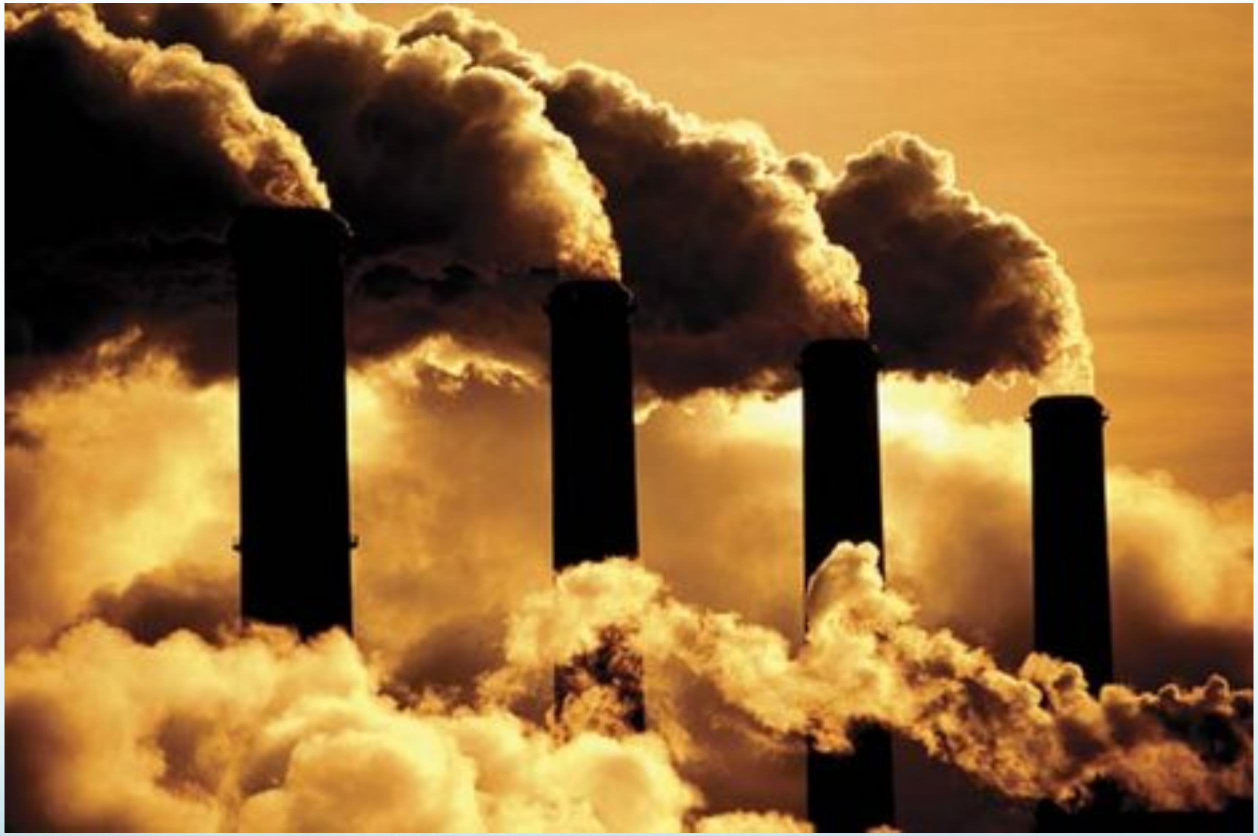
Evolution and Kantian morality

Ingela Alger (TSE, CNRS, IAST)

Jörgen Weibull (SSE, IAST)

NAG 2016

The environment affects us, directly...



The environment affects us, directly...

… the environment also affects us indirectly, through its impact on the evolution of traits

For example :

many Tibetans carry a rare gene variant that allows them to survive with little oxygen

Q: How may evolution have affected human motivation in social interactions?

Q: How may evolution have affected human motivation in social interactions?

In joint work with Jörgen Weibull we study this question by combining economics and biology models

Why should economists care about this question?

- Traditional economics models are inhabited by the selfish *Homo oeconomicus*:
  - an opportunistic creature without morality
  - most of us do not recognize ourselves in this creature

Why should economists care about this question?

- Traditional economics models are inhabited by the selfish *Homo oeconomicus*:
  - an opportunistic creature without morality
  - most of us do not recognize ourselves in this creature

- A tradition that dates back to Adam Smith (1776):

> "It is not from the benevolence of the butcher,
> the brewer, or the baker, that we expect our dinner,
> but from their regard to their own self interest."

Why should economists care about this question?

- But this view became problematic when applied to interactions between small numbers of individuals

Why should economists care about this question?

- But this view became problematic when applied to interactions between small numbers of individuals
- To explain observed behaviors: several alternative preferences have been proposed and sometimes tested

Altruism (G. Becker)
Warm glow (J. Andreoni)
Fairness/inequity aversion (M. Rabin, E. Fehr and K. Schmidt)
Conditional altruism (D. Levine)
Conformity (D. Bernheim)
Desire to avoid social stigma (A. Lindbeck, S. Nyberg, and J. Weibull)
Identity (G. Akerlof and R. Kranton)
Concern for efficiency (G. Charness and M. Rabin)
Image concerns (R. Bénabou and J. Tirole, T. Ellingsen and M. Johannesson)
Honesty (I. Alger and R. Renault)

Which preferences should we expect evolution to favor ?

Homo oeconomicus ?

If not, then what ?

And why ?

- Evolution is dynamic
- Static, stability concepts are used to gain insights

- Evolution is dynamic
- Static, stability concepts are used to gain insights
- Studies on the evolutionary stability of preferences in social interactions:

  Frank (1987)
  Güth and Yaari (1992)
  Bester and Güth (1998)
  Ok and Vega-Redondo (2001)
  Sethi and Somanathan (2001)
  Heifetz, Shannon and Spiegel (2007)
  Dekel, Ely and Yilankaya (2007)
  Alger and Weibull (2010, 2013, 2016)

- For preference evolution in decision problems: see Robson

Roadmap

A. The model

B. Results

C. Implications

D. Literature

A. The main features of our model:

- An infinite population
- Individuals interact in groups of finite size
- The behavior of an individual depends on his/her preferences
- Interactions have material consequences for each individual (reproductive success)
- We identify preferences that withstand the invasion by other traits
- Minimal restrictions on the set of possible utility functions
- Note: we do not model competition between groups (there is no group selection)

B. Our model predicts that evolution strongly favors a specific class of preferences, which may be interpreted as involving a form of Kantian morality

Immanuel Kant (1724-1804)

C. These preferences sometimes give rise to drastically different behaviors than oft-studied preferences

Immanuel Kant (1724-1804)

A. The model

The interaction

- A population that evolves in a stationary environment
- In this population individuals are matched to interact
- Examples: public goods, team work, consumption or production with environmental effects, competition and/or coordination, voting, etc.
- Preferences determine behaviors
- Behaviors determine material payoffs
- Material payoff determines reproductive success

The interaction

- An infinite population

- Individuals are randomly matched into n-player groups

- Common strategy set X

- Material payoff $\pi\left(x, \boldsymbol{y}\right)$, where $x \in X$ and $\boldsymbol{y} \in X^{n-1}$

- π is continuous

The interaction

- Each individual has a *type* θ, which defines a continuous function $u_\theta : X \times X^{n-1} \longrightarrow \mathbb{R}$

- Each individual's type is his/her *private information*

The matching process

- To define evolutionary stability of a type, consider populations with only two types present

- Consider a population state in which the share of *residents* or type θ is $1-\varepsilon$ and the share of *mutants* of type τ is ε

- For any *mutant*, let $q_m\left(\varepsilon\right)$ be the probability that there are m other mutants in his group

- Let the limit of $\boldsymbol{q}\left(\varepsilon\right)=\left(q_0\left(\varepsilon\right),...,q_{n-1}\left(\varepsilon\right)\right)$ as $\varepsilon\longrightarrow 0$ be $\boldsymbol{q}^*$

- Call $\boldsymbol{q}^*$ the *assortativity profile* of the matching process

Equilibrium strategies

- 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]

Evolutionary stability

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

Definition 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 strictly higher material payoff than individuals of type θ.

The set Θ of potential utility functions, is the set of *all* continuous functions from $X \times X^{n-1}$ to $\mathbb{R}$

In particular, Θ contains *Homo oeconomicus*: $u_E = \pi$

B. Result

Definition An individual is a **Homo moralis** with **morality profile $\boldsymbol{\mu}$** if his or her goal function u_μ satisfies

$$u_\mu\left(x, \boldsymbol{y}\right) = \mathbb{E}\left[\pi\left(x, \boldsymbol{Y}\right)\right] \quad \forall\left(x, \boldsymbol{y}\right) \in X^n.$$

where $\boldsymbol{Y}$ is a random vector such that with probability μ_m exactly m of the components of $\boldsymbol{y}$ are replaced by x, with equal probability for each subset of size m, while the remaining components of $\boldsymbol{y}$ keep their original values.

Two extremes:

$\mu_0 = 1$ gives *Homo oeconomicus*, $u_E = \pi$

$\mu_{n-1} = 1$ gives *Homo kantientis,* $u_K\left(x, \boldsymbol{y}\right) \equiv \pi\left(x, \left(x, x, ..., x\right)\right)$

Theorem *Homo moralis* with morality profile $\boldsymbol{\mu} = \boldsymbol{q}^*$is evolutionarily stable against all types τ that are behaviorally distinguishable from it. Any type $\theta \in \Theta$ that is behaviorally distinguishable from *Homo moralis* with morality profile $\boldsymbol{\mu} = \boldsymbol{q}^*$ is evolutionarily unstable.

- Intuition: *Homo moralis* with morality profile $\boldsymbol{\mu} = \boldsymbol{q}^*$ *preempts mutants*

- To see this, suppose that $n = 2$
  - Resident *Homo moralis* play some x^* satisfying

$$x^* \in \arg\max_{x \in X} q_0^* \cdot \pi\left(x, x^*\right) + q_1^* \cdot \pi\left(x, x\right)$$

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

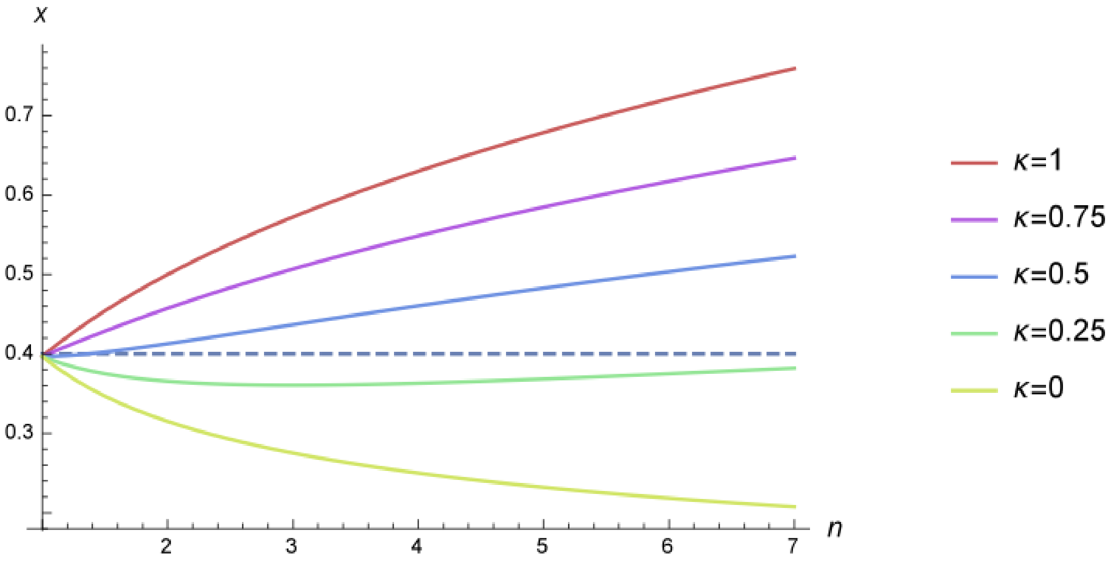
$$q_0^* \cdot \pi\left(z, x^*\right) + q_1^* \cdot \pi\left(z, z\right)$$

C. Implications

Example 1: Public goods

$$\pi\left(x_i, \boldsymbol{y}\right) = \left(x_i + \sum\nolimits_{j \neq i} y_j\right)^{1/2} - x_i^2$$

Assume that the population evolved under conditional independence in the matching process: $\boldsymbol{q}^* = Bin\left(\kappa, n-1\right)$


x
0.7
0.6
0.5
0.4
0.3
2
3
4
5
6
7
n
κ=1
κ=0.75
κ=0.5
κ=0.25
κ=0

Can these predictions help explain observations made in laboratory experiments, in which group size sometimes has a positive effect and sometimes a negative effect on individual contributions (see Nosenzo, Quercia, and Sefton, 2015, for a review)?

Example 2: Consumption of a polluting good

A continuum of identical consumers, indexed $i \in I = [0, 1]$

Two consumption goods: 1 and 2, where good 1 is environmentally neutral and good 2 is environmentally harmful

Aggregate consumptions:

$$X_1 = \int_I x_1(i)\, df \quad \text{and} \quad X_2 = \int_I x_2(i)\, df,$$

All consumers are infinitesimally small: aggregate demand is not affected by any individual's personal consumption

Material payoff to individual i: $v(x_1(i), x_2(i), X_2)$

Socially efficient consumption bundle satisfies:

$$\frac{v_2\left(x_1^*, x_2^*, X_2^*\right)}{v_1\left(x_1^*, x_2^*, X_2^*\right)} \quad = \quad p - \frac{v_3\left(x_1^*, x_2^*, X_2^*\right)}{v_1\left(x_1^*, x_2^*, X_2^*\right)},$$

In a population of *Homo oeconomicus* an (interior) equilibrium allocation satisfies:

$$\frac{v_2\left(x_1^0, x_2^0, X_2^0\right)}{v_1\left(x_1^0, x_2^0, X_2^0\right)} = p$$

A *Homo moralis* with degree of morality κ has utility:

$$u_\kappa(x,y) = v\left(x_1, x_2, (1-\kappa)\, y_2 + \kappa x_2\right)$$

In a population consisting entirely of *Homo moralis* with the same degree of morality κ, an interior equilibrium allocation satisfies:

$$\frac{v_2\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}{v_1\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)} \quad = \quad p - \kappa \cdot \frac{v_3\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}{v_1\left(x_1^\kappa, x_2^\kappa, x_2^\kappa\right)}.$$

For any positive degree of morality κ each individual refrains somewhat from consuming the environmentally harmful good, compared to *Homo oeconomicus*, although each individual—knowing that she is negligible—is fully aware that her own consumption has *no* effect on the overall quality of the environment!

D. Literature

- Preference evolution under incomplete information: Ok & Vega-Redondo (2001) and Dekel, Ely & Yilankaya (2007) assume uniform random matching

- Then: *Homo oeconomicus* prevails

- But assortativity is positive as soon as there is a positive probability that both parties in an interaction have inherited their preferences or moral values from a common "ancestor" (genetic or cultural)

- Homophily, education, geography, networks...

- Assortativity is common and uniform random matching rare

- A large literature in biology: kinship, population genetics, structured populations

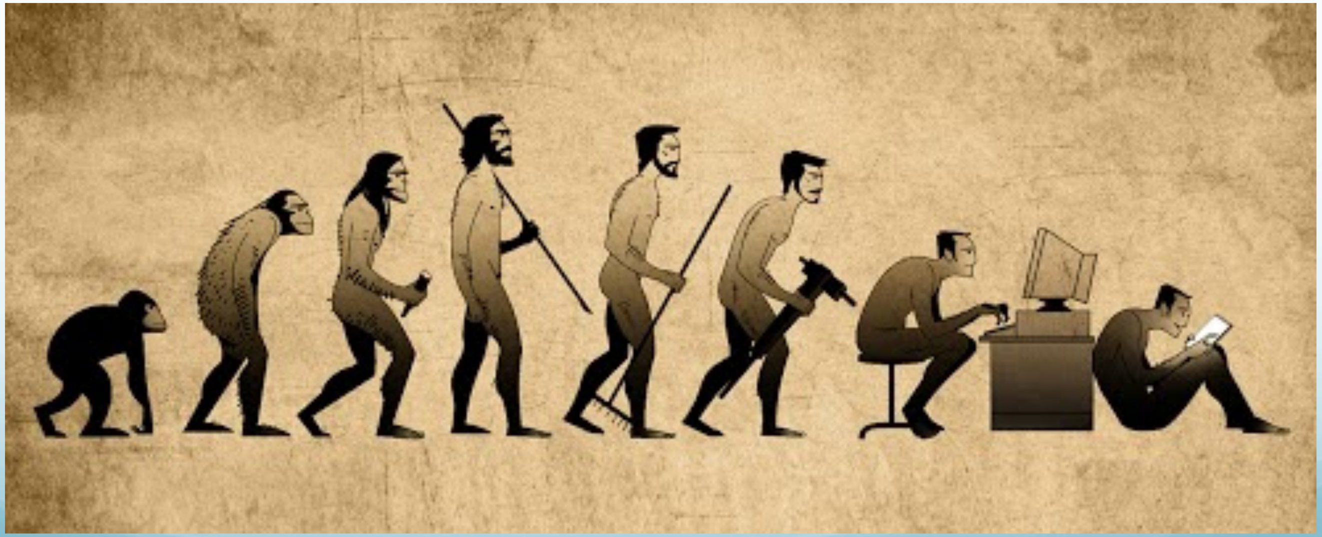
- Wright (1921, 1922, 1931), Hamilton (1964), Hines & Maynard Smith (1979), Grafen (1979, 2006), Bergstrom (1995, 2003, 2009, 2012), Rousset (2004), Lehmann & Rousset (2011)...

Conclusion

- Evolutionary logic can help us understand the ultimate causes behind human behavior
- When minimal restrictions are imposed, a particular preference class emerges from evolutionary stability:
  - HM preferences connect with moral philosophy
  - HM preferences are new to economics – it sprung out from the math, we did not invent them
  - HM preferences are distinct from social preferences
- Recall: no group selection

Conclusion

- Evolutionary logic may also help us predict how changes in the economic environment may change our preferences in the long run!



Merci !

cnrs
dépasser les frontières

Toulouse
School
of Economics

IAST
INSTITUTE for
ADVANCED
STUDY in
TOULOUSE

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR