall E\geq 0)(\forall i)(X^i(E,E,...,E)=\frac{G(E^S)}{n}) \ . \tag{2.6}$$

Of course, the proportional rule $X^{\Pr}$ is a symmetric rule. Then:

Proposition 2.1 *The SKE for any concave production economy and any symmetric allocation rule is Pareto efficient*.

Proof:

The typical producer maximizes $u(\frac{1}{n}G(nE),E)$, using the definition (2.6), and the characterizing F.O.C. is (2.5). ■

Another historically important allocation rule for hunter-gather societies was the equal-division rule, defined by:

$$X^{ED,i}(E^1,...,E^n)=\frac{G(E^S)}{n} \ . \tag{2.7}$$

Since the equal-division rule is symmetric, it follows that the *SKE* for hunting societies – which often used this rule – is Pareto efficient. Again, the Nash equilibrium of the game defined by X^{ED} is Pareto inefficient. But the tragedy is of a different sort from that in the fisher economy: this time, hunters hunt too *little* at the Nash equilibrium. The characterizing condition for an interior Nash equilibrium is:

$$-\frac{u_2[i]}{u_1[i]}=\frac{1}{n}G'(E^S) \ . \tag{2.8}$$

As long as $n>1$, the MRS for each player is less than the MRT. Note this is, as well, the case when *G* is linear, and in this sense the tragedy is deeper than in the fishing economy.

In sum, Kantian optimization can resolve inefficiencies that plague autarkic optimization in simple fishing and hunting economies. Did the producers in some such economies, in ancient times, learn to think in the Kantian manner, leading to greater

success of their communities? Is it possible that Kantian thinking became a meme, passed down through the generations, so that the individual fitness of the members of these groups was greater than of groups using the Nash optimization protocol? Can we see today, in hunting and fishing economies that remain, indications of Kantian reasoning?

2.6 Four models

I propose a 2 x 2 typology of modeling.

preferences → / optimization ↓	Self-interested	Altruistic, Complex
Nash	classical	behavioral econ
Kant	this book, most chapters	this book, chapter 5

The northwest cell in the matrix is the classical model. Behavioral economists alter the column of the matrix by proposing non-classical preferences but retaining Nash optimization; my proposal is to change the row. (The southeast cell of the matrix studies Kantian equilibrium with altruistic preferences.) To entertain this proposal one must of course relax one's belief that autarkic optimization is the *only rational* way of thinking in a game. While this may be a correct statement for a decision problem, it is not obviously so for a game. Those of us who have been schooled in Nash equilibrium tend to view many examples of successful cooperation as irrational. Would it not perhaps be more modest to think that we have not properly characterized rationality in games? In some social situations, at least, people may adopt the Kantian protocol, resolving free rider problems.

2.7 Literature notes

Jean-Jacques Laffont (1975) wrote:

> To give substance to the concept of a new ethics, we postulate that a typical agent assumes (according to Kant's morals) that the other agents will act as he does, and he

> maximizes his utility function under this new constraint … Our proposition is then equivalent to a special assumption of others' behavior. It is clear that the meaning of 'the same action' will depend on the model and will usually mean 'the same *kind* of action.'

Not only does Laffont deserve credit for the general idea argued here, but his recognition that a Kantian must generally think in terms of the same *kind* of action will become clear in chapter 3.

Ted Bergstrom (1995), in a discussion of selective adaptation, defines the Kantian golden rule for asexual siblings as "Act toward your siblings as would be in your own best interest if your siblings' action would mimic your own."

Robert Sugden (1982) discusses philanthropy and argues, with empirical evidence, that "the Nash assumption" (that donors take the contributions of others as given) is not empirically verified. He writes : "Or suppose that each person, instead of having Nash conjectures, believes that if he gives a certain minimum sum of money, everyone else will do the same, but he gives less, everyone else will give nothing." This is his Kantian premise.

Tim Feddersen (2004) offers a 'group-based ethical model' to explain the voting paradox. He writes, "First, ethical agents evaluate alternative behavioral rules in a Kantian manner by comparing the outcomes that would occur if everyone who shares their preferences were to act according to the same rule."

Brekke, Kverndokk and Nyborg (2003) propose that in a symmetric contribution game with a public good, agents define the moral action as the simple Kantian equilibrium (not those words). But they then introduce a penalty term in utility, which decreases utility to the extent that the player deviates from the Kantian action, so that it becomes a Nash equilibrium to play the SKE. From my viewpoint, this a gratuitous move: why say players pay a 'cost' for deviating from the Kantian action, rather than just saying they play the action they think is moral? Is not the latter simpler, although heretical from the classical viewpoint?

Chapter 3: Heterogeneous preferences: Multiplicative and additive Kantian optimization

3.1 Fishing and hunting economies

We now suppose that the fishers have arbitrary concave preferences over consumption and effort represented by utility functions $\{u^i \mid i = 1,...,n\}$. In the fishing economy, $X^{\Pr}$ continues to be the allocation rule: each fisher keeps her catch. A simple Kantian equilibrium will not exist: that is, each fisher would choose a different effort vector on the diagonal of $\Re_+^n$ as the common level of effort. Even if we relax the definition, and define $E^{*,i}$ as the effort that fisher *i* would like all to expend, the vector $(E^{*1},...,E^{*n})$ will not be Pareto efficient.

Suppose, at an effort allocation $(E^1,...,E^n)$ a fisher thinks: 'I'd like to increase my fishing time by 10%. But I only should do this if I would be happy if all were to increase their fishing time by 10%.' Do not (at this point) ask where this thought comes from, but let's define an equilibrium with respect to this kind of thinking.

Definition 3.1 A *multiplicative Kantian equilibrium* in a game $\{V^i\}$ is an effort vector $(E^1,..,E^n)$ such that *nobody* would prefer to alter *everybody's* fishing time by *any* non-negative factor. Formally:

$$(\forall i)(\forall r \geq 0)(V^i(E^1,...,E^n) \geq V^i(rE^1,...,rE^n)) \ . \qquad (3.1)$$

We denote such an allocation as a $K^{\times}$ equilibrium.

The fishing game is a strictly monotone decreasing game: if anyone else increases his effort, my catch decreases, because the productivity of the lake decreases. We generalize Proposition 2.1:

Proposition 3.1 *Let* $\mathbf{E} = (E^1,E^2,...,E^n)$ *be a strictly positive multiplicative Kantian equilibrium in a strictly monotone (increasing or decreasing) game. Then it is Pareto efficient in the game.*

Proof:

1. Let the game be strictly monotone decreasing. Suppose **E** were Pareto dominated by an effort vector $\mathbf{E}_* = (E_*^1,...,E_*^n)$. Let k be an index such that $\frac{E_*^i}{E^i}$ is minimized. Define $r = \frac{E_*^k}{E^k}$. Note that $rE^k = E_*^k$ and for $j \neq k$, $rE^j \leq E_*^j$, by definition of *r*. Furthermore, for at least one *j*, $rE^j < E_*^j$. For otherwise, $\mathbf{E}_* = r\mathbf{E}$, and since $\mathbf{E}_*$ Pareto dominates **E,** at least one agent would prefer $r\mathbf{E}$ to $\mathbf{E}$, which contradicts the fact that **E** is a multiplicative Kantian equilibrium. It follows that

$$V^k(r\mathbf{E}) > V^k(\mathbf{E}_*) \geq V^k(\mathbf{E}) \ , \qquad (3.2)$$

where the first inequality follows because the game is strictly monotone decreasing, and the second follows because $\mathbf{E}_*$ Pareto dominates **E.** But (3.2) contradicts the fact that $\mathbf{E}$ is a multiplicative Kantian equilibrium – Mr. *k* would advocate changing the scale of $\mathbf{E}$ by a factor of *r*. This contradicts the supposition that $\mathbf{E}$ is not Pareto efficient in the game.

2. If the game is strictly monotone increasing, then we define *k* to an index that *maximizes* $\frac{E_*^i}{E^i}$. The positivity of the vector **E** guarantees that this number is not infinite. The proof proceeds as above. ■

Now let the game $\{V^i\}$ be the fishing game; that is:

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

The fact that multiplicative Kantian equilibrium is Pareto efficient *in the game* $\{V^i\}$ does not imply it is Pareto efficient in the *economy*. The *game* requires allocations to be proportional, but there may be some *non-proportional* allocation in the economy (requiring transfers among fishers) that Pareto dominates the $K^\times$ equilibrium. The next proposition shows that this is not the case.

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

Proof:

By concavity, the first-order condition is sufficient to establish multiplicative Kantian equilibrium:

$$(\forall i)(\frac{d}{dr}\Big|_{r=1} u^i(\frac{rE^i}{rE^S}G(rE^S), rE^i) = 0. \tag{3.3}$$

Compute that (3.3) reduces to:

$$u_1^i \cdot (\frac{E^i}{E^S}G'(E^S)E^S) + u_2^i E^i = 0; \tag{3.4}$$

dividing through by the *positive* number E^i , and rearranging, we have:

$$(\forall i) \quad -\frac{u_2^i[i]}{u_1^i[i]} = G'(E^S) \ , \tag{3.5}$$

which proves the claim. ■

Recall Laffont's comment: 'It is clear that the meaning of 'the same action' will depend on the model and will usually mean 'the same *kind* of action.' In the fishing game, the same *kind* of action means 'changing all efforts by a scale factor.' This is , admittedly, more complex than 'taking the action I'd like all to take.' The efficiency result (Prop. 3.1) suggests that successful fishing communities may have discovered $K^{\times}$ reasoning through cultural evolution (see Boyd and Richerson [1985]).

The reader should note the formal similarity between $K^{\times}$ and Nash equilibrium. Both use ordinal preferences only. Each considers a counterfactual: with Nash reasoning the counterfactual is that only I change my strategy, while in Kantian reasoning, we all change our strategies in a prescribed way. An equilibrium, in either case, is a strategy profile which dominates all permissible counterfactual profiles. An optimizing agent in both cases evaluates the counterfactual profile using his own preferences only. Other similarities will appear in the discussion of existence and dynamics (chapter 7).

Now let us consider hunting economies, which use the allocation rule X^{ED} . Hunters fan out into the bush searching for game, and after several days they return

to camp, dividing the capture equally. At an effort allocation $(E^1,...,E^n)$, a hunter thinks, 'I'd like to take a two hour nap under that tree. But I should do this only if I would be happy if all hunters took a two hour nap.' This time, the *kind* of action that the Kantian contemplates is additive rather than multiplicative.

Definition 3.2 An *additive Kantian equilibrium* (K^+) is an allocation such that nobody would prefer to add any constant to all efforts. That is:

$$(\forall i)(\forall r \geq -\min_j E^j)(V^i(E^1,...,E^n) \geq V^i(E^1+r,...,E^n+r)). \quad (3.6)$$

The analogue of Proposition 3.2 continues to hold –except this time, we need not require that the equilibrium allocation be positive.

Proposition 3.3 *Any* K^+ *equilibrium is Pareto efficient (in the economy).*

Proof:

The F.O.C. that characterizes an interior K^+ equilibrium is:

$$\left.\frac{d}{dr}\right|_{r=0} u^i(\frac{G(E^S+nr)}{n}, E^i+r) = 0 \ , \quad (3.7)$$

which expands to:

$$u_1^i \cdot \frac{G'(E^S)}{n} n + u_2^i = 0 \ , \quad (3.8)$$

which immediately reduces to $MRS^i = MRT$. ■

So Kantian reasoning resolves the tragedy of the commons in both hunting and fishing communities with heterogeneous preferences – but the *kind* of action that a producer contemplates universalizing changes with the allocation rule. I can think of no normative argument for why fishers might be led to think multiplicatively and hunters additively; if these communities discovered the right kind of Kantian counterfactual, this was due to chance cultural mutation, as in biological evolution.

What is the relationship between $K^\times, K^+$ and simple Kantian equilibrium? We have:

Proposition 3.4 *In a production game where all players have the same preferences u*

and the allocation rule X is symmetric, any positive SKE is both a $K^{\times}$ and a K^{+} *equilibrium.*

Proof:

Let the SKE be E^{*} . Define the share functions θ^{i} by $\theta^{i}(E^{1},...,E^{n})G(E^{S}) = X^{i}(E^{1},...,E^{n})$. To show that E^{*} is a $K^{\times}$ equilibrium, we need to show that:

$$\left.\frac{d}{dr}\right|_{r=1} u(\theta^{i}(rE^{*},...,rE^{*})G(rE^{S}),rE^{*}) = 0, \tag{3.9}$$

which reduces to:

$$u_{1}\cdot(\theta^{i}(E^{*},...,E^{*})G'(nE^{*})nE^{*} + \nabla\theta^{i}\bullet\mathbf{E}^{*}) + u_{2}E^{*} = 0 \tag{3.10}$$

where $\nabla\theta^{i}$ is the gradient vector of θ^{i} at $\mathbf{E}^{*}$ where $\mathbf{E}^{*} = (E^{*},...,E^{*})$. Because $E^{*} > 0$, we may rewrite (3.10) as:

$$-\frac{u_{2}[i]}{u_{1}[i]} = \frac{\theta^{i}G'(nE^{*})nE^{*} + \nabla\theta^{i}\bullet\mathbf{E}^{*}}{E^{*}} \ ; \tag{3.11}$$

the right-hand side of (3.11) reduces to the MRT $G'(nE^{*})$ if:

$$\frac{\theta^{i}(\mathbf{E}^{*})nE^{*}}{E^{*}} = 1 \quad \text{and} \quad \nabla\theta^{i}(\mathbf{E}^{*})\bullet\mathbf{E}^{*} = 0 \ .$$

The first condition is true, because, by symmetry, $\theta^{i}(\mathbf{E}^{*}) = 1/n$, and the second condition is likewise true by symmetry, for it says the directional derivative of θ^{i} at $\mathbf{E}^{*}$ in the direction $\mathbf{E}^{*}$ is zero – and this is true, because θ^{i} is constant at $1/n$ along that path. Therefore, a positive SKE is a $K^{\times}$ equilibrium.

The demonstration that an SKE is a K^{+} equilibrium is left to the reader. ■

Thus, multiplicative and additive Kantian equilibria are true generalizations of the natural concept of simple Kantian equilibrium to the case of heterogeneous preferences.

3.2 Incentive compatibility

Let us look more carefully at the Kantian equilibria in the fishing and hunting economies for the special case, canonical in optimal taxation theory, in which workers have the same preferences over consumption and labor time (*L*), but their skills are

different. Suppose everyone has preferences over consumption and labor expended represented by a concave utility function $u(x,L)$, but productivities, *w*, are distributed according to some distribution function *F*, so that utility functions expressed as functions of consumption and efficiency units of labor are given by:

$$u^w(x,E) = u(x,\frac{E}{w}) \ . \tag{3.12}$$

(That is, $E^i = w^i L^i$.) So, although workers share the common preferences *u*, the differential skills they possess make this a case of heterogeneous preferences when we express labor in efficiency units.

We can ask whether the fishing and hunting equilibria are incentive compatible in the sense that, at the equilibrium, utility increases with skill. In the production economies studied above, the condition for Pareto efficiency is that $u_1^i G'(E^S) + u_2^i = 0$ for all *i*. For the special case of (3.12), this becomes:

$$(\forall w) \quad u_1 w G'(E^S) + u_2 = 0 \, , \tag{3.13}$$

where *u* is evaluated at the argument $(x, E/w)$.

Example: Quasi-linearity in a continuum economy

Assume $u(x,L) = x - \frac{1}{2}L^2$, and let skill levels *w* be distributed according to a distribution function *F*. Let $G(E) = 2\sqrt{E}$. In the continuum economy, we replace E^S with $\bar{E} = \int E(w)dF(w)$, and E^i with $E(w)$. Thus (3.13) becomes:

$$\bar{E}^{-1/2} w = \frac{E(w)}{w} \text{ or } \bar{E}^{-1/2} w^2 = E(w) \ . \tag{3.14}$$

Integrating this equation gives us

$$\bar{E}^{-1/2} \mu_2 = \bar{E} \text{ and so } \bar{E} = (\mu_2)^{2/3} \ , \tag{3.15}$$

where $\mu_2 = \int w^2 \, dF(w)$ is the second moment of the real wage distribution. It now follows from (3.14) that $E(w) = \frac{w^2}{(\mu_2)^{1/3}}$. By recalling that the consumption of

individual w is $\frac{E(w)}{\bar{E}}G(\bar{E})$ in the multiplicative Kantian equilibrium, we can compute that w's utility at the solution is given by:

$$u^{K^\times}[w] = u(\frac{E(w)}{\bar{E}}G(\bar{E}), \frac{E(w)}{w}) = \frac{3}{2}\frac{w^2}{(\mu_2)^{2/3}} \ . \tag{3.16}$$

Hence utility[1] is indeed increasing in w.

Let us compute the Nash equilibrium of this fishing game. The first-order condition for Nash equilibrium is $\frac{d}{dE}u(E\frac{G(\bar{E}^N)}{\bar{E}^N}, \frac{E}{w}) = 0$, or:

$$\frac{2(\bar{E}^N)^{1/2}}{\bar{E}^N} = \frac{E^N(w)}{w^2} \tag{3.17}$$

or $E^N(w) = \frac{2w^2}{\sqrt{\bar{E}^N}}$, which integrates to give:

$$(\bar{E}^N)^{3/2} = 2\mu_2 \text{ or } \bar{E}^N = (2\mu_2)^{2/3} \ , \tag{3.18}$$

and it follows that $E^N(w) = \frac{2w^2}{(2\mu_2)^{1/3}}$. Compute that utilities at the Nash equilibrium are given by:

$$u^N[w] = \frac{w^2}{\mu_2}2(2\mu_2)^{1/3} - \frac{1}{2}(\frac{2w}{(2\mu_2)^{1/3}})^2 = 2^{1/3}\frac{w^2}{(\mu_2)^{2/3}} \ . \tag{3.19}$$

Comparing (3.19) with (3.16), we see that all players are strictly better off in the Kantian equilibrium, because $\frac{3}{2} > \sqrt[3]{2}$.

It follows from (3.16) that utility indeed increases with w at the $K^\times$ equilibrium.

Let us now compute the utilities at the additive Kantian equilibrium for this example – which means at the equal-division solution. Because of the quasi-linear structure, the values of $E(w)$ are the same for all Pareto efficient allocations. We

[1] In particular, we have shown the existence and uniqueness of a $K^\times$ equilibrium for this economy.

therefore know that $E(w) = \frac{w^2}{(\mu_2)^{1/3}}$ in the hunting economy. The only change from the multiplicative Kantian equilibrium is that consumption for all agents is $G(\bar{E}) = 2(\mu_2)^{1/3}$. It follows that utilities in the K^+ equilibrium of the X^{ED} economy are given by

$$u^{K^+}[w] = 2(\mu_2)^{1/3} - \frac{w^2}{2(\mu_2)^{2/3}} \quad . \tag{3.20}$$

The first-order condition for Nash equilibrium in the equal-division economy is $\frac{d}{dE} u(G(\hat{E}^N), \frac{E}{w}) = 0$, where $\hat{E}^N$ is the average efficiency units of effort expended in this Nash equilibrium. This first-order condition reduces to $\frac{u_2}{w} = 0$, so all efforts are zero in the Nash equilibrium of the equal-division economy. From (3.20), it follows that the K^+ equilibrium is *not* incentive compatible – utility is strictly decreasing in *w*.

How disturbing or relevant is this? I question the relevance of the result. In ancient hunting economies, young men, who were the hunters, acquired their skills during youth and adolescence, when there was doubtless praise and respect showered on those who developed high skill by their elders. Hunters in the bush had their reputations to maintain, so utility is probably not properly represented by functions like *u*. In modern times, we think of the more radical kibbutzim in Israel, which used, more or less, an equal-division allocation rule. Some members, with high earning power who worked outside the kibbutz, contributed more than others to the common pool of consumption goods. In the presence of a cooperative ethos, incentive incompatibility is not a death knell, although we can expect that ethos is more difficult to maintain if the variance in skills is high. Moreover, the incentive incompatible nature of the equal-division solution suggests that the K^+ equilibrium will be harder to maintain than the proportional $K^\times$ equilibrium.

3.2 Sustainability in a dynamic setting

The fishing game is a very simple example of the tragedy of the commons. More realistically, one should examine the nature of stationary states where a

common pool renewable resource is exploited by a community. Here we modify a model proposed by Richter and Grasman (2013).

Consider a community that exploits a renewable resource, such as a fishery. At any point in time, the harvest will be proportional to the total extraction effort of the community, where the factor of proportionality is itself proportional to the total amount of the resource; that is:

$$H(t) = qX(t)E^S(t) \tag{3. 21}$$

where X and E^S are total supply of the resource (the fish population in the lake) and total effort of extraction, and H is the harvest at time t. Think of (3.21) as follows: in unit time, an amount proportional to the total fish population can be extracted, $qX(t)$ -- we view qX as a measure of the density of the fish in the lake. We now assume that there are constant returns in effort, at least for efforts that are not too large relative to X. Formulation (3.21) is standard in resource economics.

The law of motion of the renewable resource is given by:

$$X(t+1) = X(t) + rX(t)(1 - \frac{X(t)}{K}) - H(t) \ , \tag{3.22}$$

where K is the maximum possible population of fish, or the capacity constraint of the lake. In other words, the fish population renews itself at a rate that is decreasing as the resource approaches the maximum capacity. It follows that the *stationary states* are given by:

$$H = rX(1 - \frac{X}{K}) \tag{3.23}$$

or, using (3.21):

$$qE^S = r(1 - \frac{X}{K}) \ . \tag{3.24}$$

Suppose that the utility function of producer i is given by:

$$u^i(x,E) = x - \nu^i E^2 \tag{3.25}$$

where x is consumption of the resource and E is extraction effort. The community wishes to choose among possible *sustainable extraction rules*: that is, it wishes to choose

a stationary state (X, E^S) as defined by (3.24). As well as choosing the stationary state, it must choose the individual efforts E^i so that $E^S = \sum E^i$.

Production is carried out by individuals: thus, each keeps the resource he harvests. Since each producer is equally likely to extract a unit of the resource with the application of a unit of labor (in efficiency units), the total harvested resource is allocated in proportion to the efforts expended.

We examine a *multiplicative Kantian equilibrium* for such a problem. Imagine that the community is considering a particular stationary state (X, E^S) . Suppose everyone were to multiply his effort by a positive number ρ: then a new stationary fish population X^ρ would ensue, where this quantity is defined by:

$$q\rho E^S = r(1 - \frac{X^\rho}{K}), \text{ or } X^\rho = K(1 - \frac{q\rho E^S}{r}) . \qquad (3.26)$$

It is assumed for simplicity that producers will exert the same effort at every date, forever. Thus, they will quickly converge to a stationary state once the total effort is fixed. (This follows from an examination of (3.22).) Implicitly, producers are maximizing a discounted sum of their period utilities, and we ignore the issue of transition to a stationary state. Since maximizing the present value of a constant stream of utilities is equivalent to maximizing the single-period utility, we need not further consider the discounted sum, although it is their looking into the future that motivates the fishers to study the stationary (sustainable) states. Either because they have sufficiently low discount rates, or because they care about future generations of producers, they limit their search to sustainable states.

Now a Kantian equilibrium in such a situation is a vector of effort levels $\mathbf{E} = (E^1, ..., E^n)$, inducing a total effort E^S, and a stationary state via (3.24), such that no producer would advocate changing *all* effort levels by *any* constant factor, passing to the associated new stationary state. In other words, **E** has the following property:

$$(\forall i)(1 = \arg\max_{\rho}[\frac{E^i}{E^S} qX^{\rho}\rho E^S - \nu^i(\rho E^i)^2]) \ . \tag{3.27}$$

To understand (3.18), note that, at (X, E^S) , the amount of the resource (fish) that agent *i* gets is equal to his fraction of the total extraction time multiplied by the total harvest, which is qXE^S. So if *i* were to advocate multiplying all efforts by ρ, her new resource harvest would be $\frac{E^i}{E^S} qX^{\rho}\rho E$, and her new utility would be the expression in square brackets in (3.18). Thus, (3.18) is the condition for the effort vector's being a $K^{\times}$ equilibrium.

Substituting for X^{ρ} from (3.17), the above maximization is:

$$\max_{\rho} E^i q\rho K(1 - \frac{q\rho E^S}{r}) - \rho^2 \nu^i (E^i)^2 \ , \tag{3.28}$$

which is concave in ρ, and hence we examine the first-order condition for the solution to (3.28) at $\rho = 1$, which reduces to:

$$E^i = \frac{qK(1 - \frac{qE^S}{r}) - \frac{q^2 KE^S}{r}}{2\nu^i} \ . \tag{3.29}$$

Adding equations (3.20) over all *i* gives us an equation in E^S, which solves to give:

$$E^S = \frac{\Omega}{2} qK(1 - \frac{2qE^S}{r}) \ , \tag{3.30}$$

where $\Omega \equiv \sum \frac{1}{\nu^i}$, which in turn gives:

$$E^S = \frac{\Omega qKr}{2(r + \Omega q^2 K)} \ . \tag{3.31}$$

Now the value of *X* follows from (3.26), and the individual effort levels are given by substituting E^S into (3.29); they turn out to be:

$$E^i = \frac{qK}{2\nu^i}\frac{r}{r+\Omega q^2 K} \quad . \tag{3.32}$$

Unsurprisingly, the individual efforts are inversely proportional to the disutilities of effort (ν^i). They are also increasing in *r*, the regeneration rate of the resource.

We next ask about the welfare properties of this solution to the commons problem. If the society limits itself to sustainable solutions (in the sense of (3.26)), what are the Pareto efficient allocations of the resource and effort? In other words, we seek to characterize the Pareto efficient allocations of the resource and effort, subject to sustainability. To solve this problem, we maximize the utility of an arbitrary agent *i* subject to placing lower bounds on the utilities of all other agents, and restricting ourselves to sustainable solutions. The problem is:

$$\begin{aligned}
&\max x^i - \nu^i (E^i)^2 \\
&\text{subj. to} \\
&(\forall j \neq i)\; x^j - \nu^j (E^j)^2 \geq k_j && (\lambda_j) \\
&\sum_{all\ \mathrm{j}} E^j = E^S && (b) \\
&qXE^S \geq \sum_{\mathrm{j}} x^j && (a) \\
&qE^S = r(1-\frac{X}{K}) && (c)
\end{aligned} \tag{3.33}$$

where I have listed Lagrangian multipliers to the right of the constraints. Constraints (*b*) and (*a*) are feasibility constraints, and (*c*) is the sustainability constraint. The variables that must be chosen are $\{x^j, E^j, E^S, X\}$. The Kuhn-Tucker conditions for a solution to this problem are:

$$
\begin{aligned}
&(\partial x^i)\ a = 1 \\
&(\partial x^j, j \neq i) \quad \lambda_j = a \\
&(\partial E^i)\ E^i = \frac{b}{2\nu^i} \\
&(\partial E^j, j \neq i)\ E^j = \frac{b}{2\lambda_j \nu^j} = \frac{b}{2\nu^j} \\
&(\partial X)\ qE^S = \frac{cr}{K} \\
&(\partial E^S)\ qX = b + cq
\end{aligned}
\tag{3.34}
$$

It immediately follows that $a = 1 = \lambda_j$ and $E^S = \frac{b\Omega}{2}$. From the last condition, $b = q(X - c) = (X - q\frac{E^S K}{r})$ and so $E^S = \frac{\Omega}{2} q(X - q\frac{E^S K}{r})$, which solves to give:

$$E^S = \frac{\Omega q X r}{2r + \Omega q^2 K} \ . \tag{3.35}$$

Now, substituting into (3.35) from the last constraint in (3.33), we compute:

$$E^S = \frac{\Omega q r K}{2(r + \Omega q^2 K)} \ . \tag{3.36}$$

But this is identical to the total effort in the Kantian equilibrium , see (3.31). Moreover, the individual efforts E^i are also identical to those of the Kantian equilibrium: this is obvious, since we note from (3.34) that the individual efforts are inversely proportional to the ν^i as well, and must add up to the same total effort. *Any* allocation of the harvested resource among the producers generates a Pareto efficient solution – the Kantian equilibrium picks out the allocation where 'each keeps his catch'.

To complete the KT analysis, we must check the sign of the shadow prices. $c = \frac{qKE^S}{r} > 0$ from the (∂X) condition. It remains only to check that $b \geq 0$. Now $b = q(X - c)$, so we need check that $X \geq c$, that is, $rX \geq qKE^S$. Using constraint (*c*) in (3.33), this becomes $E \leq^{?} \frac{r}{2q}$, or $\frac{\Omega q r K}{2(r + \Omega q^2 K)} \leq \frac{r}{2q}$ or $\frac{\Omega q^2 K}{(r + \Omega q^2 K)} \leq 1$ which is true. This completes the argument.

Thus, multiplicative Kantian optimization is a protocol for solving the problem of efficient, sustainable exploitation of a renewable resource.

3.3 Oligopolistic collusion[2]

Consider *n* producers in an oligopolistic market, who face a demand curve $D(p)$, where we assume D^{-1} is a concave function, and where producer *i* has a convex cost function $c^i(y)$. The oligopolist game, where firms choose quantities, is given by the payoff functions:

$$V^i(y^1,...,y^n) = D^{-1}(y^S)y^i - c^i(y^i) \ . \qquad (3.37)$$

Because D^{-1} is a decreasing function, the game so defined is strictly monotone decreasing, and hence by Proposition 2.1, the $K^\times$ equilibrium is Pareto efficient (for the producers).

3.4 Strikes

A group of workers is contemplating a strike. Each worker's strategy is the probability that he will join the strike, π^i. If he does not join the strike, he scabs. Workers are of types *i,* with n^i workers of type *i*. The probability that the strike wins is $p(m)$ where $m = \sum \pi^i n^i$, the number of strikers, and p is monotone increasing.

The utilities of a striker are:

A^i if the strike wins, B^i if it loses where $A^i > B^i$

and scabs earn, in addition $C(m)$, where *C* is a decreasing function of *m*. Thus the utility of a worker (striker or scab) at a profile $\pi = (\pi^1,...,\pi^n)$ is:

$$EU^i = p(m)A^i + (1-p(m))B^i + (1-\pi^i)C(m) \ . \qquad (3.38)$$

Note that scabs enjoy the outcome of the strike whatever it is. This is not a monotone game, since *C* is decreasing in the strategies of the other players, while the first two terms are in sum increasing, since $A^i - B^i > 0$.

[2] I thank Luis Corchon for this example.

We first compute the $K^{\times}$ equilibrium, which involves solving:

$$\max_r p(rm)(A^i - B^i) + B^i + (1 - r\pi^i)C(rm) \ .$$

The F.O.C. for the solution at $r = 1$ is:

$$p'(m)m(A^i - B^i) + (1-\pi^i)C'(m)m - \pi^i C(m) = 0 \text{ or}$$

$$\forall i \quad p'(m)(A^i - B^i) + (1-\pi^i)C'(m) = \pi^i \frac{C(m)}{m} \ . \tag{3.39}$$

Now we solve this program to characterize interior Pareto efficient allocations:

$$\begin{aligned} &\max p(m)(A^i - B^i) + (1-\pi^i)C(m) \\ &s.t. \\ &(\forall j \neq i) \quad p(m)(A^j - B^j) + (1-\pi^j)C(m) \geq k^j \quad (\lambda^j) \end{aligned} \tag{3.40}$$

The KT conditions for an interior solution (that is, $0 < \pi^i < 1$) can be written as follows. Let $\lambda^i = 1$. Then :

$$(\forall j = 1,...,n)\,(C(m)\lambda^j = \sum_{all\ k} \lambda^k n^j (p'(m)(A^k - B^k) + C'(m)(1-\pi^k)) \ . \tag{3.41}$$

Define $K = \sum_{all\ k} \lambda^k (p'(m)(A^k - B^k) + C'(m)(1-\pi^k))$, and so $\lambda^j = \frac{n^j K}{C(m)}$ holds for all *j*.

Substituting these values into the r.h.s. of (3.41) gives:

$$n^j K = n^j \sum n^k \frac{K}{C(m)} \left(p'(m)(A^k - B^k) + C'(m)(1-\pi^k) \right)$$

or

$$C(m) = \sum n^k \left(p'(m)(A^k - B^k) + C'(m)(1-\pi^k) \right) \ . \tag{3.42}$$

Notice that the $\lambda^j \geq 0$ if and only if *K* is non-negative which occurs when

$$\sum_{all\ k} n^k (p'(m)(A^k - B^k) + C'(m)(1-\pi^k)) \geq 0 \ . \tag{3.43}$$

But this is true, since the l.h.s. of (3.43) is simply $C(m)$. Thus a vector of probabilities $(\pi^1,...,\pi^n)$ is ex ante Pareto efficient if (3.42) holds.

Eqn. (3.43) can be written:

$$p'(m)\sum_{all\ k} n^k(A^k - B^k) \geq -C'(m)\sum\ (n^k - n^k\pi^k)),$$

which says that the expected marginal gain to all workers, where the margin is increasing participation in the strike must be at least equal to the expected marginal loss to the scabs[3].

Now multiply equations (3.39) by n^i and add over *i,* giving:

$$\sum\ n^i(p'(m)(A^i - B^i) + (1-\pi^i)C'(m)) = C(m),$$

which is eqn. (3.42). Thus, multiplicative Kantian equilibria, where the strategies are probabilities of striking, are (ex ante) Pareto efficient. If strikers choose their probabilities of participation in a Kantian manner, the strike is efficient.

3.5 Lindahl equilibrium for a public-good economy

Individuals in a society have utility functions u^i defined over arguments (y,E^i) where *y* is the value of a public good, E^i is *i*'s contribution to the public good, and the cost function is $C(y) = E$. The production function *G* is the inverse of the cost function.

The payoff function of individual *i* is $u^i(G(E^S),E^i)$. The $K^\times$ equilibrium (if it exists) is characterized by:

$$\left.\frac{d}{dr}\right|_{r=1} u^i(G(rE^S),rE^i) = 0 \text{ or } u_1^iG'(E^S)E^S + u_2^iE^i = 0\ ,$$

which can be written:

$$\frac{-u_1^i}{u_2^i} = \frac{E^i}{E^S}\frac{1}{G'(E^S)}\ . \qquad (3.44)$$

Now $\frac{1}{G'(E^S)} = C'(y)$, and so adding (3.44) over all *i* gives:

$$\sum\frac{1}{MRS^i} = C'(y) \qquad (3.45)$$

[3] If there is no interior solution in the probabilities to (3.42), then ex ante Pareto efficiency will require , for some *i,* $\pi^i \in \{0,1\}$.

which is the Samuelson condition for efficiency in the public good economy.

Definition 3.3 A *linear cost-share equilibrium* is a vector of shares $(b^1,...,b^n) \in [0,1]^n$ such that $\sum b^i = 1$, a contribution vector $(E^1,..,E^n)$ and a public good level y, which is feasible, such that:

$(\forall i)(E^i = b^i C(y)$ and y maximizes $u^i(y, b^i C(y))$.

(See Mas-Colell and Silvestre (1989).)

A linear-cost-share equilibrium is special case of a Lindahl equilibrium. Suppose the $K^\times$ equilibrium characterized by (3.35) exists. Define $b^i = \frac{E^i}{E^S}$. Then the linear-cost-share equilibrium for the vector *b* solves:

for all *i*, y maximizes $u^i(y, b^i C(y))$.

The F.O.C.s for this problem are:

for all $i \quad u_1^i + u_2^i b^i C'(y) = 0$. (3.46)

But these equations are identical to (3.44), and so Kantian optimization decentralizes the Lindahl equilibrium. Mas-Colell and Silvestre (1989) prove such an equilibrium exists, and therefore the multiplicative Kantian equilibrium exists as well.

3.6 Affine taxation in a linear economy

Consider an economy which produces a good, which is redistributed through an affine tax scheme (t,b) where *t* is the constant marginal tax rate on income and *b* is the demogrant received by all. Utilities are $u^i(y^i, E^i)$ defined on income and labor. Production is linear in efficiency units of labor: $y = aE^S$. The real wage for an efficiency unit of labor, in a competitive economy, equals its marginal product, which is *a*.

Let $t \in [0,1]$ be the tax rate. Then the payoff function for individual *i* is

$$u^i((1-t)aE^i + \frac{taE^S}{n}, E^i) \ ,$$

because the demogrant will equal taE^S / n . The Nash equilibrium of this game (where the strategies are the labor supplies) is , of course, inefficient for $t > 0$: this is the deadweight loss of taxation. Let us examine the K^+ equilibrium of this game. It is characterized by:

$$\left.\frac{d}{dr}\right|_{r=0} u^i((1-t)a(E^i + r) + \frac{ta(E^S + nr)}{n}, E^i + r) = 0$$

which expands to:

$$u_1^i \cdot (a(1-t) + at) + u_2^i = 0 \text{ or } -\frac{u_2^i}{u_1^i} = a\,. \qquad (3.47)$$

But this says , *regardless of the tax rate*, $MRS^i = MRT$ for all *i*, and so the allocation is Pareto efficient.

In other words, for a linear economy, additive Kantian optimization resolves the deadweight loss of affine taxation. It allows us to completely separate distribution from efficiency. Why does this occur? Because, with additive Kantian optimization, the individual does not take the demogrant as fixed, even if he is only one of very many taxpayers. Contemplating a reduction in his labor supply makes him contemplate a similar reduction in *everyone's* labor supply thus reducing the size of the demogrant. The rule of thumb the additive Kantian optimizer uses is: reduce my labor supply only if my marginal rate of substitution is greater than the marginal rate of transformation (and analogously for increasing the labor supply): this is the verbal interpretation of (3.38). Thus, stability of labor supply occurs at the efficient allocation.

It is interesting to see what happens if production is concave but not linear. Then the condition for K^+ equilibrium becomes (recalling that the wage equals the marginal product):

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

or $$u_1^i \cdot ((1-t)E^i G''(E^S)n + (1-t)G'(E^S) + tG'(E^S)) + u_2^i = 0$$

or $$-\frac{u_2^i}{u_1^i} = (1-t)nE^i G''(E^S) + G'(E^S)\ , \qquad (3.48)$$

and so the MRS is always less than the MRT when $G'' < 0$. (The Kantian optimizer does not assume the wage is fixed, but recognizes the responsiveness of the wage to the effort supply.) Notice, however, that if $t = 1$, we again have Pareto efficiency, even for concave production – but this is because $t = 1$ is simply the equal-division (hunting) economy! So, in a sense that is not precise, the deadweight loss *diminishes* as the tax rate approaches unity in K^+ equilibrium.

3.7 Gift exchange

In a well-known paper, Akerlof (1982) explains the fact that in some firms, workers work more than a stipulated, required minimum, and firms pay more than the market wage, as a gift exchange.

Here is a model of Akerlof's observation. The firm's profit function is $P(w,e)$, which is concave, increasing in effort of workers *e*, and increasing in *w,* the wage, for sufficiently small *w* but decreasing in *w* thereafter. We interpret the wage as the weekly income of the worker, independent of her effort *e*. The existence of a region in which $P_1 > 0$ is explained by the fact that increasing the wage induces low turnover of workers by increasing the opportunity cost of quitting, which is of greater value to the firm than the increased cost of labor, as long as the wage is not too high. The worker's utility function is $u(w,e)$, concave, increasing in *w* and decreasing in *e.*

Normal firms specify a minimal acceptable effort level e_m , and the equilibrium in a normal labor-firm relationship is a Nash equilibrium where *w* is the firm's strategy and *e* is the worker's. The unique Nash equilibrium in the firm-worker game is given by:

$$e^N = e_m, \quad P_1(w^N, e_m) = 0 \ . \tag{3.49}$$

However, Akerlof observes that there are other firms where workers offer more effort than e_m , employers pay a wage greater than w^N , and presumably both workers' utility and firm profits are greater than in the normal firm. Akerlof explains this by a gift relationship: the workers provide a gift to the employer, by working harder than necessary, and in return the employer offers a gift to the workers, of a higher than normal wage.

I will propose an alternative explanation to Akerlof's, which is that the players in the game are playing a multiplicative Kantian equilibrium. A $K^{\times}$ equilibrium is a pair (w,e) such that:

$$1 = \arg\max_r P(rw,re) \text{ and } 1 = \arg\max_r u(rw,re) \ . \qquad (3.50)$$

Notice this is a strictly increasing monotone game: each player's payoff is strictly increasing in the other player's strategy. It follows by Proposition 3.1 that the solution, if it exists, is Pareto efficient. Typically, the Nash equilibrium in the game will not be Pareto efficient: so it is certainly possible that the Kantian equilibrium Pareto dominates the Nash equilibrium, and the other observations made above hold – that $e^K > e_m$ and $w^K > w^N$. We cannot, however, deduce these inequalities without more structure.

Consider this example:

$$u(w,e) = w - \alpha e, \quad P(w,e) = w - \frac{\beta}{2}w^2 + \gamma e$$

where (α,β,γ) are positive numbers. Kantian equilibrium , the solution of (3.50), is given by:

$$w = \alpha e, \quad (1-\beta w)w + \gamma e = 0$$

which solves to :

$$w^K = \frac{1}{\beta}(1+\frac{\gamma}{\alpha}), \quad e^K = w^K/\alpha, \quad u(w^K,e^K) = 0, \quad P(w^K,e^K) = \frac{1}{2\beta}(1+\frac{\gamma}{\alpha})^2 \ . \qquad (3.51)$$

On the other hand, the Nash equilibrium is given by:

$$w^N = \frac{1}{\beta}, \quad e^N = e_m, \quad u(w^N,e^N) = \frac{1}{\beta} - \alpha e_m, \quad P(w^N,e^N) = \frac{1}{2\beta} + \gamma e_m \ .$$

One can compute that both players do better in the Kantian equilibrium than the Nash equilibrium if and only if :

$$\frac{1}{2\beta\alpha} < e_m < \frac{1}{2\beta}\frac{1}{\alpha}(2+\frac{\gamma}{\alpha}) \ . \qquad (3.52)$$

The wage is always greater in the Kantian equilibrium, and effort is greater if and only if

$$\frac{1}{\alpha\beta}(1+\frac{\gamma}{\alpha}) > e_m \ . \qquad (3.53)$$

Check that (3.53) is implied by (3.52). It follows that all the features of the observed characteristics of the normal and 'gifting' firms hold precisely when (3.52) is true.

So there are certainly environments in which the phenomenon Akerlof observes is explained by Kantian optimization. Many firms are caught in a non-cooperative Nash equilibrium, and some have achieved cooperation, in the sense that the worker and employer are optimizing in the Kantian manner. Both gift and Kantian explanations are based upon trust: for Akerlof, each side trusts that the other side will make a gift if it does, and in my explanation, each side trusts the other will optimize in the Kantian manner if it does. It may be very difficult to decide if one explanation is better than the other. Indeed, the 'gift' explanation may just be another way – but an imprecise one -- of stating Kantian optimization.

The advantage of the Kantian approach is that it gives an exact solution to the game. Akerlof's explanation is incomplete, for it does not determine how large the gifts will be.

3.8 Summary thoughts

Kantian optimization provides microfoundations for the efficient solution of many phenomena involving public goods and bads: achieving efficiency in common-pool resource problems, explaining collusion among oligopolists, decentralizing Lindahl equilibrium, resolving the voting paradox, Akerlovian gift exchange, and strikes. It also suggests how certain problems might be resolved that at present have not been, such as reducing the deadweight loss from taxation. The virtue of the approach is that it gives a precise solution to many games (modulo the existence question), a solution that does not depend upon parameterizing the role of 'exotic' arguments that behavioral economists insert into preferences. Preferences in all the examples I have given are classical. With heterogeneous preferences, we have introduced Kantian protocols which, mathematically, sound a lot like Nash optimization: we have simply chosen the counterfactual to which the agent compares the present strategy profile in a different manner from Nash.

Major drawbacks seem to be that there will be difficulty in generalizing the approach to more complex strategy spaces than intervals of real numbers, we have no explanation for *which* Kantian protocol a person chooses in a given situation, and there

remains the skepticism of the Nash advocate that Kantian optimization is not rational. I will have something (though not a lot) to say about the first problem in chapter 10; my answer to the second question is that the right protocol is discovered, if it is, through random cultural mutation; my answer to the third question is that morality does not disappear with the Kantian approach, but rather it enters in a different place: in the choice of how to optimize rather than as an argument of preferences[4].

Perhaps this is an appropriate place to expand on the morality of Kantian optimization, although I can hardly do better than Immanuel Kant. When a fisher believes he must justify an expansion in his own labor supply by 10% by asking how *she would feel* if others similarly expanded their labor supplies, she is internalizing the negative externality her labor expansion imposes on others (through reducing the lake's productivity). But she does not internalize this by contemplating how the reduction in the lake's productivity will hurt others (that would be altruism) : rather, she asks how *similar* behavior by others would impact upon *her.* This approach to moral thinking has several advantages: first, it does not require that she know the preferences of others, and second, it does not require her to care about others. We use the same technique in teaching our children not to litter (we ask the child how *he* would feel if others were to litter the way he is doing). Our practice with littering children suggests to me that appealing to the categorical imperative is more persuasive than appealing to altruism.

3.9 Literature notes

In Roemer and Silvestre (1993), we proved that in fishing economies, more general than the ones defined here, allocations exist in which consumption is proportional to labor expended *and* which are Pareto efficient. We viewed this is a canonical 'socialist allocation': it adjoins to the socialist principle of proportionality of consumption to labor, Pareto efficiency, strangely ignored in the socialist tradition. We called this allocation a 'proportional solution.' In Roemer (1996), I noted that the

[4] The inquisitive reader may ask whether the choice of a Kantian optimization protocol is somehow equivalent to optimizing à la Nash but with altered preferences. This is the topic of chapter 6 below.

proportional solution possesses the multiplicative Kantian property, and so named it a Kantian equilibrium.

In Roemer (2006), I applied multiplicative Kantian equilibrium to solve the free-rider problem of donors contributing to a political party in an efficient manner.

At some point during the last decade, Joaquim Silvestre suggested varying the Kantian protocol to 'additive.' He noted that one advantage of additive Kantian equilibrium is that it eliminates one embarrassing (multiplicative) Kantian equilibrium in the fishing game, where all efforts are zero. In general, one need not specify positivity in K^+ equilibrium to guarantee Pareto efficiency, while one does, for $K^\times$ equilibrium.

The proof presented here that positive multiplicative Kantian equilibria in strictly monotone games with heterogeneous preferences is Pareto efficient is due to Colin Stewart.

The presentation in this book is non-chronological. I first discovered multiplicative Kantian equilibrium, and only much later, saw the simpler idea of simple Kantian equilibrium in symmetric games, to which I credit Brekke, Kverndokk and Nyborg (2003).