Int J Game Theory
DOI 10.1007/s00182-016-0548-4

The preferences of Homo Moralis are unstable under evolving assortativity

Jonathan Newton[1]

Accepted: 20 July 2016

Abstract Differing degrees of assortativity in matching can be expected to have both genetic and cultural determinants. When assortativity is subject to evolution, the main result of Alger and Weibull (Econometrica 81:2269–2302 2013) on the evolution of stable other-regarding preferences does not hold. Instead, both non-Nash and Pareto inefficient behavior are evolutionarily unstable.

Keywords Evolution · Moral values · Assortative matching

JEL Classification C73

1 Introduction

Alger and Weibull (2013) show that, under an exogenously given matching protocol, a population consisting of types (genotypes) whose behavior (phenotype) is determined by a particular utility function, *homo hamiltonensis*, is robust to invasion by types whose behavior differs from that of *homo hamiltonensis*. The level of other-regarding behavior exhibited by *homo hamiltonensis* depends directly on the level of assortativity in matching shown by small populations of invading mutants. Specifically, in a population of whom a proportion $1-\varepsilon$ are *homo hamiltonensis* and a proportion ε are

Come rain or come shine, I can be reached at jonathan.newton@sydney.edu.au, telephone +61293514429. This work was completed while the author was supported by a Discovery Early Career Researcher Award (DE130101768) funded by the Australian Research Council. Sincere thanks are given to the associate editor and an anonymous referee.

✉ Jonathan Newton
jonathan.newton@sydney.edu.au

[1] School of Economics, University of Sydney, New South Wales, Australia

Published online: 26 July 2016

some invading type τ, the level of other-regarding behavior by *homo hamiltonensis* is given by $\lim_{\varepsilon \to 0} Pr[\tau|\tau, \varepsilon]$, where $Pr[\tau|\tau, \varepsilon]$ is the probability that an agent matches with a τ-type given that he himself is of type τ and that there are ε τ-types in the population. That is, *homo hamiltonensis*' behavior depends directly on the behavior, as manifested via the degree of assortativity, of an invading type.

It is assumed by Alger and Weibull (2013) that the degree of assortativity is type independent. That is, $Pr[\tau|\tau, \varepsilon]$ does not depend on τ. This is a very strong assumption, whether $Pr[\tau|\tau, \varepsilon]$ is considered to be biologically or culturally determined. In fact, a deep and interesting literature exists that looks at the evolution of assortative behavior, in which mutants can exhibit higher or lower degrees of assortativity.[1] In addition, factors that indirectly lead to greater or lesser assortativity, such as the predilection to roam far from home or habitat location and size, are subject to evolutionary pressures.[2] Cultural determinants of assortativity also differ as social groups vary in degree of hostility to outsiders and openness to external influence.[3]

Considering the above, it is important to include assortativity in the possible behaviors determined by evolution. That is, to consider $Pr[\tau|\tau, \varepsilon]$ dependent on τ. Following this change, the predictions of Alger and Weibull (2013) no longer hold. If there exists a *selfish rover* type, τ_r, whose strategic behavior is determined solely by individual fitness considerations, and for whom $\lim_{\varepsilon \to 0} Pr[\tau|\tau, \varepsilon] \to 0$, then non-Nash equilibrium behavior is evolutionarily unstable. Moreover, if there exists a *Kantian parochial* type, τ_p, whose strategic behavior maximizes fitness from symmetric strategy profiles, and for whom $Pr[\tau|\tau, \varepsilon] = 1$, then Pareto inefficient behavior is also evolutionarily unstable. The only exception to these negative results arises when the incumbent type is Kantian parochial. Such types will not interact with mutants, and without such interaction, mutants cannot do better than Kantian parochial incumbents.

Thus, unless there exists a symmetric Nash equilibrium that is efficient amongst all symmetric strategy profiles, or incumbents are perfectly isolated from any mutant invasion, no single form of other-regarding preferences can monopolize the population; we should expect to observe a variety of self-regarding and other-regarding behavior by humans. People will behave differently, even when they face identical situations. That is, we anticipate diverse preferences, rather than a specific type of *homo moralis*.

2 Model and result

Consider a population whose individuals are randomly matched into pairs to engage in a symmetric interaction with the common strategy set X. An individual playing strategy x against an individual playing strategy y receives a payoff, representing biological fitness, $\pi(x, y)$, where $\pi : X^2 \to \mathbb{R}$. The pair $\langle X, \pi \rangle$ is the *fitness game*. X is a nonempty, compact and convex set in a topological vector space and π is continuous. Each individual is characterized by a type $\theta \in \Theta$ which defines a continuous utility

[1] See, for example, (Servedio 2010; Cara, M.A.R.d., Barton, N.H., Kirkpatrick, M., 2008; Dieckmann and Doebeli 1999; Otto et al. 2008; Matessi et al. 2002; Pennings et al. 2008).

[2] Dyson-Hudson and Smith (1978), Bearhop et al. (2005), López-Sepulcre and Kokko (2005).

[3] Cashdan (2001), Choi and Bowles (2007), Fry and Söderberg (2013).

function $u_\theta : X^2 \to \mathbb{R}$ and an index of assortativity $\sigma_\theta \in [0, 1]$. An individual's type is his private information.

Consider a population with two types present and define a *population state* $s = (\theta, \tau, \varepsilon)$, where $\theta, \tau \in \Theta$ are the two types and $\varepsilon \in (0, 1)$ is the population share of type τ.

The random matching process is such that firstly a share σ_θ of individuals of type θ are matched amongst themselves, and likewise a share σ_τ of individuals of type τ are matched amongst themselves. The remaining individuals of type θ and τ are then uniformly matched. Let $Pr[\theta|\theta, \varepsilon]$, $Pr[\tau|\theta, \varepsilon]$ be the probabilities that a given individual of type θ is matched with an individual of type θ, τ respectively. Let $Pr[\theta|\tau, \varepsilon]$, $Pr[\tau|\tau, \varepsilon]$ be the probabilities that a given individual of type τ is matched with an individual of type θ, τ respectively. The above description of the matching process yields

$$Pr[\tau|\tau, \varepsilon] = \frac{(1 - \sigma_\tau)\varepsilon + (1 - \sigma_\theta)\sigma_\tau(1 - \varepsilon)}{(1 - \sigma_\theta)(1 - \varepsilon) + (1 - \sigma_\tau)\varepsilon}.$$

Consequently, if $\sigma_\theta < 1$, then $\lim_{\varepsilon \to 0} Pr[\tau|\tau, \varepsilon] = \sigma_\tau$, and if $\sigma_\theta = 1$, then $\lim_{\varepsilon \to 0} Pr[\tau|\tau, \varepsilon] = 1$. In either case, the balancing condition for heterogeneous matchings implies that $\lim_{\varepsilon \to 0} Pr[\theta|\theta, \varepsilon] = 1$.

For a state $s = (\theta, \tau, \varepsilon)$, strategies $x \in X$ used by type θ and $y \in X$ used by type τ, the average fitness of each type is

$$\Pi_\theta(x, y, \varepsilon) = Pr[\theta|\theta, \varepsilon] \cdot \pi(x, x) + Pr[\tau|\theta, \varepsilon] \cdot \pi(x, y), \tag{1}$$

$$\Pi_\tau(x, y, \varepsilon) = Pr[\theta|\tau, \varepsilon] \cdot \pi(y, x) + Pr[\tau|\tau, \varepsilon] \cdot \pi(y, y). \tag{2}$$

It is assumed that the strategies chosen by individuals of both types are a (Bayesian) Nash equilibrium.

Definition 2.1 In any state $s = (\theta, \tau, \varepsilon)$, a strategy pair $(x^*, y^*) \in X^2$ is a (Bayesian) Nash Equilibrium, $(x^*, y^*) \in B^{\mathrm{NE}}(s)$, if

$$\begin{cases} x^* \in \arg\max_{x \in X} Pr[\theta|\theta, \varepsilon] \cdot u_\theta(x, x^*) + Pr[\tau|\theta, \varepsilon] \cdot u_\theta(x, y^*), \\ y^* \in \arg\max_{y \in X} Pr[\theta|\tau, \varepsilon] \cdot u_\tau(y, x^*) + Pr[\tau|\tau, \varepsilon] \cdot u_\tau(y, y^*). \end{cases} \tag{3}$$

This definition defines, for fixed types θ, τ, an equilibrium correspondence $B^{\mathrm{NE}}(\theta, \tau, \cdot) : (0, 1) \rightrightarrows X^2$ that maps mutant population shares to equilibria. Letting $Pr[\cdot|\cdot, 0] = \lim_{\varepsilon \to 0} Pr[\cdot|\cdot, \varepsilon]$, the domain of $B^{\mathrm{NE}}(\theta, \tau, \cdot)$ can be extended to $[0, 1)$.

The same definition of evolutionary instability as Alger and Weibull (2013) is used.[4]

[4] The requirement that an incumbent strategy be strictly beaten by an invader in this definition makes it more akin to neutral stability (Maynard Smith 1982) than to typical notions of evolutionary stability (Maynard Smith and Price 1973; Taylor and Jonker 1978). For an elegant discussion of the relation between concepts of evolutionary stability and asymptotic and Lyapunov stability under the replicator dynamics, the reader is referred to Bomze and Weibull (1995).

Definition 2.2 A type $\theta \in \Theta$ is *evolutionarily unstable* if there exists a type $\tau \in \Theta$ and $\bar{\varepsilon} > 0$ such that $\Pi_\theta(x^*, y^*, \varepsilon) < \Pi_\tau(x^*, y^*, \varepsilon)$ in all Nash equilibria (x^*, y^*) in all states $s = (\theta, \tau, \varepsilon)$ with $\varepsilon \in (0, \bar{\varepsilon})$.

Two types are now defined, the first of which will guarantee that non-Nash behavior is unstable, the second of which will guarantee that Pareto inefficient behavior is unstable. Note that, replacing a universal value of σ by type specific σ_τ, both of these types are varieties of *homo hamiltonensis* (Alger and Weibull 2013).

Definition 2.3 The *selfish rover* type τ_r satisfies $u_{\tau_r}(x, y) = \pi(x, y); \quad \sigma_{\tau_r} = 0$.
The *Kantian parochial* type τ_p satisfies $u_{\tau_p}(x, y) = \pi(x, x); \quad \sigma_{\tau_p} = 1$.

Type τ_r is *selfish* in that individuals with this type have preferences that are aligned perfectly to individual fitness. They are *rovers* in that they have no proclivity to have a disproportionate share of their interactions with individuals of the same type as themselves. Type τ_p is *Kantian* in the sense of Alger and Weibull (2013) who argue that the maximization of symmetric payoffs is akin to Kant's (1785) categorical imperative to "act only on that maxim whereby thou canst at the same time will that it should become a universal law."[5] They are *parochial* in that they interact only with individuals of the same type.

For each type $\theta \in \Theta$, let $\beta_\theta : X \rightrightarrows X$ denote the best response correspondence, $\beta_\theta(y) = \arg\max_{x \in X} u_\theta(x, y)\ \forall y \in X$, and $X_\theta \subseteq X$ the set of fixed points under β_θ,

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

Note that X_{τ_r} corresponds to the set of symmetric Nash equilibria when selfish individuals maximize their own fitness. In contrast, X_{τ_p} corresponds to the set of Pareto efficient symmetric strategy profiles.

Theorem 2.1 (a) *If $X_\theta \cap X_{\tau_r} = \emptyset$, $\sigma_\theta < 1$, $\tau_r \in \Theta$ and $X_\theta = \{\tilde{x}\}$, $\beta_{\tau_r}(\tilde{x})$ are singletons, then θ is evolutionarily unstable.*
(b) *If $X_\theta \cap X_{\tau_p} = \emptyset$ and $\tau_p \in \Theta$, then θ is evolutionarily unstable.*
(c) *τ_p is not evolutionarily unstable.*

Proof Let $s = (\theta, \tau_r, \varepsilon)$, $(x^*, y^*) \in B^{\mathrm{NE}}(\theta, \tau_r, 0)$. Note that $Pr[\theta|\theta, 0] = 1$ and, as $\sigma_\theta < 1$, $Pr[\tau_r|\tau_r, 0] = 0$. From (3) it follows that

$$x^* \in \arg\max_{x \in X} u_\theta(x, x^*) \tag{4}$$

which implies $x^* \in X_\theta$. Also from (3),

$$y^* \in \arg\max_{y \in X} u_{\tau_r}(y, x^*) = \arg\max_{y \in X} \pi(y, x^*). \tag{5}$$

[5] For this Kantian interpretation, fitnesses $\pi(.,.)$ must be in some sense absolute and not just relative, otherwise there would be no reason to desire any given society-wide symmetric strategy profile over another. There is a separate debate that can be had on why or whether one would desire a universal law that maximizes the reproductive fitness of society.

If $x^* \in \arg\max_{y\in X} \pi(y, x^*)$, then $x^* \in X_{\tau_r}$, contradicting $X_\theta \cap X_{\tau_r} = \emptyset$. Therefore, $x^* \notin \arg\max_{y\in X} \pi(y, x^*)$, which implies that $\pi(y^*, x^*) > \pi(x^*, x^*)$, hence $\Pi_{\tau_r}(x^*, y^*, 0) > \Pi_\theta(x^*, y^*, 0)$.

The assumption that $X_\theta = \{\tilde{x}\}$ and $\beta_{\tau_r}(\tilde{x}) = \arg\max_{y\in X} \pi(y, \tilde{x})$ are singletons implies that $X_\theta = \{x^*\}$ and $\arg\max_{y\in X} \pi(y, x^*) = \{y^*\}$. That is, (x^*, y^*) is uniquely determined by (4) and (5). Continuity of Π_{τ_r}, Π_θ implies that $\Pi_{\tau_r}(x, y, \varepsilon) > \Pi_\theta(x, y, \varepsilon)$ for all (x, y, ε) in some neighborhood $U \subset X^2 \times [0, 1)$ of $(x^*, y^*, 0)$. By upper-hemicontinuity of $B^{\mathrm{NE}}(\theta, \tau_r, \cdot)$ (Lemma 1 of Alger and Weibull 2013), for small enough ε we have that $B^{\mathrm{NE}}(\theta, \tau_r, \varepsilon) \subset U$. This proves (a).

Now, let $s = (\theta, \tau_p, \varepsilon)$, $(x^*, y^*) \in B^{\mathrm{NE}}(\theta, \tau_p, \varepsilon)$. Note that $Pr[\theta|\theta, \varepsilon] = 1$, $Pr[\tau_p|\tau_p, \varepsilon] = 1$. (4) continues to hold so $x^* \in X_\theta$. Now,

$$y^* \in \arg\max_{y\in X} u_{\tau_p}(y, y^*) = \arg\max_{y\in X} \pi(y, y). \tag{6}$$

If $x^* \in \arg\max_{y\in X} \pi(y, y)$, then $x^* \in X_{\tau_p}$, contradicting $X_\theta \cap X_{\tau_p} = \emptyset$. Therefore, $x^* \notin \arg\max_{y\in X} \pi(y, y)$, which implies that $\pi(y^*, y^*) > \pi(x^*, x^*)$, hence $\Pi_{\tau_p}(x^*, y^*, \varepsilon) > \Pi_\theta(x^*, y^*, \varepsilon)$. This holds for any ε, $(x^*, y^*) \in B^{\mathrm{NE}}(\theta, \tau_p, \varepsilon)$, which proves (b).

Reversing the positions of θ and τ_p in the preceding paragraph so that $s = (\tau_p, \theta, \varepsilon)$ and τ_p is the incumbent type, we have that $x^* \in \arg\max_{y\in X} \pi(y, y)$, so $\pi(x^*, x^*) \geq \pi(y^*, y^*)$ and $\Pi_{\tau_p}(x^*, y^*, \varepsilon) \geq \Pi_\theta(x^*, y^*, \varepsilon)$ for any invading type θ, therefore τ_p is not evolutionarily unstable. □

Note that the condition in Theorem 2.1(a) that X_θ be a singleton mirrors the similar condition in Theorem 1 of Alger and Weibull (2013). It stems from the extremely strict definition of evolutionary instability adopted by the authors of the cited work (and hence here), whereby there must exist an invading strategy that, for small enough ε, can outperform the incumbent strategy in every equilibrium. It is not enough that the invading type eventually outperforms the incumbent type on any sequence of equilibria converging to any given $(x^*, y^*) \in B^{\mathrm{NE}}(\theta, \tau_r, 0)$ as $\varepsilon \to 0$.

Further note that the argument of Theorem 2.1(b) together with continuity of π implies that for a given incumbent type θ and sequence of equilibria converging to any given $(x^*, y^*) \in B^{\mathrm{NE}}(\theta, \tau_p, 0)$ as $\varepsilon \to 0$, for τ_p to eventually outperform θ it is not necessary that $\sigma_{\tau_p} = 1$, only that σ_{τ_p} be sufficiently close to 1. In contrast, Theorem 2.1(a) suggests that the qualitative implications of Theorem 2.1(c) are not robust to less than perfect segregation, as incumbent Kantians with σ strictly less than 1 are vulnerable to invasion by type τ_r. Finally, in a setting with unobservable types, it seems highly unlikely that an entire incumbent population will never interact with any mutant type. Consequently, the message the reader should take from this Theorem is that we can expect preferences to be unstable when they are conditioned only on an unobservable type.

3 Discussion

Theorem 2.1 shows that when both assortativity and preferences evolve, it is highly unlikely that either non-Nash behavior or inefficient behavior will persist indefinitely

in the long run. This is not particularly troublesome: the world, after all, is a dynamic and changing place. However, it is relevant to ask under what conditions the results of Alger and Weibull (2013) will pertain. It is clear that a necessary condition is that changes in assortativity take place on a much longer timescale than changes in preferences.[6] Unfortunately, such situations are difficult to conceive and would not seem to be common. The examples in the paper under discussion do not help here. The example of "Kin" (p. 2286, op.cit.) is uncontroversial but not germane as past behavior of close relatives would likely be observable, from which their (pheno)type could be inferred. Besides, in very small groups, there is a significant chance of a mutation attaining fixation via genetic drift, regardless of the direction of selection.

The example of "Geography, Homophily and Business Partnerships" (p. 2287, op.cit.) is one where assortativity would be expected to vary and be subject to selection on a similar timescale to preferences. The cited example considers N groups, each containing n individuals, with mutation affecting individuals within a single group. As mutations are restricted to a single group, the share of mutants in the population, ε, can be taken to zero by letting the number of groups N go to infinity. In the limit, the probability of an individual matching within his own group is denoted $p^*(n)$. If we assume that every member of the mutant-containing group is a mutant[7], then $p^*(n)$ is exactly the limiting probability of a mutant being matched with another mutant. That is, $Pr[\tau|\tau, 0] = \sigma_\tau = p^*(n)$. Now, let mutants, rather than having σ_τ fixed and equal to $p^*(n)$, differ in their proclivity to match within their own group. Consequently, if $p^*(n) < 1$, then $Pr[\tau|\tau, 0]$ will vary with σ_τ and Theorem 2.1(a), (b) will apply. If $p^*(n) = 1$, then Theorem 2.1(b), (c) will apply.

Finally, note that even if assortativity and preferences are determined at different genetic loci and simultaneous mutation is rare, the implications of Theorem 2.1 still hold. Given any fixed choice behavior in a population, there is no selection for or against different degrees of assortativity, so genetic drift will create clusters of rovers and clusters of parochials, thus providing hospitable environments for the invasion of selfish or Kantian behavior.

References

Alger I, Weibull JW (2013) Homo moralis-preference evolution under incomplete information and assortative matching. Econometrica 81:2269–2302

Bearhop S, Fiedler W, Furness RW, Votier SC, Waldron S, Newton J, Bowen GJ, Berthold P, Farnsworth K (2005) Assortative mating as a mechanism for rapid evolution of a migratory divide. Science 310:502–504

Bomze IM, Weibull JW (1995) Does neutral stability imply Lyapunov stability? Games Econ Behav 11:173–192

Cara MARd, Barton NH, Kirkpatrick M (2008) A model for the evolution of assortative mating. Am Naturalist 171:580–596

Cashdan E (2001) Ethnocentrism and xenophobia: a crosscultural study. Curr Anthropol 42:760–765

Choi JK, Bowles S (2007) The coevolution of parochial altruism and war. Science 318:636–640

[6] This is not required for similar work (Wilson and Dugatkin 1997) where behavioral type is observed and therefore individuals can intentionally assort by type.

[7] If the mutants must be a strict subset of the group membership, then this gives an upper bound on $Pr[\tau|\tau, \varepsilon]$.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. Nature 400:354–357
Dyson-Hudson R, Smith EA (1978) Human territoriality: an ecological reassessment. Am Anthropol 80:21–41
Fry DP, Söderberg P (2013) Lethal aggression in mobile forager bands and implications for the origins of war. Science 341:270–273
Kant I (1785) Fundamental principles of the metaphysic of morals, translation by Abbott, Thomas Kingsmill, 1829–1913. Project Gutenberg (2004)
López-Sepulcre A, Kokko H (2005) Territorial defense, territory size, and population regulation. Am Naturalist 166:317–325
Matessi C, Gimelfarb A, Gavrilets S (2002) Long-term buildup of reproductive isolation promoted by disruptive selection: how far does it go? Selection 2:41–64
Maynard Smith J (1982) Evol Theor Games. Cambridge University Press, Cambridge
Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18
Otto SP, Servedio MR, Nuismer SL (2008) Frequency-dependent selection and the evolution of assortative mating. Genetics 179:2091–2112
Pennings PS, Kopp M, Meszéna G, Dieckmann U, Hermisson J (2008) An analytically tractable model for competitive speciation. Am Naturalist 171:E44–E71
Servedio MR (2010) Limits to the evolution of assortative mating by female choice under restricted gene flow. Proceedings of the Royal Society B, Biological Sciences
Taylor PD, Jonker LB (1978) Evolutionary stable strategies and game dynamics. Math Biosci 40:145–156
Wilson DS, Dugatkin LA (1997) Group selection and assortative interactions. Am Naturalist 149:336–351