Preference evolution

Ingela Alger and Jörgen Weibull

May 8, 2014

1 Introduction

- Preferences $\Rightarrow$ behaviors $\Rightarrow$ material payoff consequences $\Rightarrow$ evolutionary selection pressure on preferences [indirect evolution, Güth and Yaari (1992)]

- Question: how could preferences that differ from material payoff maximization survive?

- Literature on preference evolution has so far shown that there are two mechanisms whereby evolution by way of natural selection leads to non-selfish preferences

- First mechanism: *effect of own preferences on others' behaviors* [Schelling (1960)]
  - Inequity-averse responders do well in ultimatum bargaining

- Preference evolution under complete information [Fershtman and Judd (1987), Bester & Güth (1998), Bolle (2000), Possajennikov (2000), Koçkesen, Ok & Sethi (2000), Sethi & Somanathan (2001), Heifetz, Shannon and Spiegel (2007)]: *non-selfish preferences*

- Preference evolution under incomplete information [Ok & Vega-Redondo (2001), Dekel, Ely & Yilankaya (2007)]: *selfish preferences*

- Second mechanism: *assortative matching*

- A long-standing tradition in biology [Hamilton (1964), Hines and Maynard Smith (1979), Grafen (1979), Bergstrom (1995, 2003)]

- Literature on preference evolution [Alger (2010), Alger and Weibull (2010, 2012, 2013)]

- Result: preferences that induce non-selfish behaviors are selected for, and selfish preferences are selected against

- Assortativity is positive as soon as there is a positive probability that interacting parties have inherited their preferences or moral values from a common "ancestor" (genetic or cultural)

- In biology: genetics, kinship and "inclusive fitness" (Hamilton, 1964)

- In social science: culture, education, ethnicity, geography, networks, customs and habits

- Homophily [McPherson, Smith-Lovin, and Cook (2001), Ruef, Aldrich, and Carter (2003), Currarini, Jackson, and Pin (2009, 2010), Bramoullé and Rogers (2009)]

- This morning:

1. Evolutionary stability of *strategies* in a population where individuals are *uniformly randomly* matched into *pairs* to interact

2. Evolutionary stability of preferences (within the parametric class of altruistic preferences) in a population where *siblings* interact in *pairs*; in sibling interactions there is *assortativity*: a mutant is more likely than a resident to interact with a mutant

- What's next?

A general model of evolutionary stability of *traits* in a population where individuals are *randomly* (but perhaps *assortatively*) matched into n-*player groups* to interact $+$ applications

2 The general model

- A continuum population
- Individuals are randomly (but not necessarily uniformly) matched into n-player groups
- Each group plays a symmetric game in material payoffs
- Material payoff from playing $x_i \in X$ against $\mathbf{x}_{-i} \in X^{n-1}$: $\pi(x_i, \mathbf{x}_{-i})$
- Normal form (material) game $\Gamma = \langle X, \pi, n \rangle$

- Each individual carries some heritable *trait* $\theta \in \Theta$ which determines his/her behavior in the material game

- For our stability analysis we consider populations with at most two types present, θ and τ, in arbitrary proportions $1 - \varepsilon$ and ε

- If ε is small and positive, θ is called the *resident* trait and τ the *mutant* trait

- We study the type distribution's robustness to small and rare random shocks

- The matching process is exogenous and random

- For a given population *state* $s = (\theta, \tau, \varepsilon)$:
  - let $\Pr(\theta|\theta, \varepsilon)$ be the probability that, for a given **resident**, another group member (uniformly randomly drawn from the group) is a resident
  - let $\Pr(\theta|\tau, \varepsilon)$ be the probability that, for a given **mutant**, another group member (uniformly randomly drawn from the group) is a resident

- Let $\phi(\varepsilon) = \Pr[\theta|\theta, \varepsilon] - \Pr[\theta|\tau, \varepsilon]$ and call ϕ the *assortment function*

- Let $\lim_{\varepsilon \to 0} \phi(\varepsilon) = \sigma$, for some $\sigma \in [0, 1]$, the *index of assortativity*
  - Uniform random matching $\Rightarrow \sigma = 0$
  - Sibling interactions when types are inherited from parents $\Rightarrow \sigma = 1/2$
  - "Cultural parents" and homophily: $\sigma \in (0, 1)$

- Statistical issue for $n > 2$: potential *conditional dependence* (given the type of the individual at hand, between pairs of other members)

- We assume that conditional dependence vanishes in the limit as $\varepsilon \longrightarrow 0$

- Thus, for a mutant, the type distribution among the other $n - 1$ players converges to $Bin\left(\sigma, n-1\right)$ as $\varepsilon \longrightarrow 0$

- Assume: an individual's trait uniquely determines her average material payoff

- Let $F(\theta, \tau, \varepsilon)$ and $G(\theta, \tau, \varepsilon)$ denote the average material payoff to an individual with trait θ and trait τ, respectively

- Assume: $F(\theta, \tau, \cdot)$ and $G(\theta, \tau, \cdot)$ are continuous

Definition 1 *A trait $\theta \in \Theta$ is* **evolutionarily stable against a trait** $\tau \in \Theta$ *if there exists an $\bar{\varepsilon}_\tau > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_\tau)$:*

$$F(\theta, \tau, \varepsilon) > G(\theta, \tau, \varepsilon).$$

θ is an **evolutionarily stable trait (EST)** *if it is evolutionarily stable against all traits $\tau \neq \theta$ in Θ.*

A *sufficient* condition for $\theta \in \Theta$ to be an EST is that, for all $\tau \neq \theta$,

$$\lim_{\varepsilon \to 0} F(\theta, \tau, \varepsilon) > \lim_{\varepsilon \to 0} G(\theta, \tau, \varepsilon) \tag{1}$$

Let $H : \Theta^2 \to \mathbb{R}$ be the function defined by

$$H(\tau, \theta) = \lim_{\varepsilon \to 0} G(\theta, \tau, \varepsilon)$$

$H(\theta, \theta) = \lim_{\varepsilon \to 0} G(\theta, \theta, \varepsilon) = \lim_{\varepsilon \to 0} F(\theta, \theta, \varepsilon) = \lim_{\varepsilon \to 0} F(\theta, \tau, \varepsilon)$ implies that (1) may be written:

$$H(\theta, \theta) > H(\tau, \theta)$$

Proposition *For $\theta \in \Theta$ to be an EST, (θ, θ) must be a Nash equilibrium of the two-player game in which the common strategy set is Θ and the payoff function is H. A sufficient condition is that (θ, θ) is a strict Nash equilibrium of this game.*

2.1 Strategy evolution

An individual's strategy depends only on his/her trait; formally, let $\Theta = X$

If x is the resident strategy and y the mutant strategy,

$$G(x,y,\varepsilon) = \sum_{m=1}^{n} \binom{n-1}{m-1} [\Pr(y|y,\varepsilon)]^{m-1} [\Pr(x|y,\varepsilon)]^{n-m} \cdot \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right)$$

and

$$H(y,x) = \sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right)$$

For $n = 2$:

$$H(y, x) = (1 - \sigma) \cdot \pi(y, x) + \sigma \cdot \pi(y, y)$$

For $n = 3$:

$$H(y, x) = (1 - \sigma)^2 \cdot \pi(y, x, x) + 2\sigma \cdot (1 - \sigma) \cdot \pi(y, y, x) + \sigma^2 \cdot \pi(y, y, y)$$

Proposition *Suppose that π is continuously differentiable and that X is an open set. Then, if $\hat{x} \in X$ is an evolutionarily stable strategy,*

$$\pi_1(\hat{\mathbf{x}}) + \sigma \cdot (n - 1) \cdot \pi_n(\hat{\mathbf{x}}) = 0,$$

where $\hat{x}$ is the n-dimensional vector whose components all equal $\hat{x}$.

A canonical public-goods situation ($\gamma \in (0,1]$ and $c > 0$):

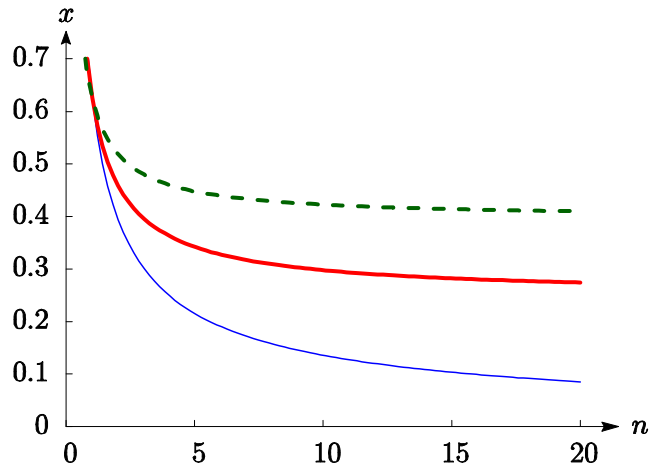
$$\pi\left(x_i, \mathbf{x}_{-i}\right) = \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^{\gamma} - \frac{c}{2}x_i^2$$

$$\begin{aligned} H\left(y,x\right) &= \sum_{m=1}^{n}\left(\begin{array}{c} n-1 \\ m-1 \end{array}\right)\sigma^{m-1}\left(1-\sigma\right)^{n-m} \\ &\quad \cdot\left[\frac{m}{n}y + \left(1-\frac{m}{n}\right)x\right]^{\gamma} - \frac{c}{2}y^2 \end{aligned}$$

$H_1\left(y,x\right)_{|y=x} = 0$ is necessary and sufficient for x to be an ESS

Proposition *The unique ESS is:*

$$\hat{x} = \left[\frac{\sigma\gamma + \frac{1}{n}(1-\sigma)\gamma}{c} \right]^{\frac{1}{2-\gamma}}$$



2.2 Preference evolution under complete information

- Each trait $\theta \in \Theta$ uniquely determines a *utility function* $u_\theta : X^n \to \mathbb{R}$

- Letting $\Pi^{(n)}(\tau, \theta, m/n)$ be the equilibrium material payoff to a τ-individual in a group with a share m/n of τ-individuals:

$$G(\theta, \tau, \varepsilon) = \sum_{m=1}^{n} \binom{n-1}{m-1} [\Pr(\tau|\tau, \varepsilon)]^{m-1} [\Pr(\theta|\tau, \varepsilon)]^{n-m} \cdot \Pi^{(n)}(\tau, \theta, m/n)$$

and

$$H(\tau, \theta) = \sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} \Pi^{(n)}(\tau, \theta, m/n)$$

2.2.1 Altruism

- Trait: degree of altruism

- Utility for an individual i with degree of altruism α:

$$u_\alpha\left(x_i, \mathbf{x}_{-i}\right) = \pi\left(x_i, \mathbf{x}_{-i}\right) + \alpha \sum_{j \neq i} \pi\left(x_j, \mathbf{x}_{-j}\right)$$

- Set of potential traits: $\Theta = [-1, 1]$

- Let α be the resident trait and β the mutant trait:

$$H\left(\beta, \alpha\right) = \sum_{m=1}^{n} \left(\begin{array}{c} n-1 \\ m-1 \end{array} \right) \sigma^{m-1} \left(1-\sigma\right)^{n-m} \Pi^{(n)}\left(\beta, \alpha, m/n\right)$$

The public goods example again:

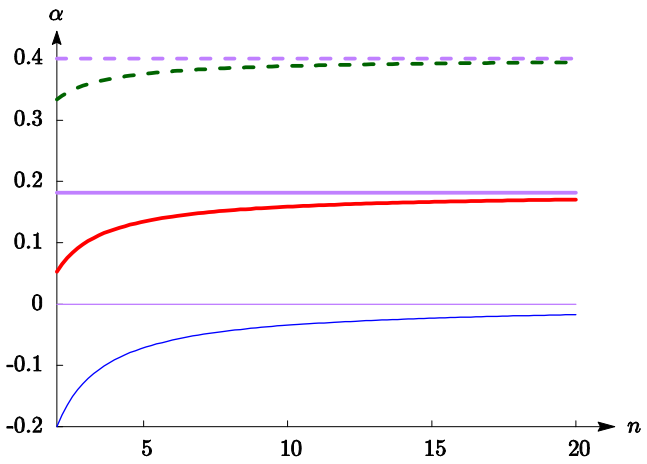
$$\begin{aligned} u_\alpha\left(x_i, \mathbf{x}_{-i}\right) &= \pi\left(x_i, \mathbf{x}_{-i}\right) + \alpha \cdot \sum_{j \neq i} \pi\left(x_j, \mathbf{x}_{-j}\right) \\ &= \left[1 + (n-1)\,\alpha\right] \cdot \left(\frac{1}{n} x_i + \frac{1}{n} \sum_{j \neq i} x_j\right)^\gamma \\ &\quad -\frac{c}{2} \cdot \left(x_i^2 + \alpha \sum_{j \neq i} x_j^2\right) \end{aligned}$$

If there are m β-altruists and $(n-m)$ α-altruists, a Nash eq. strategy profile is a n-dimensional vector with m components equal to y and $n-m$ components equal to x, where (x, y) solves:

$$\begin{cases} \gamma\left[\frac{1}{n} + \left(1 - \frac{1}{n}\right)\alpha\right] \cdot \left[\frac{m}{n} y + \left(1 - \frac{m}{n}\right) x\right]^{\gamma-1} - cx = 0 \\ \gamma\left[\frac{1}{n} + \left(1 - \frac{1}{n}\right)\beta\right] \cdot \left[\frac{m}{n} y + \left(1 - \frac{m}{n}\right) x\right]^{\gamma-1} - cy = 0 \end{cases}$$

Proposition *The unique locally evolutionarily stable degree of altruism is*

$$\hat{\alpha} = \frac{\sigma - \frac{1}{n}(1-\gamma)(1-\sigma)}{1 + \left(1 - \frac{1}{n}\right)(1-\gamma)(1-\sigma)}$$



2.3 Preference evolution under incomplete information

- Each trait $\theta \in \Theta$ uniquely determines a *utility function* $u_\theta : X^n \to \mathbb{R}$

Definition 2 *In any state* $s = (\theta, \tau, \varepsilon) \in S$*, the (assumed unique) (***Bayesian) Nash Equilibrium** *is the strategy pair* $(x^*, y^*) \in X^2$ *satisfying*

$$\begin{cases} x^* \in \arg\max_{x \in X} \ U_\theta \\ y^* \in \arg\max_{y \in X} \ U_\tau \end{cases}$$

where

$$U_\theta = \sum_{m=0}^{n-1} \binom{n-1}{m} \left[\Pr\left(\theta|\theta,\varepsilon\right)\right]^{n-m-1} \left[\Pr\left(\tau|\theta,\varepsilon\right)\right]^{m} u_\theta\left(x, \mathbf{y}^{*(m)}, \mathbf{x}^{*(n-m-1)}\right)$$

$$U_\tau = \sum_{m=1}^{n} \binom{n-1}{m} \left[\Pr\left(\theta|\tau,\varepsilon\right)\right]^{n-m} \left[\Pr\left(\tau|\tau,\varepsilon\right)\right]^{m-1} u_\tau\left(y, \mathbf{y}^{*(m-1)}, \mathbf{x}^{*(n-m)}\right)$$

- Given $s = (\theta, \tau, \varepsilon)$, let $\left(x^*_{(\varepsilon)}, y^*_{(\varepsilon)} \right)$ denote the unique BNE. Then:

$$G(\theta, \tau, \varepsilon) = \sum_{m=1}^{n} \left(\begin{array}{c} n-1 \\ m-1 \end{array} \right) [\Pr(\tau|\tau, \varepsilon)]^{m-1} [\Pr(\theta|\tau, \varepsilon)]^{n-m} \cdot \pi \left(y^*_{(\varepsilon)}, \mathbf{y}^{*\,(m-1)}_{(\varepsilon)}, \mathbf{x}^{*\,(n-m)}_{(\varepsilon)} \right)$$

$$H(\tau, \theta) = \sum_{m=1}^{n} \left(\begin{array}{c} n-1 \\ m-1 \end{array} \right) \sigma^{m-1} (1-\sigma)^{n-m} \pi \left(y^*_{(0)}, \mathbf{y}^{*\,(m-1)}_{(0)}, \mathbf{x}^{*\,(n-m)}_{(0)} \right)$$

- Let $\beta_\theta : X \rightrightarrows X$ denote the the best-reply correspondence,

$$\beta_\theta(y) = \arg\max_{x \in X} u_\theta\left(x, \mathbf{y}^{(n-1)}\right) \quad \forall y \in X$$

and $X_\theta \subseteq X$ the set of fixed points under β_θ,

$$X_\theta = \{x \in X : x \in \beta_\theta(x)\}$$

- Let Θ_θ be the set of behavioral clones:

$$\Theta_\theta = \left\{\tau \in \Theta : \exists\, x \in X_\theta \text{ such that } (x, x) \in B^{NE}(\theta, \tau, 0)\right\}$$

Theorem *Suppose the behavior of homo moralis, in the absence of mutants, is uniquely determined. Then:*

(a) Homo moralis with degree of morality σ is evolutionarily stable against all types that are not its behavioral clones.

(b) All types that are not its behavioral clones are evolutionarily unstable if the type set is rich.

- So, what, exactly, is a *homo moralis*?

- For each $\mathbf{x} \in X^n$ and $\kappa \in [0, 1]$, and any player i, let $\tilde{\mathbf{x}}_{-i}$ be a random vector with statistically independent components $\tilde{x}_j$ $(j \neq i)$ where

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

Definition 3 *A* **homo moralis** *is an individual with utility function*

$$u_\kappa\left(x_i, \mathbf{x}_{-i}\right) = \mathbb{E}_\kappa\left[\pi\left(x_i, \tilde{\mathbf{x}}_{-i}\right)\right] \quad \forall \mathbf{x} \in X^n.$$

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

- For $0 < \kappa < 1$, the individual's goal is to choose a strategy x_i that, if used with probability κ by other players, would maximize her material payoff. (How would it be if, with probability κ, each individual would do what I do?)

- For $n = 2$:

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

- For $n = 3$:

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

Corollary *Any BNE strategy x^* in a monomorphic population of homo moralis with $\kappa = \sigma$ is also an ESS. Moreover, if a strategy is a ESS for some σ, it is also a BNE strategy in a monomorphic population of homo moralis with degree of morality $\kappa = \sigma$.*

- Evolutionarily stable strategies may be viewed as emerging from preference evolution when individuals are not programmed to strategies but are rational and play equilibria under incomplete information.

3 Implications

- Applications to
  - environmental economics
  - moral hazard, principal-agent relations

    (Alger and Ma (2003), Alger and Renault (2006,2007)
  - bargaining
  - participation and voting in elections

4 Conclusions

- Our analysis suggests that selfishness is evolutionarily stable only in special circumstances, while *homo moralis* with degree of morality equal to the index of assortativity is always evolutionarily stable.

- Moral preferences may thrive, even under incomplete information and even in very large groups

- Lots of new challenges: extensions, applications, tests in laboratory experiments...

THE END