Homo Moralis

PREFERENCE EVOLUTION UNDER INCOMPLETE INFORMATION AND ASSORTATIVE MATCHING*

Ingela Alger† and Jörgen W. Weibull‡

February 12, 2013. This version: May 14, 2013.

Abstract

What preferences will prevail in a society of rational individuals when preference evolution is driven by the resulting payoffs? We show that when individuals' preferences are their private information, a convex combination of selfishness and morality stands out as evolutionarily stable. We call individuals with such preferences *homo moralis*. At one end of the spectrum is *homo oeconomicus*, who acts so as to maximize his or her own payoff. At the opposite end is *homo kantiensis*, who does what would be "the right thing to do," in terms of payoffs, if all others would do likewise. We show that the stable degree of morality - the weight placed on the moral goal - is determined by the degree of assortativity in the process whereby individuals are matched to interact.

Keywords: evolutionary stability, preference evolution, moral values, incomplete information, assortative matching.

JEL codes: C73, D03.

---

*We thank the editor and three anonymous referees for helpful comments. Earlier versions of this manuscript have been presented at an NBER workshop on culture and institutions, at Tel Aviv University, Stockholm School of Economics, IMEBE 2012, University of Southern California, UC Santa Barbara, UC Riverside, UC San Diego, the Becker-Friedman Institute conference "Biological Basis of Preferences and Behavior," University of Warwick, Ecole Polytechnique (Paris), Lancaster University (UK), University of Oxford, Institute for International Economic Studies (Stockholm), Toulouse School of Economics, GAMES 2012, ITAM, Frankfurt University, University of York, and University of Zürich. We thank Rajiv Sethi, Immanuel Bomze, Avinash Dixit, Tore Ellingsen, Jens Josephson, Wolfgang Leininger, Klaus Ritzberger, François Salanié, and Giancarlo Spagnolo for comments. This research received financial support from the Knut and Alice Wallenberg Research Foundation. Ingela Alger is grateful to Carleton University and SSHRC for financial support and to the Stockholm School of Economics for its hospitality.

†Toulouse School of Economics (LERNA, CNRS) and Institute for Advanced Study in Toulouse

‡Stockholm School of Economics and Institute for Advanced Study in Toulouse

1 Introduction

Most of contemporary economics is premised on the assumption that human behavior is driven by self-interest. However, in the early history of the profession it was common to include moral values as part of human motivation, see, e.g., Smith (1759) and Edgeworth (1881), and, for more recent examples, Arrow (1973), Laffont (1975), Sen (1977) and Tabellini (2008).[1] Furthermore, in recent years many economists have begun to question the predictive power of pure selfishness in certain interactions, and turned to alternative preferences such as altruism (Becker, 1976), warm glow (Andreoni, 1990), inequity aversion (Fehr and Schmidt, 1999), reciprocal altruism (Levine, 1998), sense of identity (Akerlof and Kranton, 2000, Bénabou and Tirole, 2011), preference for efficiency (Charness and Rabin, 2002), and desire for social esteem (Bénabou and Tirole, 2006, Ellingsen and Johannesson, 2008). Our goal here is to clarify the evolutionary foundation of human motivation, by asking from first principles what preferences and moral values humans should be expected to have.

It is well-known from theoretical biology that evolution favors altruistic behaviors—behaviors that benefit others at a cost to oneself—between relatives. This insight was formally developed by Hamilton (1964a,b); see also Grafen (1979, 2006), Hines and Maynard Smith (1979), Bergstrom (1995), and Day and Taylor (1998). While the genetics is often complex, the intuition is simple; a gene in an individual has a high probability, depending on the degree of kinship, to be present in his or her relatives. In particular, if this gene expresses itself in behaviors helpful to relatives, the reproductive success of said gene is enhanced, as long as the behavior is not too costly for the actor. While kinship altruism evidently cannot explain altruistic behaviors among non-kin, it has been recognized in the literature that any mechanism that brings about assortativity in the matching process can even favor altruistic behaviors among unrelated individuals;[2] a prime example of such a mechanism is geographic dispersion. In this literature, the unit of selection is behaviors (strategies) rather than, as here, preferences or moral values. Nevertheless, if one were to interpret the evolved behaviors as resulting from utility maximization, then this literature would point to two distinct classes of preferences: (a) altruistic preferences, whereby individuals attach positive weight to the well-being or fitness of others, and (b) moral preferences, whereby individuals instead are concerned with what is "the right thing to do".[3] Clearly, these two motivations may give

[1] See Binmore (1994) for a game-theoretic discussion of ethics, and Bolle and Ockenfels (1990), Sugden (1995, 2011), Bacharach (1999), Brekke, Kverndokk, and Nyborg (2003), Alger and Ma (2003), Alger and Renault (2006, 2007), Bénabou and Tirole (2011), Huck, Kübler and Weibull (2012), and Roemer (2010) for alternative models of moral motivation.

[2] See, e.g., Hamilton (1971, 1975), Boorman and Levitt (1980), Eshel and Cavalli-Sforza (1982), Toro and Silio (1986), Frank (1987,1988), Wilson and Dugatkin (1997), Sober and Wilson (1998), Rousset (2004), Nowak (2006), and Bergstrom (2003, 2009).

[3] The idea that moral values may have been formed by evolutionary forces can be traced back to at least

rise to different behaviors. However, this literature is silent as to whether either altruistic or moral preferences would in fact arise if evolution were to operate on preferences—as a way for nature to delegate the choices of concrete actions to the individual in any given situation. It is our goal to fill this gap.

There are several challenges associated with raising the domain of the analysis from behaviors to preferences. We show that these difficulties can be dealt with in a general model with minimal assumptions on the class of interactions and potential preferences. More exactly, we analyze the evolution of preferences in a large population where individuals are randomly and pairwise matched to interact. We follow the indirect evolutionary approach, pioneered by Güth and Yaari (1992), by assuming that individual behavior is driven by (subjective) utility maximization, while evolutionary success is driven by some (objective) payoff. A large body of research has shown that natural selection leads to preferences that deviate from objective-payoff-maximization when individuals who interact know each other's preferences.[4] We focus here instead on the case when each individual's type (preferences or moral values) is her private information. Moreover, we relax the commonly made assumption that all matches are equally likely (uniform random matching) and ask whether assortativity in the process whereby people are matched to interact affects the preferences that natural selection favors. Indeed, as we argue in Section 5, assortativity arises in many human interactions for a variety of different reasons.

We impose few assumptions on the set of admissible preferences that are subject to evolutionary selection. In particular, these may be altruistic, moral, selfish, driven by inequity aversion or commitment to particular behaviors, etc. Our analysis applies to symmetric interactions and to asymmetric interactions with *ex ante* symmetry, that is, when each individual is just as likely to be in one player role as in the other. For asymmetric interactions, then, evolution selects preferences behind a veil of ignorance regarding which role the individual will eventually play. The matching process is exogenous, and, building on Bergstrom (2003), we identify a single parameter, the *index of assortativity*, as a key parameter for the population-statistical analysis. We generalize the standard definition of evolutionary stabil-

---

Darwin (1871). More recent, but informal, treatments include, to mention a few, Alexander (1987), Nichols (2004) and de Waal (2006). The latter claims that moral codes also exist in other primates.

[4]See Robson (1990), Güth and Yaari (1992), Ockenfels (1993), Ellingsen (1997), Bester and Güth (1998), Fershtman and Weiss (1998), Koçkesen, Ok and Sethi (2000), Bolle (2000), Possajennikov (2000), Ok and Vega-Redondo (2001), Sethi and Somanathan (2001), Heifetz, Shannon and Spiegel (2006, 2007) Dekel, Ely and Yilankaya (2007), Alger (2010), and Alger and Weibull (2010, 2012a). As observed already by Schelling (1960), it may be advantageous in strategic interactions to be committed to certain behaviors, even if these appear to be at odds with one's objective self-interest. Indeed, certain other-regarding preferences such as altruism, spite, reciprocal altruism, or inequity aversion, if known or believed by others, may be strategically advantageous (or disadvantageous). For example, a manager of a firm in Cournot competition, with complete information about managers' contracts, will do better, in terms of equilibrium profits, if the contract rewards both profits and sales, rather than only profits (a literature pioneered by Fershtman and Judd, 1987).

ity, due to Maynard Smith and Price (1973), to allow for arbitrary degrees of assortativity and apply this to preference evolution when each matched pair plays a (Bayesian) Nash equilibrium of the associated game under incomplete information.[5]

Our main result is that natural selection leads to a certain one-dimensional family of moral preferences, a family that springs out from the mathematics. This family consists of all convex combinations of selfishness ("maximization of own payoff") and morality ("to do the right thing"). We call individuals with such preferences *homo moralis* and call the weight attached to the moral goal the *degree of morality*. A special case is the familiar *homo oeconomicus*, who attaches zero weight to morality. At the other extreme one finds *homo kantiensis* who attaches unit weight to morality. We show that evolution selects that degree of morality which equals the index of assortativity of the matching process. Such preferences in a resident population provide the most effective protection against mutants, since the residents' behavior is the behavior that would maximize the expected payoffs to mutants (when rare). It is as if *homo moralis* with degree of morality equal to the index of assortativity *preempts* mutants; any rare mutant can at best match the payoff of the residents.[6]

We also establish evolutionary instability of all preferences that induce other behaviors than those of *homo moralis* with degree of morality equal to the index of assortativity. A population consisting of individuals that behave differently would be vulnerable to invasion of mutants with other preferences. This instability result has dire consequences for *homo oeconomicus*, who is selected against in a large class of interactions that are strategic in the sense that a player's payoff depends on the other player's strategy, whenever there is a positive index of assortativity in the matching process. A sufficient condition for this is that the behavior of *homo oeconomicus* (when resident) be uniquely determined and different from that of individuals with degree of morality equal to the (positive) index of assortativity.

Our work establishes a link between two strands of literature, one (mostly biological) dealing with strategy evolution under assortative matching and another (in economics) dealing with preference evolution under uniform random matching. In the first strand, the most closely related work is that of Bergstrom (1995) who analyzes evolutionarily stable strategies in symmetric interactions between siblings. Bergstrom provides a moral interpretation of the

[5]Since the matching process is exogenous and an individual's preferences are her private information, there is no possibility for partner choice or mimickry. Alternative approaches would let individuals choose partners (see e.g. Frank, 1987 and 1988) or allow individuals to quit a partner and rematch (see e.g. Jackson and Watts, 2010). However, these approaches would add informational, strategic and matching-technological elements beyond the scope of this study.

[6]In a related literature, on cultural evolution, parents are assumed to be altruistic (or interested in their future treatment by their children) and, at some cost, they can influence their childrens' preferences and values; see, e.g., Bisin and Verdier (2001), Hauk and Saez-Martí (2002), Bisin, Topa and Verdier (2004), and Lindbeck and Nyberg (2006).

resulting behaviors, which he calls "semi-Kantian" (here corresponding to the behavior of *homo moralis* with degree of morality one half). In a similar spirit, Bergstrom (2009) provides game-theoretic interpretations of several existing moral maxims and relates these to evolutionarily stable strategies under assortative matching. In the second strand, the most closely related work is that of Ok and Vega-Redondo (2001) and Dekel, Ely and Yilankaya (2007). Their main result for interactions under incomplete information is that *homo oeconomicus* will prevail, a result that is corroborated in our analysis in the special case when the index of assortativity is zero.

In classic evolutionary game theory, evolutionary stability is a property of (pure or mixed) strategies and is usually applied to interactions in which individuals are "programmed" to strategies (Maynard Smith and Price, 1973). As a side result in this study, we obtain a new perspective on evolutionarily stable strategies, namely, that these behaviors are precisely those used in Nash equilibrium play when evolution operates at the level of preferences under incomplete information. Hence, evolutionary stability of strategies need not be interpreted in the narrow sense that individuals are "programmed" to a given strategy; the same behavior emerges if they are rational and play optimally under correct population-statistical beliefs about each other. This sharpens and generalizes the result in Dekel, Ely and Yilankaya (2007) that preference evolution under incomplete information and uniform random matching in finite games implies Nash equilibrium play, as defined in terms of the underlying payoffs, and is implied by strict Nash equilibrium (again in terms of payoffs).[7]

The rest of the paper is organized as follows. The model is set up in the next section. In Section 3 we establish our main result and show some of its implications. Section 4 is devoted to finite games. In Section 5 we study a variety of matching processes. Three topics are discussed in Section 6: asymmetric interactions, the difference between morality and altruism, and ways to test empirically the existence of *homo moralis*. Section 7 concludes.

2 Model

Consider a population where individuals are randomly matched into pairs to engage in a symmetric interaction with the common strategy set, X. While behavior is driven by (subjective) utility maximization, evolutionary success is determined by the resulting payoffs. An individual playing strategy x against an individual playing strategy y gets *payoff*, or *fitness increment*, $\pi(x, y)$, where $\pi : X^2 \to \mathbb{R}$. We will refer to the pair $\langle X, \pi \rangle$ as the *fitness game*. We assume that X is a non-empty, compact and convex set in a topological vector

[7] Strategies used in symmetric strict Nash equilibria are evolutionarily stable, and any evolutionarily stable strategy playing against itself makes a symmetric Nash equilibrium.

space and that π is continuous.[8] Each individual is characterized by his or her *type* $\theta \in \Theta$, which defines a continuous *(utility) function*, $u_\theta : X^2 \to \mathbb{R}$. We impose no relation between a utility function u_θ and the payoff function π. A special type is *homo oeconomicus*, by which we mean individuals with the utility function $u = \pi$. An individual's type is her private information.

For the subsequent analysis, it will be sufficient to consider populations with two types present. The two types and the respective population shares together 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 set of population states is thus $S = \Theta^2 \times (0, 1)$. If ε is small, we will refer to θ as the *resident* type and call τ the *mutant* type. The matching process is random and exogenous, and it may be assortative. More exactly, in a given state $s = (\theta, \tau, \varepsilon)$, let $\Pr[\tau|\theta, \varepsilon]$ denote the probability that a given individual of type θ is matched with an individual of type τ, and let $\Pr[\theta|\tau, \varepsilon]$ denote the probability that a given individual of type τ is matched with an individual of type θ. In the special case of uniform random matching, $\Pr[\tau|\theta, \varepsilon] = \Pr[\tau|\tau, \varepsilon] = \varepsilon$ for all $\varepsilon \in (0, 1)$.

For each state $s = (\theta, \tau, \varepsilon) \in S$ and any strategy $x \in X$ used by type θ and any strategy $y \in X$ used by type τ, the resulting average payoff, or fitness, to each type is:

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

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

As for the choices made by individuals, a (Bayesian) Nash equilibrium is a pair of strategies, one for each type, where each strategy is a best reply to the other in the given population state:

Definition 1 *In any state $s = (\theta, \tau, \varepsilon) \in S$, a strategy pair $(x^*, y^*) \in X^2$ is a **(Bayesian) Nash Equilibrium (BNE)** if*

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

We define evolutionary stability under the assumption that the resulting payoffs are determined by this equilibrium set.[9] With potential multiplicity of equilibria, one may require the resident type to withstand invasion in some or all equilibria. We have chosen the most stringent criterion.

---

[8] To be more precise, we assume X to be a locally convex Hausdorff space, see Aliprantis and Border (2006). For example, the game $\langle X, \pi \rangle$ may be a finite two-player extensive-form game, where X is the set of mixed or behavior strategies, see Section 6.1.

[9] This is in line with the literature on "indirect evolution" — see e.g. Güth and Yaari (1992), Huck and Oechssler (1999), and Dekel *et al.* (2007) — and can be interpreted as an adiabatic process in which preferences change on a slower time scale than actions, see Sandholm (2001).

Definition 2 *A type $\theta \in \Theta$ is **evolutionarily stable against a type** $\tau \in \Theta$ if there exists an $\bar{\varepsilon} > 0$ 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 $\varepsilon \in (0, \bar{\varepsilon})$. A type θ is **evolutionarily stable** if it is evolutionarily stable against all types $\tau \neq \theta$ in Θ.*

This definition formalizes the notion that a resident population with individuals of a given type would withstand a small-scale "invasion" of individuals of another type. It generalizes the Maynard Smith and Price (1973) concept of evolutionary stability, a property they defined for mixed strategies in finite and symmetric two-player games under uniform random matching. However, in a rich enough type set Θ, no type is evolutionarily stable against all other types, since for each resident type there then exist mutant types who behave like the residents and thus earn the same average payoff. A definition of such "behavioral clones" is given in Section 3.

We introduce a stringent notion of instability by requiring that there should exist some mutant type against which the resident type achieves strictly less payoff in every equilibrium in all population states where the mutant is arbitrarily rare:

Definition 3 *A type $\theta \in \Theta$ is **evolutionarily unstable** if there exists a type $\tau \in \Theta$ and an $\bar{\varepsilon} > 0$ 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 $\varepsilon \in (0, \bar{\varepsilon})$.*

The next subsection describes the algebra of assortative encounters introduced by Bergstrom (2003). This facilitates the analysis and clarifies the population statistics.

2.1 Algebra of assortative encounters

For given types $\theta, \tau \in \Theta$, and a population state $s = (\theta, \tau, \varepsilon)$ with $\varepsilon \in (0, 1)$, let $\phi(\varepsilon)$ be the difference between the conditional probabilities for an individual to be matched with an individual with type θ, given that the individual him- or herself either has type θ as well or, alternatively, type τ:

$$\phi(\varepsilon) = \Pr[\theta|\theta, \varepsilon] - \Pr[\theta|\tau, \varepsilon]. \tag{4}$$

This defines the *assortment function* $\phi : (0, 1) \to [-1, 1]$. Using the following necessary balancing condition for the number of pairwise matches between individuals with types θ and τ,

$$(1 - \varepsilon) \cdot \Pr[\tau|\theta, \varepsilon] = \varepsilon \cdot \Pr[\theta|\tau, \varepsilon], \tag{5}$$

one can write all conditional probabilities as functions of ε and $\phi(\varepsilon)$:

$$\begin{cases} \Pr[\theta|\theta, \varepsilon] = \phi(\varepsilon) + (1 - \varepsilon)[1 - \phi(\varepsilon)] = 1 - \Pr[\tau|\theta, \varepsilon] \\ \Pr[\theta|\tau, \varepsilon] = (1 - \varepsilon)[1 - \phi(\varepsilon)] = 1 - \Pr[\tau|\tau, \varepsilon]. \end{cases} \tag{6}$$

We assume that ϕ is continuous and that $\phi(\varepsilon)$ converges as ε tends to zero. Formally let

$$\lim_{\varepsilon \to 0} \phi(\varepsilon) = \sigma$$

for some $\sigma \in \mathbb{R}$, the *index of assortativity* of the matching process. By defining $\phi(0)$ as σ we thus extend the domain of ϕ from $(0,1)$ to $[0,1)$, and it follows from (6) that $\sigma \in [0,1]$.[10] Under uniform random matching, $\phi(\varepsilon) = 0$ for all $\varepsilon \in (0,1)$ and hence $\sigma = 0$. In pairwise interactions between siblings, $\phi(\varepsilon) = 1/2$ for all $\varepsilon \in (0,1)$ and hence $\sigma = 1/2$, see Section 5.

2.2 Homo moralis

Definition 4 *An individual is a* ***homo moralis*** *(or* ***HM****) if her utility function is of the form*

$$u_\kappa(x,y) = (1-\kappa) \cdot \pi(x,y) + \kappa \cdot \pi(x,x), \tag{7}$$

for some $\kappa \in [0,1]$*, her degree of morality.*[11]

It is as if *homo moralis* is torn between selfishness and morality. On the one hand, she would like to maximize her own payoff. On the other hand, she would like to "do the right thing," i.e., choose a strategy that, if used by all individuals, would lead to the highest possible payoff to all. This second goal can be viewed as an application of Kant's (1785) categorical imperative, to "act only on the maxim that you would at the same time will to be a universal law".[12] Torn between these two goals, *homo moralis* chooses a strategy that maximizes a convex combination of them. If $\kappa = 0$, the definition of *homo moralis* coincides with that of "pure selfishness," or *homo oeconomicus*; given any strategy y used by the other party, she will use a strategy in $\arg\max_{x \in X} \pi(x,y)$. At the opposite extreme, $\kappa = 1$, the definition of *homo moralis* coincides with that of "pure morality," or *homo kantiensis*; irrespective of what strategy the other party uses (or is expected to use), this extreme variety of *homo moralis* will use a strategy in $\arg\max_{x \in X} \pi(x,x)$.[13]

A special variety of *homo moralis* turns out to be important from an evolutionary point of view, namely *homo moralis* with degree of morality equal to the index of assortativity,

[10] This contrasts with the case of a finite population, where negative assortativity can arise for population states with few mutants (see Schaffer, 1988).

[11] We thus adopt the notational convention that types θ that are real numbers in the unit interval refer to *homo moralis* with that degree of morality.

[12] See Binmore (1994) for a critical discussion of Kant's categorical imperative.

[13] In his work on strategy evolution among siblings, Bergstrom (1995) finds that the selected strategy must be a Nash equilibrium strategy of a game in which both players have what he calls *semi-Kantian* preferences. Such preferences correspond to *homo moralis* with degree of morality $\kappa = 1/2$.

$\kappa = \sigma$:

$$u_\sigma(x,y) = (1-\sigma)\cdot\pi(x,y) + \sigma\cdot\pi(x,x). \tag{8}$$

We call this variety *homo hamiltonensis*. This terminology is a homage to the late biologist William Hamilton, who suggested that in interactions between genetically related individuals the concept of fitness should be augmented to what he called *inclusive fitness* since genes that drive the behavior of one individual are present also in the relative with some genetically determined probability (Hamilton, 1964 a,b). In interactions between individuals with genetic degree of relatedness σ (Wright, 1922), $u_\sigma(x,y)$ is the average inclusive fitness of mutants in an infinitesimally small mutant subpopulation playing x in a resident population playing y. For recent analyses of various aspects of inclusive fitness, see Rousset (2004) and Grafen (2006).

It is worth noting that the preferences of *homo moralis* differ sharply from any preferences in which the domain is the payoff distribution, such as altruism, inequity aversion, or a concern for efficiency. To see this, consider an individual who chooses a strategy in $\arg\max_{x\in X} W[\pi(x,y),\pi(y,x)]$ for some increasing (welfare) function W. This is a set that in general depends on the other party's (expected) strategy y, while *homo kantiensis* chooses a strategy in $\arg\max_{x\in X}\pi(x,x)$, a set that does not depend on the other party's strategy (see Section 6.2 for a more detailed comparison with altruism). The theory developed here also differs from models in the literature on psychological games, see, e.g., Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006).

3 Analysis

We begin with some general observations and then proceed to our main result. First, let $B^{NE}(s) \subseteq X^2$ denote the set of (Bayesian) Nash equilibria in population state $s = (\theta,\tau,\varepsilon)$, that is, all solutions (x^*,y^*) of (3). For given types θ and τ, this defines an equilibrium correspondence $B^{NE}(\theta,\tau,\cdot) : (0,1) \rightrightarrows X^2$ that maps mutant population shares ε to the associated set of equilibria. As noted above, all probabilities in (3) may be expressed in terms of the continuous assortment function ϕ, the domain of which we extended to $[0,1)$. This allows us to likewise extend the domain of $B^{NE}(\theta,\tau,\cdot)$. One may show the following by standard arguments (see Appendix for a proof):

Lemma 1 *$B^{NE}(\theta,\tau,\varepsilon)$ is compact for each $(\theta,\tau,\varepsilon)\in\Theta^2\times[0,1)$. $B^{NE}(\theta,\tau,\varepsilon)\neq\varnothing$ if u_θ and u_τ are concave in their first arguments. The correspondence $B^{NE}(\theta,\tau,\cdot):[0,1)\rightrightarrows X^2$ is upper hemi-continuous.*

Second, for each type $\theta\in\Theta$ let $\beta_\theta : X \rightrightarrows X$ denote the the best-reply correspondence,

$$\beta_\theta(y) = \arg\max_{x\in X} u_\theta(x,y) \quad \forall y\in X,$$

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

$$X_\theta = \{x \in X : x \in \beta_\theta(x)\}. \tag{9}$$

In particular, X_σ is the fixed-point set for *homo hamiltonensis*, the *Hamiltonian strategies*.

For any type $\theta \in \Theta$, let Θ_θ be the set of types τ that, as vanishingly rare mutants among residents of type θ, are behaviorally indistinguishable from the residents:

$$\Theta_\theta = \left\{\tau \in \Theta : \exists\ x \in X_\theta \text{ such that } (x,x) \in B^{NE}(\theta,\tau,0)\right\}. \tag{10}$$

Examples of such "behavioral alikes" are individuals with utility functions that are positive affine transformations of the utility function of the residents, and also individuals for whom some strategy in X_θ is dominant.[14]

Finally, the type set Θ will be said to be *rich* if for each strategy $x \in X$ there exists some type $\theta \in \Theta$ for which this strategy is strictly dominant: $u_\theta(x,y) > u_\theta(x',y)\ \ \forall x' \neq x$, $\forall y \in X$. Such a type θ will be said to be *committed* to its strategy x.

We are now in a position to state our main result:

Theorem 1 *If $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, then* homo hamiltonensis *is evolutionarily stable against all types $\tau \notin \Theta_\sigma$. If Θ is rich, $X_\theta \cap X_\sigma = \varnothing$ and X_θ is a singleton, then θ is evolutionarily unstable.*

Proof: Given any population state $s = (\theta,\tau,\varepsilon)$, the definitions (1) and (2) of the associated average payoff functions Π_θ and Π_τ may be re-written in terms of the assortment function ϕ as

$$\Pi_\theta(x,y,\varepsilon) = [1-\varepsilon+\varepsilon\phi(\varepsilon)]\cdot\pi(x,x) + \varepsilon[1-\phi(\varepsilon)]\cdot\pi(x,y) \tag{11}$$

and

$$\Pi_\tau(x,y,\varepsilon) = (1-\varepsilon)[1-\phi(\varepsilon)]\cdot\pi(y,x) + [\varepsilon+(1-\varepsilon)\phi(\varepsilon)]\cdot\pi(y,y). \tag{12}$$

Since π and ϕ are continuous by hypothesis, so are $\Pi_\theta, \Pi_\tau : X^2 \times [0,1) \to \mathbb{R}$.

For the first claim, let $\theta = \sigma$ (*homo moralis* of degree of morality σ) and suppose that $(x^*,y^*) \in B^{NE}(\sigma,\tau,0)$. Then

$$\begin{cases} x^* \in \arg\max_{x\in X} & u_\sigma(x,x^*) \\ y^* \in \arg\max_{y\in X} & (1-\sigma)\cdot u_\tau(y,x^*) + \sigma\cdot u_\tau(y,y^*). \end{cases} \tag{13}$$

Thus $x^* \in X_\sigma$ and $u_\sigma(x^*,x^*) \geq u_\sigma(y^*,x^*)$. Moreover, if $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, then the latter inequality holds strictly if $\tau \notin \Theta_\sigma$: $u_\sigma(x^*,x^*) > u_\sigma(y^*,x^*)$, or, equivalently, $\pi(x^*,x^*) > (1-\sigma)\cdot\pi(y^*,x^*) + \sigma\cdot\pi(y^*,y^*)$. By definition of Π_σ and Π_τ, we thus have

$$\Pi_\sigma(x^*,y^*,0) > \Pi_\tau(x^*,y^*,0) \tag{14}$$

[14] For example, $u_\tau(x,y) = -(x-x_\theta)^2$ for all $x,y \in X$ and some $x_\theta \in X_\theta$.

for all $(x^*, y^*) \in B^{NE}(\sigma, \tau, 0)$ and any $\tau \notin \Theta_\sigma$. By continuity of Π_σ and Π_τ, this strict inequality holds for all (x, y, ε) in a neighborhood $U \subset X^2 \times [0,1)$ of $(x^*, y^*, 0)$. Now $B^{NE}(\theta, \tau, \cdot) : [0,1) \rightrightarrows X^2$ is closed-valued and upper hemi-continuous (Lemma 1). Hence, if $(x_t, y_t) \in B^{NE}(\theta, \tau, \varepsilon_t)$ for all $t \in \mathbb{N}$, $\varepsilon_t \to 0$ and $\langle (x_t, y_t) \rangle_{t \in \mathbb{N}}$ converges, then the limit point (x^0, y^0) necessarily belongs to $B^{NE}(\theta, \tau, 0)$. Thus, for any given $\bar{\varepsilon} > 0$ there exists a T such that for all $t > T$: $0 < \varepsilon_t < \bar{\varepsilon}$ and $(x_t, y_t) \in U$, and thus $\Pi_\sigma(x_t, y_t, \varepsilon_t) > \Pi_\tau(x_t, y_t, \varepsilon_t)$, establishing the first claim.[15]

For the second claim, let $\theta \in \Theta$ be such that $X_\theta = \{x_\theta\}$ and $x_\theta \notin X_\sigma$. Then $u_\sigma(x_\theta, x_\theta) < u_\sigma(\hat{x}, x_\theta)$ for some $\hat{x} \in X$. If Θ is rich, there exists a type $\hat{\tau} \in \Theta$ committed to $\hat{x}$. Since $\hat{x}$ is dominant for $\hat{\tau}$, individuals of this type will always play $\hat{x}$. Consequently, for any $\varepsilon \in [0,1)$, $(x^*, y^*) \in B^{NE}(\theta, \hat{\tau}, \varepsilon)$ iff $y^* = \hat{x}$ and

$$x^* \in \arg\max_{x \in X} \quad [1 - \varepsilon + \varepsilon\phi(\varepsilon)] u_\theta(x, x^*) + \varepsilon [1 - \phi(\varepsilon)] u_\theta(x, \hat{x}).$$

In particular, $B^{NE}(\theta, \hat{\tau}, 0) = \{(x_\theta, \hat{x})\}$ since x_θ is the unique solution to the first condition in (13). Moreover, $u_\sigma(x_\theta, x_\theta) < u_\sigma(\hat{x}, x_\theta)$ is equivalent with

$$\pi(x_\theta, x_\theta) < (1 - \sigma) \cdot \pi(\hat{x}, x_\theta) + \sigma \cdot \pi(\hat{x}, \hat{x})$$

which in turn is equivalent with $\Pi_\theta(x_\theta, \hat{x}, 0) < \Pi_{\hat{\tau}}(x_\theta, \hat{x}, 0)$. By continuity of Π_θ and $\Pi_{\hat{\tau}}$, this strict inequality holds for all $(x, \hat{x}, \varepsilon)$ in a neighborhood $U \subset X^2 \times [0,1)$ of $(x_\theta, \hat{x}, 0)$. Now $B^{NE}(\theta, \hat{\tau}, \cdot) : [0,1) \rightrightarrows X^2$ is closed-valued and upper hemi-continuous. Therefore, if $(x_t, y_t, t) \in B^{NE}(\theta, \hat{\tau}, \varepsilon_t)$ for all $t \in \mathbb{N}$, $\varepsilon_t \to 0$ and $\langle (x_t, y_t) \rangle_{t \in \mathbb{N}}$ converges, then the limit point (x^*, y^*) necessarily belongs to $B^{NE}(\theta, \hat{\tau}, 0)$, which, in the present case is a singleton, so $(x^*, y^*) = (x_\theta, \hat{x})$. Moreover, $y_t = \hat{x}$ for all t. Thus, for any given $\bar{\varepsilon} > 0$ there exists a T such that for all $t > T$: $0 < \varepsilon_t < \bar{\varepsilon}$ and $(x_t, \hat{x}) \in U$, and thus $\Pi_\theta(x_t, \hat{x}, \varepsilon_t) < \Pi_{\hat{\tau}}(x_t, \hat{x}, \varepsilon_t)$, establishing the second claim. **Q.E.D.**

Theorem 1 establishes that *homo hamiltonensis* is favored by evolution and that certain other types are selected against. The first claim is that *homo hamiltonensis* resists "invasions" by all types that do not, as mutants, respond by playing *homo hamiltonensis'* own strategy. The intuition is that the unique "evolutionarily optimal" mutant response—in terms of the mutant population's average payoff—to a resident Hamiltonian strategy is that very same strategy. In this sense, *homo hamiltonensis* preempts mutants. The second claim is that if the type set is rich, then any type that has a unique resident strategy is vulnerable to invasion if its resident strategy is non-Hamiltonian. The uniqueness hypothesis is made for technical reasons, and it seems that it could be relaxed somewhat, but at a high price in

[15] Under the hypothesis of the theorem, it is not excluded that $B^{NE}(\sigma, \tau, 0) = \varnothing$. By upper hemi-continuity of $B^{NE}(\sigma, \tau, \cdot) : [0,1) \rightrightarrows X^2$, there then exists an $\bar{\varepsilon} > 0$ such that $B^{NE}(\sigma, \tau, \varepsilon) = \varnothing \ \forall \varepsilon \in (0, \bar{\varepsilon})$. By definition, θ is evolutionarily stable against τ in this case as well.

analytical complexity.[16] However, the intuition is clear: since the resident type does not play a Hamiltonian strategy, there exists a better reply to it in terms of *homo hamiltonensis'* preferences. Because of the nature of those preferences, such a better reply, if used by a mutant, results in higher payoff to the mutants than to the residents. Since the type space is rich, there is a mutant type that is committed to such an evolutionarily superior strategy and that will thus use it against any resident, who will then lose out in terms of payoffs.

It follows immediately from the second claim in Theorem 1 that a necessary condition for evolutionary stability of any type with a unique resident strategy is to behave like *homo hamiltonensis*:

Corollary 1 *If Θ is rich, $\theta \in \Theta$ is evolutionarily stable against all types $\tau \notin \Theta_\theta$, and $X_\theta = \{x_\theta\}$, then $x_\theta \in X_\sigma$.*

Example 1 *As an illustration of Theorem 1, consider a canonical public-goods situation. Let $\pi(x,y) = B(x+y) - C(x)$ for $B, C : [0,m] \to \mathbb{R}$ twice differentiable with $B', C', C'' > 0$ and $B'' < 0$ and $m > 0$ such that $C'(0) < B'(0)$ and $C'(m) > 2B'(2m)$. Here $B(x+y)$ is the public benefit and $C(x)$ the private cost from one's own contribution x when the other individual contributes y. Played by two* homo moralis *with degree of morality $\kappa \in [0,1]$, this interaction defines a game with a unique Nash equilibrium, and this is symmetric. The equilibrium contribution, x_κ, is the unique solution in $(0,m)$ to the first-order condition $C'(x) = (1+\kappa) B'(2x)$. Hence, $X_\kappa = \{x_\kappa\}$. We note that* homo moralis' *contribution increases from that of the selfish* homo oeconomicus *when $\kappa = 0$ to that of a benevolent social planner when $\kappa = 1$. Moreover, it is easily verified that $\beta_\kappa(y)$ is a singleton for all $y \in [0,m]$. Theorem 1 establishes that* homo hamiltonensis, *that is* homo moralis *with degree of morality $\kappa = \sigma$, is evolutionarily stable against all types that, as vanishingly rare mutants, would contribute $y \neq x_\sigma$. Moreover, if Θ is rich, and $\theta \in \Theta$ is any type that has a unique resident strategy and this strategy differs from x_σ, then θ is evolutionarily unstable.*

Note that the hypothesis in Theorem 1 that $\beta_\sigma(x)$ is a singleton can be met even by mixed strategies x when $\sigma > 0$ since the preferences of *homo hamiltonensis* are then quadratic in his or her own randomization (see section 4).

3.1 Homo oeconomicus

Theorem 1 may be used to pin down the evolutionary stability properties of *homo oeconomicus*. We immediately obtain:

[16] For a type θ that does not have a unique resident strategy, a mutant's payoff advantage when $\varepsilon = 0$ need no longer remain when $\varepsilon > 0$. However, if the correspondence is lower hemi-continuous at $\varepsilon = 0$, then we conjecture that the present proof, *mutatis mutandis*, will hold.

Corollary 2 *If $\sigma = 0$ and $\beta_0(x)$ is a singleton for all $x \in X_0$, then* homo oeconomicus *is evolutionarily stable against all types $\tau \notin \Theta_0$. If $\sigma > 0$ and Θ is rich, then* homo oeconomicus *is evolutionarily unstable if it has a unique resident strategy and this does not belong to X_σ.*

The first part of this result says that a sufficient condition for *homo oeconomicus* to be evolutionarily stable against mutants who play other strategies than those played by *homo oeconomicus* is that the index of assortativity be zero, granted *homo oeconomicus* has a unique best reply to all of its resident strategies. This result is in line with Ok and Vega-Redondo (2001) and Dekel *et al.* (2007), both of which analyze the evolution of preferences under incomplete information and uniform random matching. The second part of the corollary implies that $\sigma = 0$ is also necessary for *homo oeconomicus* to be evolutionarily stable when it has a unique resident strategy, this is not Hamiltonian and the type set is rich.

To shed more light on the stability/instability of *homo oeconomicus*, we distinguish two classes of interactions, those that essentially are decision problems and those that are truly strategic. First, consider fitness games $\langle X, \pi \rangle$ where a player's payoff does not depend on the other's strategy. For each individual in a population it is then immaterial what other individuals do, so "the right thing to do," irrespective of the index of assortativity, is simply to choose a strategy that maximizes one's own payoff. As a result, *homo oeconomicus* can thrive in such interactions even if the index of assortativity is positive, $\sigma > 0$.

Corollary 3 *Suppose that $\beta_0(x)$ is a singleton for all $x \in X_0$. If $\pi(x, y) = \pi(x, y')$ for all $x, y, y' \in X$, then* homo oeconomicus *is evolutionarily stable against all types $\tau \notin \Theta_0$ for all $\sigma \in [0, 1]$.*

In fact, in such interactions *homo moralis* with *any* degree of morality $\kappa \in [0, 1]$ is evolutionarily stable against all types who fail to maximize their own payoff.

Secondly, consider fitness games $\langle X, \pi \rangle$ that are truly strategic in the sense that a player's strategy does depend on the other player's strategy. In order to simplify the reasoning, we assume that X is one-dimensional and that π is twice differentiable, with strictly decreasing returns to the player's own strategy. Then the behavior of *homo oeconomicus* differs from that of *homo moralis* with any positive degree of morality. As a result, *homo oeconomicus* is in dire straits when the index of assortativity is positive. Letting subscripts denote partial derivatives:

Corollary 4 *Suppose that X_0 is a singleton, $\pi_{11}(x, y) < 0$ and $\pi_2(x, y) \neq 0$ for all $x, y \in X$. If Θ is rich and* homo oeconomicus *is evolutionarily stable against some type $\tau \notin \Theta_0$, then $\sigma = 0$.*

3.2 Strategy evolution

Our model differs from classic evolutionary game theory in two ways. First, that theory views strategies, not preferences or utility functions, as the replicators, the objects that are subject to evolutionary forces. Second, the background hypothesis in the standard set-up is that matching is uniform. To assume that strategies are the replicators can be formulated within the present framework as the assumption that each type is committed to some strategy and that the type set is rich. In such situations one may identify each type with a strategy and *vice versa*, and hence write $\Theta = X$.

Identifying types with strategies, our general definition of evolutionary stability applies. For any pair of strategies $x, y \in X$, hence types, and any $\varepsilon \in [0, 1)$, the average payoffs are as in equations (11) and (12), with θ being the type committed to x and τ the type committed to y. The difference between these two average payoffs, $S_{x,y}(\varepsilon) \equiv \Pi_\theta(x, y, \varepsilon) - \Pi_\tau(x, y, \varepsilon)$, is a generalization of what in standard evolutionary game theory is called the *score function* of strategy x against strategy y.[17] Applied to the present setting of strategy evolution, the stability definition in Section 2 boils down to:

Definition 5 *Let $\Theta = X$ (strategy evolution) and let matching be random with assortment function ϕ. A strategy $x \in X$ is **evolutionarily stable** if for every strategy $y \neq x$ there exists an $\bar{\varepsilon}_y \in (0, 1)$ such that $S_{x,y}(\varepsilon) > 0$ for all $\varepsilon \in (0, \bar{\varepsilon}_y)$.*

We immediately obtain from Theorem 1 (the proof follows from standard arguments and is hence omitted):[18]

Corollary 5 *Let $\Theta = X$ (strategy evolution). Every strategy $x_\sigma \in X_\sigma$ for which $\beta_\sigma(x_\sigma)$ is a singleton is evolutionarily stable. Every strategy $x \notin X_\sigma$ is evolutionarily unstable.*

In other words, every Hamiltonian strategy which is its own unique best reply is evolutionarily stable, and all non-Hamiltonian strategies are evolutionarily unstable. For certain payoff functions π, the Hamiltonian best-reply correspondence is not singleton-valued. The following characterization is a generalization of Maynard Smith's and Price's (1973) original definition and does not require singleton-valuedness. The hypothesis is instead that the degree of assortment is independent of the mutant population share, an independence property that holds in certain kinship relations, see Section 5.

---

[17] In the standard theory (Bomze and Pötscher, 1989, and Weibull, 1995), $\phi \equiv 0$, so that $S_{x,y}(\varepsilon) = (1-\varepsilon)\pi(x, x) + \varepsilon\pi(x, y) - \varepsilon\pi(y, y) - (1-\varepsilon)\pi(y, x)$.

[18] Note, however, that here *homo hamiltonensis* is not included in the type space. *Homo hamiltonensis* is instead represented by a set of committed types, one for each Hamiltonian strategy.

Proposition 1 *Let $\Theta = X$ (strategy evolution) and assume that $\phi(\varepsilon) \equiv \sigma$. A strategy $x \in X$ is evolutionarily stable if and only if (15) and (16) hold:*

$$\pi(x,x) \geq \pi(y,x) + \sigma \cdot [\pi(y,y) - \pi(y,x)] \quad \forall y \in X \tag{15}$$

$$\begin{aligned} \pi(x,x) &= \pi(y,x) + \sigma \cdot [\pi(y,y) - \pi(y,x)] \\ &\Rightarrow \pi(x,y) > \pi(y,y) + \sigma \cdot [\pi(y,y) - \pi(y,x)]. \end{aligned} \tag{16}$$

(This follows from standard arguments, so no proof is given here.)

The necessary condition (15) can be written as $x \in X_\sigma$, that is, the strategy must be Hamiltonian. Further, condition (16) may be written

$$\begin{aligned} \pi(x,x) &= \pi(y,x) + \sigma \cdot [\pi(y,y) - \pi(y,x)] \\ &\Rightarrow \pi(x,y) + \pi(y,x) - \pi(x,x) - \pi(y,y) > 0, \end{aligned}$$

a formulation that is consistent with the analysis in Hines and Maynard Smith (1979) of ESS in fitness games played by relatives; see also Grafen (1979, 2006) and Bergstrom (1995).

Remark 1 *In fitness games $< X, \pi >$ where* homo hamiltonensis *has a unique best reply to each Hamiltonian strategy, Theorem 1 and Corollary 5 establish that preference evolution under incomplete information induces the same behaviors as strategy evolution. Hence, evolutionarily stable strategies also emerge from preference evolution when individuals are not programmed to strategies but are rational and play equilibria under incomplete information.*

4 Finite games

The classic domain for evolutionary stability analyses is mixed strategies in finite and symmetric two-player games, a domain to which we now apply the above machinery. Let A be an $m \times m$ matrix that to each row $i \in I$ and column $j \in I$ assigns the payoff a_{ij} obtained when pure strategy i is used against pure strategy j, for all $i, j \in I = \{1, .., m\}$. When players are permitted to use mixed strategies, X is the $(m-1)$-dimensional unit simplex $\Delta(I) = \left\{x \in \mathbb{R}^m_+ : \sum_{i \in I} x_i = 1\right\}$, a compact and convex set in $\mathbb{R}^m$. The continuous, in fact bilinear, function $\pi : X^2 \to \mathbb{R}$ assigns the expected payoff, $\pi(x,y) = x \cdot Ay$ to each strategy $x \in X = \Delta(I)$ when used against any strategy $y \in X = \Delta(I)$.

Applying our general machinery for preference evolution under incomplete information to finite games, for each type $\theta \in \Theta$ let $u_\theta : X^2 \to \mathbb{R}$ be some continuous function where $X = \Delta(I)$. In particular, the utility function of *homo moralis*, of arbitrary degree of morality

$\kappa \in [0,1]$, is quadratic in the individual's own strategy, x, and linear in the other individual's strategy y:[19]

$$u_\kappa(x,y) = (1-\kappa)\cdot xAy + \kappa \cdot xAx = xA\left[(1-\kappa)\,y + \kappa x\right]. \tag{17}$$

A general stability analysis falls outside the scope of this study, so we here focus on the more restricted task of identifying the set of *homo-moralis* strategies in 2×2 fitness games. For this purpose, it is convenient to use the notation $x, y \in [0,1]$ for the probabilities attached to the first pure strategy. For each $\kappa \in [0,1]$, the associated set $X_\kappa \subseteq X = [0,1]$ of *homo-moralis* strategies is then the solution set to the following fixed-point condition:

$$x_\kappa \in \arg\max_{x\in[0,1]} \quad (x, 1-x)\cdot \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_\kappa + \kappa\,(x - x_\kappa) \\ 1 - x_\kappa - \kappa\,(x - x_\kappa) \end{pmatrix}. \tag{18}$$

Depending on whether the sum of the diagonal elements of A exceeds, equals or falls short of the sum of its off-diagonal elements, the utility of *homo moralis* is either strictly convex, linear, or strictly concave in his/her own strategy, so that the following result obtains:

Proposition 2 *Let*

$$\hat{x}(\kappa) = \min\left\{1, \frac{a_{12} + \kappa a_{21} - (1+\kappa)\,a_{22}}{(1+\kappa)\,(a_{12} + a_{21} - a_{11} - a_{22})}\right\}. \tag{19}$$

(a) If $\kappa > 0$ and $a_{11} + a_{22} > a_{12} + a_{21}$, then $X_\kappa \subseteq \{0,1\}$.

(b) If $\kappa = 0$ and/or $a_{11} + a_{22} = a_{12} + a_{21}$, then

$$X_\kappa = \begin{cases} \{0\} & \text{if } a_{12} + \kappa a_{21} < (1+\kappa)\,a_{22} \\ [0,1] & \text{if } a_{12} + \kappa a_{21} = (1+\kappa)\,a_{22} \\ \{1\} & \text{if } a_{12} + \kappa a_{21} > (1+\kappa)\,a_{22} \end{cases}$$

(c) If $\kappa > 0$ and $a_{11} + a_{22} < a_{12} + a_{21}$, then

$$X_\kappa = \begin{cases} \{0\} & \text{if } a_{12} + \kappa a_{21} \leq (1+\kappa)\,a_{22} \\ \{\hat{x}(\kappa)\} & \text{if } a_{12} + \kappa a_{21} > (1+\kappa)\,a_{22} \end{cases}$$

Proof: The maximand in (18) can be written as

$$\begin{aligned} &\kappa\,(a_{11} + a_{22} - a_{12} - a_{21})\cdot x^2 + (1-\kappa)\,(a_{11} + a_{22} - a_{12} - a_{21})\,x_\kappa \cdot x \\ &+ \left[a_{12} + \kappa a_{21} - (1+\kappa)\,a_{22}\right]\cdot x + (1-\kappa)\cdot(a_{21} - a_{22})\,x_\kappa + a_{22}. \end{aligned}$$

For $\kappa\,(a_{11} + a_{22} - a_{12} - a_{21}) > 0$, this is a strictly convex function of x, and hence the maximum is achieved on the boundary of $X = [0,1]$. This proves claim (a).

[19] In particular, mixed strategies may have unique best replies. For instance, if $\sigma \in (0,1]$ and $a_{22} - a_{12} < a_{21} - a_{11}$, then u_κ is strictly concave in x, for each $y \in [0,1]$.

For $\kappa(a_{11}+a_{22}-a_{12}-a_{21})=0$, the maximand is affine in x, with slope $a_{12}+\kappa a_{21}-(1+\kappa)a_{22}$. This proves (b).

For $\kappa(a_{11}+a_{22}-a_{12}-a_{21})<0$, the maximand is a strictly concave function of x, with unique global maximum (in $\mathbb{R}$) at

$$\tilde{x}=\frac{a_{12}+\kappa a_{21}-(1+\kappa)a_{22}}{(1+\kappa)(a_{12}+a_{21}-a_{11}-a_{22})}.$$

Hence, $X_\kappa=\{0\}$ if $\tilde{x}\leq 0$, $X_\kappa=\{\tilde{x}\}$ if $\tilde{x}\in[0,1]$, and $X_\kappa=\{1\}$ if $\tilde{x}>1$, which proves (c). **Q.E.D.**

As an illustration, we identify the set X_κ of *homo-moralis* strategies, for each $\kappa\in[0,1]$, in a one-shot prisoners' dilemma with payoff matrix

$$A=\begin{pmatrix} R & S \\ T & P \end{pmatrix} \tag{20}$$

where $T-R>P-S>0$. Case (c) of Proposition 2 then applies for all $\kappa>0$, and an interior solution, $\hat{x}(\kappa)\in(0,1)$, obtains for intermediate values of κ. More precisely, $X_\kappa=\{0\}$ for all $\kappa\leq(P-S)/(T-P)$, $X_\kappa=\{1\}$ for all $\kappa\geq(T-R)/(R-S)$, and $X_\kappa=\{\hat{x}(\kappa)\}$ for all κ between these two bounds. See Figure 1 below, which shows how cooperation increases as the degree of morality increases. In this example, the hypothesis in Corollary 2 is not satisfied and *homo oeconomicus* is *not* evolutionarily unstable for small $\sigma>0$. The reason is that although the behavior of *homo oeconomicus* is uniquely determined (namely to defect), it coincides with that of *homo hamiltonensis* for all $\sigma<1/4$.

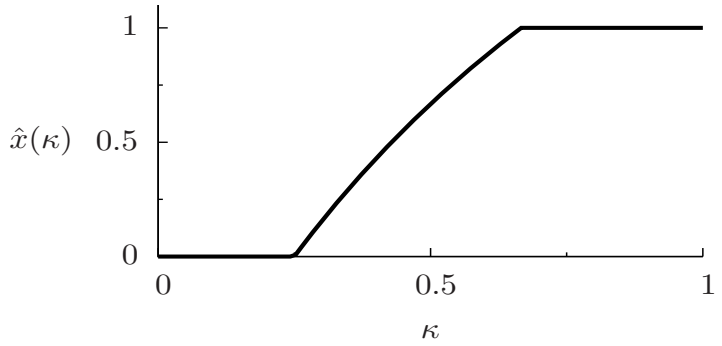

Figure 1: The (singleton) set of homo-moralis strategies for $(T,R,P,S)=(7,5,3,2)$.

As a second illustration, consider the hawk-dove game, the original example used by Maynard Smith and Price (1973) when they introduced the notion of an ESS:

$$A=\begin{pmatrix} (v-c)/2 & v \\ 0 & v/2 \end{pmatrix}$$

for $0 < v < c$ (see also Grafen, 1979). This game has a unique ESS, namely to use the first pure strategy ("hawk") with probability $x^* = v/c$. It is easy to verify that $X_\kappa = \{x_\kappa\}$, where

$$x_\kappa = \frac{1-\kappa}{1+\kappa} \cdot \frac{v}{c}.$$

The probability for the aggressive and wasteful hawk strategy thus strictly decreases in κ, from its "classic" value, v/c, when $\kappa = 0$ to zero when $\kappa = 1$.

5 Matching processes

The analysis above shows that the process whereby individuals are matched affects the stability condition for preferences, even though the latter are unobservable. This generalizes the literature on preference evolution, where the underlying assumption has been that the matching process is uniformly random.[20] Arguably, uniform random matching is unrealistic for most human interactions since it requires that there be zero correlation between the contact pattern that determines how mutations spread in society and the contact pattern that determines who interacts with whom. There are, however, many natural sources for a positive correlation between such patterns. Below we provide examples and present a simple model in which assortativity is positive in the limit as the population share of mutants tends to zero.

5.1 Kin

Our first example is inspired by the biology literature, which, as mentioned in the introduction, has devoted considerable attention to the evolution of behaviors among genetically related individuals. For interactions among kin it is straightforward to determine the assortment function ϕ and to show that it takes on a constant and positive value for all mutant population shares ε. Furthermore, it will become clear that these arguments apply equally well to traits that are culturally, rather than genetically, transmitted from parents to children.

While the following arguments can readily be adapted to interactions between other kin, we focus here on siblings. Suppose that each child inherits his or her preferences or moral values from one of her parents. The number of mutants in a pair of siblings is then a random variable, Z, that takes values in $\{0, 1, 2\}$, and whose probability distribution depends on the parents' types. Let θ be the resident preference "type" in the parent generation, represented in population share $1-\varepsilon$, and let the small residual population share ε consist of parents of some other type τ. Suppose that a given child inherits each parent's preferences

[20] Exceptions are Alger (2010) and Alger and Weibull (2010, 2012a).

with probability one half, that these random draws are statistically independent, and that parents are monogamous. The inheritance mechanism could be genetic[21] or cultural.[22] Then, the probability distribution for Z in a family where exactly one of the parents is of type τ, is $(1/4, 1/2, 1/4)$, while in a family where both parents are of type τ, the probability distribution for Z is $(0, 0, 1)$. Among families with at least one parent of type τ, the fraction $2\varepsilon(1-\varepsilon)/[2\varepsilon(1-\varepsilon)+\varepsilon^2]$ are families with exactly one parent of type τ, and the remaining fraction, $\varepsilon^2/[2\varepsilon(1-\varepsilon)+\varepsilon^2]$, consists of families where both parents are of type τ. Since on average half of the children in the first kind of family have preferences τ, the probability that a child with the mutant preference τ has two mutant parents is

$$\frac{\varepsilon^2}{2\varepsilon(1-\varepsilon)/2+\varepsilon^2}=\varepsilon.$$

Hence, $\Pr[\tau|\tau,\varepsilon]=(1-\varepsilon)\cdot 1/2+\varepsilon\cdot 1=(1+\varepsilon)/2$. Likewise, $\Pr[\theta|\theta,\varepsilon]=(1-\varepsilon)\cdot 1+\varepsilon\cdot 1/2=1-\varepsilon/2$. The function ϕ thus takes on the constant value one half, $\phi(\varepsilon)=1/2$ for all $\varepsilon\in(0,1)$, so this is the index of assortativity. Moreover, one half is the *coefficient of relatedness* between siblings (Wright, 1922).[23]

5.2 Geography, homophily and business partnerships

In our daily lives, we tend to interact more with those who live and work close to us than with people at distant locations. More generally, we tend to interact more with those who are similar to ourselves along one or more dimensions—such as language, culture, profession, religion, dress, origin. This tendency, called *homophily*, has been extensively documented by sociologists, see, e.g., McPherson, Smith-Lovin, and Cook (2001), and Ruef, Aldrich, and Carter (2003). In the economics literature, homophily has been analyzed by Currarini, Jackson, and Pin (2009, 2010) and Bramoullé and Rogers (2009). These latter studies concern race and gender-based choice of friends and meeting chances on the basis of data from U.S. high schools. Currarini, Jackson, and Pin (2009) find strong within-group biases not only in the inferred utility from meetings but also in meeting probabilities. We proceed

[21] In biological terms, we here focus on sexual reproduction in a haploid species. Thus, each child has two genetic parents, and each parent carries one set of chromosomes, and this determines heredity. Humans are a diploid species, with two sets of chromosomes, which complicates matters because of the distinction between recessive and dominant genes. For calculations of assortativeness in diploid species, see Bergstrom (1995, 2003). See further Michod and Hamilton (1980).

[22] While Bisin and Verdier (2001) focus on the case of one parent per child and assume that each parent is altruistic towards its child and makes an effort to transmit its cultural values to its child, here the probability of transmission of parents' cultural values to their children is exogenous.

[23] If the two children would have the same mother but different fathers (or the same father but different mothers), one would instead obtain $\sigma=1/4$, the coefficient of relatedness between half-siblings.

to show, by means of a simple model, how and when homophily may give rise to assortativity in the matching process, even though preferences are not observable.

Consider a finite population divided into groups of equal size $n > 1$. These groups may be defined by distinct geographical locations, languages or dialects, professions, cultures or religions etc. Initially, all individuals in the population have the same preferences, or moral values, $\theta \in \Theta$. Suddenly one group is hit by some shock, with equal probability for each group to be hit. The effect on the group in question is that a random number, Z, of its n members change their preferences or moral values from θ to some $\tau \in \Theta$, where $\mathbb{E}[Z] = \mu \in (1, n)$. If the total population size is N, the expected population share of mutants is then $\varepsilon = \mu/N$. After the population shock, pairs are randomly matched to play a symmetric fitness game as above. The matching is done as follows. First one individual is uniformly drawn from the whole population (as if looking around for a partner). With probability $p(n, N) \in [0, 1]$ the other individual in the match is uniformly drawn from the same group as the first. With the complementary probability the other individual is instead uniformly randomly drawn from the rest of the population. In both cases, the matching probabilities are "blind" to individuals' types. However, in general the population share of mutants among the matches drawn for a mutant will not be ε, not even on average. For if an individual is of the mutant type τ, then her group must be the one where the shock occurred and hence

$$\Pr[\tau|\tau, \varepsilon] = p(n, N) \cdot \mathbb{E}\left[\frac{Z-1}{n-1}\right] = p(n, N) \cdot \frac{\mu - 1}{n-1}.$$

By definition, $\phi(\varepsilon) = \Pr[\theta|\theta, \varepsilon] - 1 + \Pr[\tau|\tau, \varepsilon]$. Letting $N \to \infty$ while keeping the group size n and the mean value μ fixed: $\varepsilon = \mu/N \to 0$, $\Pr[\theta|\theta, \varepsilon] \to 1$ and

$$\sigma = \lim_{\varepsilon \to 0} \phi(\varepsilon) = \lim_{\varepsilon \to 0} \Pr[\tau|\tau, \varepsilon] = \frac{\mu - 1}{n-1} \cdot p^*(n), \tag{21}$$

where $p^*(n) = \lim_{N \to \infty} p(n, N)$. Hence, the index of assortativity is positive whenever this limit probability of intragroup matching remains positive as the number of groups tends to infinity.[24] See Rousset (2004) and Lehmann and Rousset (2010) for richer models of population structure and assortativity.

Arguably, this simple model can shed some light on the role of homophily for assortativity in matching. For it appears that "new" preferences or moral values usually arise within a single group and then spread by way of teaching or imitation within the group before they either die out, remain group specific, or begin to spread to other groups. In its initial stages, such an emergence of a new preference or moral value thus resembles a "mutation" as described in the simple model above. Under homophily, individuals have a tendency to interact with members of their own group, usually because it is easier or less costly (e.g., in

[24] An immediate extension of this simple matching model is to let both the group size and the expected number of mutants increase with the population size. If $n(N), \mu(N) \to \infty$, $\mu(N)/N \to 0$ and $\mu(N)/n(N) \to \delta$, then $\sigma = \delta \cdot p^*(n)$, where $p^*(n) = \lim_{N \to \infty} p(n(N), N)$.

terms of distance, language etc.) than interactions with members of other groups. In terms of our model, we then have $p(k,n) > 0$, and the index of assortativity will be positive if this intragroup matching probability does not tend to zero as the number of groups grows.

Sociologists Ruef, Aldrich, and Carter (2003) show that homophily is a strong factor in the formation of business partnerships. Our simple matching model also applies to this class of interactions. To see this, consider a large population of business students who after graduation set up pairwise business partnerships. Suppose that individuals' business ethics are sometimes influenced by the teaching in their school.[25] In the model above, let each graduating class be a group. Now and then, a teacher changes the teachings about business ethics, and some students are then influenced by the new material. Under the assumptions in the model, there will be positive assortativity if, on average, more than one student is influenced by the new teaching material ($\mu > 1$) and if the probability is positive for partner formation among classmates even when the number of schools is large ($p^*(n) > 0$). For example, if the average class size is one hundred, the probability for forming partnership with a classmate is one half, and if the average number of students per class who are influenced by the new teaching material is forty, then $\sigma \approx .2$.

5.3 Conditional degrees of morality

Although our stability analysis was focused on only one fitness game, the model has clear implications for the more realistic situation where each individual simultaneously engages in multiple fitness games. Indeed, insofar as individuals can distinguish the latter, the degree of morality that will be selected for is simply the index of assortativity in the matching process that corresponds to the fitness game at hand. For instance, if individuals are recurrently both engaged in some family interaction with a high index of assortativity and also in some market interaction with a low index of assortativity, then the present theory says that one and the same individual will exhibit a high degree of morality in the family interaction and be quite selfish in the market interaction. More generally, the "type" of an individual engaged in multiple interactions will be a vector of degrees of morality, one for each interaction, adapted to the matching processes in question (but independent of the payoff structure of the interaction).

For any given fitness game, an individual's degree of morality may further depend on other observable factors, such as group identity. To see this, suppose first that group membership in the matching model introduced in Section 5.2 is unobservable (to the individuals in the population); then our model of evolutionary stability of preferences predicts that all

[25] Whether or not preferences acquired in school persist throughout life is an empirical question, which we do not address here.

individuals will be *homo moralis* with degree of morality

$$\kappa = \frac{\mu - 1}{n - 1} \cdot p^* (n) .$$

However, in many real-life situations, group membership is observable. The evolutionarily stable degree of morality in an interaction will then be conditioned on group membership. To see this, suppose that all individuals in the simple matching model know their own group identity and can recognize others' group identity (but not others' preferences or moral values). Matched individuals can then condition their strategy choice on their own and their opponent's group identities. Our model of evolutionary stability of preferences can then be applied separately to every pair (i, j) of group identities.[26] In the simple matching model, where a mutation occurs only in one group, the index of assortativity in interactions between individuals from different groups is zero (since each mutant in such a pair is sure to meet a resident). By contrast, suppose that both individuals are from the same group. Then a mutant will attach probability $\Pr[\tau|\tau, \varepsilon] = (\mu - 1)/(n - 1)$ to the event that the opponent is also a mutant, while a resident who is unaware of in which group the mutation has taken place will assign probability one to the event that also the other individual is a resident in the limit as $\varepsilon \to 0$. Our model of preference stability then predicts that individuals will have zero degree of morality in all inter-group interactions and a positive degree of morality in all intragroup interactions:

$$\kappa_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ (\mu - 1)/(n - 1) & \text{if } i = j. \end{cases}$$

This suggests that we should expect individuals, in interactions where their moral values are their private information but group identity is public information, to typically show a higher degree of morality when interacting with their observational likes. Arguably important—but perhaps also controversial—applications abound. One need only think about such characteristics as language, ethnicity, nationality, religion, residential location, profession etc.

6 Discussion

6.1 Asymmetry

Symmetric games may well have asymmetric Nash equilibria that Pareto dominate the symmetric equilibria. Since evolutionary stability is concerned with strategies that are best

[26] In game-theoretic terms, each such pair defines a subgame in a game of incomplete information (as to others' types, but not group identities), and each such subgame is reached with positive probability under any strategy profile.

replies to themselves, evolutionarily stable outcomes may thus be socially inefficient. Consider, for example, symmetric 2×2-games with payoff matrix

$$A = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix} \tag{22}$$

for $a, b > 0$.[27] The unique evolutionarily stable strategy under strategy evolution and uniform random matching, the mixed strategy $x = (a/(a+b), b/(a+b))$, is Pareto dominated by each of the asymmetric equilibria; $ab/(a+b) < \min\{a, b\}$. In truly symmetric interactions, where individuals do not have any cue that assigns player roles to them, this is all that can be said. However, if there is a public randomization device, or institution, that assigns player roles 1 and 2 to the two individuals in each match, then individuals can condition their action on the assigned role. If each role assignment is also equally likely, then this defines a symmetric game in which nature first assigns player roles and then the two individuals simultaneously play in their assigned roles.[28] This enables Pareto efficient play in the above game. Indeed, play of the pure strategy equilibria is evolutionarily stable, with expected payoff $(a+b)/2$ to each participant. This reasoning can be generalized to assortative matching and to preference evolution in symmetric and asymmetric fitness games. We briefly discuss two canonical applications in the next two subsections.

6.1.1 Helping others

Situations where individuals have the opportunity to help others are common, and they can be modeled by way of a random dictator game. Assume that (a) with probability 1/2, player 1's initial wealth is w^H and 2's is $w^L \leq w^H$, (b) with probability 1/2 the players' wealth is reversed, and (c) the wealthier individual may transfer any amount of his or her wealth to the other. Let x be player 1's transfer when rich and y 2's transfer when rich, with $x, y \in X = \left[0, w^H\right]$. We may then write the payoff function in the form

$$\pi(x, y) = \frac{1}{2}\left[v\left(w^H - x\right) + v\left(w^L + y\right)\right]$$

for some differentiable function $v : [0,1] \longrightarrow \mathbb{R}$ with $v' > 0$ and $v'' < 0$. Here $v(w)$ is the fitness or well-being that results from wealth w. *Homo moralis* has the following utility function:

$$u_\kappa(x, y) = \frac{1}{2}\left[v\left(w^H - x\right) + \kappa v\left(w^L + x\right) + (1-\kappa)\, v\left(w^L + y\right)\right].$$

[27] These games are strategically equivalent to hawk-dove games, and, under relabelling of one player role's strategy set, battle-of-the-sexes games. See Alós-Ferrer and Kuzmics (2012) for an analysis of symmetry properties of games.

[28] This approach was first proposed in Selten (1980), for strategy evolution in finite games under uniform random matching.

Hence, $x_\kappa = 0$ if $v'\left(w^H\right) \geq \kappa v'\left(w^L\right)$, $x_\kappa = w^H$ if $v'(0) \leq \kappa v'\left(w^L + w^H\right)$ and otherwise $x_\kappa \in \left(0, w^H\right)$ is the unique solution to the first-order condition $v'\left(w^H - x\right) = \kappa v'\left(w^L + x\right)$. At one extreme, $\kappa = 0$, we have *homo oeconomicus* who gives nothing: $x_0 = 0$. At the opposite extreme, $\kappa = 1$, there is *homo kantiensis*, who transfers half of the initial wealth difference, $x_1 = \left(w^H - w^L\right)/2$, so that they end up with equal wealth. The (ultimate) reason why *homo moralis*, of sufficiently high degree of morality ($\kappa > v'\left(w^H\right)/v'\left(w^L\right)$), gives something to the other individual is not a concern for fairness (although this may well be an individual's proximate motivation). Instead, the ultimate reason is fitness maximization which may require some smoothing across states of nature because of the assumed concavity of fitness with respect to wealth. The "right thing to do" may thus be to give something to the poor. We note that a higher degree of morality implies more efficient risk sharing from an *ex ante* perspective.

6.1.2 Ultimatum bargaining

Consider the following scenario: (1) one monetary unit is handed either to individual 1 or to individual 2, with equal probability, (2) the party who received the monetary unit, the *proposer*, proposes a transfer $t \in [0,1]$ to the other party, (3) the other party, the *responder*, either accepts or rejects the proposal. If accept: the responder receives t and the proposer $1-t$. If reject: the monetary unit is withdrawn, so both parties receive nothing.

A monotonic pure strategy can be represented by a pair of numbers, $x = (x_1, x_2) \in X = [0,1]^2$, where the first number, x_1, is the amount to propose in the proposer role and the second number, x_2, is the smallest transfer to accept in the responder role, the "*acceptance threshold*". The monetary payoff to strategy x against a strategy y is thus

$$\pi(x,y) = v(0) + \frac{1}{2} \cdot \mathbf{1}_{\{x_1 \geq y_2\}} \cdot [v(1-x_1) - v(0)] + \frac{1}{2} \cdot \mathbf{1}_{\{y_1 \geq x_2\}} \cdot [v(y_1) - v(0)],$$

where v is as it was in the preceding subsection.[29] This payoff function is clearly not continuous. Moreover, *homo moralis'* best reply to a strategy $y \in X$ is in general not unique (if it is a best reply to accept a certain positive transfer then it is also a best reply to accept any lower transfer). By way of fairly involved but elementary calculations (see Alger and Weibull, 2012b), one can verify that, for any given degree of morality $\kappa \in [0,1]$, there is a whole continuum of *homo-moralis* strategies. Ultimatum bargaining between two *homo moralis* individuals typically admits multiple equilibrium outcomes, always including the 50/50 split which is the unique outcome when $\kappa = 1$.[30]

[29] Here $\mathbf{1}_A$ is the indicator function that takes the value 1 in the set A and zero outside it.

[30] This result is compatible with Huck and Oechssler (1999), who analyze evolutionary dynamics of strategies in an ultimatum-bargaining game under uniform random matching.

6.1.3 Repeated play

Finally, let us briefly consider a situation where no assigned player roles are given, but where the (fitness) game with payoff matrix (22) is played repeatedly. Suppose further that monitoring is perfect. Suppose, more specifically, that this game is played in rounds $t = 1, .., T$ and that payoffs from different rounds accumulate. How would *homo kantiensis* play? By definition, such an individual would use some behavior strategy y in $Y_T^* = \arg\max_{y \in X} \pi_T(y, y)$, where $\pi_T : Y_T^2 \to \mathbb{R}$ is continuous and Y_T is the non-empty and compact set of behavior strategies in this repeated game. Preliminary calculations suggest that one strategy in Y_T^* is first to randomize uniformly over the two pure actions in the initial period and to continue such (i.i.d.) randomization in each round until an asymmetric action pair has been achieved and thereafter alternate between the two pure actions in all successive rounds.[31] This alternating strategy, y_T^a, is the only element of Y_T^* when $T = 2$, while for $T > 2$ any pattern of play of the two asymmetric action pairs, once the symmetry is (randomly) broken, also belongs to Y_T^*. We conjecture that this is the unique strategy in Y_T^* if the fitness effect of repeated play is any increasing, continuous and strictly concave function of the sum of per-period payoffs, and if one requires robustness against (arbitrarily) small probabilities of termination after each round $t < T$.[32] An investigation of this topic, more broadly dealing with repeated play among *homo moralis* of arbitrary degrees of morality, falls outside the scope of this paper. Such an investigation might shed light on the empirical observation that human subjects in laboratory experiments engaged in repeated interactions seem to have a tendency to alternate between asymmetric Pareto efficient action profiles (see , e.g., Arifovic, McKelvey and Pevnitskaya, 2006).[33]

6.2 Morality vs. altruism

There is a large body of theoretical research on the evolution of altruism (e.g., Becker, 1976, Hirshleifer, 1977, Bester and Güth, 1998, Alger and Weibull, 2010, 2012a, and Alger, 2010). As noted above, the preferences of *homo moralis* differ sharply from altruism. We first show that while in some situations morality and altruism lead to the same behavior, in other situations the contrast is stark. Second, we discuss a situation where the behavior of *homo moralis* can be viewed as less "moral" than that of an altruist, or even than that of *homo*

[31] The use of early rounds of play to coordinate on future action profiles was analyzed for repeated coordination games in Crawford and Haller (1990).

[32] Arguably, even the slightest degree of concavity would favor alternating play in face of even the least risk of break-down.

[33] While our approach may explain such behavior, learning models generally fail to do so; see Hanaki *et al.* (2005) for a discussion and references, and for a model with learning among repeated-game strategies, allowing for alternation.

oeconomicus.

Altruism is usually represented by letting the altruist's utility be the sum of her own payoff and the payoff to the other individual, the latter term weighted by a factor $\alpha \in [0,1]$. In the present context:

$$u_\alpha(x,y) = \pi(x,y) + \alpha \cdot \pi(y,x), \tag{23}$$

where we will call α the *degree of altruism*.

The necessary first-order condition for an altruist at an interior symmetric equilibrium,

$$[\pi_1(x,y) + \alpha\pi_2(y,x)]_{|x=y} = 0,$$

is identical to that for a *homo moralis*,

$$[(1-\kappa)\pi_1(x,y) + \kappa\pi_1(x,x) + \kappa\pi_2(x,x)]_{|x=y} = 0,$$

if $\alpha = \kappa$. Nonetheless, the second-order conditions differ (Bergstrom, 2009). Furthermore, there is an important qualitative difference between *homo moralis* and altruists, namely, that their utility functions are in general not monotonic transformations of each other. This is seen in equations (7) and (23): for non-trivial payoff functions π and strategy sets X, and for any $\alpha, \kappa \neq 0$, there exists no function $T : \mathbb{R} \to \mathbb{R}$ such that $T[u_\alpha(x,y)] = u_\kappa(x,y)$ for all $x, y \in X$. This is seen most clearly in the case of finite games. Then u_α is linear in x while u_κ is quadratic in x:

$$\begin{cases} u_\alpha(x,y) = x \cdot Ay + \alpha y \cdot Ax \\ u_\kappa(x,y) = (1-\kappa)\, x \cdot Ay + \kappa x \cdot Ax \end{cases}$$

Consequently, the best-reply correspondence β_α of an altruist differs qualitatively, in general, from the best-reply correspondence β_κ of *homo moralis*, even when $\alpha = \kappa$. Indeed, the equilibria among altruists may differ from the equilibria among *homo moralis* also when $\alpha = \kappa$.

We further illustrate the tension between moralists and altruists, now in a finite game, an example suggested to us by Ariel Rubinstein. Take the fitness game which consists in a one-shot interaction with the payoff matrix given in (22), with $a = 2$ and $b = 1$, and consider a *homo kantiensis* ($\kappa = 1$), the "most moral" among *homo moralis*. Such an individual will play $x_{\kappa=1} = 1/2$.

Suppose now that such an individual visits a country where everyone plays the first pure strategy, and thus earns zero payoff in each encounter. When *homo kantiensis* interacts with a citizen in that society, the matched native citizen earns more than she does when interacting with other native citizens. However, if the visitor were instead a *homo oeconomicus* ($\kappa = 0$), then this new visitor would play the second pure strategy. Consequently, the other individual would earn more in this match than in a meeting with *homo kantiensis*. In fact, this lucky citizen would earn the maximal payoff in this game. Hence, citizens in this country would rather interact with *homo oeconomicus* than with *homo kantiensis*. What then if we would

instead replace *homo kantiensis* by a full-blooded altruist, someone who maximizes the sum of payoffs ($\alpha = 1$)? Given that all citizens play the first pure strategy, the best such an altruist could do would be to play the second pure strategy, just as *homo oeconomicus* would.

This example illustrates that the behavior of *homo kantiensis* is not necessarily "more moral" in an absolute sense and in all circumstances, than, say *homo oeconomicus* or an altruist. However, *homo kantiensis* is more moral in the sense of always acting in accordance with a general principle that is independent of the situation and identity of the actor (moral universalism), namely to do that which, if done by everybody, maximizes everybody's payoff.

Remark 2 *Suppose that the citizens of the country imagined above would like to achieve the highest possible payoff but are not even aware of the second pure strategy. Then* homo kantiensis *would, by his own example, show them its existence and thus show how they can increase their payoff in encounters amongst themselves. Indeed, an entrepreneurial and benevolent visitor to the imagined country could go one step further and suggest a simple institution within which to play this game, namely an initial random role allocation, at each pairwise match, whereby one individual is assigned player role 1 and the other player role 2, with equal probability for both allocations. This defines another symmetric two-player game in which each player has four pure strategies (two for each role). In this "meta-game" G',* homo kantiensis *would use any of two strategies $x'_{\kappa=1}$, each of which would maximize the payoff $\pi'(x'_{\kappa=1}, x'_{\kappa=1})$, namely to either always play the first (second) pure strategy in the original game when in player role 1 (2), or vice versa. In both cases, $\pi'(x'_{\kappa=1}, x'_{\kappa=1}) = 3/2$, a higher payoff than when* homo kantiensis *meets himself in the original game: $\pi(x_{\kappa=1}, x_{\kappa=1}) = 3/4$ (c.f. the discussion in section 6.1.3).*

6.3 Empirical testing

An interesting empirical research challenge is to find out how well *homo moralis* can explain behavior observed in controlled laboratory experiments. Consider, for example, an experiment in which (a) subjects are randomly and anonymously matched in pairs to play a two-player game (or a few different such games), (b) after the first few rounds of play, under (uniformly) random re-matching, subjects receive some information about aggregate play in these early rounds, and (c) are then invited to play the game once more (again with randomly drawn pairs). One could then analyze their behavior in these later rounds under the hypothesis that in this last round they play a (Bayesian) Nash equilibrium under incomplete information, where each individual is a *homo moralis* with an individual-specific and fixed degree of morality. How much of the observed behavior could be explained this way? How well would *homo moralis* fare in comparison with established models of social preferences? In an early experimental study of a prisoners' dilemma interaction, analyzed as a game of incomplete information, Bolle and Ockenfels (1990) show that observed behaviors are better

explained by individual-specific 'moral standards' than altruism. Similar experiments could be carried out to test the *homo moralis* hypothesis. It would also be interesting to compare empirical results for different cultures and to see if such differences can be explained in terms of assortativity differences between these cultures.

7 Conclusion

We have here tried to contribute to the understanding of ultimate causes for human motivation by proposing a theoretical model of the evolution of preferences when these preferences remain private information and when the matching process is random but may involve correlation between types in the matches. Our approach delivers new testable predictions. Although we permit all continuous preferences over strategy pairs, we find that a particular one-dimensional parametric family, the preferences of *homo moralis*, stands out in the analysis. A *homo moralis* acts as if he or she had a sense of morality: she maximizes a weighted sum of her own payoff, given her expectation of the other's action and of the payoff she would obtain if the other party were also to take the same action. We show that a certain member of this family, *homo hamiltonensis*, is particularly viable from an evolutionary perspective. The weight that *homo hamiltonensis* attaches to the moral goal is the index of assortativity in the matching process. Our theory further predicts that if one and the same individual is engaged in multiple pairwise interactions of the sort analyzed here, perhaps with a different index of assortativity associated with each interaction (say, one interaction taking place within the extended family and another one in a large anonymous market), then this individual will exhibit different degrees of morality in these interactions, adapted to the various indices of assortativity.

While the general predictive power of preferences *à la homo moralis* remains to be carefully examined, the behavior of *homo moralis* seems to be compatible with some experimental evidence (for preliminary calculations in this direction, see Alger and Weibull, 2012b). What's more, the goal function of *homo moralis* appears to be consistent with the justifications many subjects offer for their behavior in the lab, namely, saying that they wanted to "do the right thing" (see, e.g., Dawes and Thaler, 1988, Charness and Dufwenberg, 2006). While we leave analyses of the policy implications of such moral preferences for future research, we note that, in our model, the evolutionarily stable degree of morality is independent of the payoffs in the interaction at hand. Hence, the degree of morality cannot be "crowded out" in any direct sense by economic incentives. For instance, if one were to pay people for "doing the right thing" or charge a fee for "doing the wrong thing," this would change the payoffs and thus also the behavior of *homo moralis*, in an easily predictable way, but evolutionary forces would not change her degree of morality (as long as the matching process remains the same).

While the self-interested *homo oeconomicus* does well in non-strategic interactions and in

situations where there is no assortativity in the matching process, natural selection wipes out *homo oeconomicus* in large classes of other situations. Arguably, assortativity is common in human interactions. Allowing for arbitrary degrees of assortativity in the matching processes, our analysis suggests that pure selfishness, as a foundation for human motivation, should perhaps be replaced by a blend of selfishness and morality. Such a change would affect many predictions in economics.

This is but a first exploration, calling for extensions and applications in many directions and areas, such as multi-player interactions, asymmetric and repeated interactions, signals and cues about others' types, partner choice, public-goods provision, environmental policy, institution building, voting and political economy.

Appendix: proof of Lemma 1

Write (3) in terms of the assortativity function ϕ:

$$\begin{cases} x^* \in \arg\max_{x \in X} & [1-\varepsilon+\varepsilon\phi(\varepsilon)] \cdot u_\theta(x, x^*) \\ & \quad + \varepsilon[1-\phi(\varepsilon)] \cdot u_\theta(x, y^*) \\ y^* \in \arg\max_{y \in X} & (1-\varepsilon)[1-\phi(\varepsilon)] \cdot u_\tau(y, x^*) \\ & \quad + [\varepsilon+(1-\varepsilon)\phi(\varepsilon)] \cdot u_\tau(y, y^*). \end{cases} \tag{24}$$

By hypothesis, u_θ and u_τ are continuous and X is compact. Hence, each right-hand side in (24) defines a non-empty and compact set, for any given $\varepsilon \in [0,1)$, by Weierstrass's maximum theorem. For any $(\theta, \tau, \varepsilon) \in S$, condition (24) can thus be written in the form $(x^*, y^*) \in B_\varepsilon(x^*, y^*)$, where $B_\varepsilon : C \rightrightarrows C$, for $C = X^2$ and $\varepsilon \in [0,1)$ fixed, is compact-valued, and, by Berge's maximum theorem, upper hemi-continuous (see e.g. Aliprantis and Border, 2006). It follows that B_ε has a closed graph, and hence its set of fixed points, $B^{NE}(\theta, \tau, \varepsilon) = \{(x^*, y^*) \in X^2 : (x^*, y^*) \in B_\varepsilon(x^*, y^*)\}$ is closed (being the intersection of $graph(B_\varepsilon)$ with the diagonal of C^2). This establishes the first claim.

If u_θ and u_τ are concave in their first arguments, then so are the maximands in (24). Hence, B_ε is then also convex-valued, and thus has a fixed point by Kakutani's fixed-point theorem. This establishes the second claim.

For the third claim, fix θ and τ, and write the maximands in (24) as $U(x, x^*, y^*, \varepsilon)$ and $V(y, x^*, y^*, \varepsilon)$. These functions are continuous by assumption. Let $U^*(x^*, y^*, \varepsilon) = \max_{x \in X} U(x, x^*, y^*, \varepsilon)$ and $V^*(x^*, y^*, \varepsilon) = \max_{y \in X} V(y, x^*, y^*, \varepsilon)$. These functions are continuous by Berge's maximum theorem. Note that $(x^*, y^*) \in B^{NE}(\theta, \tau, \varepsilon)$ iff

$$\begin{cases} U^*(x^*, y^*, \varepsilon) - U(x, x^*, y^*, \varepsilon) \geq 0 & \forall x \in X \\ V^*(x^*, y^*, \varepsilon) - U(y, x^*, y^*, \varepsilon) \geq 0 & \forall y \in X. \end{cases} \tag{25}$$

Let $\langle \varepsilon_t \rangle_{t \in \mathbb{N}} \to \varepsilon^o \in [0,1)$ and suppose that $(x_t^*, y_t^*) \in B^{NE}(\theta, \tau, \varepsilon_t)$ and $(x_t^*, y_t^*) \to (x^o, y^o)$. By continuity of the functions on the left-hand side in (25),

$$\begin{cases} U^*(x^o, y^o, \varepsilon^o) - U(x, x^o, y^o, \varepsilon^o) \geq 0 & \forall x \in X \\ V^*(x^o, y^o, \varepsilon^o) - U(y, x^o, y^o, \varepsilon^o) \geq 0 & \forall y \in X \end{cases}$$

and hence $(x^o, y^o) \in B^{NE}(\theta, \tau, \varepsilon^o)$. This establishes the third claim.

References

Akerlof, G. and R. Kranton (2000): "Economics and Identity," *Quarterly Journal of Economics,* 115, 715-753.

Alexander, R. D. (1987): *The Biology of Moral Systems*. New York: Aldine De Gruyter.

Alger, I. (2010): "Public Goods Games, Altruism, and Evolution," *Journal of Public Economic Theory*, 12, 789-813.

Alger, I. and C-t. A. Ma (2003): "Moral Hazard, Insurance, and Some Collusion," *Journal of Economic Behavior and Organization*, 50, 225-247.

Alger, I. and R. Renault (2006): "Screening Ethics when Honest Agents Keep their Word," *Economic Theory*, 30, 291-311.

Alger, I. and R. Renault (2007): "Screening Ethics when Honest Agents Care about Fairness," *International Economic Review*, 47, 59-85.

Alger, I. and J. Weibull (2010): "Kinship, Incentives and Evolution," *American Economic Review*, 100, 1725-1758.

Alger, I. and J. Weibull (2012a): "A Generalization of Hamilton's Rule—Love Others How Much?" *Journal of Theoretical Biology*, 299, 42-54.

Alger, I. and J. Weibull (2012b): "Homo Moralis - Preference Evolution under Incomplete Information and Assortative Matching," TSE Working Paper 12-281.

Aliprantis, C.D. and K.C. Border (2006): *Infinite Dimensional Analysis*, 3rd ed. New York: Springer.

Alós-Ferrer, C. and C. Kuzmics (2012): "Hidden Symmetries and Focal Points", *Journal of Economic Theory*, forthcoming.

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100, 464-477.

Arifovic, J., R. McKelvey and S. Pevnitskaya (2006): "An Initial Implementation of the Turing Tournament to Learning in Repeated Two-person Games," *Games and Economic Behavior*, 57, 93-122.

Arrow, K. (1973): "Social Responsibility and Economic Efficiency," *Public Policy*, 21, 303-317.

Bacharach, M. (1999): "Interactive Team Reasoning: A Contribution to the Theory of Co-operation," *Research in Economics*, 53, 117-147.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," *Journal of Economic Literature*, 14, 817–826.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Eco-*

nomic Review, 96, 1652-1678.

Bénabou, R. and J. Tirole (2011): "Identity, Morals and Taboos: Beliefs as Assets," *Quarterly Journal of Economics*, 126, 805-855.

Bergstrom, T. (1995): "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review,* 85, 58-81.

Bergstrom, T. (2003): "The Algebra of Assortative Encounters and the Evolution of Cooperation," *International Game Theory Review,* 5, 211-228.

Bergstrom, T. (2009): "Ethics, Evolution, and Games among Neighbors," Working Paper, UCSB.

Bester, H. and W. Güth (1998): "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization,* 34, 193–209.

Binmore, K. (1994): *Playing Fair - Game Theory and the Social Contract.* Cambridge (MA): MIT Press.

Bisin, A., G. Topa, and T. Verdier (2004): "Cooperation as a Transmitted Cultural type," *Rationality and Society*, 16, 477-507.

Bisin, A., and T. Verdier (2001): "The Economics of Cultural Transmission and the Dynamics of Preferences," *Journal of Economic Theory,* 97, 298-319.

Bolle, F. (2000): "Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth" *Journal of Economic Behavior and Organization*, 42, 131-133.

Bolle, F., and P. Ockenfels (1990): "Prisoners' Dilemma as a Game with Incomplete Information," *Journal of Economic Psychology*, 11, 69-84.

Bomze, I.M., and B.M. Pötscher (1989): *Game Theoretical Foundations of Evolutionary Stability.* New York: Springer.

Boorman, S.A., and P.R. Levitt (1980): *The Genetics of Altruism.* New York: Academic Press.

Bramoullé, Y., and B. Rogers (2009): "Diversity and Popularity in Social Networks," Discussion Papers 1475, Northwestern University, Center for Mathematical Studies in Economics and Management Science.

Brekke, K.A., S. Kverndokk, and K. Nyborg (2003): "An Economic Model of Moral Motivation," *Journal of Public Economics*, 87, 1967–1983.

Charness, G. and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

Charness, G. and M. Rabin. (2002): "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics*, 117, 817-869.

Crawford, V. and H. Haller (1990): "Learning How to Cooperate: Optimal Play in Repeated Coordination Games", *Econometrica*, 58, 571-595.

Currarini, S., M.O. Jackson, and P. Pin (2009): "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica,* 77, 1003–1045.

Currarini, S., M.O. Jackson, and P. Pin (2010): "Identifying the Roles of Race-Based Choice and Chance in High School Friendship Network Formation," *Proceedings of the National Academy of Sciences,* 107, 4857–4861.

Darwin, C. (1871): *The Descent of Man, and Selection in Relation to Sex.* London: John Murray.

Dawes, R. and R. Thaler (1988): "Anomalies: Cooperation," *Journal of Economic Perspectives,* 2, 187-97.

Day, T., and P.D. Taylor (1998): "Unifying Genetic and Game Theoretic Models of Kin Selection for Continuous types," *Journal of Theoretical Biology*, 194, 391-407.

Dekel, E., J.C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economic Studies*, 74, 685-704.

de Waal, Frans B.M. (2006): *Primates and Philosophers. How Morality Evolved.* Princeton: Princeton University Press.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268-98.

Edgeworth, F.Y. (1881): *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences.* London: Kegan Paul.

Ellingsen, T. (1997): "The Evolution of Bargaining Behavior," *Quarterly Journal of Economics*, 112, 581-602.

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990-1008.

Eshel, I., and L.L. Cavalli-Sforza (1982): "Assortment of Encounters and Evolution of Cooperativeness," *Proceedings of the National Academy of Sciences,* 79, 1331-1335.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293-315.

Fehr, E., and K. Schmidt (1999): "A theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817-868.

Fershtman, C. and K. Judd (1987): "Equilibrium Incentives in Oligopoly," *American Economic Review*, 77, 927–940.

Fershtman, C., and Y. Weiss (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70, 53-73.

Frank, R.H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, 77, 593-604.

Frank, R.H. (1988): *Passions within Reason: The Strategic Role of the Emotions.* New York: Norton & Co.

Grafen, A. (1979): "The Hawk-Dove Game Played between Relatives," *Animal Behavior,* 27, 905–907.

Grafen, A. (2006): "Optimization of Inclusive Fitness," *Journal of Theoretical Biology,* 238, 541–563.

Güth, W., and M. Yaari (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in U. Witt. *Explaining Process and Change – Approaches to Evolutionary Economics.* Ann Arbor: University of Michigan Press.

Hamilton, W.D. (1964a): "The Genetical Evolution of Social Behaviour. I." *Journal of Theoretical Biology*, 7:1-16.

Hamilton, W.D. (1964b): "The Genetical Evolution of Social Behaviour. II." *Journal of Theoretical Biology*, 7:17-52.

Hamilton, W. D. (1971): "Selection of selfish and altruistic behaviour in some extreme models," in J.F. Eisenberg and W. S. Dillon (Eds.). *Man and Beast: Comparative Social Behavior.* Washington, DC: Smithsonian Press.

Hamilton, W. D. (1975): "Innate Social Aptitudes in Man, an Approach from Evolutionary Genetics," in R. Fox (Ed.). *Biosocial Anthropology.* London: Malaby.

Hanaki, N., R. Sethi, I. Erev, and A. Peterhansl (2005): "Learning Strategies," *Journal of Economic Behavior and Organization*, 56, 523-542.

Hauk, E., and M. Sáez-Martí (2002): "On the Cultural Transmission of Corruption," *Journal of Economic Theory*, 107, 311–35.

Heifetz, A., C. Shannon, and Y. Spiegel (2006): "The Dynamic Evolution of Preferences," *Economic Theory*, 32, 251-286.

Heifetz, A., C. Shannon, and Y. Spiegel (2007): "What to Maximize if You Must," *Journal of Economic Theory*, 133, 31-57.

Hines, W.G.S., and J. Maynard Smith (1979): "Games between Relatives," *Journal of Theoretical Biology*, 79, 19-30.

Hirshleifer, J. (1977): "Economics from a Biological Viewpoint," *Journal of Law and Economics*, 20, 1-52.

Huck, S., D. Kübler, and J.W. Weibull (2012): "Social Norms and Economic Incentives in Firms," *Journal of Economic Behavior & Organization*, 83, 173-185.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining

Fair Allocations," *Games and Economic Behavior,* 28, 13–24.

Jackson, M.O., and A. Watts (2010): "Social Games: Matching and the Play of Finitely Repeated Games," *Games and Economic Behavior,* 70, 170-191.

Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten.* [In English: *Groundwork of the Metaphysics of Morals.* 1964. New York: Harper Torch books.]

Koçkesen, L., E.A. Ok, and R. Sethi (2000): "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory,* 92, 274-299.

Laffont, J.-J. (1975): "Macroeconomic Constraints, Economic Efficiency and Ethics: an Introduction to Kantian Economics," *Economica*, 42, 430-437.

Lehmann L., and F. Rousset (2010): "How Life History and Demography Promote or Inhibit the Evolution of Helping Behaviours," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 365, 2599-2617.

Levine, D. (1998): "Modelling Altruism and Spite in Experiments," *Review of Economic Dynamics,* 1, 593-622.

Lindbeck, A. and S. Nyberg (2006): "Raising Children to Work Hard: Altruism, Work Norms and Social Insurance," *Quarterly Journal of Economics*, 121, 1473-1503.

Maynard Smith, J., and G.R. Price (1973): "The Logic of Animal Conflict," *Nature,* 246:15-18.

McPherson, M., L. Smith-Lovin, and J.M. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415-444.

Michod, R.E. and W.D. Hamilton (1980): "Coefficients of Relatedness in Sociobiology," *Nature,* 288, 694 - 697.

Nichols, S. (2004): *Sentimental Rules.* Oxford: Oxford University Press.

Nowak, M. (2006): "Five Rules for the Evolution of Cooperation," *Science*, 314, 1560-1563.

Ockenfels, P. (1993): "Cooperation in Prisoners' Dilemma—An Evolutionary Approach", *European Journal of Political Economy*, 9, 567-579.

Ok, E.A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory,* 97, 231-254.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences" *Journal of Economic Behavior and Organization*, 42, 125-129.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281-1302.

Robson, A.J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret

Handshake," *Journal of Theoretical Biology*, 144, 379-396.

Roemer, J.E. (2010): "Kantian Equilibrium," *Scandinavian Journal of Economics*, 112, 1–24.

Rousset, F. (2004): *Genetic Structure and Selection in Subdivided Populations.* Princeton: Princeton University Press.

Ruef, M., H.E. Aldrich, and N.M. Carter (2003): "The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs," *American Sociological Review*, 68, 195-222.

Sandholm, W. (2001): "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Economic Dynamics,* 4, 637-679.

Schaffer, M.E. (1988): "Evolutionarily Stable Strategies for Finite Populations and Variable Contest Size," *Journal of Theoretical Biology*, 132, 467-478.

Schelling, T. (1960): *The Strategy of Conflict.* Cambridge: Harvard University Press.

Selten, R. (1980): "A Note on Evolutionarily Stable Strategies in Asymmetric Animal Conflicts," *Journal of Theoretical Biology*, 84, 93-101.

Sen, A.K. (1977): "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy & Public Affairs*, 6, 317-344.

Sethi, R., and E. Somanathan (2001): "Preference Evolution and Reciprocity" *Journal of Economic Theory,* 97, 273-297.

Smith, A. (1759): *The Theory of Moral Sentiments.* Edited by D. D. Raphael and A. L. Macfie. Oxford: Clarendon Press; New York: Oxford University Press (1976).

Sober, E., and D.S. Wilson (1998): *Unto Others.* Cambridge: Harvard University Press.

Sugden, R. (1995): "A Theory of Focal Points," *Economic Journal*, 105, 533-50.

Sugden, R. (2011): "Mutual Advantage, Conventions and Team Reasoning," *International Review Of Economics*, 58, 9-20.

Tabellini, G. (2008): "Institutions and Culture," *Journal of the European Economic Association,* 6, 255–294.

Toro, M., and L. Silio (1986): "Assortment of Encounters in the Two-strategy Game," *Journal of Theoretical Biology,* 123, 193-204.

Weibull, J.W (1995): *Evolutionary Game Theory.* Cambridge: MIT Press.

Wilson, D.S., and L.A. Dugatkin (1997): "Group Selection and Assortative Interactions," *American Naturalist*, 149, 336-351.

Wright, S.G. (1922): "Coefficients of Inbreeding and Relationship," *American Naturalist,* 56, 330-338.