# Homo Moralis

—

# PREFERENCE EVOLUTION UNDER INCOMPLETE INFORMATION AND ASSORTATIVE MATCHING*

Ingela Alger† and Jörgen W. Weibull‡

March 1, 2012

**Abstract**

What preferences will prevail in a society of rational individuals when preference evolution is driven by their success in terms of resulting payoffs? We show that when individuals' preferences are their private information, a convex combinations of selfishness and morality stand 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 material payoff. At the opposite end is homo kantiensis, who does what would be "the right thing to do," in terms of material payoffs, if all others would do likewise. We show that the stable degree of morality - the weight placed on the moral goal - equals the index of assortativity in the matching process. The motivation of homo moralis is arguably compatible with how people often reason, and the induced behavior appear to agree with pro-social behaviors observed in many laboratory experiments.

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

JEL codes: C73, D03.

*An early version of this paper was presented at the NBER workshop on culture and institutions in November 2011. We thank Rajiv Sethi for helpful comments to that version. We also thank Immanuel Bomze, Avinash Dixit, Tore Ellingsen, Jens Josephson, 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.

†LERNA/CNRS @ Toulouse School of Economics, and Carleton University (Ottawa)

‡Stockholm School of Economics, École Polytechnique (Paris), and Royal Institute of Technology (Stockholm)

# 1 Introduction

Most of contemporary economics is premised on the assumption that human behavior is driven by self-interest. This assumption provides predictive power in many areas of economics. Moreover, economists have used it to identify conditions under which self-interest leads to the common good. However, in the early history of the profession, it was common to include moral values in 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] Moreover, in recent years many economists, particularly within behavioral and experimental economics, have begun to question the predictive power of pure selfishness in certain interactions and turned to social, or "other-regarding" preferences. Our goal is here to clarify the evolutionary foundation of human motivation. Like many others before us, but now in a more general setting, we investigate whether pure self-interest is favored by evolution. With virtually no restrictions on the class of potential preferences that may be selected for, our main result is that natural selection leads to a certain one-dimensional spectrum of moral preferences, a spectrum that sprang out from the mathematics. At one end of this spectrum is pure self-interest and at the other is pure ethical reasoning in line with Kant's categorical imperative.

It is well-known that 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 (Schelling, 1960). Likewise, 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 proposer in ultimatum bargaining may be more generous if the responder is known or believed to be inequity averse rather than solely interested in own monetary gains. This raises the further question whether evolution would favor such preferences.

In the literature addressing this issue, usually called the indirect evolutionary approach, pioneered by Güth and Yaari (1992), much work has been devoted to the case where individuals are uniformly randomly matched and know each others' preferences.[2] Usually, evolution then favors other preferences than those of *homo oeconomicus* (see Heifetz, Shannon, and Spiegel, 2007, for a particularly general such result). By contrast, when individuals are uniformly randomly matched and preferences are private information, evolution leads to the self-interested *homo oeconomicus*, see Ok and Vega-Redondo (2001), and Dekel, Ely and Yilankaya (2007). We here propose a theory for why evolution may lead to preferences that

[1] See Binmore (1994) for a game-theoretic discussion of ethics.

[2] See Frank (1987), Robson (1990), Güth and Yaari (1992), 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 and Weibull (2010, 2011), and Alger (2010).

differ from those of *homo oeconomicus* even when preferences are private information. The reason for our different result is that we permit assortativity in the matching process that brings individuals together.

More exactly, we analyze the evolution of preferences in a large population where individuals are randomly and pairwise matched to interact. In each interaction, individual behavior is driven by (subjective) utility maximization, while evolutionary success is driven by some (objective) payoff. The key element in our theory is that the matching process may be more or less *assortative*, that is, individuals with the same preferences may be more or less likely to be matched with each other. Such assortativity arises as soon as there is some probability that the individuals inherited their preferences from some common ancestor (be it a genetic or a cultural ancestor). Assortativity is arguably common, because of a tendency to interact with kin, with people in the same geographical area or from the same school, or with people with the same culture, religion, or values (see e.g. Eshel and Cavalli-Sforza, 1982). By contrast, existing models of preference evolution assume that individuals are *uniformly* randomly matched.[3] The matching process is here 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 definition of evolutionary stability to allow for arbitrary degrees of assortativity in the matching process, and apply this to preference evolution when each matched pair plays a (Bayesian) Nash equilibrium of the associated game under incomplete information, that is, as if they each knew the statistical preference distribution in their matches, but not the preferences of the other individual in the match at hand.

With a minimum of additional assumptions, this leads to a remarkable result: a certain convex combination of selfishness — "maximization of own payoff" — and a certain form of morality —"to do what would lead to maximal payoff if everybody else did likewise" — stands out as evolutionarily stable. Individuals with preferences in this one-dimensional class will be called *homo moralis* and the weight attached to the moral goal the *degree of morality.* A special case is the familiar *homo oeconomicus*, whose utility function coincides with the objective payoff function. At the other extreme one finds *homo kantiensis*, who strives to maximize the payoff that would result should everybody behave in the same way.

Our main finding is that evolution selects that degree of morality which equals the index of assortativity in the matching process. In a resident population, such preferences turn out to provide the most effective protection against rare mutants, since they induce their carriers to behave in such a way that the same behavior is then also payoff-optimal for rare mutants, should such appear. Hence, it is as if moral preferences with the right weight attached to the moral goal *preempt* mutants; a rare mutant can at best match the payoff of the resident population.

[3] Exceptions are Alger and Weibull (2010, 2011) and Alger (2010).

This result has dire consequences for *homo oeconomicus* in many situations. In particular, *homo oeconomicus* is selected against as soon as an individual's payoff depends on the other's action and the matching process has a positive index of assortativity. Arguably, these two features are common, implying that we should expect a positive degree of morality, rather than the pure selfishness of *homo oeconomicus*, to prevail in many settings.

It is beyond the scope of this paper to provide a comprehensive analysis of the behavior of *homo moralis*. It is not difficult to show, however, that *homo moralis'* behavior in interactions commonly studied in laboratory experiments is compatible with observation (as reported in, e.g., Fehr and Gächter, 2000, Gächter, Herrmann and Thöni, 2004, Henrich *et al.*, 2005, and Gächter and Herrmann, 2009). In particular, *homo moralis* gives positive amounts in dictator games, may reject positive amounts in ultimatum games, and contributes more to public goods than would be motivated by material self-interest.[4] For further analyses of the behavioral implications of preferences similar to those of *homo moralis*, see Laffont (1975), Brekke, Kverndokk, and Nyborg (2003), and Huck, Kübler and Weibull (2011).[5]

As a side result, we obtain a new interpretation of evolutionary stability of strategies (here under random matching with arbitrary degree of assortativity), namely, that these are precisely the behaviors one will observe in Nash equilibrium play under incomplete information, when evolution operates at the level of preferences, rather than directly on strategies. This sharpens and generalizes the result in Dekel *et al.* (2007) that preference evolution under incomplete information and uniform random matching in finite games implies symmetric Nash equilibrium play and is implied by strict symmetric Nash equilibrium play (in both cases with Nash equilibrium defined in terms of payoffs).

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. Section 6 discusses related literature, morality and altruism, and empirical testing. Section 7 concludes.

## 2 The model

Consider a population where individuals are 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

[4] We refer to the working paper version for this analysis. Furthermore, in Section 6 we provide a suggestion for an experimental design.

[5] See Bacharach (1999) for "team reasoning" and Roemer (2010) for an approach to "Kantian equilibrium."

that $X$ is a non-empty, compact and convex set in a topological vector space, and that $\pi$ is continuous.[6] 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 mathematical relation between a utility function $u$ and the payoff function $\pi$. A special type is *homo oeconomicus*, by which we mean individuals with the special utility function $u = \pi$.

For the subsequent analysis, it will be sufficient to consider populations with at most 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 $\tau$.[7] If $\varepsilon$ is small we will refer to $\theta$ as the *resident* type, and call $\tau$ the *mutant* type. The set of population states is thus $S = \Theta^2 \times [0,1]$.

The matching process is random and exogenous, but we allow it to 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 $\theta$ will be matched with an individual of type $\tau$, and $\Pr[\theta|\tau, \varepsilon]$ the probability that a given individual of type $\tau$ will be matched with an individual of type $\theta$. In the special case of uniform random matching, $\Pr[\tau|\theta, \varepsilon] = \Pr[\tau|\tau, \varepsilon] = \varepsilon$.

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

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

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

Turning now to the choices made by individuals in a population state, we define (Bayesian) Nash equilibrium as 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*

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

Evolutionary stability is defined under the assumption that the resulting payoffs are determined by this equilibrium set. With potential multiplicity of equilibria, one may require the resident type to withstand invasion in some or all equilibria. We have chosen the most severe criterion.[8]

[6] To be more precise, we assume $X$ to be a locally convex Hausdorff space, see Aliprantis and Border (2006). The subsequent examples are all carried out in the special but important case of Euclidean spaces.

[7] In particular, in the population state $s = (\theta, \tau, 0)$ only type $\theta$ is present.

[8] The least stringent criterion would be to replace “all Nash equilibria” by “some Nash equilibrium.”

**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 $F_\theta(x^*, y^*, \varepsilon) > F_\tau(x^*, y^*, \varepsilon)$ in all Nash equilibria $(x^*, y^*)$ in all states $s = (\theta, \tau, \varepsilon)$ with $\varepsilon \in (0, \bar{\varepsilon})$. A type $\theta$ is* ***evolutionarily stable*** *if it is evolutionarily stable against all types $\tau \neq \theta$ in $\Theta$.*

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 an evolutionarily stable strategy, a property they defined for mixed strategies in finite and symmetric two-player games under uniform random matching.

In a rich enough type space $\Theta$, no type is evolutionarily stable, since for each resident type $\theta$ there then exist mutant types $\tau$ who respond with the same strategy, in which case both types earn the same average payoff. However, many types will turn out to be vulnerable, as residents, to invasion by better-performing mutant types. We will use the following definition to describe this possibility:

**Definition 3** *A type $\theta \in \Theta$ is* ***evolutionarily unstable*** *if there exists a type $\tau \in \Theta$ such that for each $\bar{\varepsilon} > 0$ there exists an $\varepsilon \in (0, \bar{\varepsilon})$ such that $F_\theta(x^*, y^*, \varepsilon) < F_\tau(x^*, y^*, \varepsilon)$ in all Nash equilibria $(x^*, y^*)$ in state $s = (\theta, \tau, \varepsilon)$.*

This is also a stringent criterion**,** namely, there should exist some mutant type against which the resident type achieves less payoff in every equilibrium in some population states when the mutant is arbitrarily rare.

This completes the description of the model. The most closely related work on preference evolution under incomplete information, Ok and Vega-Redondo (2001), or OVR for short, and Dekel, Ely and Yilankaya (2007), or DEY for short. In all three models, the preference space is very general and interactions are symmetric. In OVR the population is finite whereas in DEY and our model it is a continuum. In OVR each interaction may involve more than two individuals, whereas in DEY and our model interactions are pairwise. In OVR, individual use pure strategies, the pure-strategy set is a continuum and the payoff function is strictly concave, whereas DEY focus on mixed strategies in finite games. We analyze both cases and do not require strict concavity. All three papers study evolutionary stability, although the definitions differ slightly from one another. A key difference between our model and OVR and DEY is that they assume uniform random matching, whereas we allow for assortative matching.

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

## 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 $\theta$, given that the individual him- or herself either also has type $\theta$, or, alternatively, type $\tau$:

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

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 $\theta$ and $\tau$,

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

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

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

We assume that $\phi$ is continuous and that $\phi(\varepsilon)$ converges to some number as $\varepsilon$ tends to zero. Formally:

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

for some $\sigma \in \mathbb{R}$, the *index of assortativity* of the matching process. This extends the domain of $\phi$ from $(0,1)$ to $[0,1]$, and it follows from (6) that $\sigma \in [0,1]$. The extreme case $\sigma = 0$ corresponds to *uniform random matching* and the opposite extreme case $\sigma = 1$ to *perfectly correlated matching* (each type only meets itself). In Section 5 we calculate the index of assortativity for a variety of matching processes.

## 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 $\kappa \in [0,1]$, the degree of morality.*[9]

It is as if *homo moralis* is torn between selfishness and morality. On the one hand, she would like to maximize her own payoff, $\pi(x,y)$. On the other hand, she would like to "do the right thing" in terms of payoffs, *i.e.*, choose a strategy $x$ that would be optimal, in terms of the payoff, if both players would choose one and the same strategy. This second goal can

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

be viewed as an application of Kant's categorical imperative to the goal of enhancement of everyone's (objective) payoff (or fitness).[10] Torn by these two goals, *homo moralis* will take an action that maximizes a convex combination of them.[11] If $\kappa = 0$, the definition of *homo moralis* coincides with that of *homo oeconomicus*.

**Remark 1** *It is clear from (7) that the behavior of* homo moralis *is unaffected by positive affine transformations of payoffs. However, if $x$ and $y$ are mixed strategies (say, in a finite game), then the preferences of homo moralis, for $\kappa > 0$, are not linear, but quadratic, in the own mixed strategy $x$. Thus* homo moralis' *preferences do not, in general, satisfy the von Neumann-Morgenstern axioms. See Section 4 for a detailed analysis.*

A special kind of *homo moralis* is *homo hamiltoniensis* (or *HH*), whose degree of morality equals the index of assortativity, $\kappa = \sigma$:

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

The following remark explains the etymology:

**Remark 2** *The late biologist William Hamilton (1964a,b) noted that for interactions between related individuals, genes driving the behavior of one individual are present in the relative with some probability, and argued that fitness had to be augmented to what he called* inclusive fitness. *$(1 - \sigma) \cdot \pi(x, y) + \sigma \cdot \pi(x, x)$ can be interpreted as the population average inclusive fitness of 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).*

# 3 Analysis

We first make three observations that will be useful and then proceed to analyze evolutionary stability properties of preferences. First, since the strategy set is non-empty and compact and each type's utility function is continuous, each type has at least one best reply to each

[10] Kant's (1785) categorical imperative can be phrased as follows: "Act only on the maxim that you would at the same time will to be a universal law." To always choose a strategy $x$ that maximizes $\pi(x, x)$ is the maxim which, if upheld as a universal law in the population at hand, leads to the highest possible payoff among all maxims that are categorical in the sense of not conditioning on any particular circumstance that would permit role-specific strategies. See Binmore (1994) for a critical discussion of Kant's categorical imperative.

[11] Note, however, that *homo moralis* is not irrational. Individuals with such preferences do have (continuous) utility functions, and these are defined over the (compact) set $X^2$ of strategy pairs.

strategy. More precisely, for each type $\theta \in \Theta$ the best-reply correspondence $\beta_\theta : X \rightrightarrows X$, defined by

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

is nonempty- and compact-valued (by Weierstrass's maximum theorem) and upper hemi-continuous by Berge's maximum theorem (see, *e.g.*, Section 17.5 in Border and Aliprantis, 2006).

Second, by Definition 1, a (Bayesian) Nash equilibrium $(x^*, y^*)$ in the limit population state $s = (\theta, \tau, 0)$ is defined by

$$\begin{cases} x^* \in \arg\max_{x \in X} & u_\theta(x, x^*) \\ y^* \in \arg\max_{y \in X} & (1-\sigma) \cdot u_\tau(y, x^*) + \sigma \cdot u_\tau(y, y^*) \,. \end{cases} \tag{9}$$

The first line in this condition requires the strategy of the resident type $\theta$ to be a best reply to itself, in terms of its own utility function. For each type $\theta \in \Theta$, let $X_\theta \subseteq X$ be the set of strategies with this fixed-point property:

$$X_\theta = \{x \in X : x \in \beta_\theta(x)\} \,. \tag{10}$$

In particular, let $X_\sigma$ be the fixed-point set of *homo hamiltoniensis*, the set of *Hamiltonian strategies.*

Third, letting $B^{NE}(s) \subseteq X^2$ denote the set of (Bayesian) Nash equilibria in population state $s = (\theta, \tau, \varepsilon)$, that is, all solutions $(x^*, y^*)$ of (3), one may show the following by standard arguments (see Appendix for a proof):

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

We henceforth assume that the type space $\Theta$ contains *homo hamiltoniensis.* Let $\Theta_\sigma^m$ be the set of types $\tau$ that, as vanishingly rare mutants, respond to a resident playing some Hamiltonian strategy by the same token:

$$\Theta_\sigma^m = \{\tau \in \Theta : (x_\sigma, x_\sigma) \text{ satisfies (9) for some } s = (\theta, \tau, 0) \text{ and } x_\sigma \in X_\sigma\} \,. \tag{11}$$

The type space will be said to be *rich* if, for each strategy there is some type for which this strategy is (strictly) dominant. Formally, for each $x \in X$ there exists some $\theta \in \Theta$ such that

$$u_\theta(x, y) > u_\theta(x', y) \quad \forall x' \neq x, y \in X. \tag{12}$$

Such a type $\theta$ will be said to be *committed* to its strategy $x$.[12]

[12] For example, $u_\theta(x', y') = -(x - x')^2$ for all $x', y' \in X$.

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

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

$$F_\theta(x, y, \varepsilon) = [1 - \varepsilon + \varepsilon\phi(\varepsilon)] \cdot \pi(x, x) + \varepsilon[1 - \phi(\varepsilon)] \cdot \pi(x, y) \tag{13}$$

and

$$F_\tau(x, y, \varepsilon) = (1 - \varepsilon)[1 - \phi(\varepsilon)] \cdot \pi(y, x) + [\varepsilon + (1 - \varepsilon)\phi(\varepsilon)] \cdot \pi(y, y) \tag{14}$$

Since $\pi$ and $\phi$ are continuous by hypothesis, so are $F_\theta, F_\tau : X^2 \times [0, 1] \to \mathbb{R}$.

For the first claim, let $(x^*, y^*)$ be a Nash equilibrium in population state $s = (\sigma, \tau, 0)$. Then $x^* \in X_\sigma$. In particular, $u_\sigma(x^*, x^*) \geq u_\sigma(y^*, x^*)$, and if $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, this inequality holds strictly if $\tau \notin \Theta_\sigma^m$: $u_\sigma(x^*, x^*) > u_\sigma(y^*, x^*)$, or, equivalently, $\pi(x^*, x^*) > (1 - \sigma) \cdot \pi(y^*, x^*) + \sigma \cdot \pi(y^*, y^*)$. By definition of $F_\sigma$ and $F_\tau$, we thus have

$$F_\sigma(x^*, y^*, 0) > F_\tau(x^*, y^*, 0) \tag{15}$$

for all $(x^*, y^*) \in B^{NE}(\sigma, \tau, 0)$ and any $\tau \notin \Theta_\sigma^m$. By continuity of $F_\sigma$ and $F_\tau$, this strict inequality holds for all $(x, y, \varepsilon)$ 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. 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 $F_\sigma(x_t, y_t, \varepsilon_t) > F_\tau(x_t, y_t, \varepsilon_t)$, establishing the first claim.[13]

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 $\Theta$ 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 $x_\theta$ is the unique solution to the first condition in (9). Moreover, $u_\sigma(x_\theta, x_\theta) < u_\sigma(\hat{x}, x_\theta)$ is equivalent with

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

[13] 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, $\theta$ is evolutionarily stable against $\tau$ also in this case.

which in turn is equivalent with $F_\theta(x_\theta, \hat{x}, 0) < F_{\hat{\tau}}(x_\theta, \hat{x}, 0)$. In other words, in the limit when $\varepsilon = 0$, the mutant $\hat{\tau}$ earns a higher payoff than the resident $\theta$. By continuity of $F_\theta$ and $F_{\hat{\tau}}$, 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. Hence, 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 $F_\theta(x_t, \hat{x}, \varepsilon_t) < F_{\hat{\tau}}(x_t, \hat{x}, \varepsilon_t)$, establishing the second claim. **Q.E.D.**

Theorem 1 establishes that *homo hamiltoniensis* is favored by evolution and that certain other types are selected against. The first claim expresses that *homo hamiltoniensis* resists "invasions" by all types who do not, as mutants, respond by playing *homo hamiltoniensis'* own strategy. The intuition is that the unique "evolutionarily optimal" mutant response (that is, in terms of the mutant population's average payoff) to a resident Hamiltonian strategy, is that same strategy. The second claim expresses that if the type space 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 somewhat relaxed, but at a high price in terms of analytical complexity.[14] 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 hamiltoniensis'* 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 who is committed to such an evolutionarily superior strategy, and thus will use it against any resident, who will then lose out in terms of payoffs.[15] 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 hamiltoniensis*:

**Corollary 1** *If $\Theta$ is rich, $\theta \in \Theta$ is evolutionarily stable and $X_\theta = \{x_\theta\}$, then $x_\theta \in X_\sigma$.*

Note that the theorem does not require existence of Nash equilibria in the population states under consideration (though existence follows from standard assumptions, see Lemma 1). Stability is defined, and proved, by imposing and verifying conditions on the payoffs in

[14] For a type $\theta$ that does not have a unique resident strategy, the Nash equilibrium correspondence may "explode" at $\varepsilon = 0$. If this happens, the resident'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 it does not explode at that point. We conjecture that the present proof, *mutatis mutandis*, will then go through.

[15] In some applications *homo hamiltoniensis* would also be a successful mutant. However, since little is known in general about equilibrium behavior of *homo hamiltoniensis* in population states where it is a rare mutant, its success as a mutant here is not guaranteed.

those Nash equilibria that do exist. Hence, if for some types $\theta, \tau \in \Theta$ there would exist no Nash equilibrium in any state $(\theta, \tau, \varepsilon)$ with $\varepsilon$ small, then $\theta$ will be deemed evolutionarily stable against $\tau$ according to our definition by "walk over." If this turns out to be a real issue in some application, one can of course restrict the type space to concave utility functions, since this guarantees the existence of Nash equilibria in all population states.

**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 $C'(2m) > 2B'(2m)$. 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, $x_\kappa$, 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 hamiltoniensis*, 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 $\Theta$ is rich, and $\theta \in \Theta$ is any type that has a unique resident strategy and this differs from $x_\sigma$, then $\theta$ is evolutionarily unstable.*

## 3.1 Homo oeconomicus

Theorem 1 may be used to pin down evolutionary stability properties of *homo oeconomicus.* The most general formulation is as follows:

**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 $\tau \notin \Theta_0^m$. If $\sigma > 0$ and $\Theta$ is rich, then* homo oeconomicus *is evolutionarily unstable if it has a unique resident strategy and this does not belong to $X_\sigma$.*

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 *homo oeconomicus* is that the index of assortativity be zero. This result is in line with Ok and Vega-Redondo (2001) and Dekel *et al.* (2007), who both analyze the evolution of preferences under incomplete information and uniform random matching.

The second part says that if the index of assortativity is positive, then *homo oeconomicus* is evolutionarily unstable when it has a unique resident strategy and this is not Hamiltonian. To further clarify the implications of this result, we distinguish two classes of interactions, according to whether or not an individual's payoff depends on the other individual's strategy.

First, consider payoff functions with no dependence on the other individual's strategy. For each individual 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, or, in other words, to act like *homo oeconomicus*. As a result, *homo oeconomicus* can thrive even if the index of assortativity is positive, $\sigma > 0$.

**Corollary 3** *Suppose that $\pi(x, y)$ is independent of $y$. Then* homo oeconomicus *is evolutionarily stable against all types $\tau \notin \Theta_0^m$ 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 types who fail to maximize their payoff. The reason is clear: such interactions, are, in effect, isolated decision problems.

Secondly, consider situations in which one's payoff does depend on the other individual's strategy and exhibits decreasing returns to one's own. 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. Assuming that $\pi$ is twice differentiable:

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

## 3.2 Strategy evolution

Our model differs from classical evolutionary game theory in two ways. First, classical evolutionary game theory views strategies, not preferences or utility functions, as the replicators, the objects that spread in populations of pairwise interacting individuals. 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 space is rich. In such situations one may identify each type with a strategy and *vice versa*, and hence write $\Theta = X$. We call this setting *strategy evolution*, since it is, in effect, as if evolution operated at the level of strategies in the underlying game in payoffs.[16]

Identifying types with strategies, our general definition of evolutionary stability, under random matching with assortment function $\phi$, 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 (1) and (2), with

[16] While classical evolutionary game theory concerns mixed strategies in finite games, it is here immaterial if the stratgies are pure or mixed, and we focus on mixed strategies in finite games in a separate section below.

$\theta$ being the type committed to $x$ and $\tau$ the type committed to $y$. The difference function $S_{x,y}(\varepsilon) \equiv F_\theta(x,y,\varepsilon) - F_\tau(x,y,\varepsilon)$ (with $\theta$ committed to $x$ and $\tau$ to $y$) 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 consider random matching with assortment function $\phi$. A strategy $x \in X$ is* ***evolutionarily stable against a strategy*** *$y \in X$ if there exists an $\bar{\varepsilon} \in (0,1)$ such that $S_{x,y}(\varepsilon) > 0$ for all $\varepsilon \in (0,\bar{\varepsilon})$. A strategy $x$ is* ***evolutionarily stable*** *if it is evolutionarily stable against all $y \neq x$ in $X$.*

We immediately obtain from Theorem 1:[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.*

**Proof:** For the first claim, let $x_\sigma \in X_\sigma$ and $y \neq x_\sigma$. In a population state $s = (x_\sigma, y, \varepsilon)$, the expected payoff to $x_\sigma$ is

$$F_{x_\sigma}(\varepsilon) = [1-\varepsilon+\varepsilon\phi(\varepsilon)] \cdot \pi(x_\sigma, x_\sigma) + \varepsilon[1-\phi(\varepsilon)] \cdot \pi(x_\sigma, y)$$

and that to $y$ is

$$F_y(\varepsilon) = [\varepsilon + (1-\varepsilon)\phi(\varepsilon)] \cdot \pi(y,y) + (1-\varepsilon)[1-\phi(\varepsilon)] \cdot \pi(y, x_\sigma)$$

(see equations (13) and (14)). It follows that $F_{x_\sigma}, F_y : [0,1] \to \mathbb{R}$ are continuous. Hence, a sufficient condition for $x_\sigma$ to be evolutionarily stable against $y$ is that $F_{x_\sigma}(0) > F_y(0)$, or, equivalently,

$$\pi(x_\sigma, x_\sigma) > (1-\sigma) \cdot \pi(y, x_\sigma) + \sigma \cdot \pi(y,y),$$

or $u_\sigma(x_\sigma, x_\sigma) > u_\sigma(y, x_\sigma)$. If $\beta_\sigma(x_\sigma)$ is a singleton, the last inequality holds for all $y \neq x_\sigma$. This establishes the first claim.

For the second claim, let $x \notin X_\sigma$. Then $u_\sigma(x,x) < u_\sigma(\hat{x}, x)$ for some $\hat{x} \in X$. Equivalently,

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

[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 that here *homo hamiltoniensis* is not included in the type space. *Homo hamiltoniensis* is instead represented by one type for each Hamiltonian strategy. The theorem nonetheless refers to the best-reply correpondence $\beta_\sigma$. This is because we allow for a very general class of assortment functions $\phi$. In Proposition 1 we dispense with the assumption of a singleton-valued $\beta_\sigma$ because we there impose more structure on $\phi$.

or, equivalently, $F_x(0) < F_{\hat{x}}(0)$. By continuity of $F_x, F_{\hat{x}} : [0,1] \to R$, this implies that $x$ is evolutionarily unstable. **Q.E.D.**

In other words, every Hamiltonian strategy which is its own unique best reply is evolutionarily stable, and all non-Hamiltonian strategies are evolutionarily unstable. In the special case of uniform random matching, $\sigma = 0$, the Hamiltonian strategies are simply those that are best replies to themselves in terms of payoffs. In the opposite extreme case, $\sigma = 1$, the Hamiltonian strategies are those that, when used by both players, result in Pareto efficiency in terms of payoffs.

**Remark 3** *For payoff functions such that* homo hamiltoniensis *has a unique best reply to all Hamiltonian strategies, Theorem 1 and Corollary 5 establish that preference evolution under incomplete information induces the same behaviors as strategy evolution.*

For certain payoff functions $\pi$, 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 population share $\varepsilon$, a property that holds in certain kinship relations, see Section 7.

**Proposition 1** *Let $\Theta = X$ (strategy evolution) and assume that the assortment function is a constant, $\phi(\varepsilon) \equiv \sigma$. A strategy $x \in X$ is evolutionarily stable if and only if*

$$\pi(x,x) \geq \pi(y,x) + \sigma \cdot [\pi(y,y) - \pi(y,x)] \quad \forall y \in X \tag{16}$$

*and*

$$\begin{aligned} \pi(x,x) &= \pi(y,x) + \sigma \cdot [\pi(y,y) - \pi(y,x)] \\ &\Rightarrow \pi(x,y) > \pi(y,y) + \sigma \cdot [\pi(y,y) - \pi(y,x)]. \end{aligned} \tag{17}$$

**Proof:** Suppose that $\phi(\varepsilon) = \sigma \in [0,1]$ for all $\varepsilon \in (0,1)$. Then

$$\begin{aligned} S_{x,y}(\varepsilon) &= (1 - \varepsilon + \varepsilon\sigma) \cdot \pi(x,x) + \varepsilon(1-\sigma) \cdot \pi(x,y) \\ &\quad - [\varepsilon + (1-\varepsilon)\sigma] \cdot \pi(y,y) - (1-\varepsilon)(1-\sigma) \cdot \pi(y,x), \end{aligned}$$

which defines $S_{x,y}$ (for given $x$ and $y \neq x$) as an affine function of $\varepsilon$. A strategy $x$ is evolutionarily stable iff $S_{x,y}(\varepsilon) > 0$ on some non-empty interval $(0, \bar{\varepsilon})$. A necessary condition for this is clearly $S_{x,y}(0) \geq 0$, or, equivalently, (16). If the latter holds with equality, then it is necessary that the slope of $S_{x,y}$ be positive, or, equivalently, (17). Conversely, if (16) and (17) both hold, then $S_{x,y}(0) \geq 0$, and, in case $S_{x,y}(0) = 0$, the slope of $S_{x,y}$ is positive, so $x$ is evolutionarily stable. **Q.E.D.**

The necessary condition (16) can be written as $x \in X_\sigma$, that is, the strategy must be Hamiltonian. Further, condition (17) 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 agrees with Hines' and Maynard Smith's (1979) analysis of ESS for games played by relatives. See also Grafen (1979, 2006).

# 4 Finite games

The classical 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 thus $A$ be an $m \times m$ matrix, that to each row $i \in S$ and column $j \in S$ assigns the payoff $a_{ij}$ obtained when pure strategy $i$ is used against pure strategy $j$, for all $i, j \in S = \{1, .., m\}$. Permitting players to use mixed strategies, $X$ is now the $(m-1)$-dimensional unit simplex $\Delta(S) = \left\{x \in \mathbb{R}_+^m : \sum_{i \in S} x_i = 1\right\}$, 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 $x \in X = \Delta(S)$ when used against any strategy $y \in X = \Delta(S)$.

Applying our general machinery for preference evolution under incomplete information to finite games, for each type $\theta \in \Theta$ let $u_\theta : X^2 \to R$ be some continuous function, where $X = \Delta(S)$. 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 $y$:

$$u_\kappa(x,y) = (1-\kappa) \cdot xAy + \kappa \cdot xAx = xA[(1-\kappa)y + \kappa x]. \tag{18}$$

For $\kappa > 0$, the utility functions permitted here generically violate the expected-utility hypothesis — which requires linearity with respect to probability distributions. Hence, when we below examine the stability of preferences, this is not only against preferences that meet the von Neumann-Morgenstern axioms, but against preferences in a much wider class. A general stability analysis appears to be a daunting task, while insights and technical difficulties may appear already in simple fitness games, so we here focus on the more restrictive task of identifying the set of *homo-moralis* strategies in $2 \times 2$ 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:

$$x_\kappa \in \arg\max_{x \in [0,1]} \quad (x, 1-x) \cdot \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_\kappa + \kappa(x - x_\kappa) \\ 1 - x_\kappa - \kappa(x - x_\kappa) \end{pmatrix}. \tag{19}$$

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*

$$\hat{x}(\kappa) = \min \left\{ 1, \frac{a_{12} + \kappa a_{21} - (1+\kappa) a_{22}}{(1+\kappa)(a_{12} + a_{21} - a_{11} - a_{22})} \right\}. \tag{20}$$

*(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 (19) can be written as

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

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).

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 minimum (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.**

**Remark 4** *It is seen in (20) that the set of homo-moralis strategies is invariant under positive affine transformations of payoffs. More generally, if $A$ is any $n \times n$ payoff-matrix and $A^* = \lambda A + \gamma E$, where $\lambda$ is any positive scalar, $\gamma$ any positive or negative scalar, and $E$ is the $n \times n$-matrix with all entries equal to 1, then the best-reply correspondence $\beta_\kappa$ associated with $u_\kappa$ in (18) is unaffected if $A$ is replaced by $A^*$.*

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

$$A = \begin{pmatrix} R & S \\ T & P \end{pmatrix} \tag{21}$$

where we assume that $T - R > P - S > 0$. 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 $\kappa$. 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 $\kappa$ between these two bounds. See Figure 1 below, which shows how co-operation increases as the degree of morality increases.

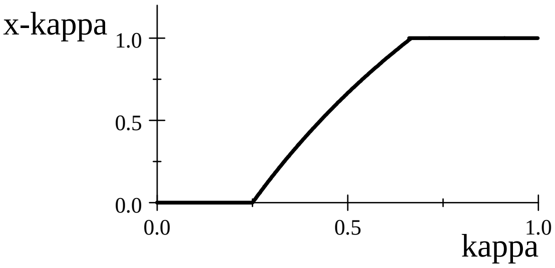

Figure 1: The (singleton) set of homo-moralis strategies for $(T, R, P, S) = (7, 5, 3, 2)$.

# 5 Matching processes

Most of the literature on preference evolution has focused on the case of when all matchings are equally likely.[19] Under such uniform random matching, $\Pr[\theta|\theta,\varepsilon] = \Pr[\theta|\tau,\varepsilon] = 1-\varepsilon$, and hence there is no assortativity: $\phi(\varepsilon) = 0$ for all $\varepsilon \in (0,1)$. Arguably, positive assortativity arises naturally in many, if not most, human interactions, due to socioeconomic population structure, limited social or geographical mobility, habitat preferences, local customs and cultures etc., see Eshel and Cavalli-Sforza (1982) and Bergstrom (1995, 2003). In this section we explore a variety of such possibilities, and identify the associated index of assortativity, $\sigma$.

Broadly speaking, whether preferences are transmitted genetically or culturally, assortativity is positive as soon as there is a positive probability that both parties in an interaction have inherited their preferences (or moral values) from a common "ancestor" (genetic or cultural).

[19] Exceptions are Alger and Weibull (2010, 2011), and Alger (2010), who allow for assortativeness when analyzing the evolution of preferences under complete information. Hines and Maynard Smith (1979), Grafen (1979), Bergstrom (1995, 2003) and Day and Taylor (1998) allow for assortativeness in models of strategy evolution. See Gardner and West (2004) for an account of how negative assortativity may arise.

## 5.1 Interactions between kin

While the following arguments can readily be adapted to interactions between other kin (see Alger and Weibull, 2011), here we study pairwise interactions between siblings, for which preferences are not gender specific. Consider a population of grown-ups where a proportion $1-\varepsilon$ have preferences of type $\theta \in \Theta$ and the residual proportion has preferences of type $\tau \in \Theta$, the same proportion among men as among women, and suppose that couples are formed randomly with respect to their preference types — "random mating". We here show how the index of assortativity between siblings then depends on whether a child inherits his/her type from the parents or from others in society, and if the former case, whether the siblings have the same parents.

### 5.1.1 Vertical transmission

Assume that each child is equally likely to inherit each parent's preference type (and these random draws are statistically independent). The inheritance mechanism could be genetic or cultural.[20] Suppose, first, that siblings have the same parents.

**Proposition 3** *Under random mating and monogamy, $\phi(\varepsilon) = 1/2$ for all $\varepsilon \in (0,1)$.*

**Proof:** Consider a population where a proportion $\varepsilon \in (0,1)$ of the adult population carries type $\tau$, and a proportion $1-\varepsilon$ carries type $\theta$. A $\theta$-child necessarily has at least one $\theta$-parent. With probability $1-\varepsilon$, the other parent also has type $\theta$, in which case both siblings must have type $\theta$. If the other parent has type $\tau$, which happens with probability $\varepsilon$, the other sibling has type $\theta$ with probability $1/2$ and type $\tau$ with probability $1/2$. Hence, the probability that a $\theta$-child's sibling also is a $\theta$-child is $\Pr[\theta|\theta,\varepsilon] = 1-\varepsilon/2$. Similarly, a $\tau$-child has at least one $\tau$-parent. With probability $1-\varepsilon$ the other parent has type $\theta$, in which case the probability that the sibling has type $\theta$ is $1/2$. Hence, the probability that a $\tau$-child's sibling has type $\theta$ is $\Pr[\theta|\tau,\varepsilon] = (1-\varepsilon)/2$. We obtain $\phi(\varepsilon) = \Pr[\theta|\theta,\varepsilon] - \Pr[\theta|\tau,\varepsilon] = 1/2$. **Q.E.D.**

Note that $\sigma = 1/2$ is the *coefficient of relatedness* between siblings (Wright, 1922).

Second, suppose, still, that there is random matching among parents, but that siblings always have different fathers.[21]

[20] 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).

[21] The same result holds under polygyny, that is, when (half-)siblings always have the same father but different mothers.

**Proposition 4** *Under random mating and polyandry, $\phi(\varepsilon) = 1/4$ for all $\varepsilon \in (0,1)$.*

(A proof is given in a working paper available on our personal web sites.) As in the case with full siblings, here $\sigma = 1/4$ is the *coefficient of relatedness* between half-siblings (Wright, 1922).

These results can be combined to calculate the index of assortativity between siblings in a population where some parents remain monogamous while others divorce after the first child and get a new child with their new partner.

**Proposition 5** *Suppose that a fraction $\lambda \in [0,1]$ of couples divorce. Then $\phi(\varepsilon) = 1/2 - \lambda/4$ for all $\varepsilon \in (0,1)$.*

**Proof:** For children whose parents did not divorce, $\phi(\varepsilon) = 1/2$, see Proposition 3. For children whose parents did divorce, $\phi(\varepsilon) = 1/4$, see Proposition 4. On average, then, in the child population: $\phi(\varepsilon) = (1-\lambda)/2 + \lambda/4 = 1/2 - \lambda/4$. **Q.E.D.**

### 5.1.2 Oblique transmission

The above examples apply to both genetic and cultural transmission from parents to children. But "cultural parents" (or role models) other than the parents can also be encompassed in the framework, along the lines suggested by Bisin and Verdier (2001). This may happen when social values spread through institutions other than the family, such as schools, media, religious institutions etc.

**Proposition 6** *Assume monogamy, but suppose now that each child inherits a parent's preferences with probability $\rho \in [0,1]$, that otherwise the child takes on the preferences of a uniformly randomly drawn grown-up, from the population at large, and that the siblings' choices of role model are statistically independent. Then $\sigma = \rho^2/2$.*

**Proof:** Let $\varepsilon \in (0,1)$. For a child born into a family where both children inherited their preferences from their parents, $\phi(\varepsilon) = 1/2$, as in Proposition 3. This is the case for the fraction $\rho^2$ of all sibling pairs. For a child born into a family where at least one child's type was drawn from a random grown-up in the population, $\phi(\varepsilon) \approx \varepsilon$. **Q.E.D.**

## 5.2 Interactions among non-kin

For a variety of interactions among non-kin, education, population structure, social structure, culture, ethnicity, geography, networks, customs and habits, may all lead to various deviations from uniform randomness in the pairwise matchings.

### 5.2.1 Education

Assortativity may arise in business partnerships. To see this, consider a large population engaged in pairwise business partnerships, represented by a symmetric fitness game $\langle X, \pi \rangle$. Suppose that each individual acquires her preferences concerning business strategies in school and enters a new two-person business partnership upon finishing school. Now and then, a school changes the business values they teach.

**Proposition 7** *Let $\upsilon \in [0,1]$ be the probability that a newly minted graduate forms a business partnership with a former schoolmate, and suppose that otherwise the graduate forms a partnership with a graduate uniformly randomly drawn from the whole pool of newly minted graduates in society at large. Then $\sigma = \upsilon$.*

**Proof:** Consider a large collection of schools, where the proportion $1-\varepsilon$ teach a business value system $\theta \in \Theta$ and a proportion $\varepsilon$ teaches a business value system $\theta$. Let $\varepsilon \to 0$. For graduates pairing up with a former schoolmate, $\sigma = 1$. For all other graduates, $\sigma = 0$. Since there is a fraction $\upsilon$ of graduates who pair up with a schoolmate, on average, then, in the population of newly minted graduates, $\sigma = \upsilon$. **Q.E.D.**

### 5.2.2 Migration

In pre-historic societies, assortativity may have come about as a result of migration patterns. Consider a society in which individuals grow up in small communities, where each community has a hunting team consisting of two men from the community, where each community teaches some values system for how to act in a hunting team to their youngsters, and that some young men migrate from one community to another after their training but before they become members of a hunting team (for example because of marriage).

**Proposition 8** *Suppose that a fraction $\gamma \in [0,1]$ of the young men migrate from their native community to a uniformly randomly drawn community in society at large, while the others remain in their native community. Then $\sigma = 1 - \gamma$.*

**Proof:** Consider a large collection of communities, where the proportion $1 - \varepsilon$ teaches hunting values $\theta$ and the proportion $\varepsilon$ teaches hunting values $\tau$. For men who remained in their native community, $\sigma = 1 - \gamma$, while for men who moved, $\sigma = 0$. Since there is a fraction $1 - \gamma$ of young men who remain in their native community, on average, then, in the population of young men, $\sigma = (1 - \gamma)$. **Q.E.D.**

In traditional societies, migration is often linked to marriage, and typically tradition dictates whether the bride or bridegroom migrates. Thus, our model suggests that differences in marriage customs may have affected the evolution of preferences within each gender.

# 6 Discussion

## 6.1 Related literature

The idea that moral values may have been formed by evolutionary forces is evidently not new, and there is a substantial literature on this theme. The idea can be traced back to at least Darwin (1871). More recent treatments include, to mention a few, Alexander (1987), Nichols (2004) and de Waal (2006). The latter claims that there is evidence that moral codes also exist in other primates. In this literature, mathematical analyses are rare, however. An exception is the work by Bergstrom (1995, 2009). Bergstrom (1995) studies strategy evolution in sibling interactions, where strategies are genetically transmitted from parents to children. He finds that evolution favors strategies that are as though individuals had Kantian preferences under asexual reproduction, and "semi-Kantian preferences" under sexual diploid reproduction with recessive genes. This is exactly in line with our findings, since $\sigma = 1$ under asexual reproduction and $\sigma = 1/2$ under sexual diploid reproduction with recessive genes. Bergstrom (2009) extends this reasoning to arbitrary degrees of relatedness, and also provides a thought-provoking discussion and interpretation of various moral maxims.[22]

## 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, Bolle, 2000, Possajennikov, 2000, Alger and Weibull, 2010, 2011, and Alger, 2010). Altruism towards another individual is often represented in parametric form by letting the utility function of the altruist be the sum of two terms, where the first term is his or her own payoff and the other term is the other individual's payoff multiplied by a factor $\alpha \in [0,1]$. In the present context,

$$u_\alpha(x,y) = \pi(x,y) + \alpha\pi(y,x), \tag{22}$$

for some *degree of altruism* $\alpha \in [0,1]$. By contrast, our *homo moralis* has preferences of the form

$$u_\kappa(x,y) = (1-\kappa)\pi(x,y) + \kappa\pi(x,x) \tag{23}$$

for some *degree of of morality* $\kappa \in [0,1]$. Hence, while an altruist cares about the other's payoff, *homo moralis* cares about what is the "right thing to do," irrespective of what the other party actually does or is expected to do. We first show that while in some situations, morality and altruism lead to the same behavior, in others 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 oeconomicus.*

[22] Unlike us, however, Bergstrom (1995, 2009) bases his analysis on pure strategies in finite games, rather than on mixed strategies, as we here do.

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

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

is identical with that for a *homo moralis*,

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

if $\alpha=\kappa$. Nonetheless, 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 (22) and (23): for non-trivial payoff functions $\pi$ and strategy sets $X$, and for any $\alpha,\kappa\neq 0$, there exists no function $T:\mathbb{R}\to\mathbb{R}$ such that $T\left[u_\alpha(x,y)\right]=u_\kappa(x,y)$ for all $x,y\in X$. This is seen most clearly in the case of finite games. Then $u_\alpha$ is linear in $x$ while $u_\kappa$ 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 $\beta_\alpha$ of an altruist in general differs qualitatively from the best-reply correspondence $\beta_\kappa$ 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. Let

$$A=\begin{pmatrix}\delta & 2\\ 1 & 0\end{pmatrix}$$

for some $\delta\in(0,1)$. Consider a *homo kantiensis* ($\kappa=1$), the "most moral" among *homo moralis*. Such a creature will always play

$$x_{\kappa=1}=\frac{3}{6-2\delta}.$$

Suppose now that such an individual visits a country where everyone always plays the first pure strategy, thus earning payoff $\delta$ in each encounter with each other. When *homo kantiensis* interacts with a citizen in that society, the matched native earns more than when interacting with other natives. However, if the visitor instead were a *homo oeconomicus* ($\kappa=0$), then this new visitor would play the second pure strategy. Consequently, the other individual in the match would earn more 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 be even more delighted to interact with *homo oeconomicus* than with *homo kantiensis*. Should then *homo oeconomicus* be deemed "less moral" than *homo kantiensis* in this situation? What 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 always 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 *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 5** *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 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 $\pi'\left(x'_{\kappa=1}, x'_{\kappa=1}\right)$, 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, $\pi'\left(x'_{\kappa=1}, x'_{\kappa=1}\right) = 3/2$, a higher payoff than when* homo kantiensis *meets himself in the original game: $\pi\left(x_{\kappa=1}, x_{\kappa=1}\right) = (2 - 2\delta/3)^{-1} \cdot 3/2$.*

## 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 some two-player game in monetary payoffs (or a few different such games), (b) after the first few rounds of play, under random re-matching, subjects receive some information about aggregate play in these early rounds, and (c) are then invited to play some more rounds (again with randomly drawn opponents). One could then analyze their behavior in these later rounds as if they played a (Bayesian) Nash equilibrium under incomplete information, where each individual is a *homo moralis* with an individual-specific and presumed fixed degree of morality (presumably given from that individual's background, experience and personality). How much of the observed behavior could be explained this way? If one were to embed the simple preferences of homo moralis in a more general class of other-regarding preferences, how much more explanatory power would then be gained? Supposing that the subjects behave as they usually do in similar real-world interactions, one could compare estimates of the degrees of morality in different subject pools and see if this seems to map relevant cultural and socioeconomic differences, in line with *homo hamiltoniensis*.[23]

**Remark 6** *Although in the model above individuals play only one game, the model has clear implications for the more realistic situation where each individual engages in multiple interactions. Indeed, the degree of morality that will be selected for will simply correspond to the index of assortativity in the matching process for the interaction 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 above theory says that one and the same individual will tend to 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 nature of the interaction).*

# 7 Concluding remarks

Economic analysis is based upon assumptions about human motivation. Presumably, higher predictive power can be achieved the better the assumptions reflect the actual motivation. In order to enhance the predictive power of economic models, a deeper understanding of both proximate and ultimate causes of human motivation is necessary. Our research contributes to the understanding of ultimate causes, by proposing a theoretical model of the evolution of preferences.[24] We follow a long tradition in the literature by asking whether evolution will select preferences whereby individuals selfishly maximize their individual fitness payoff.[25] So far, the leading theory for why deviations from such preferences may survive evolution is that, if known or believed by others, such preferences may give its carrier a strategic commitment advantage in terms of payoff consequences. By contrast, we here show that

[23] If there is uniform random matching in the lab, then presumably many individuals will gradually, perhaps quickly for some and slowly (or not at all) for others, change their behavior, during long sessions in the lab, towards one where they behave more like *homo oeconomicus*, as it will in our theory if $\sigma = 0$.

[24] Our theory is, however, silent as to which proximate causes may come into play, and whether it applies only to humans. Proximate causes may include culture (Gächter and Herrmann, 2009), genes (Cesarini et al., 2008), and hormones and neural circuitry (Eisenegger et al, 2010, Fehr and Camerer, 2007, Harbaugh et al., 2007, Moll et al., 2006, Rilling et al., 2002). Pro-social behavior has been observed in other animals, such as rats (Bartal et al., 2011).

[25] In a related literature, on cultural evolution, altruistic parents can, at some cost, 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). In our model, evolution is an exogenously given process, and parents (be they cultural of genetic parents), do not need to be altruistic for non-selfish preferences to be favored by evolution.

deviations from selfish preferences will typically be evolutionarily stable also when preferences are private information, as long as the matching process that governs interactions involves some assortativity.[26] Our theory thus delivers new, testable predictions, regarding human motivation.

Although we permit virtually any preferences (as long as they can be represented by continuous functions), 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 own payoff, given her expectation of the other's action, and the payoff that she would obtain if both individuals were to take the same action. A certain member of this family, *homo hamiltoniensis*, is particularly viable from an evolutionary perspective. The weight that *homo hamiltoniensis* attaches to the second goal is the index of assortativity in the matching process. The viability of *homo hamiltoniensis* stems from the fact that the best a mutant can do, in order to "invade" such a resident population, is to choose the same strategy. Moreover, any resident type that does not play a Hamiltonian strategy is vulnerable to invasion by mutants.

These results have important implications regarding *homo oeconomicus.* We show that *homo oeconomicus*, who seeks to maximize his own payoff, does well in situations where there is no assortativity in the matching process and where each individual's payoff does not depend on others' behavior. By contrast, natural selection wipes out *homo oeconomicus* in all other situations.[27]

As is common in the literature, in our model evolutionary success is determined by behavior in a symmetric interaction. As illustrated by our analyses of the dictator and ultimatum-bargaining games, the symmetry assumption does not need to apply in a literal sense, however. The model applies to interactions where there is some asymmetry between the individuals' situations (say, helping interactions where one individual happens to be rich and the other poor), as long as from an *ex ante* perspective it is not known which individual will be in which situation. In such cases, evolution selects preferences that favors the individual, behind the veil of ignorance regarding which situation the individual will eventually end up in.

While the predictive power of preferences *à la homo moralis* remains to be analyzed carefully, we argue that at first glance the behavior of *homo moralis* seems to be broadly compatible with experimental evidence. What's more, *homo moralis* may explain why many

[26] This paper complements our work in Alger and Weibull (2010, 2011), and Alger (2010), where we studied evolutionary stability of preferences under assortative matching and complete information. In that setting, other factors in the environment (in particular the structure of the payoffs in the interactions driving evolution) may also affect the evolution of preferences.

[27] It is important to point out that maximizing own fitness payoff involves acting in an other-regarding manner if breeding is cooperative or reproduction involves mate competition. For analysis of this, see Weibull and Salomonsson (2006).

subjects justify their behavior in the lab by saying that they wanted to "do the right thing" (see, e.g., Dawes and Thaler, 1988, Charness and Dufwenberg, 2006). While we leave theoretical analyses of the policy implications of such moral preferences for future research, we note that in our model the degree of morality, that is, the weight attached to the second goal, is independent of the interaction at hand. Hence, the degree of morality cannot be "crowded out" in any direct sense by economic incentives or laws. For instance, if one were to change an interaction (for example public goods provision) by way of paying people for "doing the right thing" (say, contributing the socially optimal amount), or by way of punishing them for doing otherwise, this would change the payoff function, and thus also the behavior of *homo moralis*, but in an easily predictable way, since *homo moralis* cares neither about other's opinion of her, nor about other's actual payoffs. Moreover, our theory 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.

Some new ground has been covered here, but many deep and important questions about ultimate causes behind human motivation remain unanswered. For instance, what are the effects of group size? Can the theory be generalized so as to explain stable preference heterogeneity in populations? What happens in truly asymmetric interactions, such as between parents and children, men and women, or individuals in hierarchies?

# 8 Appendix: Proof of Lemma 1

By hypothesis, $u_\theta$ and $u_\tau$ are continuous and $X$ is compact. Hence, each right-hand side in (3) 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 (3) 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. It follows that $B_\varepsilon$ 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 $u_\theta$ and $u_\tau$ are concave in their first arguments, then so are the maximands in (3). Hence, $B_\varepsilon$ 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 $\theta$ and $\tau$, and write the maximands in (3) 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

$$\begin{cases} U^*(x^*, y^*, \varepsilon) - U(x, x^*, y^*, \varepsilon) \geq 0 & \forall x \in X \\ V^*(x^*, y^*, \varepsilon) - U(y, x^*, y^*, \varepsilon) \geq 0 & \forall y \in X. \end{cases} \tag{24}$$

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 (24),

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

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

# References

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

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

Alger, Ingela, and Jörgen W. Weibull. 2010. "Kinship, Incentives, and Evolution," *American Economic Review*, 100:1725-1758.

Alger, Ingela, and Jörgen W. Weibull. 2011. "A Generalization of Hamilton's Rule—Love Others how much?" *Journal of Theoretical Biology*, forthcoming.

Arrow, Kenneth. 1973. "Social responsibility and economic efficiency," *Public Policy*, 21:303-317.

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

Bartal, Inbal, Jean Decety, and Peggy Mason. 2011. "Empathy and Pro-Social Behavior in Rats," *Science*, 334:1427-1430.

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

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

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

Bergstrom, Theodore C. 2009. "Ethics, Evolution, and Games among Neighbors," Working Paper, UCSB.

Bester, Helmut, and Werner Güth. 1998. "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization,* 34:193–209.

Binmore, Ken. 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.

Bomze, Immanuel M., and Benedikt M. Pötscher. 1989. *Game Theoretical Foundations* *of Evolutionary Stability.* New York: Springer.

Border, Kim C. and Charalamros D. Aliprantis. 2006. *Infinite Dimensional Analysis.* 3rd ed. New York: Springer.

Brekke, Kjell Arne, Snorre Kverndokk, and Karine Nyborg. 2003. “An Economic Model of Moral Motivation,” *Journal of Public Economics*, 87:1967–1983.

Cesarini, D., C. T. Dawes, J. Fowler, M. Johannesson, P. Lichtenstein, and B. Wallace. 2008. “Heritability of Cooperative Behavior in the Trust Game,” *Proceedings of the National Academy of Sciences*, 105:3271-3276.

Charness, Gary and Martin Dufwenberg. 2006. “Promises and Partnership,” *Econometrica*, 74:1579–1601.

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

Dawes, Robyn and Richard Thaler. 1988. “Anomalies: Cooperation,” *Journal of Economic Perspectives*, 2:187-97.

Day, Troy, and Peter D. Taylor. 1998. “Unifying Genetic and Game Theoretic Models of Kin Selection for Continuous types.” *Journal of Theoretical Biology*, 194:391-407.

Dekel, Eddie, Jeffrey C. Ely, and Okan 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.

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

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., and E. Fehr. 2010. “Prejudice and Truth about the Effect of Testosterone on Human Bargaining Behaviour,” *Nature,* 463:356-359.

Ellingsen, Tore. 1997. “The Evolution of Bargaining Behavior,” *Quarterly Journal of Economics*, 112:581-602.

Eshel, Ilan, and Luigi Luca Cavalli-Sforza. 1982. “Assortment of Encounters and Evolution of Cooperativeness,” *Proceedings of the National Academy of Sciences,* 79:1331-1335.

Fehr, E. and C.F. Camerer. 2007. “Social Neuroeconomics: The Neural Circuitry of Social Preferences” *TRENDS in Cognitive Sciences*, 11:419-426.

Fehr, E. and S. Gächter. 2000. “Cooperation and Punishment in Public Goods Experiments” *American Economic Review*, 90:980-994.

Fershtman, Chaim, and Yoram Weiss. 1998. “Social Rewards, Externalities and Stable Preferences,” *Journal of Public Economics*, 70:53-73.

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

Gächter, S. and B. Herrmann. 2009. "Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment" *Philosophical Transactions of the Royal Society B*, 364:791-806;

Gächter, S., B. Herrmann, and C. Thöni. 2004. "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence" *Journal of Economic Behavior and Organization*, 55:505-531.

Gardner, Andy, and Stuart A. West. 2004. "Spite and the Scale of Competition," *Journal of Evolutionary Biology*, 17:1195–1203.

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

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

Güth, Werner, and Menahem 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, William D. 1964a. "The Genetical Evolution of Social Behaviour. I." *Journal of Theoretical Biology*, 7:1-16.

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

Harbaugh, W.T., U. Mayr, and D.R. Burghart. 2007. "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations" *Science*, 316:1622-1625.

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

Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. 2006. "The Dynamic Evolution of Preferences," *Economic Theory*, 32:251-286.

Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. 2007. "What to Maximize if You Must," *Journal of Economic Theory*, 133:31-57.

Hines, W. Gord S., and John Maynard Smith. 1979. "Games between Relatives," *Journal of Theoretical Biology*, 79:19-30.

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

Huck, Steffen, Dorothea Kübler, and Jörgen W. Weibull. 2011. "Social Norms and Economic Incentives in Firms," Working Paper, Stockholm School of Economics.

Huck, Steffen, and Jörg Oechssler. 1999. "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior,* 28:13–24.

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

Koçkesen, Levent, Efe A. Ok, and Rajiv Sethi. 2000. "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory,* 92:274-299.

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

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, John. 1974. "The Theory of Games and the Evolution of Animal Conflicts", *Journal of Theoretical Biology,* 47, 209-221.

Maynard Smith, John, and George R. Price. 1973. "The Logic of Animal Conflict," *Nature,* 246:15-18.

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

Moll, J., F. Krueger, R. Zahn, M. Pardini, R. de Oliveira-Souza, and J. Grafman. 2006. "Human Fronto-mesolimbic Networks Guide Decisions about Charitable Donation" *Proceedings of the National Academy of Sciences*, 103:15623-15628.

Nichols, Shaun. 2004. *Sentimental Rules.* Oxford: Oxford University Press.

Ok, Efe A., and Fernando 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.

Rilling, J.K., D.A. Gutman, T.R. Zeh, G. Pagnoni, G.S. Berns, and C.D. Kilts. 2002. "A Neural Basis for Social Cooperation" *Neuron*, 35:395–405.

Robson, Arthur J. 1990. "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144:379-396.

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

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

Schelling, Thomas. 1960. *The Strategy of Conflict.* Cambridge: Harvard University Press.

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

Sethi, Rajiv, and E. Somanathan. 2001. “Preference Evolution and Reciprocity” *Journal of Economic Theory,* 97:273-297.

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

Tabellini, Guido. 2008. “Institutions and Culture,” *Journal of the European Economic Association,* 6:255–294.

Weibull, Jörgen W. 1995. *Evolutionary Game Theory.* Cambridge: MIT Press.

Weibull, Jörgen W. and Marcus Salomonsson. 2006. “Natural Selection and Social Preferences” *Journal of Theoretical Biology*, 239:79-92.

Wright, Sewall G. 1922. “Coefficients of Inbreeding and Relationship,” *American Naturalist,* 56:330-338.