Lectures for the Canezi Winter School on Inequality, January 2016

How we do and could cooperate:

A Kantian Approach

John E. Roemer

Yale University

john.roemer@yale.edu

1. **Cooperation unique in humans** (among great apes)

* M. Tomasello's experiments with chimps & human babies. Chimps do not cooperate; babies do.

+ cooperating to get food

+ playing with adult humans

* only humans point

* only humans mime.

* only humans have sclera (white of the eyes)

* Social evidence: living in large cities, fraction of GDP in state taxation, decrease in

violence (S. Pinker) . (25% of young men died in warfare in HG socieites)

* Evolution of language could only have occurred because humans are cooperative. Because if non-cooperative, why trust what the other person is telling you? He's out to max his own interests and will deceive you.

So language doesn't get off the ground without trust based on cooperation.

Tomasello: Language probably developed to coordinate hunting activities. Other great apes do not cooperate in hunting!

2. **Economics has a thin theory of cooperation**

*Nash Eq is non-cooperative (autarkic), as is Walrasian eq.

*'cooperative' game theory is a misnomer.

* what theories of cooperation we have employ Nash equilibrium, typically with infinitely repeated games and trigger-like strategies. Is this parsimonious? Cooperation is only an equilibrium if the game has an infinite or unknown number of stages. The key is that those who are supposed to punish deviators are themselves punished if they fail

to do so. Is this realistic? Do you believe that cooperation occurs because punishers of violators are themselves kept in line by the threat of punishment by others? Who in turn…

3. **Behavioral Economics (BE)**

* strategy is to explain results in experimental games as *Nash* equilibria of players with *exotic preferences*.

* That is, the non-cooperative nature of the *play* remains unchanged. BE attempts to get observed results by injecting altruism or fairness considerations –or fears of ostracism, punishment -- into preferences.

* dictator games (altruism), ultimatum games (fairness).

* tradition goes back to Mancur Olson. How does O. explain strikes? Unionization? Change payoffs (with side payments) to

make it a *Nash* equilibrium to join the union. With strikes, it's negative side payments.

4. Source of cooperation

* I think: it's neither altruism nor even fairness at the most basic level. It is *solidarity*.

"…a community experiences solidarity just in case they have *common interests,* and must work together to address them."

I.e. "We all hang together, or we each hang separately…."

We don't cooperate to help *others,* but to help *ourselves*. But we recognize that cooperation will make *each of us* (in particular, *me*) better off.

Now we may evolve to having a *sense of fairness* which says "When all are in the same boat, we should act in unison, and not take advantage of each other." In fact, I think we *have* evolved this sense of fairness. Or we could call it 'common sense.' But at its basis is the fact (that I'll demonstrate) that if we all *do* act in unison, we are better off than if we each act autarkically.

5. Symmetric games

* All players have identical preferences. I propose each asks the question: **"If we were all to play the same strategy, what would I like it to be?"** I denote the answer that all (identically) give to this question to be the *simple Kantian equilibrium (SKE).*

* Prisoners' dilemma. Strategy space $\{C,D\}$. I'd like us all to play C, if we are all to play the same strategy. Notice the equilibrium is Pareto efficient.

Random dictator game

1. Nature chooses who is dictator
2. Dictator makes offer in [0,1]

End of game.

Suppose both players have concave vNM utility functions over money lotteries *u,v*.

The Kantian equilibrium: choose x to maximize $\frac{1}{2}u(x)+\frac{1}{2}u(1-x)$. The solution is $x=\frac{1}{2}$. Recognizing 'we're all in the same boat' means *before the game begins,* before Nature moves. This involves a commitment not to reneg and view the game as a stage game once Nature has moved.

One might say *fairness* is involved: the randomness of Nature's move makes it only *fair* that we choose our strategies before we know the outcome of Her choice. But I choose to view this fairness as inducing the optimization protocol; it is not modeled as an argument of preferences.

The trust game

A popular game in the literature, which is a public-good game, used by behavioral economists. There are two players.Nature chooses one to be the first mover. 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 sub-game perfect Nash equilibrium is $x = y = 0$ if the players have self-interested preferences.

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

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

s.t.

$$0 \le x \le M$$

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

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

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

This is the SKE.

What we observe in the lab when this game is played *as a repeated game* is that many agents play the SKE in the first round, but if they are paired with a Nash player, they stop playing it. We will address this behavior later.

Random ultimatum game

* The SKE is to give one-half if chosen to be the ultimator and to accept one-half if one is the recipient.

Suppose both players have vNM preferences u. Then the Kantian says: choose (x,z) to maximize:

$$\frac{1}{2}u(x)+\frac{1}{2}u(z)$$
$$s.t. \quad z \le 1-x$$

Again, the solution is $x = z = \frac{1}{2}$.

Notice *preferences* are classical in this formulation. No fairness considerations or altruism in preferences.. I evaluate *with my*

own classical preferences what strategy I'd like each of us to play, assuming we all play the *same* strategy.

* 'I play that strategy that I would like to see *universalized*.'

* But does this idea generalize to games where players are different? Have different preferences? The answer is yes, at least for important classes of games.

Before turning to this, let's see some examples.

6. Examples: real life

1. Recycling. There's effectively no penalty for playing Nash. Nobody sees you. But many people recycle. It's a Kantian equilibrium with symmetric prefs.

2. Instructing children: "Don't throw your candy wrapper on the sidewalk. How would you feel if everyone did so?" Notice invoking the categorical imperative induces the child to *internalize the negative externality,* but *not* by asking the child to think about how *others feel* when she litters. That would be exploiting her altruism. The negative externality is made salient by asking

the child how *her* welfare would be impacted by the littering behavior of *others*.

3. Paying taxes. Most people pay, *not* because they fear the fine if they don't, but because they would like this action to be universalized. I pay b/c this is the action I'd like everyone to choose (using my own self-regarding preferences). This is also a case of our 'all being in the same boat.'

4. Courageous soldiers in battle. I fight to defend my comrades, b/c that's the action I'd like them to take on my account. This is *not* altruism. (There may *also* be altruism – that I care about my comrades. Here is a case

wehre both altruism and cooperation are active.)

5. The British in WWII: 'doing my bit.'

6. Voting. This has been a bugaboo for the rational-choice model, with Nash reasoning. I vote b/c it's the action I'd like everyone to take. Voting is a SKE.

7. Charity: I give b/c I'd like all others in my situation to give. *Different* from the 'veil of ignorance' explanation: "There but for the grace God go I…" Also *different* from 'warm glow,' which is *fiddling with preferences* (Andreoni)

8. Politeness norms. Do we follow them for Kantian reasons (we'd like others to follow

them) or for Nash reasons (afraid of social ostracism if we violate)?

9. Obeying the law. (same as paying taxes).

One venue which provides *machinery* for Kantian optimization is elections. I vote for the tax rate I would like *all* of us to pay. So why is the equilibrium not Pareto efficient (deadweight loss of taxation)? Because when we optimize our labor supply against the tax rate *we do so in an autarkic manner.* We *vote* as Kantians, we *decide our labor supply* as Nash players.

Also: This is not a symmetric game. Voters have different preferences due to differential income.

Nevertheless, deciding communally on the tax rate surely generates a better equilibrium than if we each gave voluntarily to the government observing what all others were giving. I.e. the Nash Equilibrium would be awful.

The warm glow

Andreoni proposes that people get a warm glow from cooperating. You have a argument in your utility function that ‘turns

on' when you do the right thing, and adds to utility. It becomes a NE to do the right thing.

I agree that the warm glow exists. But Andreoni has it backwards. I get the warm glow *because* I've done the right thing – it's not that I do the right thing to get the warm glow. Think of helping your daughter with her algebra homework… Andreoni has reversed cause and effect.

7. Common pool resource problems

* Economic environment: n agents, utility fcns $u^i(x,E)$, production fcn $G(E^S), E^S = \sum E^i$. u^i, G concave.

* Example: Fishing on a lake with decreasing returns in effort. The allocation of fish is:

$$x^i = \frac{E^i}{E^S} G(E^S) \ .$$

* Pareto efficiency requires:

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

$$i.e. \quad MRS^i = MRT$$

* The Nash Eqm of game where efforts are strategies satisfies:

$$(\forall i)(MRS^i = G'(E^S)\frac{E^i}{E^S} + \frac{G(E^S)}{E^S}(1 - \frac{E^i}{E^S})$$

* So are n becomes large, MRS^i converges to the *average* product, not the marginal product. This is simplest example of the 'tragedy of the commons.'

* Now suppose all fishers have the same u. Each asks, "What effort would I like all of us to play?"

maximize $u(\frac{1}{n}G(nE), E)$: FOC is

$$\frac{d}{dE}u(\frac{1}{n}G(nE), E) = \frac{1}{n}nG'(nE)u_1 + u_2 = 0$$

$$\Rightarrow G'(nE) = -\frac{u_2}{u_1}!$$

The simple Kantian equilibrium is Pareto efficient.

8. General result on Pareto efficiency

Let P^i be the payoff functions for $i = 1,...,n$ of a symmetric game with strategy space *S*. The game is *monotone increasing (decreasing)* if for all *i*, $P^i(s^i, s^{-i})$ is increasing (decreasing) in s^{-i}. The TC is a monotone decreasing game.

Proposition. *The SKE of a monotone game (inc or dec) is Pareto efficient.*

Proof:

Let $(s^*,...,s^*)$ be the SKE of a montone increasing game. Suppose it is Pareto dominated by $(s^1,...,s^n)$. Let $s^i = \max_{1 \le j \le n} s^j$.

Of course, $P^i(s^1,...,s^n) \ge P^i(s^*,...,s^*)$. But therefore:

$$P^i(s^i, s^i, ..., s^i) > P^i(s^1, ..., s^n) \geq P^i(s^*, ..., s^*) ,$$

contradicting the fact that s^* is the simple Kantian equilibrium. qed

The PD is a monotone decreasing game where $S = \{0,1\}$ and 1 is Cooperate, 0 is Defect. Each person's payoff is monotone increasing in the other's strategy. The trust game is a monotone increasing game. The fisher's game (common pool resource problem) is a montone *decreasing* game.

Note. Although it immediately follows that the common-pool resource game's SKE is

Pareto efficient *in the game*, it does not follow it is Pareto efficient *in the economy*. The *game* fixes all allocations to be proportional. The economy makes requires such restriction. Above, we showed the SKE is indeed PE in the *economy* as well as the game.

9. Generalization: Heterogeneous preferences

Consider the common pool resource problem (fishing) with arbitrary $\{u^i\}$ and concave G. At an allocation $(E^1,.....,E^i,....,E^n)$, I think: "I'd like to expand my fishing effort by 10%. But I *should do so* only if I'd be happy if *everyone* expanded her effort by 10%." That is, if my deviation were *universalized.*

*** A multiplicative Kantian equilibrium is an effort allocation** $(E^1,.....,E^i,....,E^n)$ **such that *nobody* would like *everybody* to multiply his effort by *any* factor *r*.**

What are the payoff functions?

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

$(E^1,...,E^n)$ is a multiplicative Kantian equilibrium if the value of *r* that maximizes

$$u^i(\frac{rE^i}{rE^S}G(rE^S),rE^i)$$

for all *i* is ONE! *Nobody* would advocate re-scaling *everybody's* labor by *any* non-neg. factor.

The FOC is:

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

Expand:

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

$$\frac{E^i}{E^S}G'(E^S)u_1^iE^S + u_2^iE^i \Rightarrow$$

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

Multiplicative Kantian Equilibrium ($K^{\times}$) is Pareto Efficient! Solves Tragedy of the Commons.

Notice each fisher uses *classical, self-regarding preferences.* No altruism. What has changed is the *optimization protocol*. Fairness resides in the optimization protocol: I only increase my labor supply by 10% if I'd be happy were all to do so.

Can you get this result (PE) with altruism or by adding 'exotic' arguments to utility functions? To be addressed below.

10. Varying the optimization protocol

Since Nash, or even earlier, economists have used a *fixed* optimization protocol: autarkic. The counterfactual is that everyone else is inert, and only I optimize. Hence to explain 'strange' observations, our only modeling *variable* is preferences. Hence, behavioral economics fiddles with preferences.

But I claim it's more parsimonious to leave preferences classical, and vary the *optimization protocol*. It is not a stretch in the symmetric situation for each to think in the Kantian way. Yes, fairness is important:

but the concept of fairness induces agents to optimize in the Kantian way. We do *not* model fairness as an argument of preferences. I'll argue below why the Kantian approach is superior to the BE approach.

This leads to a four-fold choice of models:

Preferences / Optim'z'n Prot.	Self-regarding	Altruistic
Nash	classical	Behavior econ
Kantian	this talk	another talk

My view is that attempting to explain all non-classical behavior as falling in the blue box is like drawing Ptolemaic epicycles but keeping

the Earth at the center, Earth being Nash Equilibrium. It's much simpler to move to the red box, varying the optimization protocol.

11. M.Olson and E.Ostrom

* Olson: workers join unions because of side-payments. They stay on strike because of fear of penalties if they scab.

* Isn't it more reasonable to say "I join the union because that's what I'd like everybody to do." "I stay on strike b/c that's what I'd like of us all to do."

* E. Ostrom studied 'fishing economies' and found they solve the TC. She proposed: do this by using fines, ostracism, etc., for those who fish more than allowed. In other words, she *changed payoffs of game* to make the PE solution a Nash Equilibrium.

As I said, if penalties are used to control bad guys, the only time they work is if the game has an *infinite or unknown horizon*. In a finite horizon game, any Nash player will not apply a penalty to a non-cooperator. So again, we have a *Ptolemaic* attempt to harness Nash Equilibrium to explain coop.

A mixture of types

More realistically, we have *a mixture of types*. Some Kantians, some Nash. I think the penalties are needed to control the Nash players. But only Kantian players will apply them! I also think there are many Kantian

optimizers. My view is that cooperation would fall apart if *everyone* were a Nash optimizer.

Is it reasonable to think a community could maintain the Efficient Solution to the fishing problem if *everyone* were a Nash optimizer, even with penalties?

12. A hunting economy

Hunters in pre-agricultural societies shared the catch equally, not proportionally.

Allocation rule:

$$x^i = \frac{G(E^S)}{n} \ .$$

The Nash Equilibrium of the Hunting Game is characterized by:

$$MRS^i = \frac{MRT}{n} \ .$$

Only get PE solution when $n = 1!$.

Now suppose a hunter says, at an allocation $(E^1,...,E^i,...,E^n)$. "I'd like to take a nap for two hours. But I *should only do so* if

I'd be happy if everybody reduced his effort by two hours."

An *additive Kantian equilibrium* (K^+) is an allocation such that nobody would like to add a constant to *all* players' actions.

The equilibrium in this case satisfies:

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

You may compute this FOC reduces to:

$$MRS^i = MRT \ .$$

The *additive Kantian equilibrium* is PE for the Hunting Game.

Lesson: the kind of *Kantian variation* that is needed for Pareto efficiency depends on the allocation rule (that maps effort vectors into consumption vectors). Kantian optimization is *context dependent.*

Additive and multiplicative conceptions of Kantian optimization are complex. Is it reasonable to suppose they are used, or is this a mathematical curiosum?

First, let us note that Nash equilibrium is a complex process as well. It is difficult to rationalize. Think of the dynamic process that leads to NE in well-behaved games:

iterated best responses. At each state in the dynamic process, agents choose their best response assuming that others' are fixed, an assumption which is immediately falsified in the dynamic. So justifying NE is difficult. But we assume that, in many contexts, players find the NE.

Let us consider the fishing game. Granted, the multiplicative K. protocol is complex. But let's suppose through experience, fishers have identified two behaviors with regard to fishing times – the 'right' behavior and the 'wrong one.' With this highly simplified strategy space, the

common-pool resource game becomes a prisoners' dilemma! So it suffices for Kantians to play the SKE: Kantians all choose the 'right' move.

The results on Pareto efficiency with $K^{\times}$ and K^{+} optimization show that the concept of 'all making the same move' generalizes to games with heterogeneous preferences and continuous strategy spaces. But it's probably the reduced game on the simple, discrete strategy space and its SKE that explains actual solutions in reality to the tragedy of the commons.

13. Evolutionary story

* Suppose there are many fishing tribes, each with their lake. A clever priest (Archimedes) in one of them proposes multiplicative Kantian optimization…..

*Remember, if all fishers have the same preferences, the protocol is *very simple*

* This is a story of *group selection* through cultural adaptation. Kantian optimization is a meme. Cultural evolution: Richerson and Boyd (1985). Group selection is suspect for genes, but it is quite compelling for culture. So Kantian reasoning may well have

survived through group selection and the Pareto efficiency that it entails.

14. Can Kantians survive in competition with Nashers?

Consider the class of 2 x 2 symmetric games of the form:

	X	**Y**
X	(1,1)	(*a*,*b*)
Y	(*b*,*a*)	(0,0)

where $a,b \in \Re$. This is the generic class of all 2 x 2 symmetric games.

Suppose there is a large population of players, some of whom are Kantians and some Nashers, and they are randomly paired at each date to play a fixed (*a,b*) game. The game is one with mixed strategies: so the

strategy space for each player is [0,1]. There is a well-defined SKE for each game: it is the mixed strategy that each player would like both to play (to maximize his expected payoff). There is also at least one NE in each (a,b) game.

Suppose the fraction of Kantians in the population is ν. When two players meet, they cannot identify the type of the other player. Nashers always play the NE (suppose, for now, that the NE is unique). Kantians always play the SKE. This gives rise to an average payoff for the Kantians, denoted $V^K(\nu;a,b)$ and an average payoff for the Nashers, denoted $V^N(\nu;a,b)$. If $V^K > V^N$,

then the fraction of Kantians increases, and if $V^N > V^K$ then the fraction of Nashers increases.

Definition. A frequency ν^* of Kantians in the population is *evolutionarily stable* if:

(i) $V^N(\nu^*;a,b) = V^K(\nu^*;a,b)$,and

(ii) $\left.\frac{d(\Delta\nu)}{d\nu}\right|_{\nu=\nu^*} < 0$.

The second condition says any slight displacement from the stationary frequency is self-correcting… we return to the stationary state.

Definition. An (*a,b*) game is a *coordination game* if has possesses two pure-strategy Nash equilibria which are Pareto-ranked. (I.e., (X,X) and (Y,Y).)

Theorem

For the class of (a,b) *games, the following conditions are equivalent:*

a. $a < 0$ and $b < 1$.

b. Kantian players drive Nash players to extinction as long as Nashers do not play the Kantian strategy as well.

c. The game is one of pure coordination.

d. The game is supermodular and a mixed-strategy Nash equilibrium exists.

In words, in all other (a,b) games (including the PD games) Nashers drive Kantians to extinction.

Supermodularity means that

$$\frac{\partial^2 V(p,q)}{\partial p \partial q} > 0 \ .$$

This means, roughly speaking, that cooperation begets more cooperation. That is, the rate at which my payoff increases as I cooperate more (increase p) increases as you cooperate more (increase q).

Here is an example of a pure-coordination game from Tomasello (2016). A male and a female are competing for food. The issue is whether to **Grab** the food for oneself or to **Share** it. Each player has some interest in the other player's health, because they might be future mates. So the payoff matrix is:

	Share	Grab
Share	(1,1)	(-1,0.5)
Grab	(0.5,-1)	(0,0)

This is a pure coordination game: both (1,1) and (0,0) are Nash equilibria. If this game is the one being played in the evolutionary context, then Kantians will drive Nashers to extinction – unless Nashers learn to play the (Share, Share) NE almost always.

The PD is not a pure coordination game. If this is the game being played, then Nashers drive Kantians to extinction.

15. Can Nashers and Kantians co-exist?

(due to B. Unveren)

A corollary of the last theorem is that – unless the game is one where the NE and SKE strategies are the same, one type drives the other to extinction. But in real life, we observe both types of player. How can this be explained?

Suppose there are two games being played: with probability φ Nature sets the game as a pure-coordination game (a,b) and with probability $1-\varphi$, she sets the game at a PD game (α,β).

Proposition *If* φ *is sufficiently close to one, there is an evolutionarily stable equilibrium in which both types of player co-exist*.

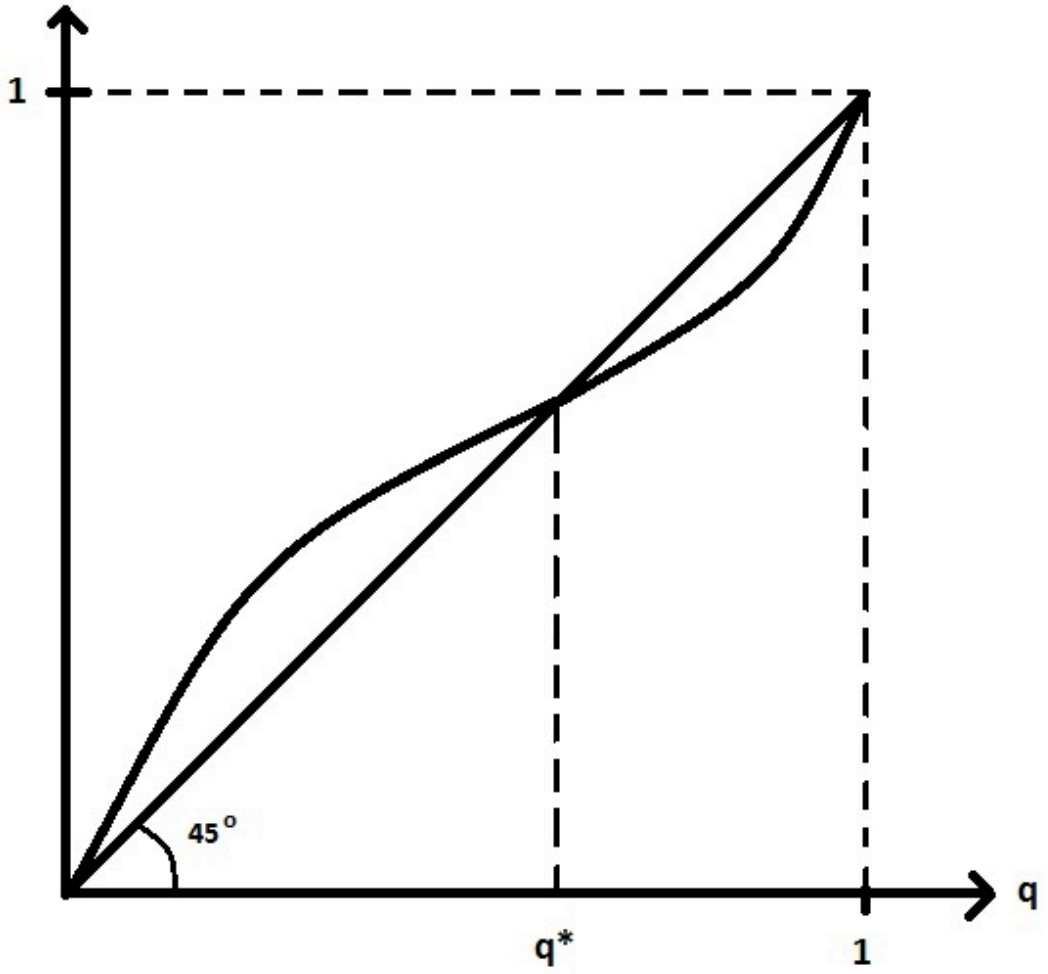
This seems to be a nice resolution; neither type is driven to extinction. It explains why we see both types in the world.

16. More on mixture of types

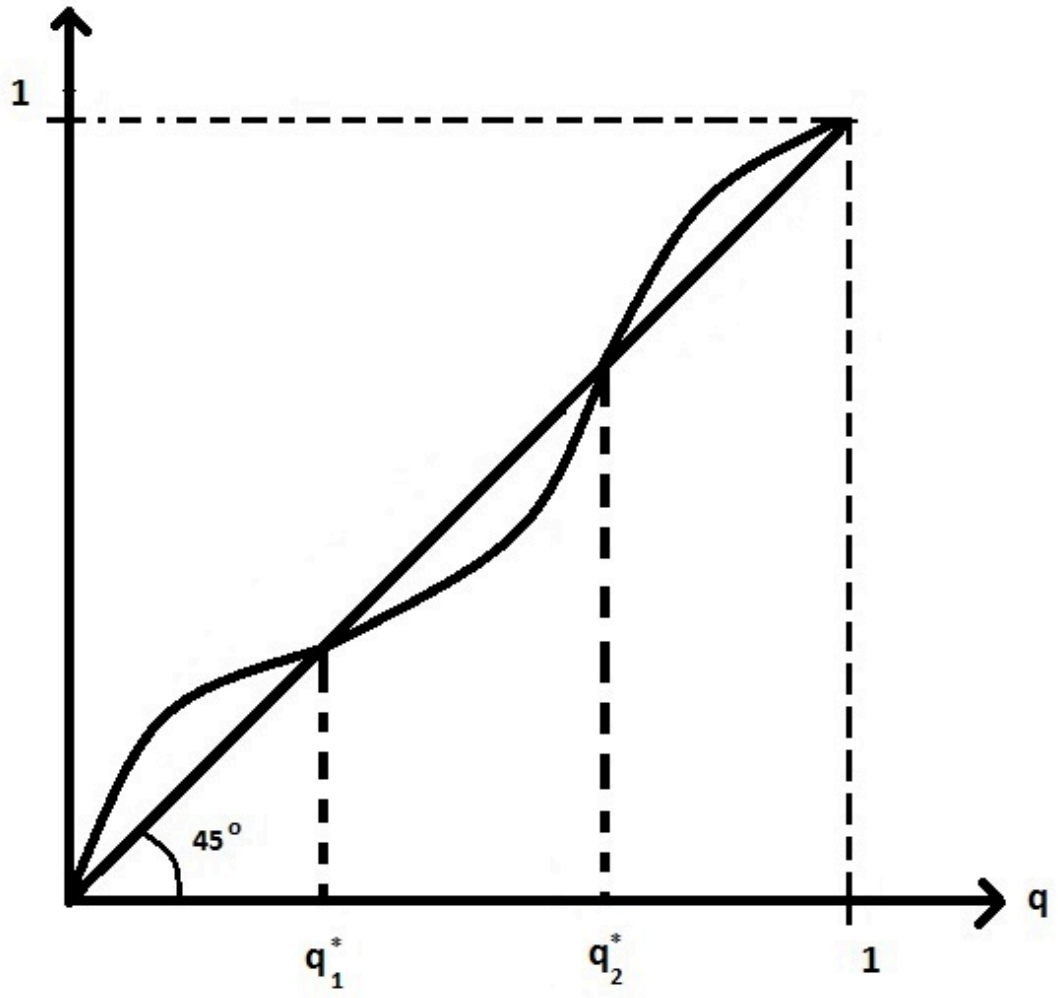
Recycling game. Each person *i* has a threshold: *i* will recycle if and only if she sees that fraction at least q^i are recycling. There is a distribution function *Q* of q^i . Those for whom $q^i = 0$ are *unconditional Kantians*. Those for whom $q^i = 1$ are unconditional Nash. Most people are *conditional Kantians:* that is, $0 < q < 1$.

What's the recycling equilibrium? Depends upon shape of *Q*.

In the case below, the only stable Kantian equilibrium is at q^* .

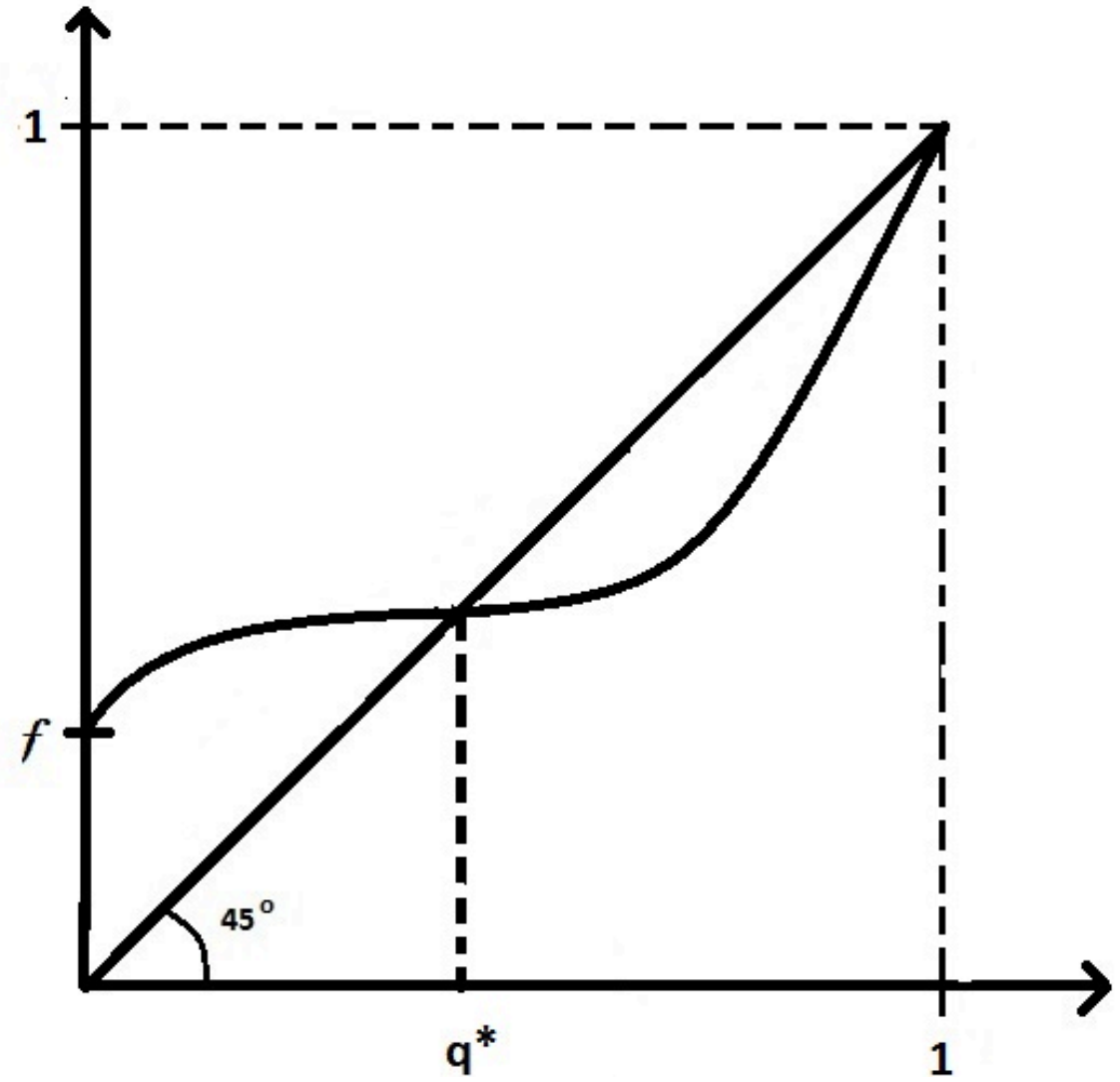

1
45°
q
q*
1

In this case, there are two stable Kantian equilibria, at q^* and 1 .



Finally here is a case where a group of people form a club and all agree to start the

ball rolling:



The **consequence is an equilibrium at** q^* **.**

17. Altruism

Suppose agents have altruistic preferences of the form:

$$U^i(u^1,...,u^n) = u^i(x^i,E^i) + \alpha^i S(u^1,u^2,...,u^n)$$

where $\alpha^i \geq 0$.

Suppose this is a fishing game, so allocation rule is

$x^i = \frac{E^i}{E^S} G(E^S)$. Then:

Theorem. *The Kantian equilibria of this game are* identical *to the Kantian equilibria of the game with* $\alpha^i = 0$.

In other words, the games with altruism and without it are *observationally equivalent.* It is therefore superfluous to suppose people

are being altruistic when they achieve cooperative solutions. The full benefits of cooperation are realized from Kantian optimization without altruism. This again points to the importance of distinguishing cooperation from altruism.

Tomasello believes that 'sympathy' [altruism] develops first, and then cooperation. I am skeptical. Once cooperation exists (in the sense of KE) it seems that altruism adds no new equilibria.

18. Distinguishing Nash equilibrium with altruism from Kantian equilibrium without it

A Nash theorist, skeptical of this analysis, could say, "Well, you can always rationalize a Kantian equilibrium as a Nash equilibrium of a game where each player has preferences over all players' welfare. So you've no reason to take recourse in this Kantian protocol, which is a distraction."

I ask two questions here: (1) Is this true? and, if so (2) Is it convincing?

To define the question precisely, consider the common-pool resource problem of the fishing game. Consider two players, with

utility functions u^1, u^2 over fish and labor. Thus their classical payoff functions are:

$$P^i(E^1, E^2) = u^i(\frac{E^i}{E^S} G(E^S), E^i) \ .$$

Now suppose that $u^1 = u^2$. Suppose each player instead desires to maximize the *sum* of utilities. So the payoff functions are:

$$\hat{P}^i(E^1, E^2) = u(\frac{E^1}{E^S} G(E^S), E^1) + u(\frac{E^2}{E^S} G(E^S), E^2) \ .$$

It is easy to prove that *the Nash equilibrium of the game* $(\hat{P}^1, \hat{P}^2)$ *is the SKE of the game* (P^1, P^2) .

So the answer to the first question, in this case , is YES. Whether players are playing the simple Kantian equilibrium in the game with self-regarding preferences, or the Nash

equilibrium in the game with altruistic preferences cannot be decided by observation.

In fact, in virtually all laboratory games, the players are endowed with identical utility functions. *So these experiments are incapable of distinguishing between 'altruism' and Kantian reasoning* as an explanation of what appears to be a Kantian equilibrium in many experiments.

Does this generalize? We have:

Proposition. *There are exactly two cases when the multiplicative Kantian equilibrium of the fisher game is identical to the Nash equilibrium of the game with extended*

'altruistic' preferences: when the utility functions are identical, or when they are each quasi-linear ($u^i(x,E) = x - h^i(E)$.)

Furthermore, the altruism comes with a caveat. This is a *cardinal* result. Suppose the two players have the same *preferences* but different utility functions are chosen to represent them. It no longer follows that the NE of the game in which players maximize the sum of utilities is the Kantian equilibrium.

In lab games, this issue does not arise, because players are usually assigned cardinal payoffs – and it would be natural to add them up. But it *is* a problem for a general rationalization.

Does this result generalize to arbitrary preferences – not identical, and not quasi-linear? We have:

Theorem *Let* G *be any differentiable concave production function. Let*

$$u^1(x,E) = x(1-E)^m, u^2(x,E) = x(1-E)^n \text{ ,}$$

$0 < m < n < \infty$ *be Cobb-Douglas utility functions for two fishers. Then there exists a differentiable social welfare function* **V** *such that in the game induced by the economy* **(V,V,G)**, *where the preferences of each player are given by* $V(u^1(\cdot,\cdot),u^2(\cdot,\cdot))$ *, the Nash equilibrium is the multiplicative Kantian equilibrium of the game* (u^1,u^2,G) .

So the answer to the first question is YES.

Indeed, the theorem is true for any differentiable concave utility functions, not just Cobb-Douglas. *It is therefore formally possible to represent Kantian equilibria in the common-pool resource problem as Nash equilibria of games with extended preferences,* in which each player is maximizing some social welfare function.

However, is this *convincing?* That is, can it be an explanation of how players solve the tragedy of the commons? I say no: because the function V is in general very complicated. For instance, in the Cobb-Douglas case, we cannot compute V in closed form. It is only in the two cases mentioned earlier (identical

or quasi-linear utility functions) that $V(u^1,u^2)=u^1+u^2$ works.

It is decidedly *not* the case that the same social welfare function works for any pair of utility functions u^1,u^2. So if you wish to argue for this Nash rationalization, you must accept the fact that the SWF adopted by players depends upon their preferences! This sharply weakens the attraction of the 'Nash justification.'

Again: in almost all lab games, players are endowed with the same cardinal utility functions, and so in these experiments we cannot distinguish observationally between Kantian optimization or maximization of the

sum of payoffs.

19. More on Kantian protocol vs. Behavioral Econ

I do not object to calling Kantian optimization a species of behavioral economics. But for pedagogical purposes, I prefer to characterize BE, as it has been thus far practiced, as the move of *altering preferences from classical ones to ones with exotic preferences so that the Nash equilibrium of the altered game conforms to observation.* Kantian optimization, to repeat, maintains classical preferences, but alters the optimization protocol.

One strong argument against the BE approach is that one generally has to know *in*

advance what the cooperative equilibrium is. One then builds into preferences the view that playing that action is *fair*. If this is properly done, then the NE of the new game will be the fair allocation.

But the Kantian protocol does not have to identify the fair action *a priori.* In the fishing or hunting games (u^1, u^2, G), the strategy space is a real interval of labor choices: it is not *a priori* obvious what the cooperative (i.e., Pareto efficient) solution is. But Kantian optimization finds it. In this sense, Kantian equilibrium is a mathematical cousin of Nash equilibrium. It locates an equilibrium in a large class of games where it

is not *a priori* obvious what the cooperative solution is. BE jimmies the preferences in order to produce, as an equilibrium, the behavior that everyone knows is the cooperative solution in advance.

20. Summary

* We should not try to explain successful cooperation as Nash equilibria. This can only *convincingly* be done by inserting 'exotic' arguments into preferences, and in *very simple* games.

* In simple games there is usually a clear conception of 'the right thing to do.' But it's the right thing because *if we all do it, we are better* off than if we all *don't* do it. (Eg, play C in PD)

* But for complicated games the 'right thing' to do is not obvious. So we need a way of *decentralizing* cooperation.

* All the games in lab experiments are symmetric, and here the right thing to do is usually clear. But when people have different preferences, the 'right thing to do' is unclear. Various kinds of K-equilibrium locate it.

* Important distinction between cooperation and altruism. Cooperation: "We all hang together or we each hang separately." This is a different motivation from altruism. These two concepts are confused in the behavioral econ. literature. I believe cooperation is a more fundamental principle than altruism for relations between non-kin.

**Morality* still exists in K-equilibrium: But it is in the choice of optimization protocol, not an argument of preferences. This is a parsimonious modeling choice.

* Where did we learn this morality? M. Tomasello tells us it is inborn in our species… a product of evolution…. something none of the other great ape species possess. It may well be the *precursor* to language: for language is only necessary if the members of the species perform complex projects together -- in other words, if they cooperate. Language, Tomasello claims, would die out if it started to evolve in chimpanzees, because, with their level of

cooperation, it is unnecessary. (Proof: Chimps do not even point or mime to communicate.)

* Tomasello believes that the key evolutionary invention that engendered cooperation among humans was the mental construct of 'joint intentionality,' the understanding that we each have, that if we perform certain actions in tandem with another, we can achieve a goal desired by each of us. His experimental work consists in demonstrating that human infants possess joint intentionality, but JI is lacking in chimps, bonobos , gorillas and orang-outans.

* I view the theory I've outlined here as providing a micro-foundation for turning JI into equilibria --- micro in the sense of being rationalized at the level of individual decision making.

* If the view that the major achievements of HS are due to cooperation, this is a hugely neglected area of economic analysis.