Evolutionary Foundations of Morality and Other-regard
— Recent Advances —

Ingela Alger

Toulouse School of Economics, CNRS, IAST

ASSA 2024

San Antonio, 5th January 2024

Introduction

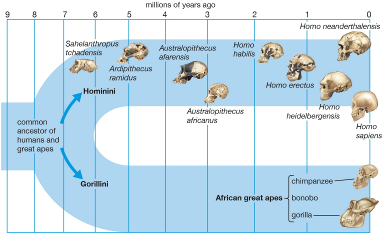
$$\max_{\boldsymbol{x} \in F(\Omega)} \quad u\left(\boldsymbol{x}, \cdot\right)$$

Introduction

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

- 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



- Evolution: competition for survival and reproduction
- Evolutionary logic: those alive today have ancestors who were successful at surviving and reproducing
  - our preferences should reflect this!
- Theory of preference evolution [Frank 1987, Güth and Yaari 1992]

Introduction

Evolution is a 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

Introduction

- Evolutionary logic $\rightarrow$ *reproductive success* is the name of the game:
  - Q1: do we simply maximize own reproductive success ?

Introduction

- Evolutionary logic $\rightarrow$ *reproductive success* is the name of the game:
  - Q1: do we simply maximize own reproductive success ?
    - No! (except in special cases)

Introduction

- Evolutionary logic $\rightarrow$ *reproductive success* is the name of the game:
  - Q1: do we simply maximize own reproductive success ?
    - No! (except in special cases)
- Social scientists rarely work with data on reproductive success:
  - Q2: predictions for preferences over material payoffs ?

Introduction

- Evolutionary logic → *reproductive success* is the name of the game:
  - Q1: do we simply maximize own reproductive success ?
    - No! (except in special cases)
- Social scientists rarely work with data on reproductive success:
  - Q2: predictions for preferences over material payoffs ?
    - Yes!

Introduction

Group structure is a key factor in our evolutionary past:

- our ancestors (last 2 MY) lived in small groups (5-150 grown-ups)
- limited migration between the groups
- part of the environment of evolutionary adaptedness of the human lineage [van Schaik (2016)]

Roadmap

- Theoretical predictions
  - Model
    - A – Group structure not explicitly modeled
    - B – Group structure explicitly modeled
- Experimental evidence

Model

- A large (continuum) population
- Individuals are randomly matched into pairs
- Each pair has a symmetric interaction, with strategy set X
- Each individual has a *preference type* $\theta \in \Theta$, which defines a utility function u_θ: $X^2 \to \mathbb{R}$
- $w(x, y)$: *reproductive success* from playing x against y

Model

- Consider a population with some *resident preference type* $\theta \in \Theta$
- Inject some individuals with some *mutant preference type* $\tau \in \Theta$
- Posit an information structure and evaluate reproductive success at Nash equilibrium strategy profile(s)
- θ withstands the invasion of τ if the average reproductive success of residents exceeds that of mutants, when the mutants are rare
- θ is then *evolutionarily stable*

Model

- Consider a population with some *resident preference type* $\theta \in \Theta$
- Inject some individuals with some *mutant preference type* $\tau \in \Theta$
- Posit an information structure and evaluate reproductive success at Nash equilibrium strategy profile(s)
- θ withstands the invasion of τ if the average reproductive success of residents exceeds that of mutants, when the mutants are rare
- θ is then *evolutionarily stable*
- Today:
  - Θ: the set of all continuous functions $u : X^2 \to \mathbb{R}$
  - incomplete information

Model A: group structure not modeled explictly

Model A: group structure not modeled explictly

- Abstract modeling of group structure:
  - $\Pr[\tau|\tau, \varepsilon]$ may be greater than ε, the share of mutants

Model A: group structure not modeled explictly

- Abstract modeling of group structure:
  - $\Pr[\tau|\tau,\varepsilon]$ may be greater than ε, the share of mutants
  - Write r for $\lim_{\varepsilon\to 0}\Pr[\tau|\tau,\varepsilon]$
    - Uniform random matching $\Rightarrow r = 0$
    - Interactions between full siblings $\Rightarrow r = 1/2$
  - r is the *coefficient of relatedness* [Wright, 1931]

Model A: group structure not modeled explictly

- Abstract modeling of group structure:
  - $\Pr[\tau|\tau,\varepsilon]$ may be greater than ε, the share of mutants
  - Write r for $\lim_{\varepsilon\to 0}\Pr[\tau|\tau,\varepsilon]$
    - Uniform random matching $\Rightarrow r = 0$
    - Interactions between full siblings $\Rightarrow r = 1/2$
  - r is the *coefficient of relatedness* [Wright, 1931]
- Alger and Weibull [*Econometrica* 2013, *Games and Economic Behavior* 2016, 2023]

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 own strategy x and opponent's strategy y
- $w(x,x)$: own reproductive success if—hypothetically—own strategy x was universalised

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 have semi-Kantian concerns

Model A: group structure not modeled explictly

Result

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.*

Model A: group structure not modeled explictly

Result

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* with $\kappa = r$ 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

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

Model B: group structure modeled explictly

Model B: group structure modeled explictly

- Group structure modeled explicitly:
  - a long-standing tradition in biology [the island model, Wright 1931]
  - collaborations with evolutionary biologist Laurent Lehmann

Model B: group structure modeled explictly

- Group structure modeled explicitly:
  - a long-standing tradition in biology [the island model, Wright 1931]
  - collaborations with evolutionary biologist Laurent Lehmann
- Alger, Weibull, and Lehmann [*Journal of Economic Theory* 2020]

Model B: group structure modeled explictly

- An infinite number of *groups* of size n
- Evolution takes place perpetually over discrete time
- Each *demographic time period* consists of two phases:

1. *Phase 1*: the n adults in each island interact (X, π)
2. *Phase 2*: realized material payoffs $\rightarrow$ each adult's survival and fecundity; following reproduction, offspring may migrate from their native island to other islands (probability $m > 0$); following migration, individuals compete for available spots

- This determines each adult i's *reproductive success* $\tilde{w}\left(\pi_i, \boldsymbol{\pi}_{-i}, \bar{\pi}^*\right)$: the expected number of i's *immediate descendants* who have secured a "breeding spot" in the next demographic time period

Model B: group structure modeled explictly

Result 1

Theorem

Evolutionary stability requires residents to play some strategy satisfying:

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

where $r(x_i, x^*)$ is the probability for a randomly drawn mutant playing x_i that his neighbor is also a mutant, when residents play x^*.

- For preferences expressed in terms of reproductive success: a semi-Kantian concern [as in Alger and Weibull 2013, 2016]
- But now relatedness depends on group structure

Model B: group structure modeled explictly

Result 2

- Now let material payoffs affect reproductive success marginally and

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

Theorem

Under weak selection, v is evolutionarily stable:

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

- For preferences expressed in terms of material payoffs: a semi-Kantian concern combined with other-regard (spite if $\lambda > 0$, altruism if $\lambda < 0$

Model B: group structure modeled explictly

Three canonical scenarios: Genes

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \bar{\pi}^*\right) \;=\; s\left(\pi_i\right) \;+\; m \cdot \left[1 - s\left(\bar{\pi}^*\right)\right] n \cdot \frac{f\left(\pi_i\right)}{nf\left(\bar{\pi}^*\right)}$$

$$+\;\left(1-m\right) \cdot \left(n - \sum_{j=1}^{n} s\left(\pi_j\right)\right) \cdot \frac{f\left(\pi_i\right)}{\left(1-m\right)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\bar{\pi}^*\right)}$$

$s\left(\pi_i\right)$: probability that i survives to the next demographic time period

$f\left(\pi_i\right) > 0$: i's expected number of offspring

Model B: group structure modeled explictly

Three canonical scenarios: Genes

Suppose that $s(\pi_i) = s_0$ and $f(\pi_i) = f_0 \cdot \exp(\delta \cdot \pi_i)$. Then:

$$r = \frac{(1-m)^2 + (1+m^2)\, s_0}{n - (n-1)(1-m)^2 + (1-(n-1)m^2)\, s_0}$$

$$\lambda = \frac{(n-1)(1-m)^2}{n-(1-m)^2}$$

In this scenario, $\lambda > 0$: the model predicts a combination of material self-interest, a semi-Kantian concern, and spite.

Model B: group structure modeled explictly

Three canonical scenarios: Guns

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \bar{\pi}^*\right) = \left[(1-\rho) + 2\rho v\left(\boldsymbol{\pi}, \bar{\pi}^*\right)\right] \cdot \Bigg[m \cdot \frac{f\left(\pi_i\right)}{f\left(\bar{\pi}^*\right)} + (1-m)\, n \cdot \frac{f\left(\pi_i\right)}{(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\bar{\pi}^*\right)}\Bigg]$$

ρ: probability that any given island is drawn into war

$v\left(\boldsymbol{\pi}, \bar{\pi}^*\right)$: probability that an island, in which material payoff profile $\boldsymbol{\pi} \in \mathbb{R}^n$ obtains, wins a war when the average payoff in the rest of the population is $\bar{\pi}^*$

Model B: group structure modeled explictly

Three canonical scenarios: Guns

If $f\left(\pi_i\right) = f_0 \cdot \exp\left(\delta \cdot \pi_i\right)$ and $v_n\left(\boldsymbol{\pi}, \bar{\pi}^*\right) = \frac{\exp(\delta \cdot n\bar{\pi})}{\exp(\delta \cdot n\bar{\pi}) + \exp(\delta \cdot n\pi^*)}$, then:

$$r = \frac{\left(1-m\right)^2}{n - \left(n-1\right)\left(1-m\right)^2}$$

$$\lambda = \frac{\left(n-1\right)\left(1-m\right)^2 - \rho\left(n-1\right)n/2}{n - \left(1-m\right)^2 + \rho n/2}$$

In this scenario, $\lambda > 0$ if ρ is small, but $\lambda < 0$ if ρ is large: the model predicts a combination of material self-interest, a semi-Kantian concern, and either spite or altruism, depending on the frequency of wars.

Model B: group structure modeled explicitly

Three canonical scenarios: Culture

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \bar{\pi}^*\right) = s\left(\pi_i\right) + \; m \cdot \left[1 - s\left(\bar{\pi}^*\right)\right] \cdot \frac{f\left(\pi_i\right)}{f\left(\bar{\pi}^*\right)}$$

$$+ \; \left(1-m\right) \cdot \left(n - \sum_{j=1}^{n} s\left(\pi_j\right)\right) \cdot \frac{f\left(\pi_i\right)}{\sum_{j=1}^{n} f\left(\pi_j\right)}$$

$s\left(\pi_i\right)$: probability that i's child emulates i's trait

$f\left(\pi_i\right)$: attractiveness of the trait used by i

Model B: group structure modeled explictly

Three canonical scenarios: Culture

Suppose that $f(\pi_i) = f_0 \cdot \exp(\delta \cdot \pi_i)$ and $s(\pi_i) = s$. Then:

$$r = \frac{(1-m)\left[2s_0 + (1-m)(1-s_0)\right]}{n(1+s_0) - (1-m)(n-1)\left[2s_0 + (1-m)(1-s_0)\right]}$$

$$\lambda = \frac{(n-1)(1-m)}{n-(1-m)}$$

In this scenario, $\lambda > 0$: the model predicts a combination of material self-interest, a semi-Kantian concern, and spite.

Experimental evidence

Experimental evidence

- Van Leeuwen and Alger [forthcoming *JPE Microeconomics*]
- Participants play 18 different sequential game protocols
  - 6 (mini) Trust Games (TG)
  - 6 (mini) Ultimatum Games (UG)
  - 6 Sequential Prisoner's Dilemma's (SPD)

Experimental evidence

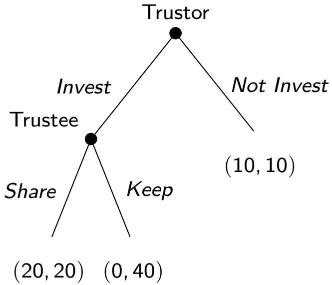
- We posit this utility function:

$$
\begin{aligned}
u_i(x,y) &= \pi(x,y) \\
&- (\alpha_i + q\delta_i)\cdot \max\{0, \pi(y,x) - \pi(x,y)\} \\
&- (\beta_i + p\gamma_i)\cdot \max\{0, \pi(x,y) - \pi(y,x)\} \\
&+ \kappa_i \cdot [\pi(x,x) - \pi(x,y)]
\end{aligned}
$$

  material self-interest
  attitude towards being behind (augmented by negative reciprocity)
  attitude towards being ahead (augmented by negative reciprocity)
  Kantian concern

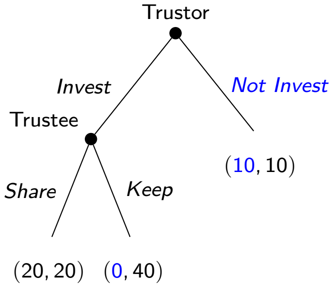
- We estimate each subject i's preference "type" $(\alpha_i, \beta_i, \kappa_i, \delta_i, \gamma_i)$, and their consistency with the posited utility function.
- We also examine whether estimation of a small number of preference "types" is sufficient to capture observed behavior.

Experimental evidence



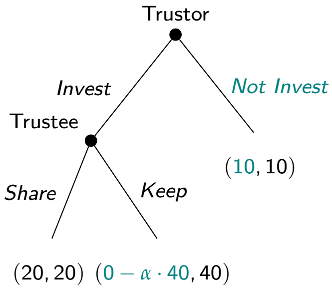
- suppose Trustor believes Trustee will *Keep*
- will Trustor *Invest* or *Not Invest*?

Experimental evidence



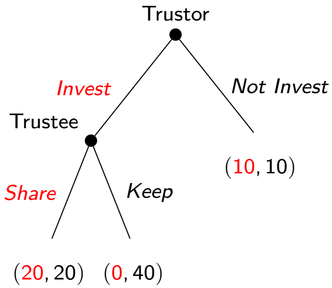
- suppose Trustor believes Trustee will *Keep*
- will Trustor *Invest* or *Not Invest*?
- it depends on his/her preferences:
  - material self-interest

Experimental evidence



- suppose Trustor believes Trustee will *Keep*
- will Trustor *Invest* or *Not Invest*?
- it depends on his/her preferences:
  - material self-interest
  - behindness aversion

Experimental evidence



- suppose Trustor believes Trustee will *Keep*
- will Trustor *Invest* or *Not Invest*?
- it depends on his/her preferences:
  - material self-interest
  - behindness aversion
  - Kantian concern

Experimental evidence

- Main findings:
  - heterogeneity in estimated preference types
  - most subjects' behavior is consistent with a combination of material self-interest, a semi-Kantian concern, and other-regard (altruism or spite)

Concluding remarks

- Theoretical models of preference evolution:
  - impact of environment on preferences?
  - discovery of preference classes that are novel to economics [Alger and Weibull 2013, and Alger, Weibull, and Lehmann 2020]
  - in particular: group structure $\rightarrow$ preferences which combine a concern for own material payoff, a semi-Kantian concern, as well as altruism or spite
- Recent surveys:
  - Alger and Weibull [*Annual Review of Economics* 2019]
  - Alger [*Philosophical Transactions B* 2023]

Concluding remarks

- Experimental evidence of semi-Kantian concerns
  [Capraro and Rand 2018, Miettinen, Kosfeld, Fehr and Weibull 2020, Levine et al. 2020, Van Leeuwen and Alger (forthc.), Alger and Rivero Wildemauwe (WiP)]
- Theoretical predictions under semi-Kantian concerns
  [Laffont 1975, Bergstrom 1995, Alger and Weibull (2017), Sarkisian (2017, 2021), Roemer (2019), Norman (2020), De Donder et al. (2021), Eichner and Pethig (2021, 2022), Ayoubi and Thurm (2022), Muñoz (2022), Alger and Laslier (2022), Salonia (2023), Juan Bartroli and Karagözoğlu (2023), Juan Bartroli (2023)]

Thanks

Funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 789111 - ERC EvolvingEconomics) is gratefully acknowledged.