EVOLUTIONARY GAME THEORY

Evolutionary foundations of morality and other-regarding preferences

Jörgen Weibull

Delhi Winter School, December 2017

Today's lecture is based on results from an ongoing research project with

Ingela Alger

(Toulouse School of Economics & Institute for Advanced Study)

1 Introduction

- Behavioral economists usually postulate specific forms of other-regarding preferences, such as
  - Altruism (Becker)
  - Inequity aversion (Fehr-Schmidt)
  - Conditional concern for welfare (Charness-Rabin)
  - Conditional altruism (Levine)

- We instead ask what preferences, if any, survive evolutionary pressures based on material-payoff performance
  - under complete information (interacting individuals know each others' preferences)
  - under incomplete information (preferences are private information)

- Today, I will give an overview of our results so far. You can find details in our publications:
  - "Kinship, incentives and evolution" (2010)
  - "Homo moralis – Preference evolution under incomplete information and assortativity" (2013)
  - "Evolution and Kantian morality" (2016)
  - "Morality: Evolutionary foundations and policy implications" (2017a)
  - "Strategic behavior of moralists and altruists" (2017b)

- I will also briefly report preliminary results from joint experimental work with Ernst Fehr, Topi Miettinen, and Michael Kosfeld:
  - "Revealed preferences in a sequential prisoners' dilemma: a horse-race between five utility functions" (2017)

2 Evolutionary foundations under complete information

[Alger and Weibull, 2010]

Imagine:

- Pairs of more or less altruistic siblings who know each other's degree of altruism, α_1 and α_2, where

$$\begin{cases} u_1(x,y) = \pi_1(x,y) + \alpha_1 \pi_2(x,y) \\ u_2(x,y) = \pi_2(x,y) + \alpha_2 \pi_1(x,y) \end{cases}$$

- Each sibling independently makes a productive effort that determines the probability distribution of his or her output

- After both siblings' outputs have materialized, they may make voluntary transfers to each other

- Given their degrees of altruism, there exists a unique subgame-perfect equilibrium for their production efforts and conditional transfers

- This equilibrium induces a probability distribution over their material payoffs

Evolutionary scenario:

- Each sibling inherits his or her degree of altruism from one parent, with equal probability for each parent

Definition:

- A degree of altruism is *evolutionarily stable*, if a mutation in the parent population (in a small population share $\varepsilon > 0$) to any other degree of altruism leads to lower expected material payoff for the mutants

Main result:

- There exists an evolutionarily stable degree of sibling altruism ("strength of family ties") and this is lower than 1/2. It depends on the "harshness of the production environment" in such a way that altruism is lower in harsher climates ("Sweden") than in milder climates ("Italy")

3 Evolutionary foundations under incomplete information

[Alger and Weibull, 2013, 2016]

Imagine:

- Individuals in a large population who are now and then randomly matched into groups of size $n > 1$ to interact with each other

- The random matching may be assortative, that is, an individual's conditional matching probability distribution may depend on the individual's type

- The interaction takes the form of a (normal-form) game in material payoffs

- Material payoff functions are *aggregative symmetric* in the sense that a participant's material payoff depends only on own strategy and on other group members' strategies, with permutation invariance among others' strategies:

$$\pi\left(x, \boldsymbol{y}\right) \quad \text{where} \quad \boldsymbol{y} = \left(y_1, ..., y_{n-1}\right)$$

- No other restrictions on the game: it may be multi-stage, involve moves by "nature", allow for cooperation, competition, signalling, punishment etc.

$$\pi : X^n \to \mathbb{R}$$

- Each individual has a *goal function*, or *utility function*, the expected value of which he or she seeks to maximize ("Savage rationality") under her probabilistic beliefs

- Any continuous utility function $u : X^n \to \mathbb{R}$ is allowed. For example, pure material self-interest ($u = \pi$), altruism, spite, inequity aversion, preference for fairness or morality, preferences that do not depend on material payoffs, "crazy" preferences

- Each individual's utility function is his or her *private information*

3.1 Definitions

We generalize Maynard-Smith's & Price's (1973) notion of an *evolutionarily stable strategy (ESS)* in symmetric 2-player games, to evolutionary stability of utility functions in aggregative symmetric n-player games:

- A utility function is *evolutionarily stable* if, when almost all individuals in the population have this utility function, and a small population share $\varepsilon > 0$ have some other utility function, the "incumbents" outperform the "mutants", in all (Bayesian) Nash equilibria under incomplete information

- A utility function is *evolutionarily unstable* if there exists another utility function such that, no matter how small its population share $\varepsilon > 0$, there is *some* (Bayesian) Nash equilibrium in which the mutants materially outperform the incumbents

We generalize from uniform random matching to potentially assortative random matching:

- Given any population share ε of mutants:
  - let $p_m(\varepsilon)$ be the probability that exactly $m = 0, 1, ...n-1$ of the other players in *an incumbent's* group are mutants, $\boldsymbol{p}(\varepsilon) \in \boldsymbol{\Delta}$
  - let $q_m(\varepsilon)$ be the probability that exactly $m = 0, 1, ..., n-1$ of the other players in *a mutant's* group are mutants, $\boldsymbol{q}(\varepsilon) \in \boldsymbol{\Delta}$

- The *assortativity profile* of the random matching is the limit probability distribution, $\boldsymbol{a} \in \boldsymbol{\Delta}$, for the number of other mutants in a mutant's group, as the population share of mutants tend to zero:

$$\boldsymbol{a} = \lim_{\varepsilon \to 0} \boldsymbol{q}(\varepsilon)$$

- Uniform random matching: $\boldsymbol{a} = (1, 0, 0, ..., 0)$
- Pairwise random matching among siblings under genetic transmission: $\boldsymbol{a} = (1/2, 1/2)$
- Random matching conditioned on geography, language, culture etc.: $\boldsymbol{a} = (a_0, a_1, ..., a_{n-1})$

Formally:

Definition 3.1 *A utility function u is* **evolutionarily stable** *against a utility function v if there exists an $\bar{\varepsilon} > 0$ such that individuals with utility function u earn a higher material payoff than individuals of with utility function v in all (Bayesian) Nash equilibria in all population states (u, v, ε) with $0 < \varepsilon \in \bar{\varepsilon}$.*

Definition 3.2 *A utility function u is* **evolutionarily unstable** *if there exists a utility function v such that for every $\bar{\varepsilon} > 0$ there exists a population state (u, v, ε) with $0 < \varepsilon \in \bar{\varepsilon}$ and a (Bayesian) Nash equilibrium in which individuals with utility function v earn a higher material payoff than those with utility function u.*

Definition 3.3 *In any population state* (u, v, ε)*, a (Bayesian)* **Nash equilibrium** *is a strategy pair* $(\hat{x}, \hat{y}) \in X^2$ *such that*

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot u\left(x, \hat{\boldsymbol{y}}^{(m)}\right) \\ \hat{y} \in \arg\max_{y \in X} \ \sum_{m=0}^{n-1} q_m(\varepsilon) \cdot v\left(y, \hat{\boldsymbol{y}}^{(m)}\right) \end{cases}$$

where $\hat{\boldsymbol{y}}^{(m)} \in X^{n-1}$ *has* m *components* $\hat{y}$ *and the rest* $\hat{x}$*.*

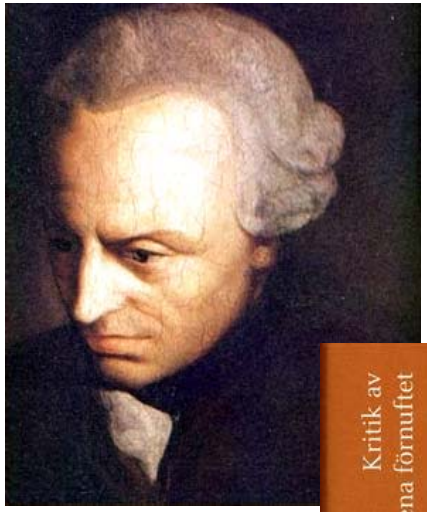
3.2 Main result

- Natural selection, as modelled here, turns out to favor a particular class of utility functions, the carriers of which we call *Homo moralis:*

Theorem 3.1 (Alger & Weibull, 2013 & 2016) Homo moralis *with morality profile* $\mu = a$ *is evolutionarily stable. Any preferences that are behaviorally distinct from these are evolutionarily unstable.*

- A *Homo moralis* places some weight on his or her expected material payoff and some weight on "the right thing to do" if, hypothetically, some or all other individuals in her group would act like him or her. The *morality profile* $\mu \in \Delta$ of a *Homo moralis* is the vector of probabilities that she places on the events that $m = 0, 1, ..., n-1$ others in the group would act likewise.

"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]



Immanuel Kant

(1724 – 1804)

- In pairwise interactions, a *Homo moralis* with *degree of morality* $\kappa \in [0, 1]$ has utility function

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

where $x \in X$ is own strategy, $y \in X$ the other group member's strategy.

  - *Homo oeconomicus*: $\kappa = 0$, someone who cares *only* about his or her own material payoff
  - *Homo kantiensis:* $\kappa = 1$, someone who cares *only* about "the right thing to do"
  - A continuum of *Homo moralis* preferences between these two extremes, with $\kappa = a_1$ being evolutionarily stable

- For groups of arbitrary group size n:

Definition 3.4 *A* **Homo moralis** *with* **morality profile** $\boldsymbol{\mu} \in \Delta$ *is an individual with goal function*

$$u(x, \boldsymbol{y}) = \mathbb{E}[\pi(x, \mathbf{Y})]$$

where $\boldsymbol{y} \in X^{n-1}$ *is the vector of other group members' strategies, and* $\mathbf{Y} \in X^{n-1}$ *is a random vector such that with probability* μ_m *exactly* m *of the* $n-1$ *components of* $\boldsymbol{y}$ *are replaced by strategy* x*, with equal probability for each subset of size* m*, while the remaining components of* $\boldsymbol{y}$ *keep their original values.*

- Why are these preferences, for $\mu = a$, evolutionarily stable and other preferences unstable?

  - *Homo moralis* of the right degree of morality *preempts entry* of mutants: the "best" any mutant can do, in order to gain material payoffs in a *Homo moralis* population, is to mimic *Homo moralis*

  - For any other, behavioral distinct preference, if incumbent: there exist a utility function that can profitably "invade" in a small scale (earn higher material payoffs). For example, rare mutants who are "committed to" a particular strategy (that is, who find it strictly dominant)

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

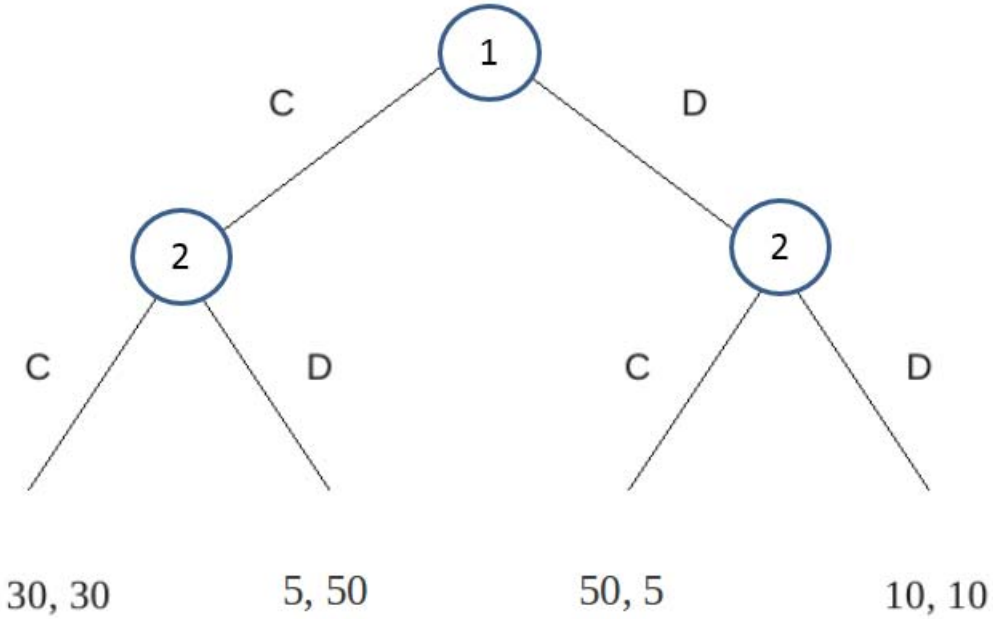
- Is there any empirical evidence for their existence?

- How does *Homo moralis* behave?

4 Preliminary experimental evidence

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

- Anonymous pairwise 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?

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, have just been carried out in Tilburg

5 Economic implications

[Alger and Weibull (2017a,b)]

- What if *Homo oeconomicus* is replaced by (the more general) *Homo moralis* in standard economic interactions, how do behavioral predictions then change?
  - in trust games
  - in environmental economics
  - in coordination games
  - in repeated games

6 Conclusion

1. Human motivation is richer than narrow self-interest, as represented by *Homo oeconomicus.*

2. The literature on behavioral and experimental economics proposes different "other-regarding" (social) preferences, such as altruism, inequity aversion, fairness, welfare concern, warm glow etc.

3. In this project, we explore implications from theoretical evolutionary principles, and find a combination of self-interest and (Kantian) morality, what we call *Homo moralis* (having *Homo oeconomicus* as a special case)

4. It appears, preliminarily, that this preference class may have good predictive power

5. One of the criticisms of economics is its reliance on the assumption of selfishness of economic agents. Arguably, economics would be viewed more favorably by non-economists if it instead was based on a more general class of motivations, allowing for some social concerns and/or morality. The degree of selfishness or morality would then be an empirical question.