Evolution leads to Kantian morality

Ingela Alger (Toulouse School of Economics and IAST)

Jörgen Weibull (Stockholm School of Economics and IAST)

Oslo, 6 October 6, 2015

Introduction

- For decades economics was based on *homo oeconomicus*, a species motivated by material self-interest

- Focus on:
  - extrinsic incentives to work, to pay taxes, etc
  - the role of institutions in mitigating information problems

- Then came behavioral economics

- Focus on preferences:
  - altruism (Becker)
  - warm glow (Andreoni)
  - fairness, inequity aversion (Rabin, Fehr and Schmidt)
  - reciprocal altruism (Levine)
  - honesty (Alger and Ma)
  - esteem (Bénabou and Tirole, Ellingsen and Johannesson)

- Moral values sometimes included by economists:
  - Smith (1759) and Edgeworth (1881)
  - Arrow (1973), Laffont (1975), Sen (1977)
  - Brekke, Kverndokk and Nyborg (2003), Tabellini (2008), Bénabou and Tirole (2011)

- But how can differences between countries be explained?

- Could the Nordic model have arisen in, say, South America?

- Since institutions are built by people, preferences should matter

- How are preferences formed in the first place?
  - study evolutionary foundations of human motivation!
  - "behavioral ecology" approach

- What preferences and/or moral values should we expect humans to have from first principles?

- Evolutionary logic:

1. Human populations have evolved under scarcity of resources

2. Not all who are born survive and not all who survive reproduce

3. Darwinian logic: those alive today had ancestors who were successful at surviving and reproducing; we have inherited their traits

- The approach in today's paper:
  - preferences $\Rightarrow$ behaviors $\Rightarrow$ material payoff consequences $\Rightarrow$ evolutionary selection pressure on preferences
  - place minimal restrictions on the set of possible preferences
  - which preferences are evolutionarily stable?

- This sounds too general to give anything...

- But ... the math leads to a new class of social preferences *cum* moral values: those of *homo moralis*

The model

- A continuum population
- Individuals are randomly matched into n-player groups
- Strategy set: X
- Each individual has a *preference type* $\theta \in \Theta$, which defines a continuous function $u_\theta : X^n \to \mathbb{R}$
- Each individual's type is his/her private information
- *Material payoff* from using strategy $x \in X$ against $x_{-i} \in X^{n-1}$: $\pi(x, x_{-i})$

- Each randomly matched group of n individuals play some (Bayesian Nash) equilibrium under incomplete information (as if individuals would know the type-distribution they meet, but not the types of the other individuals in their group)

Definition 1 *A type $\theta \in \Theta$ is* **evolutionarily stable against type** $\tau \in \Theta$ *if, for all sufficiently small $\varepsilon > 0$, individuals of type θ on average earn a higher material payoff than individuals of type τ in all equilibria under incomplete information.*

Definition 2 *A type θ is* **evolutionarily stable** *if it is evolutionarily stable against every type $\tau \neq \theta$.*

Definition 3 *A type θ is* **evolutionarily unstable** *if there exists a type τ such that, irrespective of how small $\varepsilon > 0$ is, there exists an equilibrium in which individuals of type τ earn a higher material payoff than individuals of type θ.*

Main result

For a given *population state* $s = (\theta, \tau, \varepsilon)$:

- Let $\Pr(\tau|\tau, \varepsilon)$ be probability that, for a given mutant, another group member, uniformly randomly sampled from the group, is also a mutant
- Let $\sigma = \lim_{\varepsilon \to 0} \Pr(\tau|\tau, \varepsilon)$ and call this the *index of assortativity*

Theorem 1 *Assume conditionally independent random matching, with index of assortativity* $\sigma \in [0, 1]$*.*

(i) Homo moralis *with degree of morality* $\kappa = \sigma$ *is evolutionarily stable against all types* τ *that are not its behavioral alikes.*

(ii) A type $\theta \in \Theta$ *is evolutionarily unstable if its carrier does not behave like* homo moralis *with degree of morality* $\kappa = \sigma$*.*

- So what is a homo moralis?

Homo moralis

- For any player i, any $\kappa \in [0, 1]$, and any strategy profile (x_i, x_{-i}), let $\tilde{x}_{-i}$ be a random vector with statistically independent components $\tilde{x}_j$ (for all $j \neq i$) such that

$$\Pr\left[\tilde{x}_j = x_i\right] = \kappa \text{ and } \Pr\left[\tilde{x}_j = x_j\right] = 1 - \kappa$$

Definition 4 *A* **homo moralis** *is an individual with utility function*

$$u_\kappa\left(x_i, x_{-i}\right) = \mathbb{E}_\kappa\left[\pi\left(x_i, \tilde{x}_{-i}\right)\right]$$

for some $\kappa \in [0, 1]$*, the individual's* **degree of morality**.

$n = 2$:

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

$n = 3$:

$$\begin{aligned} u_\kappa(x, y, z) &= (1 - \kappa)^2 \cdot \pi(x, y, z) \\ &\quad + \kappa(1 - \kappa) \cdot [\pi(x, x, z) + \pi(x, y, x)] \\ &\quad + \kappa^2 \cdot \pi(x, x, x) \end{aligned}$$

Remarks:

- *Homo moralis* with degree of morality $\kappa = 0$ is purely self-interested, while *homo moralis* with degree of morality $\kappa = 1$ always "does the right thing" according to *Kant's categorical imperative:* "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law" [*Grundlegung zür Metaphysik der Sitten, 1785*]

- For intermediate degrees of morality, $0 < \kappa < 1$, *homo moralis* chooses a strategy that would maximize her expected material payoff if others were to choose that same strategy with probability κ

Intuition for the theorem:

(i) *HM* with $\kappa = \sigma$ preempts mutants.

For instance, for $n = 2$: a resident population of *HM* with $\kappa = \sigma$ play some

$$x_\sigma \in \arg\max_{x \in X} (1-\sigma) \cdot \pi(x, x_\sigma) + \sigma \cdot \pi(x, x),$$

while a vanishingly rare mutant type, who plays some $z \in X$, obtains expected material payoff

$$(1-\sigma) \cdot \pi(z, x_\sigma) + \sigma \cdot \pi(z, z)$$

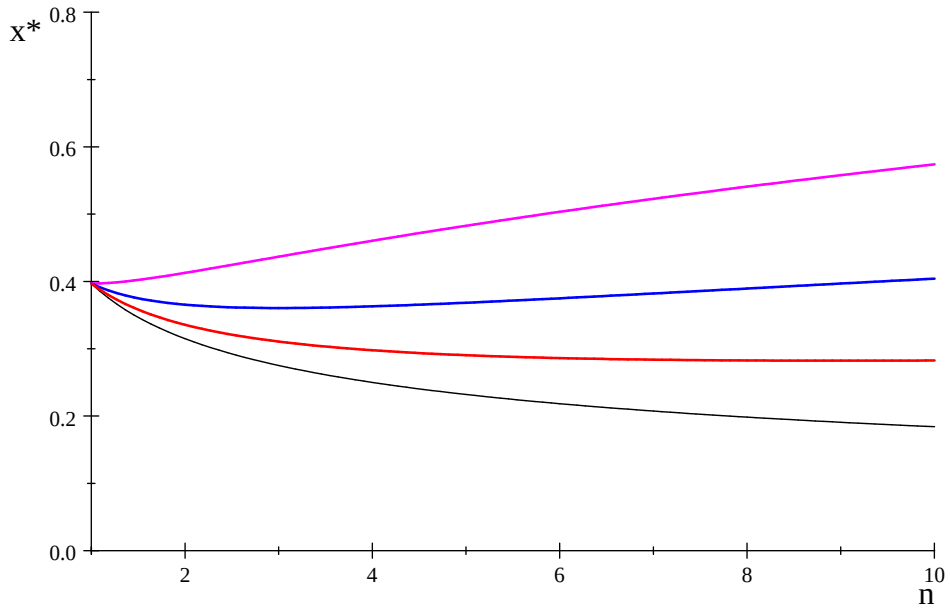
(ii) Any resident type θ that does not behave like *HM* with $\kappa = \sigma$ can be invaded by a mutant that is "committed" to a strategy that would maximize the mutant's material payoff in the limit as the mutant population share $\varepsilon \to 0$

How does homo moralis behave? An example

A public goods game

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = \left[\sum_{i=1}^{n} x_i\right]^{1/2} - x_i^2$$

$$x^* = n^{1/3} \cdot \left[\frac{1 + \sigma\left(n-1\right)}{4}\right]^{2/3}$$



$\sigma = 0.5$ (pink), $\sigma = 0.25$ (blue), $\sigma = 0.1$ (red), $\sigma = 0$ (black)

Discussion

- A new class of preferences, that springs out of stability under natural selection
- This preference class connects with moral philosophy
- Implications of *homo moralis* preferences for economics?
  - principal-agent relations
  - bargaining
  - participation and voting in elections
  - taxation
  - environmental economics

The Nordic model?

- Small degrees of morality can lead to high levels of contributions to public goods even in large groups

- Our theory predicts that high degrees of morality should be expected where assortativity is high

The Nordic model?

- The size of the welfare system
- Alger and Weibull (*AER* 2010) "Kinship, incentives and evolution"
- Main finding: natural selection leads to weaker family ties in harsher environments

- "The great achievement of ... the ethical and ascetic sects of Protestantism was to *shatter the fetters of the sib*. These religions established ... a common ethical way of life in opposition to the community of blood, even to a large extent in opposition to the family." (*Max Weber: The Religion of China*)

- Evolution by way of natural selection may explain Weber's observation about the "fetters of the sib" without recourse to Protestantism as a cause: perhaps "nature" selects family ties, and families select religions that fit their values

The Nordic model?

- Family structure
- Alger (WP 2015) "How many wives? On the evolution of preferences over polygyny rates"
- Main finding: natural selection leads to preference for monogamy in harsher environments
- Implications for incentives for males to fight to accumulate resources...