HOW DO ALTRUISTS AND MORALISTS BEHAVE?

Jörgen Weibull

MSE seminar
September 22, 2017

This talk is based on joint work with *Ingela Alger* (TSE and IAST)

1. "Homo Moralis – Preference Evolution under Incomplete Information and Assortativity," *Econometrica* (2013)

2. "Evolution and Kantian Morality," *Games and Economic Behavior* (2016)

3. "Morality: Evolutionary Foundations and Policy Implications," chapter in forthcoming *MIT Press book* (2017)

4. "Strategic behavior of altruists and moralists", *Games* (2017)

and in part on

5. Miettinen, Kosfeld, Fehr, and Weibull: "Revealed preferences in a sequential prisoners' dilemma: a horse-race between five utility functions", CESifo WP (2017)

1 Introduction

- There is overwhelming experimental evidence against predictions based on *Homo oeconomicus*, pure material self interest

- It thus appears relevant to ask whether and how richer motivations affect outcomes in standard economic interactions

- This is the question I will address in this seminar, focusing on *altruism* and (*Kantian*) *morality*

- It is commonly believed that if an element of altruism or morality were added to economic agents' self-interest, then "the world would be a better place", or, more specifically, economic outcomes would improve

- Presumably, people would
  - cheat less and be more trusting
  - work hard even when not monitored or paid fixed wages
  - contribute more to public goods
  - more easily maintain long-run cooperation

- While this has certainly proved to be right in many interactions, this belief is not generally valid

  - Lindbeck and Weibull (1988) show that *altruism can diminish welfare* among strategically interacting individuals, engaged in intertemporal decision-making. The reason is the lack of commitment to not help each other. (Mitigated by compulsory pension schemes.)

  - Bernheim and Stark (1988) show that *altruism may be harmful to long-run cooperation*: although altruism diminishes the temptation to defect (since defecting harms others), it also diminishes the severity and credibility of punishments after deviations (a loving parent has difficulty to credibly threaten a child's misbehavior by punishment)

- The aim of today's discussion is to examine strategic interactions between altruists, as well as between moralists, in order to shed light on

the complex and non-trivial effects of altruism and morality on equilibrium behavior and implications for welfare

- By 'altruism' we here mean that an individual cares not only about own material welfare but also about the material welfare of others, in line with Becker (1974,1976), Andreoni (1988), Bernheim and Stark (1988), and Lindbeck and Weibull (1988).

- As for 'morality' we rely on recent results in the literature on preference evolution, results which show that a certain preference class, *Homo moralis*, is favored by natural selection (Alger and Weibull, 2013, 2016). Such preference give some weight to own material welfare but also to "what is the right thing to do" if others would act likewise.

- The main difference between altruism and morality that while the first is purely consequentialistic, the second is partly deontological

2 Modelling altruists and moralists

- We consider n-player normal-form games in which each player has the same strategy set X, and where $\pi(x, \boldsymbol{y})$ is the *material payoff* from strategy $x \in X$, when used against strategy profile $\boldsymbol{y} = (y_1, y_2, ..., y_{n-1}) \in X^{n-1}$ used the others
  - By 'material payoff' we mean the resulting consumption utility for the invidual in question

- We assume π to be *aggregative* in the sense that $\pi(x, \boldsymbol{y})$ is invariant under permutation of the components of $\boldsymbol{y}$

- The strategy set X is any non-empty, compact and convex set in any normed vector space

- In particular, the interaction may be dynamic, involve imperfect information, etc.

- We say that an individual i is a *Homo oeconomicus* if his utility function is

$$u\left(x_i, \boldsymbol{x}_{-i}\right) = \pi\left(x_i, \boldsymbol{x}_{-i}\right) \quad \forall \left(x_i, \boldsymbol{x}_{-i}\right) \in X^n.$$

- An individual i is an *altruist* with degree of altruism $\alpha_i \in [0, 1]$ if her utility function is

$$v\left(x_i, \boldsymbol{x}_{-i}\right) = \pi\left(x_i, \boldsymbol{x}_{-i}\right) + \alpha_i \cdot \sum_{j \neq i} \pi\left(x_j, \boldsymbol{x}_{-j}\right) \quad \forall \left(x_i, \boldsymbol{x}_{-i}\right) \in X^n.$$

  - For $\alpha_i = 0$, this is *Homo oeconomicus*

  - For $\alpha_i = 1$, this is the *full-blood altruist* who gives equal weight to everybody's material payoff ("who loves others as much as herself")

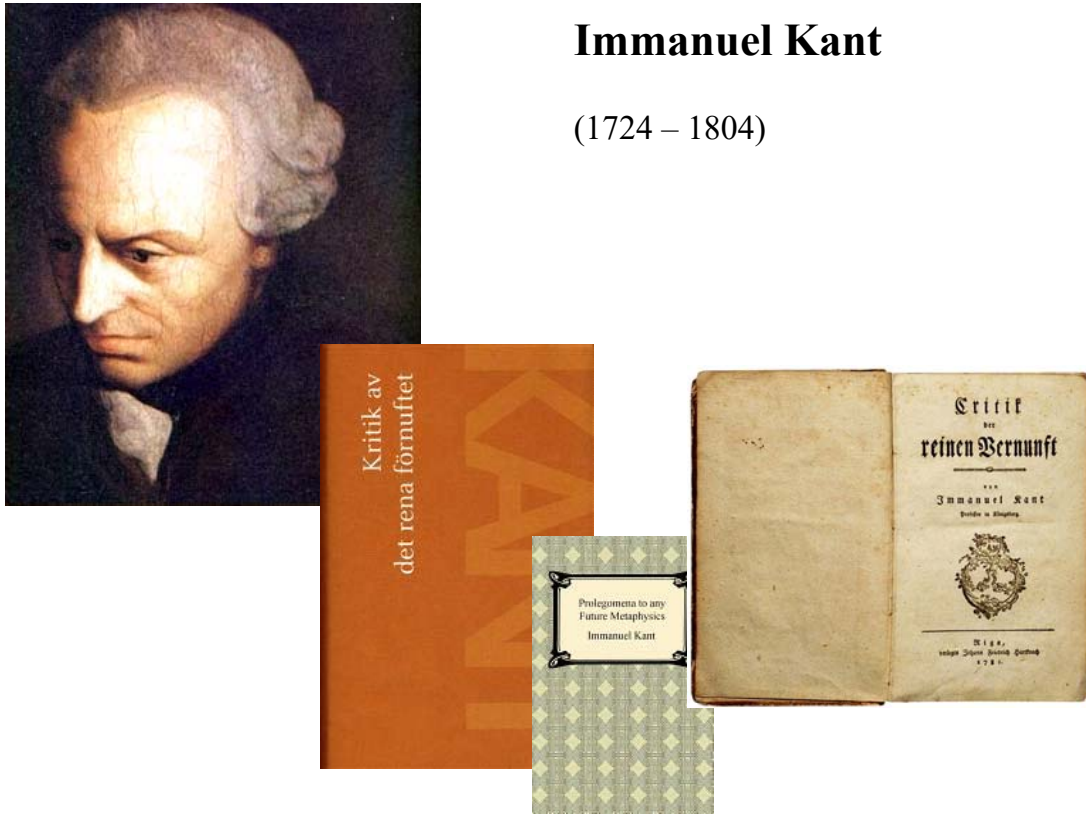
- An individual i is a *Homo moralis* with degree of morality $\kappa \in [0,1]$ if his utility function is

$$w\left(x_i, \boldsymbol{x}_{-i}\right) = \mathbb{E}\left[\pi\left(x_i, \tilde{\boldsymbol{x}}_{-i}^{m}\right)\right] \quad \forall\left(x_i, \boldsymbol{x}_{-i}\right) \in X^n \tag{1}$$

  where $\tilde{\boldsymbol{x}}_{-i}^{m}$ is a random $(n-1)$-vector such that with probability $\kappa_i^m \left(1-\kappa_i\right)^{n-m-1}$ exactly m of the $n-1$ components of $\boldsymbol{x}_{-i}$ are replaced by x_i

  - For $\kappa_i = 0$, this is (again) *Homo oeconomicus*
  - For $\kappa_i = 1$, this is *Homo kantiensis*, who chooses a strategy that, if used by everyone, would maximise everybody's material payoff

"Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."
[Immanuel Kant, *Groundwork of the Metaphysics of Morals*, 1785]



- In pairwise interactions ($n = 2$):
  - the *altruist:*

$$v(x, y) = \pi(x, y) + \alpha \cdot \pi(y, x)$$

  - the *moralist*:

$$w(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x)$$

- To the best of our knowledge, *Homo moralis* preferences have not been analyzed, or even known, before

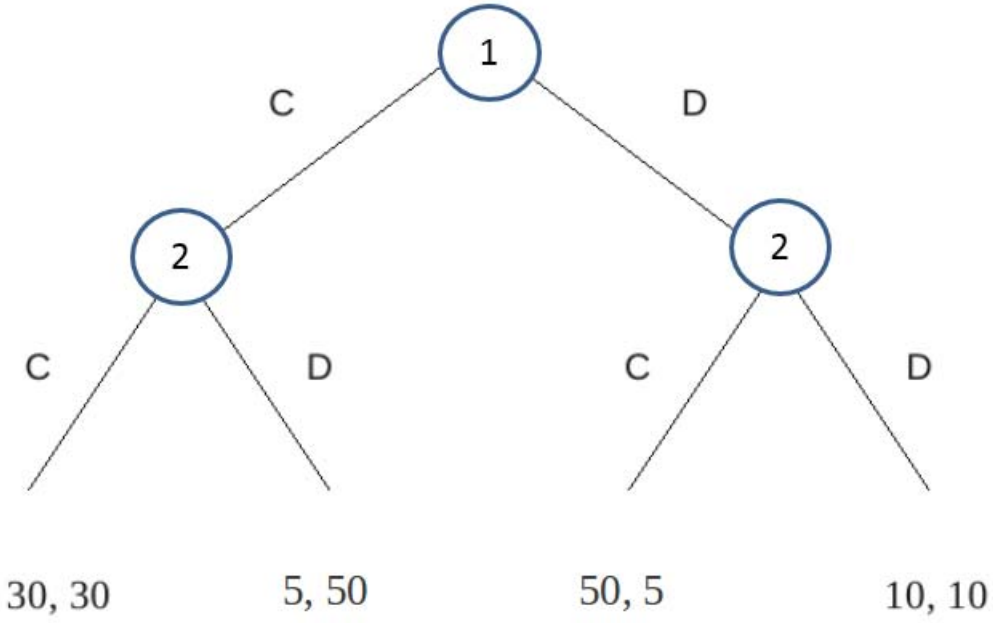
- Is there any empirical evidence for their existence?

- What is the predictive power of *Homo moralis*, in comparison with altrusim and other social preferences in behavioral economics?

3 Experimental evidence

[Miettinen, Kosfeld, Fehr & Weibull (2017)]

- Anonymous random matching of 98 master students from ETH and Zürich University to play *a sequential prisoners' dilemma* in material payoffs


1
C
D
2
2
C
D
C
D
30, 30
5, 50
50, 5
10, 10

- What percentage of the subjects behave in accordance with
  - Homo oeconomicus?
  - Altruism (Becker)?
  - Inequity aversion (Fehr-Schmidt)?
  - Conditional concern for welfare (Charness-Rabin)?
  - Homo moralis (Alger-Weibull)?

PRELIMINARY RESULTS

Model	**hit rate**	**parameters**
Homo oeconomicus	28%	0
Altruism	44%	1
Inequity aversion	60%	2
Conditional welfare	82%	2
Homo moralis	83%	1

- More comprehensive experiments, in joint work with Ingela Alger & Boris van Leeuwen (Tilburg), have recently been carried out

4 How altruists and moralists behave

4.1 Public goods

Let

$$\pi\left(x_i, \boldsymbol{x}_{-i}\right) = B\left(\sum_{j=1}^{n} x_j\right) - C\left(x_i\right),$$

where $x_i \geq 0$ is i's contribution to the public good, and B and C are "well behaved" production and cost functions, respectively.

- Played by altruists with common degree of altruism α: there exists a unique Nash equilibrium. Each participant contributes $x^A\left(\alpha\right)$

- Played by moralists with common degree of morality κ: there exists a unique Nash equilibrium. Each participant contributes $x^M\left(\kappa\right)$

- If $\alpha = \kappa$, then $x^A(\alpha) = x^M(\kappa)$
  - Hence, the behavioral effects of morality and altruism are here indistinguishable

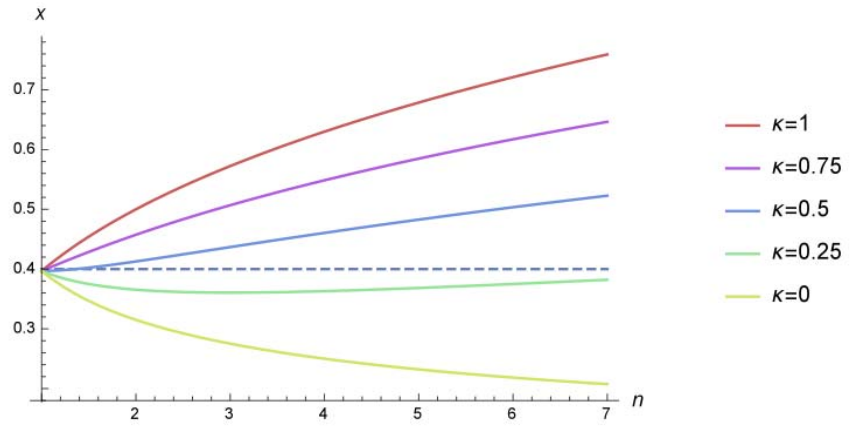


Figure 1: The unique Nash equilibrium contribution in the public-goods game for different degrees of morality.

4.2 Infinitely repeated prisoners' dilemmas

4.2.1 Altruists

Consider a prisoners' dilemma played by two equally altruistic and patient individuals:

	C	D
C	$(1+\alpha)\,R$	$S+\alpha T$
D	$T+\alpha S$	$(1+\alpha)\,P$

where $S < P < R < T$.

- Cooperation, (C, C), is a Nash equilibrium iff $\alpha \geq \alpha^*$

$$\alpha^* = \frac{T-R}{R-S}. \tag{2}$$

- Grim trigger, if used by both players, is a subgame-perfect equilibrium that sustains perpetual cooperation in the infinitely repeated game with discount factor $\delta \in (0,1)$ iff

$$(1+\alpha)\,R \;\geq\; (1-\delta)\cdot(T+\alpha S) \;+\; \delta\,(1+\alpha)\,P \tag{3}$$

and

$$\alpha \leq \alpha^{**} = \frac{P-S}{T-P}. \tag{4}$$

  - The first inequality makes one-shot deviations from cooperation (C,C) unprofitable

  - The second inequality makes a one-shot deviation from mutual punishment (D,D) unprofitable

- Note that (3) holds iff $\delta \geq \delta_A$, where

$$\delta_A = \frac{T-R-\alpha\,(R-S)}{T-P-\alpha\,(P-S)}. \tag{5}$$

- In sum, perpetual cooperation can be sustained if
  - altruism is strong enough, $\alpha \geq \alpha^*$ (irrespective of δ)

    or

  - players are selfish enough to credibly punish defection, $\alpha \leq \min\{\alpha^*, \alpha^{**}\}$, and patient enough to prefer the long-term benefits of cooperation over the immediate reward from defection, $\delta \geq \delta_A$

- In the intermediate case, $\alpha^{**} < \alpha < \alpha^*$, cooperation is not sustainable for any discount factor δ

4.2.2 Moralists

Consider the same prisoners' dilemma played by two equally moral and patient individualsT

	C	D
C	R	$(1-\kappa)\,S+\kappa R$
D	$(1-\kappa)\,T+\kappa P$	P

- Cooperation, (C,C), is a Nash equilibrium iff $\kappa \geq \kappa^*$, where

$$\kappa^* = \frac{T-R}{T-P}.$$

- We note that

$$\begin{cases} \alpha^* < \kappa^* & \text{if } R-S > T-P \\ \alpha^* = \kappa^* & \text{if } R-S = T-P \\ \alpha^* > \kappa^* & \text{if } R-S < T-P \end{cases}$$

- Grim trigger sustains perpetual cooperation between two equally moral individuals as a subgame perfect equilibrium outcome if $\delta \geq \delta_K$, where

$$\delta_K = \frac{T - R - \kappa(T - P)}{T - P - \kappa(T - P)} \tag{6}$$

and

$$\kappa \leq \kappa^{**} = \frac{P - S}{R - S} \tag{7}$$

- In sum, perpetual cooperation can be sustained if
  - morality is sufficiently strong, $\kappa \geq \kappa^*$ (irrespective of δ)

    or

  - they are sufficiently selfish, $\kappa \leq \min\{\kappa^*, \kappa^{**}\}$, and patient, $\delta \geq \delta_K$.

4.2.3 Comparing morality with altruism

Consider a pair of equally altruistic players with $\alpha \in [0, 1]$ with a pair of equally moral players with degree of morality $\kappa = \alpha$

- First, suppoes that $T - R = P - S$. Then

$$\alpha^{**} = \kappa^{**} = \alpha^{*} = \kappa^{*}$$

- Second, suppose that $T - R > P - S$. Then $\kappa^{*} < \alpha^{*}$. If $\alpha \in (\kappa^{*}, \alpha^{*})$, then (C, C) is an equilibrium of the stage game between moralists but not of the stage game between altruists. Moreover,

$$\kappa^{*} < \kappa^{**} \quad \text{and} \quad \alpha^{**} < \alpha^{*} \quad \text{and} \quad \alpha^{**} < \kappa^{**}$$

  - hence, there are values of α for which altruists are not able to sustain cooperation for any discount factor δ, whereas a pair of moralists with any degree of morality $\kappa = \alpha$ can sustain perpetual cooperation if sufficiently patient

- Third, suppose that $T - R < P - S$. Then the opposite holds

4.3 Coordination

- n individuals have to simultaneously choose between action A and B
- write $s_i \in S = \{0, 1\}$ for the choice of individual i, where $s_i = 1$ means that i chooses A, and $s_i = 0$ that instead B is chosen
- write $\boldsymbol{s}_{-i} \in S^{n-1}$ for the strategy profile of all other individuals
- the material payoff to an individual from choosing A when n_A others choose action A is $n_A \cdot a$, and it is $n_B \cdot b$ from choosing B when n_B others choose B.
- Suppose that $0 < b < a$. Think of A as a socially efficient norm and B as a socially inefficient norm

- The utility function of a *Homo oeconomicus* individual is then

$$u_i\left(s_i, \boldsymbol{s}_{-i}\right) = a s_i \cdot \sum_{j \neq i} s_j + b\left(1 - s_i\right) \cdot \sum_{j \neq i} \left(1 - s_j\right) \tag{8}$$

- For an altruist it is

$$v_i\left(s_i, \boldsymbol{s}_{-i}\right) = u_i\left(s_i, \boldsymbol{s}_{-i}\right) + \alpha_i \cdot \sum_{j \neq i} u_j\left(s_j, \boldsymbol{s}_{-j}\right) \tag{9}$$

- Evidently the efficient norm A, can always be sustained as a Nash equilibrium for arbitrarily altruistic individuals

- But also the inefficient norm B is a Nash equilibrium for arbitrarily altruistic individuals

- The utility function of a *Homo moralis* is

$$w_i\left(s_i, \boldsymbol{s}_{-i}\right) = \mathbb{E}_{\kappa_i}\left[u_i\left(s_i, \tilde{\boldsymbol{s}}_{-i}^m\right)\right], \tag{10}$$

where $\tilde{\boldsymbol{s}}_{-i}^m$ is a random vector in S^{n-1} such that with probability $\kappa_i^m\left(1-\kappa_i\right)^{n-m-1}$ exactly $m \in \{0, ..., n-1\}$ of the $n-1$ components of $\boldsymbol{s}_{-i}$ are replaced by s_i, while the remaining components of $\boldsymbol{s}_{-i}$ keep their original values.

- Thanks to the linearity of the material payoffs: the efficient social norm A can clearly always be sustained as a Nash equilibrium, since when all the others are playing A, individual i gets utility $(n-1)\,a$ from taking action A, and $b\,(n-1)\,\kappa_i$ from deviating to B

- The inefficient social norm cannot be sustained for all degrees of morality, since if all the others play B, then individual i gets utility $(n-1)\,b$ from also playing B, and would get utility $a\,(n-1)\,\kappa_i$ from deviating to A. Hence, the inefficient norm B can be sustained in Nash equilibrium iff an only if

$$\kappa_i \leq \frac{b}{a} \quad \forall i$$

4.4 Environmental economics

- Consider the behavior of *Homo moralis* in an otherwise classical economics model of consumption with external effects [Musgrave (1959), Arrow (1970)]

- Imagine a *continuum* population of consumers, 2 consumption goods, where good 1 is environmentally neutral and good 2 is environmentally harmful

- Let p denote the relative price of good 2

- The environment depends on the population's *total* consumption of good 2

- Each consumer derives utility from own consumption of both goods, and also from the quality of the environment: $u(x_1, x_2, X_2)$

- What will happen?
  - What will a *Homo oeconomicus* do?
  - What will a *Homo kantiensis* do?
  - What will a *Homo moralis* with intermediate degree of morality do?
  - What will an *altruist* or *environmentalist* do?

- Necessary first-order condition for a *Homo moralis* with arbitrary degree of morality μ:

$$\frac{u_2\left(x_1, x_2, X_2\right)}{u_1\left(x_1, x_2, X_2\right)} \quad = \quad p \; - \; \mu \cdot \frac{u_3\left(x_1, x_2, X_2\right)}{u_1\left(x_1, x_2, X_2\right)}.$$

$\mu = 0$: *Homo oeconomicus*

$\mu = 1$: *Homo kantiensis*

$0 < \mu < 1$: *Homo moralis* with intermediate morality, reduces her consumption of the harmful good somewhat, as if it were more expensive

5 Conclusion

1. Behavioral and experimental economics, other social and behavioral sciences, everyday observation and introspection suggest that human motivation is richer, and more complex, than narrow self-interest

2. We here explore behavioral implications of altruism and of morality