Evolutionary Foundations of Morality and Altruism

—

Recent Advances

Ingela Alger (CNRS, TSE, IAST)

Atkinson Memorial Lecture

Global Priorities Institute, Oxford, June 13, 2022

World Bank, Washington DC, July 13, 2016

Introduction

- Economists formulate policy recommendations.
- Goal: maximize human welfare, given available resources.

Introduction

First theorem of welfare economics:

If (a) markets are complete, (b) there is perfect competition, and (c) information is complete, then any market equilibrium is Pareto-efficient.

Economic policies seek to mitigate inefficiencies stemming from incomplete markets, market power, and/or information asymmetries and imperfections.

Introduction

"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." *[A. Smith, The Wealth of Nations, 1776]*

"How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it."
[A. Smith, The Theory of Moral Sentiments, 1759]

"[B]etween the frozen pole of egoism and the tropical expanse of utilitarianism [there is] (...) the position of one for whom in a calm moment his neighbour's utility compared with his own neither counts for nothing, nor 'counts for one', but counts for a fraction. "
[F.Y. Edgeworth, Mathematical Psychics, 1881]

Introduction

The complexity of human motivations was pushed to the background.

$$\max_{(\boldsymbol{c},\ell)\in F(\Omega)} \quad u\left(\boldsymbol{c},\ell\right)$$

First theorem of welfare economics:

If (a) markets are complete, (b) there is perfect competition, (c) information is complete, and (d) individuals care only about consumption and leisure (i.e., are Homo oeconomicus), then any market equilibrium is Pareto-efficient.

Introduction

The complexity of human motivations was pushed to the background.

$$\max_{(\boldsymbol{c},\ell)\in F(\Omega)} \quad u\left(\boldsymbol{c},\ell\right)$$

First theorem of welfare economics:

If (a) markets are complete, (b) there is perfect competition, (c) information is complete, and (d) individuals care only about consumption and leisure (i.e., are Homo oeconomicus), then any market equilibrium is Pareto-efficient.

Policy recommendations based on material incentives.

Introduction

The complexity of human motivations was pushed to the background.

$$\max_{(\boldsymbol{c},\ell)\in F(\Omega)} \quad u\left(\boldsymbol{c},\ell\right)$$

First theorem of welfare economics:

If (a) markets are complete, (b) there is perfect competition, (c) information is complete, and (d) individuals care only about consumption and leisure (i.e., are Homo oeconomicus), then any market equilibrium is Pareto-efficient.

Policy recommendations based on material incentives.

An accurate account of human behaviour is necessary to identifiy policies that are *desirable* and *effective*.

Introduction

Is *Homo sapiens* really a *Homo oeconomicus*?

- 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 concerns (G. Akerlof and R. Kranton)
- Efficiency concerns (G. Charness and M. Rabin)
- Image concerns (R. Bénabou and J. Tirole, T. Ellingsen and M. Johannesson)
- Honesty concerns (U. Gneezy, I. Alger and R. Renault)

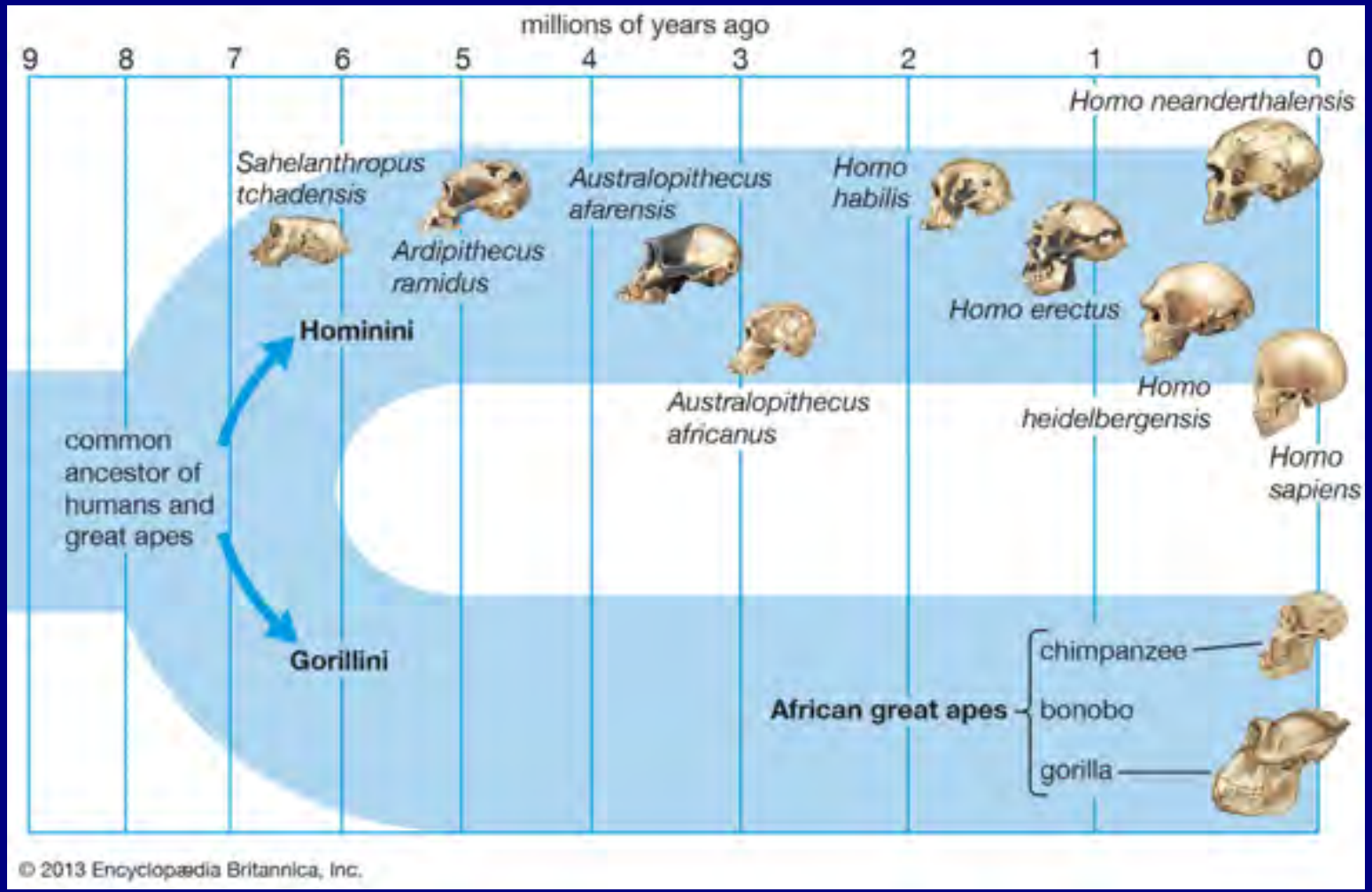
Introduction

Is *Homo sapiens* really a *Homo oeconomicus*?

- There is no consensus...
- One step beyond: which preferences should we expect, from first principles?
- Ideally, such a theory would shed light on:
  - which preferences are more plausible than others
  - why

Introduction

Is *Homo sapiens* really a *Homo oeconomicus*?



Introduction

Evolutionary logic

- Evolution: competition for survival and reproduction
- Not all who are born survive and not all who survive reproduce
- Darwinian logic:
  - those alive today have ancestors who were successful at surviving and reproducing
  - our preferences should reflect this

Introduction

Evolutionary logic

- Evolution: competition for survival and reproduction
- Not all who are born survive and not all who survive reproduce
- Darwinian logic:
  - those alive today have ancestors who were successful at surviving and reproducing
  - our preferences should reflect this
- Use this to develop a theory for the evolutionary foundations of preferences

Introduction

Evolutionary logic

According to evolutionary logic, *reproductive success* is the name of the game.

Shouldn't humans simply be expected to be equipped with traits that make them maximize own reproductive success then?

Main theoretical challenge: answer this question, and understand mechanisms.

Roadmap

- framework
- two insights + implications
- a third insight
- concluding remarks

Framework

Process of *mutation* and *selection* in a population:

1. a sequence of generations
2. in each generation there is a certain distribution of preferences
3. sometimes a novel (mutant) preference type appears
4. individuals are somehow matched together to interact
5. preferences guide behavior
6. behavior results in material payoffs
7. material payoffs determine reproductive success

- NB: transmission can be biological or cultural

Framework

- Goal 1: determine which preferences this process leads to [Frank (1988), Güth and Yaari (1992)]
- Goal 2: understand how features of the *environment in which a population evolves* affects the evolutionary viability of preferences
- Many modeling choices:
  - how are individuals matched?
  - informational assumptions?
  - set of potential preferences?

Insight #1

Evolution by natural selection may favor

weaker intra-family altruism in harsher environments

Interactions within the family

Evolution of altruistic preferences under complete information

- Interactions between members of the same family
- Interactions under complete information
- Each individual has some *degree of altruism* $\alpha \in (-1, 1)$ towards the other:

$$u_\alpha(x, y) = w(x, y) + \alpha \cdot w(y, x)$$

- Alger and Weibull (2010, 2012)
- See also Heifetz, Shannon, and Spiegel (2007)

Interactions within the family

Application: production and sharing within the family

- Interaction:

1. A pair of siblings simultaneously choose productive efforts
2. Each sibling's random output is realized, $Y_i \in \left\{ Y^L, Y^H \right\}$. It depends probabilistically on own effort.
3. The siblings observe the outputs, and make transfers to each other.

Interactions within the family

Application: production and sharing within the family

- $Y^L = \lambda Y^H$, where $\lambda < 1$ measures *output variability*
- $p(x) = 1 - e^{-\theta x}$, where $\theta > 0$ is a *return to effort* parameter

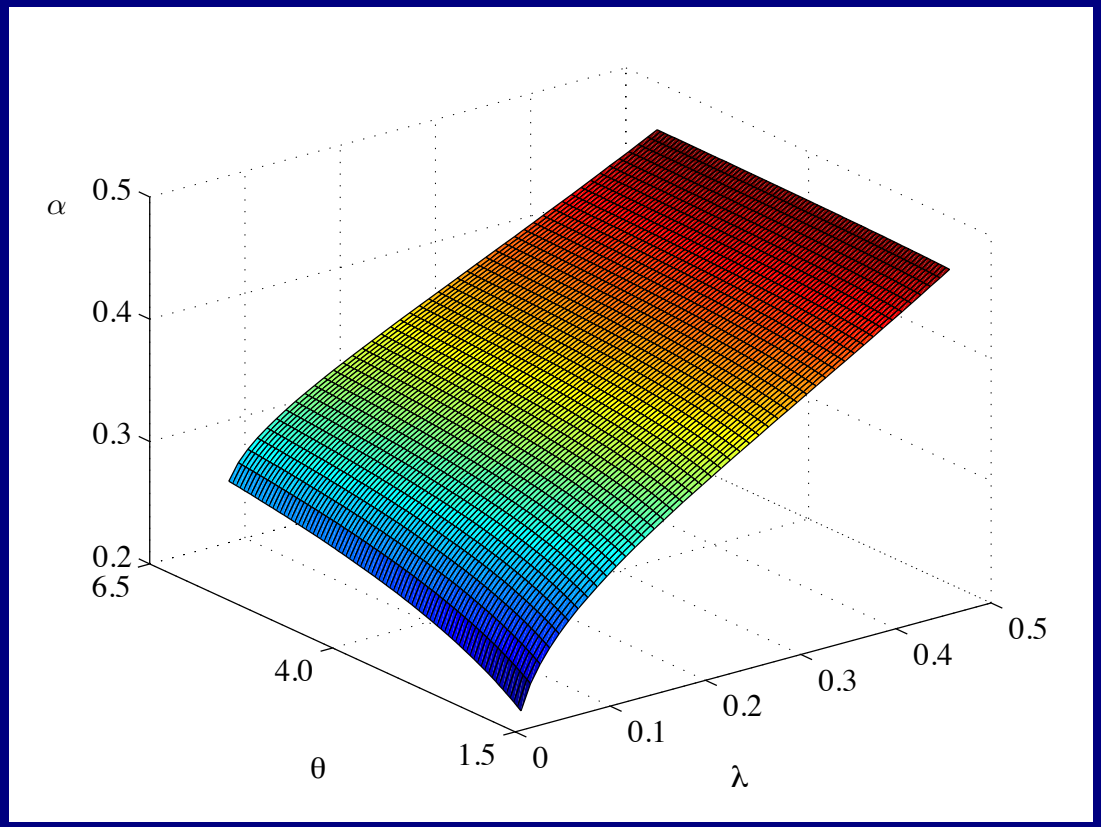
Interactions within the family

Application: production and sharing within the family

- $Y^L = \lambda Y^H$, where $\lambda < 1$ measures *output variability*
- $p(x) = 1 - e^{-\theta x}$, where $\theta > 0$ is a *return to effort* parameter
- (λ, θ) is the environment
- An environment (λ', θ') is *harsher* than another environment (λ, θ) if the output variability is more pronounced ($\lambda' \leq \lambda$), and the marginal return to effort is smaller ($\theta' \leq \theta$)

Interactions within the family

Application: production and sharing within the family



Interactions within the family

Application: production and sharing within the family

- In sum:
  - For pre-industrial times in agricultural societies: our model predicts weaker altruism in harsher climates
  - In line with rise of individualism in northwestern Europe
  - Max Weber (*Religion of China*, 1915):

    > *the great achievement of [...] the ethical and ascetic sects of Protestantism was to shatter the fetters of the sib [the extended family]. 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.*

Interactions within the family

Application: production and sharing within the family

- More generally:
  - Our model predicts that the strength of intra-family altruism depends on the environment
  - Is the model of individual utility maximization even relevant when intra-family altruism is strong enough?
  - Implications for economic development?

Insight #2

Evolution by natural selection favors

a concern for universalisation

Interactions beyond the family

Homo moralis

- Interactions between "strangers"
- For interactions beyond the family: observability pf preferences is questionable
- Interactions under incomplete information
- Each individual has some continuous utility function that describes his/her preferences over the strategies played by self and other

- Alger and Weibull (2013, 2016)
- See also Ok and Vega-Redondo (2001) and Dekel, Ely, and Yilankaya (2007)

Interactions beyond the family

Homo moralis

Definition

An individual is a *Homo moralis* with degree of morality $\kappa \in [0, 1]$ if her utility function is of the form

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

- $w(x, y)$: own reproductive success, given that self plays x and other plays y
- $w(x, x)$: own reproductive success if—hypothetically—own strategy x was universalised

Interactions beyond the family

Homo moralis

Definition

An individual is a *Homo moralis* with degree of morality $\kappa \in [0,1]$ if her utility function is of the form

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

- Kant (*Grundlegung zür Metaphysik der Sitten*, 1785):
  "Act only according to that maxim whereby you can [...] will that it should become a universal law."
- *Homo moralis* can be said to:
  "act according to that maxim whereby (s)he can will that others should do likewise with probability κ."

Interactions beyond the family

Homo moralis

Theorem

(a) Homo moralis with degree of morality $\kappa = r$ is evolutionarily stable against all behaviorally distinguishable types.

(b) Any type which is behaviorally distinguishable from Homo moralis of degree of morality $\kappa = r$ is evolutionarily unstable.

Interactions beyond the family

Homo moralis

Theorem

(a) Homo moralis with degree of morality $\kappa = r$ is evolutionarily stable against all behaviorally distinguishable types.

(b) Any type which is behaviorally distinguishable from Homo moralis of degree of morality $\kappa = r$ is evolutionarily unstable.

- r is the *coefficient of relatedness* [Wright (1931)]
- the probability that interactants have a common ancestor in a not too distant past

Interactions beyond the family

Homo moralis

Theorem

(a) Homo moralis with degree of morality $\kappa = r$ is evolutionarily stable against all behaviorally distinguishable types.

(b) Any type which is behaviorally distinguishable from Homo moralis of degree of morality $\kappa = r$ is evolutionarily unstable.

- Intuition: *HM* preempts mutants
- A resident population of *HM* play some x_r such that

$$x_r \in \arg\max_{x \in X} (1 - r) \cdot w(x, x_r) + r \cdot w(x, x)$$

- A vanishingly rare mutant type, who plays some $z \in X$, obtains average reproductive success

$$(1 - r) \cdot w(z, x_r) + r \cdot w(z, z)$$

Interactions beyond the family

Homo moralis

- How does a positive coefficient of relatedness r arise?
- The tendency for individuals sharing a common ancestor to interact arises in populations structured into groups, with limited migration between them.
- Our ancestors (last two million years) lived in small groups (5-150 grown-ups), extending beyond the nuclear family [Grueter, Chapais, and Zinner (2012), Malone, Fuentes, and White (2012), van Schaik (2016), Layton et al. (2012)]
- Part of the environment of evolutionary adaptedness (EEA) of the human lineage [van Schaik (2016)].

Interactions beyond the family

Homo moralis

	C	D
C	$10, 10$	$-2, 15$
D	$15, -2$	$2, 2$

Homo oeconomicus: (D, D) is the unique equilibrium

Altruists: (C, C) is the unique equilibrium for α high enough

Homo moralis: (C, C) is the unique equilibrium for κ high enough

Interactions beyond the family

Homo moralis

	G	B
G	$10, 10$	$0, 0$
B	$0, 0$	$2, 2$

Homo oeconomicus: (G, G) and (B, B) are equilibria

Altruists: (G, G) and (B, B) are equilibria

Homo moralis: (G, G) is the unique equilibrium for κ high enough

Interactions beyond the family

Homo moralis

In public goods settings where each individual's real impact on the aggregate outcome is negligible (climate change, pollution, voting):

Homo oeconomicus: do not contribute/vote

Altruists: do not contribute/vote

Homo moralis: do contribute/vote for κ high enough

Interactions beyond the family

Homo moralis

- The *Homo moralis* preference class is novel to economics.
- Although: see Laffont (1975) and Bergstrom (1995)
- Incentive contracts in firms [Sarkisian (2017)]
- Climate policies [Eichner and Pethig (2020)]
- Evolution of fiat money [Norman (2020)]
- Voluntary contributions to public goods and optimal taxation [Muñoz (WiP)]
- Voting [Alger and Laslier (2022 + WiP)]

Interactions beyond the family: a closer look

Other-regarding Kantians

Insight #3

Evolution by natural selection favors

Kantian concerns at the reproductive success level...

but Kantian concerns mixed with spite or altruism

at the material payoff level

Interactions beyond the family: a closer look

Other-regarding Kantians

- Lehmann, Alger, and Weibull (2015), Alger, Weibull, and Lehmann (2020)

Interactions beyond the family: a closer look

Other-regarding Kantians

Theorem

Uninvadability requires residents to play some strategy satisfying:

$$x^* \in \arg\max_{x \in X} \left[1 - r\left(x_i, x^*\right)\right] \cdot \tilde{w}\left(x_i, x_j, x^*\right) + r\left(x_i, x^*\right) \cdot \tilde{w}\left(x_i, x_i, x^*\right).$$

- Kantian concern at the reproductive success level

Interactions beyond the family: a closer look

Other-regarding Kantians

- Weak selection (material payoffs affect RS marginally)

Theorem

Under weak selection, v is uninvadable:

$$\begin{aligned} v\left(x_i, x_j\right) &= (1-r)\cdot\left[\pi\left(x_i, x_j\right) - \lambda\cdot\pi\left(x_j, x_i\right)\right] \\ &\quad + r\cdot\left[\pi\left(x_i, x_i\right) - \lambda\cdot\pi\left(x_i, x_i\right)\right] \end{aligned}$$

where λ is the coefficient of reproductive success interdependence:

$$\lambda = \left(-\frac{\partial w\left(\bar{\pi}_i, \bar{\pi}_j, \bar{\pi}^*\right)}{\partial \bar{\pi}_j}\right) / \left(\frac{\partial w\left(\bar{\pi}_i, \bar{\pi}_j, \bar{\pi}^*\right)}{\partial \bar{\pi}_i}\right).$$

- A mix of self-interest, a Kantian concern at the material payoff level, and a comparison with other's material payoff: *other-regarding Kantians*

Concluding remarks

- Theory helps us understand how evolutionary forces may have shaped the preferences of *Homo sapiens*:
  - impact of environment on preferences?
  - discovery of novel preference classes
- To formulate desirable and effective policy recommendations:
  - necessary to assess theoretical implications
  - necessary to assess empirical relevance
    [some experimental evidence: Capraro and Rand (2018), Levine et al. (2020), van Leeuwen and Alger (2021)]

Recent surveys

- Preference evolution in strategic interactions: Alger and Weibull (*Annual Review of Economics* 2019)
- Evolutionary game theory: Newton (2017)
- For preference evolution in decision problems: see the work of Arthur Robson

Thanks

Laurent Lehmann

Jörgen W. Weibull

Thanks

Merci !

Toulouse School of Economics

IAST
INSTITUTE for ADVANCED STUDY in TOULOUSE