Games and Economic Behavior 98 (2016) 56–67

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Evolution and Kantian morality ☆

Ingela Alger [a,b,c,*], Jörgen W. Weibull [d,c,e,a]

[a] Toulouse School of Economics, France
[b] CNRS, France
[c] Institute for Advanced Study in Toulouse, France
[d] Stockholm School of Economics, Sweden
[e] KTH Royal Institute of Technology, Sweden

ARTICLE INFO

Article history:
Received 28 April 2015
Available online 9 June 2016

JEL classification:
C73
D01
D03

Keywords:
Preference evolution
Evolutionary stability
Assortativity
Morality
Homo moralis
Social preferences

ABSTRACT

What kind of preferences should one expect evolution to favor? We propose a definition of evolutionary stability of preferences in interactions in groups of arbitrary finite size. Groups are formed under random matching that may be assortative. Individuals' preferences are their private information. The set of potential preferences are all those that can be represented by continuous functions. We show that a certain class of such preferences, that combine self-interest with morality of a Kantian flavor, are evolutionarily stable, and that preferences resulting in other behaviors are evolutionarily unstable. We also establish a connection between evolutionary stability of preferences and a generalized version of Maynard Smith's and Price's (1973) notion of evolutionary stability of strategies.

1. Introduction

Economics traditionally takes individuals' motivations—their preferences—as given. Hence, the predictive power of economic models depends on the assumptions made regarding these motivations. But if preferences are inherited from past generations, a theory of their evolutionary foundation is called for. In particular, one may ask what preferences have survival value in a given society. Should we expect pure self-interest, altruism (Becker, 1976), warm glow (Andreoni, 1990), reciprocal altruism (Levine, 1998), inequity aversion (Fehr and Schmidt, 1999), self-image concerns (Bénabou and Tirole, 2006), moral motivation (Brekke et al., 2003), or something else? This question is at the heart of the literature on preference evolution, initiated by Güth and Yaari (1992). In a recent contribution to this literature (Alger and Weibull, 2013), we showed that, for pairwise interactions evolution favors a class of social preferences, the carriers of which we called *Homo moralis*. However,

☆ Support by Knut and Alice Wallenberg Research Foundation and by ANR-Labex IAST is gratefully acknowledged. We also thank Agence Nationale de la Recherche for funding (Chaire d'Excellence ANR-12-CHEX-0012-01 for I. Alger, and Chaire IDEX ANR-11-IDEX-0002-02 for J. Weibull). We are grateful for comments from an anonymous referee and from seminar participants at EconomiX, WZB, GREQAM, ETH Zürich, Nottingham, OECD, Séminaire Roy (PSE), IGIER Bocconi, Bern, École Polytechnique, University of Amsterdam, Banco de México, and from participants at the 2nd Toulouse Economics and Biology Workshop, the 2014 EEA meeting, the 2014 ASSET meeting (Aix-en-Provence), the Conference in Honor of Peyton Young (Oxford), the 15th SAET Conference (Cambridge), the ESOP workshop on "Work motivation and the Nordic model" (Oslo), and the "Panorama of Mathematics" conference at the Hausdorff Center for Mathematics (Bonn) in October 2015.

* Corresponding author at: Toulouse School of Economics, France.
E-mail address: ingela.alger@tse-fr.eu (I. Alger).

http://dx.doi.org/10.1016/j.geb.2016.05.006
0899-8256/© 2016 Published by Elsevier Inc.

many interactions involve more than two persons. Can the definitions, methods and results for pairwise interactions be generalized to interactions in groups of arbitrary size? Like in Alger and Weibull (2013), we here let the mathematics show the way from evolutionary stability conditions to the preferences that evolution favors.

More precisely, we develop a theoretical model for analysis of the evolutionary foundations of human preferences in strategic interactions in arbitrarily large groups, of given finite size, under assortative random matching. Individuals in an infinite population are randomly matched in groups of a given size to play a symmetric and aggregative game in material payoffs (or fitness).[1] Each individual is endowed with a (subjective) goal function that expresses his or her preferences over strategy profiles in the interaction at hand. Goal functions are private information. We allow for a wide range of symmetric aggregative interactions, including complex intertemporal interactions of arbitrary length. The only requirement is that each player's strategy set be a compact and convex set in some normed vector space. We define a goal function to be *evolutionarily stable* against another goal function if, in every (Bayesian) Nash equilibrium in every population state where the latter goal function is sufficiently rare, individuals endowed with the former goal function materially outperform those with the latter. Conversely, a goal function is *evolutionarily unstable* if there exists another goal function such that, no matter how small its population share, there is some (Bayesian) Nash equilibrium in which the latter goal function materially outperforms the former one.

A key feature of our model is that it allows the random matching to be assortative in the sense that individuals who are of a vanishingly rare ("mutant") type may face a positive probability of being matched with others of their own type, even in the limit as their type vanishes. While such matching patterns may at first appear counter-intuitive or even impossible, it is not difficult to think of reasons for why they can arise. First, while distance is not explicitly modeled here, geographic, cultural, linguistic and socio-economic distance imposes (literal or metaphoric) transportation costs, which imply that (1) individuals tend to interact more with individuals in their (geographic, cultural, linguistic or socio-economic) vicinity,[2] and (2) cultural or genetic transmission of types (say, behavior patterns, preferences or moral values) from one generation to the next also has a tendency to take place in the vicinity of where the type originated.[3] Taken together, these two tendencies may generate the assortativity that we here allow for. We formalize the assortativity of a random matching process in terms of a vector we call the *assortativity profile*. This is the probability vector for the events that none, some, or all the individuals in a (vanishingly rare) mutant's group also are mutants. This generalizes Bergstrom's (2003) definition of the index of assortativity for pairwise encounters.[4]

Our analysis delivers three main results. First, although we impose virtually no restrictions on potential goal functions, hence allowing for an infinite-dimensional type space, evolution favors a particular finite-dimensional class of goal functions. Individuals with preferences in this class attach some weight to their own material self-interest but also to what can be interpreted as a probabilistic version of Kantian morality.[5] To be more precise; in his *Grundlegung zur Metaphysik der Sitten* (1785), Immanuel Kant wrote "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." Similarly, *Homo moralis* attaches some weight to the goal of "acting according to that maxim whereby you can, at the same time, will that others should do likewise with some probability." More precisely, a *Homo moralis* individual in a group of arbitrary size $n > 1$ maximizes a weighted average of equally many terms, indexed $k = 0, 1, \ldots, n-1$, where each term is the material payoff that she would obtain if, hypothetically, she could replace the strategies of so many other individuals in the group by her strategy. We call the vector of these weights the individual's *morality profile*. *Homo moralis* holds correct expectations about others' actions and yet partly evaluates her own actions in this (hypothetical) probabilistic Kantian sense. In other words, she is concerned with the morality of her own acting, irrespective of what others do.[6] What action would she prefer if, hypothetically, also others would choose the same action in her situation? We show that, in order for such preferences to be evolutionarily stable, the morality profile must equal the assortativity profile.[7]

Our second main result is that any preferences that induce equilibrium behaviors that differ from those of *Homo moralis* with the stable morality profile are evolutionarily unstable. In particular, then, our results imply that *Homo oeconomicus*—pure material self-interest—is evolutionarily unstable under any random matching process in which the probability is positive that at least one other group member of a vanishingly rare mutant's group is also is a mutant.

[1] The notion of aggregative games is due to Dubey et al. (1980). See also Corchón (1996). The key feature is that the payoff to a player is a function of the player's own strategy and some aggregate of all others' strategies, see e.g. Acemoglu and Jensen (2013). For models closer to ours, see Haigh and Cannings (1989) and Koçkesen et al. (2000a, 2000b).

[2] Homophily has been documented by sociologists (e.g., McPherson et al., 2001, and Ruef et al., 2003) and economists (e.g., Currarini et al., 2009, 2010, and Bramoullé and Rogers, 2009).

[3] In biology, the concept of assortativity is known as *relatedness*, and the propensity to interact with individuals locally is nicely captured in the infinite island model, originally due to Wright (1931); see also Rousset (2004). In Lehmann et al. (2015) we carry out an analysis in this vein.

[4] See Bergstrom (2013) and Alger and Weibull (2013) for further discussions of assortativity under pairwise matchings.

[5] For a discussion of several ethical principles, see Bergstrom (2009). See also Bergstrom (1995).

[6] *Homo moralis* preferences are distinct from the equilibrium concept "Kantian equilibrium" proposed by Roemer (2010). Moreover, *Homo moralis* is not subject to "magical reasoning" or the "transparent disposition fallacy," see discussion in Binmore (1994, Ch. 3).

[7] Clearly, no goal function is evolutionarily stable against all other goal functions, since distinct functions may generate the same behavior. However, *Homo moralis* with the right morality profile is evolutionarily stable against all goal functions that are not its "behavioral alikes," i.e., goal functions that induce behaviors that would also be rational for a *Homo moralis* with the right morality profile.

The third main result in this study is that the equilibrium strategies used by *Homo moralis* with morality profile equal to the assortativity profile coincides with the set of evolutionarily stable strategies, as defined by Maynard Smith and Price (1973) for pairwise uniform random matching. Hence, evolutionarily stable strategies in the classical sense can be interpreted as the result of free choice of rational individuals whose preferences have emerged from natural selection. We establish operational first- and second-order conditions for the equilibrium behavior of *Homo moralis* for the special case of differentiable material payoff functions and conditionally independent random matching.

The rest of the paper is organized as follows: the model is specified in Section 2, the main result is presented in Section 3, strategy evolution is analyzed in Section 4, first- and second-order conditions are given in Section 5, the literature is reviewed in Section 6, and Section 7 concludes.

2. Model

Maynard Smith and Price (1973) defined evolutionary stability as a property of (pure or mixed) strategies in finite and symmetric two-player games. With X denoting each player's set of mixed strategies and $\pi(x, y) \in \mathbb{R}$ denoting the payoff to strategy x against strategy y, they called a strategy x *evolutionarily stable* if it is a best reply to itself and a strictly better reply to all its alternative best replies than these are to themselves. In an infinitely large population of uniformly randomly matched paired individuals who play a finite game, this definition is equivalent to the existence of a positive invasion barrier against every "mutant" strategy $y \neq x$, such that if the population share of y is below its invasion barrier, then it earns a lower average payoff than x.

We generalize this stability criterion to obtain a notion of evolutionary stability of (subjective) *goal functions* for players in symmetric n-player games with general strategy spaces X, under random matching that may be assortative. To be more precise, we consider n-player games (for any $n \geq 1$) in which each player has the same set X of (pure or mixed) strategies, and $\pi(x, \boldsymbol{y}) \in \mathbb{R}$ is the *material payoff* to strategy $x \in X$ when used against strategy profile $\boldsymbol{y} = (y_1, \ldots, y_{n-1}) \in X^{n-1}$. If the players are firms the material payoff is *profit,* if the players are individuals it is the *fitness effect* upon the individual. We assume π to be continuous, and *aggregative* in the sense that $\pi(x, \boldsymbol{y})$ is invariant under permutation of the components of $\boldsymbol{y}$. The strategy set X is taken to be a non-empty, compact and convex set in some normed vector space. This generality of the nature of the strategy set allows for a large variety of interactions, and evidently generalizes that of standard evolutionary game theory, where $n = 2$ and X is the unit simplex of mixed strategies in a finite and symmetric two-player game (and hence X is a compact and convex set in a Euclidean space).

We will henceforth consider as given such a game in material payoffs. Let F denote the associated set of continuous and aggregative functions $f : X^n \to \mathbb{R}$.[8] An individual's *goal function* is such a function $f \in F$, to also be called the individual's *utility function* or *type*. An example is $f = \pi$, that is, an individual whose goal function coincides with his or her material payoff, a type we call *Homo oeconomicus*.

We analyze interactions in which each individual's type is his or her private information. Consequently, an individual's strategy choice cannot be conditioned on the types of the others with whom (s)he has been matched to play. However, an individual's strategic behavior may be adapted to the statistical frequency of types in her random matches, and such an adaptation may be used to define evolutionary stability of a type, a topic to which we now turn.

2.1. Matching

For the purpose of defining evolutionary stability it is sufficient, as in the original set-up of Maynard Smith and Price (1973), to consider populations in which only one or two types are present. For any types $f, g \in F$, and any $\varepsilon \in [0, 1]$, let $s = (f, g, \varepsilon)$ be the *population state* in which the two types are represented in population shares $1 - \varepsilon$ and ε, respectively. Let $S = F^2 \times [0, 1]$ denote the *state space*, the set of possible population states. We are particularly interested in states $s = (f, g, \varepsilon)$ in which ε is positive and small. By convention, we then call f the *resident* (or *incumbent*) type and g, being rare, the *mutant* type.

The matching process is exogenous. In any population state $s = (f, g, \varepsilon) \in S$, the number of mutants—individuals of type g—in a group that is about to play the material-payoff game is a random variable that we denote T. For any *resident* drawn at random from the population let $p_m(\varepsilon)$ be the conditional probability $\Pr[T = m \mid f, s]$ that the number of mutants in the resident's group is m, for $m = 0, 1, .., n - 1$.[9] Likewise, for any mutant, also drawn at random from the population, let $q_m(\varepsilon)$ be the conditional probability $\Pr[T = m + 1 \mid g, s]$ that the number of *other* mutants in his or her group is m, again for $m = 0, .., n - 1$. Let $\boldsymbol{p}(\varepsilon) = (p_0(\varepsilon), \ldots, p_{n-1}(\varepsilon))$ and $\boldsymbol{q}(\varepsilon) = (q_0(\varepsilon), \ldots, q_{n-1}(\varepsilon))$ be the so defined probability distributions, both belonging to the $((n-1)$-dimensional) unit simplex Δ^{n-1} in $\mathbb{R}^n$. We assume that $\boldsymbol{p}(\varepsilon)$ and $\boldsymbol{q}(\varepsilon)$ are continuous in the mutant population share $\varepsilon \in (0, 1)$ and that they converge to $\boldsymbol{p}^* \in \Delta^{n-1}$ and $\boldsymbol{q}^* \in \Delta^{n-1}$, respectively, as $\varepsilon \to 0$.

---

[8] Extensions to more general goal functions are conceivable but appear to call for a considerably more involved analysis, perhaps leading to the same results as we now have. In order to define non-aggregative goal functions some or all group members would have to be exogenously assigned distinct player roles and allowed to value others' strategy choice differently depending on player roles.

[9] The first random draw cannot, technically, be uniform, in an infinite population. The reasoning in this section is concerned with matchings in finite populations in the limit as the total population size goes to infinity. We refer the reader to the appendix for a detailed example.

In order to get a grip on these limiting probabilities, we use the *algebra of assortative encounters* developed by Bergstrom (2003) for pairwise interactions. For a given population state $s = (f, g, \varepsilon)$, let $\Pr[f|f, \varepsilon]$ denote the conditional probability for an individual of the resident type f that another, uniformly randomly drawn member of his or her group also is of the resident type. Likewise, let $\Pr[f|g, \varepsilon]$ be the conditional probability for an individual of the mutant type g that another, uniformly randomly drawn member of his or her group is of the resident type f. Let $\phi(\varepsilon)$ be the difference between these conditional probabilities:

$$\phi(\varepsilon) = \Pr[f|f, \varepsilon] - \Pr[f|g, \varepsilon]. \tag{1}$$

This defines the *assortment function* $\phi : (0, 1) \to [-1, 1]$. We assume that this function is continuous and that it has a limit value as the mutant share tends to zero:

$$\lim_{\varepsilon \to 0} \phi(\varepsilon) = \sigma, \tag{2}$$

for some $\sigma \in \mathbb{R}$, the *index of assortativity* of the matching process (Bergstrom, 2003). Moreover, by setting $\phi(0) = \sigma$ we continuously extend ϕ from its domain $(0, 1)$ to the domain $[0, 1)$.

The following equation is a necessary balancing condition for all $\varepsilon \in (0, 1)$:

$$(1 - \varepsilon) \cdot [1 - \Pr[f|f, \varepsilon]] = \varepsilon \cdot \Pr[f|g, \varepsilon]. \tag{3}$$

Each side of the equation equals the probability for the following event: draw at random an individual from the population at large and then draw at random another individual from the first individual's group, and observe that one individual is a resident and the other a mutant. Equations (1) and (3) together give

$$\begin{cases} \Pr[f|f, \varepsilon] = \phi(\varepsilon) + (1 - \varepsilon)[1 - \phi(\varepsilon)] \\ \Pr[f|g, \varepsilon] = (1 - \varepsilon)[1 - \phi(\varepsilon)]. \end{cases} \tag{4}$$

Now let $\varepsilon \to 0$. From (3) we obtain $\Pr[f|f, \varepsilon] \to 1$, and hence $\boldsymbol{p}(\varepsilon) \to \boldsymbol{p}^* = (1, 0, 0, \ldots, 0) \in \Delta^{n-1}$. In other words, residents virtually never meet mutants when the latter are vanishingly rare. It follows from (4) that $\sigma \in [0, 1]$.[10]

Turning now to the limit vector $\boldsymbol{q}^* \in \Delta^{n-1}$, which we call the *assortativity profile* of the matching process, we note that in the special case of pairwise interactions ($n = 2$), $\boldsymbol{q}^* = (1 - \sigma, \sigma)$, since then:

$$q_1^* = \lim_{\varepsilon \to 0} \Pr[g|g, \varepsilon] = 1 - \lim_{\varepsilon \to 0} \Pr[f|g, \varepsilon] = 1 - \left[1 - \lim_{\varepsilon \to 0} \phi(\varepsilon)\right] = \sigma.$$

However, for $n > 2$ there remains a statistical issue, namely whether or not the types of *other* members of a mutant's group are statistically dependent or not (in the given population state). Later on, we will consider the case of conditional statistical independence. However, in the general analysis we make no assumption about dependence or independence of the types of other group members.

Just as with the assortativity function ϕ, we continuously extend the domain also of $\boldsymbol{p}(\cdot)$ and $\boldsymbol{q}(\cdot)$ from $(0, 1)$ to $[0, 1)$ by setting $\boldsymbol{p}(0) = \boldsymbol{p}^*$ and $\boldsymbol{q}(0) = \boldsymbol{q}^*$. *Uniform* random matching is the special case when $\boldsymbol{q}^* = \boldsymbol{p}^*$, that is, when the limiting conditional matching-probabilities for mutants are the same as for residents.

2.2. Equilibrium play

In any given population state $s = (f, g, \varepsilon) \in F^2 \times (0, 1)$, a type-homogeneous (Bayesian) Nash equilibrium is a *pair* of strategies, one for each type, such that each type's strategy is a best reply, in terms of the goal function for any individual of that type in the given population state. Let $(\hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}})$ denote any vector $\boldsymbol{y} \in X^{n-1}$ with m components equal to $\hat{y}$ and the remaining components equal to $\hat{x}$. Using our continuous extensions of the matching probabilities to include the limit case when $\varepsilon = 0$, we may define equilibrium for all population states with a non-negative share of mutants:

Definition 1. A strategy pair $(\hat{x}, \hat{y}) \in X^2$ is a type-homogeneous (Bayesian) **Nash Equilibrium** in population state $s = (f, g, \varepsilon) \in F^2 \times [0, 1)$ if

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot f\left(x, (\hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}})\right) \\ \hat{y} \in \arg\max_{y \in X} \sum_{m=0}^{n-1} q_m(\varepsilon) \cdot g\left(y, (\hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}})\right). \end{cases} \tag{5}$$

Let $B^{NE}(s) \subseteq X^2$ denote the set of such Nash equilibria in population state s. For any given types f and g, this defines an *equilibrium correspondence* $B^{NE}(f, g, \cdot) : [0, 1) \rightrightarrows X^2$ that maps every mutant population share $\varepsilon \in [0, 1)$ to the associated set of equilibria. We note that $B^{NE}(s)$ is compact (by compactness of X and continuity of f and g). Moreover, if f

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

and g are concave in their first argument (for any given vector of others' strategies), then $B^{NE}(s)$ is non-empty, by standard arguments.[11] We also note that if $\varepsilon = 0$, then the first equation in (5) is equivalent with (type-homogeneous) Nash equilibrium play among residents in the absence of mutants (and is hence independent of the mutant type g). Formally, $\varepsilon = 0 \Rightarrow \hat{x} \in \beta_f(\hat{x})$, where, for any goal function $f \in F$, the best-reply correspondence $\beta_f : X \rightrightarrows X$ is defined by

$$\beta_f(x) = \arg\max_{y \in X} f(y, (x, x, \ldots, x)). \tag{6}$$

We call the set $X(f) = \{x \in X : x \in \beta_f(x)\}$ the set of *residential equilibrium strategies* associated with the goal function f.

By a slight generalization of the arguments in the proof of Lemma 1 in Alger and Weibull (2013) one obtains that the equilibrium correspondence $B^{NE}(f, g, \cdot) : [0, 1) \rightrightarrows X^2$ is upper hemi-continuous, a fact we will use when analyzing evolutionary stability.

Let $(\hat{x}, \hat{y}) \in B^{NE}(s)$ for some population state. The associated average *equilibrium material payoffs* to residents and mutants are

$$\Pi_R(\hat{x}, \hat{y}, \varepsilon) = \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot \pi\left(\hat{x}, \left(\hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}}\right)\right) \tag{7}$$

and

$$\Pi_M(\hat{x}, \hat{y}, \varepsilon) = \sum_{m=0}^{n-1} q_m(\varepsilon) \cdot \pi\left(\hat{y}, \left(\hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}}\right)\right), \tag{8}$$

respectively. Both quantities are continuous in $(\hat{x}, \hat{y}, \varepsilon) \in X^2 \times [0, 1)$ by virtue of the assumed continuity of the material-payoff function and the matching probabilities.

2.3. Evolutionary stability

The following definitions of evolutionary stability and instability are generalizations of the definitions in Alger and Weibull (2013).

Definition 2. A goal function $f \in F$ is **evolutionarily stable** against a goal function $g \in F$ if there exists an $\bar{\varepsilon} > 0$ such that $\Pi_R(\hat{x}, \hat{y}, \varepsilon) > \Pi_M(\hat{x}, \hat{y}, \varepsilon)$ for all $\varepsilon \in (0, \bar{\varepsilon})$ and all $(\hat{x}, \hat{y}) \in B^{NE}(f, g, \varepsilon)$. A goal function $f \in F$ is **evolutionarily unstable** if there exists a goal function $g \in F$ such that for all $\bar{\varepsilon} > 0$ there is a population state $(f, g, \varepsilon) \in F^2 \times (0, 1)$ with $\varepsilon < \bar{\varepsilon}$ and a strategy pair $(\hat{x}, \hat{y}) \in B^{NE}(f, g, \varepsilon)$ such that $\Pi_M(\hat{x}, \hat{y}, \varepsilon) > \Pi_R(\hat{x}, \hat{y}, \varepsilon)$.

In other words, the requirement for stability is that the residents should earn a higher material payoff in all Nash equilibria for all sufficiently small population shares of the mutant type. By contrast, the requirement for instability is that there should exist at least one mutant type that would earn a higher material payoff in some Nash equilibrium when mutants are arbitrarily rare. Note that we do not require that populations should always play Nash equilibria. For stability we simply require that *if* this were the case, *then* residents must outperform mutants materially. Likewise, for instability we require that there exists a mutant type and a sequence of Nash equilibria, with smaller and smaller population shares of this mutant, such that *if* one of these Nash equilibria were played, then the mutants would outperform residents materially.

Clearly, no goal function is evolutionarily stable against all other goal functions in F, for the simple reason that (infinitely many) goal functions result in precisely the same behavior in all situations, and hence in precisely the same material payoffs. For example, consider $g = \lambda f + c$ for some $\lambda > 0$ and $c \in \mathbb{R}$. More generally, for any $f \in F$, let $A(f) \subseteq F$ be the set of "behavioral alikes" of f. This is the set of mutant goal functions $g \in F$ such that for *some* Nash equilibrium among the residents there is strategy that is also a best reply for a mutant. Formally,

$$A(f) = \left\{g \in F : \exists (\hat{x}, \hat{y}) \in B^{NE}(f, g, 0) \text{ where } \hat{y} \in \beta_f(\hat{x})\right\}.$$

In particular, $f \in A(f)$ for every goal function $f \in F$, and $g \in A(f)$ if $g = \lambda f + c$ for some $\lambda > 0$ and $c \in \mathbb{R}$. Other examples are goal functions g of the form $g(x, \mathbf{y}) \equiv -\|x - \hat{x}\|$ for some $\hat{x} \in X(f)$, that is, individuals for whom one of f's residential equilibrium strategies is strictly dominant. Then clearly $g \in A(f)$.

Definition 3. $f \in F$ is **evolutionarily stable** if it is evolutionarily stable against all $g \notin A(f)$.[12]

[11] This follows from Berge's maximum theorem combined with the Fan–Glicksberg–Kakutani fixed-point theorem, theorems that hold when $X \neq \varnothing$ is a compact and convex set in a locally convex Hausdorff vector space (hence in particular in a normed vector space), see e.g. Aliprantis and Border (2006).

[12] This weaker definition than that in Alger and Weibull (2013) allows a less involved stability statement in the theorem.

2.4. *Homo moralis*

We need only one more ingredient before we can state and prove our main result, namely, a generalized definition of "*Homo moralis*" (Alger and Weibull, 2013).

Definition 4. An individual is a **Homo moralis** with **morality profile** $\boldsymbol{\mu} \in \Delta^{n-1}$ if his or her goal function f satisfies

$$f(x, \boldsymbol{y}) = \mathbb{E}[\pi(x, \boldsymbol{Y})] \quad \forall (x, \boldsymbol{y}) \in X^n, \tag{9}$$

where $\boldsymbol{Y}$ is a random $(n-1)$-vector such that with probability μ_m exactly $m \in \{0, \dots, n-1\}$ of the $n-1$ components of $\boldsymbol{y}$ are replaced by x, with equal probability for each subset of size m, while the remaining components of $\boldsymbol{y}$ keep their original values.

Clearly, *Homo oeconomicus* ($f = \pi$) is a *Homo moralis* goal function, namely, the one with morality profile $\boldsymbol{\mu} = (1, 0, \dots, 0)$, so that $\Pr[\boldsymbol{Y} = \boldsymbol{y}] = 1$. At the opposite extreme of the spectrum of *Homo moralis* we find *Homo kantientis*, the variety of *Homo moralis* that has morality profile $\boldsymbol{\mu} = (0, \dots, 0, 1)$, so that $\Pr[\boldsymbol{Y} = (x, x, \dots, x)] = 1$. We write f_K for this goal function, defined by $f_K(x, \boldsymbol{y}) = \pi(x, (x, x, \dots, x))$ for all $x \in X$ and $\boldsymbol{y} \in X^{n-1}$. Individuals of this "pure Kantian" type always choose a strategy that, if hypothetically adopted by everyone in the group, would maximize all group members' material payoffs. As mentioned in the introduction, such reasoning is in line with Immanuel Kant's categorical imperative. The behavior of all other varieties of *Homo moralis* lies between these two extremes. *Homo moralis* with morality profile $\boldsymbol{\mu} \in \Delta^{n-1}$ behaves as if she followed a probabilistic version of Kant's categorical imperative; she evaluates the strategies at her disposal in the light of what would happen in the hypothetical scenario in which others would probabilistically use her strategy, according to the probability distribution $\boldsymbol{\mu}$.[13]

3. Main result

We are now in a position to state and prove our main result, namely, that *Homo moralis* with the assortativity profile of the matching process as its morality profile is evolutionarily stable against all types that are not its behavioral alikes, and that any type that does not behave like this particular variety of *Homo moralis* when resident is unstable. Write $f^* \in F$ for the goal function of a *Homo moralis* with morality profile $\boldsymbol{\mu} = \boldsymbol{q}^*$.

Theorem 1. *f^* is evolutionarily stable. Any $f \in F$ with $X(f) \cap X(f^*) = \varnothing$ is evolutionarily unstable.*

Proof. For the first claim, let $f = f^*$ and $g \notin A(f)$, and suppose that $(x, y) \in B^{NE}(f, g, 0)$. Since $g \notin A(f)$, $f(x, (x, \dots, x)) > f(y, (x, \dots, x))$. By definition of f^*, the last inequality is equivalent with $D(x, y) > 0$, where $D : X^2 \to \mathbb{R}$ is defined by $D(x, y) = \Pi_R(x, y, 0) - \Pi_M(x, y, 0)$. By continuity of Π_M and Π_R, D is continuous. Since $B^{NE}(f, g, 0) \subseteq X^2$ is compact and $D(x, y) > 0$ on $B^{NE}(f, g, 0)$, there exists, by Weierstrass' maximum theorem, a $\delta > 0$ such that $D(x, y) \geq \delta$ for all $(x, y) \in B^{NE}(f, g, 0)$. Again by continuity of Π_R and Π_M, there exists a neighborhood $U \subseteq X^2 \times [0, 1)$ of the compact set $B^{NE}(f, g, 0) \times \{0\}$ such that $\Pi_R(x, y, \varepsilon) - \Pi_M(x, y, \varepsilon) > \delta/2$ for all $(x, y, \varepsilon) \in U$. Moreover, since $B^{NE}(f, g, \cdot) : [0, 1) \rightrightarrows X^2$ is compact-valued and upper hemi-continuous, there exists an $\bar{\varepsilon} > 0$ such that $B^{NE}(f, g, \varepsilon) \times [0, \varepsilon] \subset U$ for all $\varepsilon \in [0, \bar{\varepsilon})$. Thus $\Pi_R(x, y, \varepsilon) - \Pi_M(x, y, \varepsilon) > \delta/2$ for all $\varepsilon \in [0, \bar{\varepsilon})$ and $(x, y) \in B^{NE}(f, g, \varepsilon)$. This establishes the first claim.

For the second claim, let $f \in F$ be such that $X(f) \cap X(f^*) = \varnothing$, and let $\hat{x} \in X(f)$. Then $f^*(\tilde{x}, (\hat{x}, \dots, \hat{x})) > f^*(\hat{x}, (\hat{x}, \dots, \hat{x}))$ for some $\tilde{x} \in X$. Since F is the set of all continuous (aggregative) functions, there exists a type $g \in F$ for which $\tilde{x}$ is a strictly dominant strategy (for example $g(x, \boldsymbol{y}) \equiv -\|x - \tilde{x}\|$), so individuals of that type will always play $\tilde{x}$. By definition of f^*,

$$\Pi_M(\hat{x}, \tilde{x}, 0) = f^*(\tilde{x}, (\hat{x}, \dots, \hat{x})) > f^*(\hat{x}, (\hat{x}, \dots, \hat{x})) = \Pi_R(\hat{x}, \tilde{x}, 0).$$

Let $\langle \varepsilon_t \rangle_{t \in \mathbb{N}}$ be any sequence from $(0, 1)$ such that $\varepsilon_t \to 0$. By upper hemi-continuity of $B^{NE}(f, g, \cdot)$ there exists a sequence $\langle x_t, y_t \rangle_{t \in \mathbb{N}}$ from X^2 such that $(x_t, y_t) \in B^{NE}(f, g, \varepsilon_t)$ for all $t \in \mathbb{N}$ and $x_t \to \hat{x}$ for some $\hat{x} \in X(f)$. By definition of the mutant type g, $y_t = \tilde{x}$ for all $t \in \mathbb{N}$. Since Π_R and Π_M are continuous, there exists a $T > 0$ such that $\Pi_M(x_t, \tilde{x}, \varepsilon_t) > \Pi_R(x_t, \tilde{x}, \varepsilon_t)$ for all $t > T$. Hence, f is evolutionarily unstable. □

The theorem establishes that evolutionary stability favors *Homo moralis* preferences with a morality profile that precisely reflects the assortativity of the matching process. If there is assortativity in the sense that $q_0^* < 1$, this result implies some morality in the sense that the random vector $\boldsymbol{Y}$ in the definition of *Homo moralis* satisfies $\Pr[\boldsymbol{Y} = \boldsymbol{y}] < 1$. The intuition for this result is that in a population that consists almost solely of *Homo moralis* with the "right" morality profile, these individuals use some strategy that would maximize the average material payoff to a vanishingly rare mutant who would

[13] These intermediate cases are most easily seen in the case of pairwise interactions. For $n = 2$, equation (9) boils down to $f(x, y) = \mu_0 \cdot \pi(x, y) + \mu_1 \cdot \pi(x, x)$, a convex combination of selfishness and morality, with weight $\mu_1 \in [0, 1]$ to the latter. In Alger and Weibull (2013), we call μ_1 the *degree of morality*.

enter this population. Stated differently, a population consisting of such *Homo moralis* preempts entry by rare mutants, rather than doing what would be best, in terms of their own material payoff, for the residents in the absence of mutants.[14] Importantly, then, both the result and its intuition is very different from that of group selection.

The theorem is built upon a definition of evolutionary stability that takes as the relevant set of material payoffs those that arise in Nash equilibria of the game under incomplete information. The first claim in the theorem—about the stability of *Homo moralis* with morality profile equal to the assortativity profile—evidently holds for any refinement of Nash equilibrium, such as perfect or proper equilibria, or sequential or perfect Bayesian equilibria in the extensive form of the game in question. Likewise, the second claim in the theorem—about the evolutionary instability of behaviorally distinct other goal functions—evidently holds for any coarsening of Nash equilibrium, such as rationalizability. Moreover, it is not difficult to verify that this second claim, about instability, in fact holds for any upper hemi-continuous refinement of the set of Nash equilibria. We also note that the proof of the theorem only concerns strategy profiles with at most two distinct components. Hence, the theorem remains valid if one were to accordingly weaken the definition of *Homo moralis*.

The result in Alger and Weibull (2013) provided insight about whether the assumption of selfishness, which is common in economics, has an evolutionary justification when individuals interact in pairs. Our present result generalizes this insight to interactions with $n \geq 2$ players.[15] In a nutshell, the theorem says that if preferences are unobservable, selfishness, that is, the goal function $f_E = \pi$, is evolutionarily stable (modulo behavioral alikes) if and only if there is no assortativity at all in the matching process, i.e., $q_0^* = 1$. Furthermore, and importantly, our result gives the exact form in which morality may be expected to operate—and why—in groups of arbitrary size.

To see why the two-player case did not provide a clear indication as to how morality would appear in larger groups, recall that for $n = 2$, (9) boils down to

$$f(x, y) = \mu_0 \cdot \pi(x, y) + \mu_1 \cdot \pi(x, x).$$

From this expression it is not clear whether, when $n > 2$ individuals should maximize a convex combination of selfishness, $\pi(x, \mathbf{y})$, and the categorical imperative, $\pi(x, (x, x, \ldots, x))$, or some other function. Our result shows that unless $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_{n-1}) \in \Delta^{n-1}$ is such that $\mu_0 + \mu_{n-1} = 1$, individuals will also attach weights to the material payoffs that would arise should, hypothetically, subsets of the others use the same strategy.

In general, *Homo moralis* may have a utility function that appears to be fairly involved. However, this need not be the case. Consider, for instance, the case of conditional independence in the random matching. By this we mean that, for a given mutant who has just been matched, the types of any two *other* members in her group are statistically independent, at least in the limit as the mutant becomes rare. Clearly, this restriction on the nature of the matching process is vacuous in the case of pairwise matching. However, for matchings into larger groups, it renders the family of *Homo moralis* preferences one-dimensional. To see this, let $\sigma \in [0, 1]$ be as defined in (2). Conditional independence in the limit as the mutant goes extinct gives a binomial limit distribution,

$$q_m^* = \lim_{\varepsilon \to 0} \binom{n-1}{m} (\Pr[g|g, \varepsilon])^m (\Pr[f|g, \varepsilon])^{n-m-1} = \binom{n-1}{m} \sigma^m (1-\sigma)^{n-m-1}, \tag{10}$$

for any $n \geq 2$ and $m = 0, 1, \ldots, n-1$.[16]

Hence, under conditional independence, evolution favors *Homo moralis* preferences with a particularly simple morality profile, $\boldsymbol{\mu} = Bin(n-1, \sigma)$, one that (for a given group size) can be described with a single parameter, $\sigma \in [0, 1]$. The goal of such an individual is to maximize his or her expected material payoff if, hypothetically, each other member of his or her group would statistically independently switch to her strategy with probability σ. From a mathematical viewpoint, the *Homo moralis* family then defines a homotopy (see e.g. Munkres, 1975) parametrized by σ, between pure selfishness $\sigma = 0$ and pure Kantian morality, $\sigma = 1$. In this one-dimensional case, we will refer to σ as the *degree of morality*.

4. Strategy evolution

Here we adopt the assumption in the original formulation of evolutionary stability (Maynard Smith and Price, 1973), namely, that an individual's type is a strategy that she always uses. A question of particular interest here is whether strategy evolution gives guidance to the behaviors that result under preference evolution.

Formally, we can embed strategy evolution as a special case of preference evolution by restricting the type space from the set F of all goal functions to a subset of goal functions that all render a particular strategy strictly dominant, and such that there for each strategy exists at least one such goal function. The simplest subset of goal functions with this property is, arguably,

$$\hat{F} = \left\{ f \in F : f(x, \mathbf{y}) \equiv -\left\| x - \hat{x} \right\| \text{ for some } \hat{x} \in X \right\}.$$

[14] See also Alger and Weibull (2013) and Robson and Szentes (2014) for a similar observation.

[15] Note that, by contrast to Alger and Weibull (2013), here the theorem does not require uniqueness of the best response of *Homo moralis*. This is because here our definition of the set of behavioral alikes is wider.

[16] In the appendix we present a matching process with this conditional statistical independence property.

Thus, in a population state $s = (f, g, \varepsilon)$ where $f, g \in \hat{F}$, some strategy $\hat{x} \in X$ is always played by the residents and some strategy $\hat{y} \in X$ is always played by the mutants. Hence, the associated set $B^{NE}(s)$ of type-homogeneous (Bayesian) Nash equilibria is then the singleton set $\{(\hat{x}, \hat{y})\}$, and the above machinery applies. In particular, a strategy $x \in X$ is *evolutionarily stable* against a strategy $y \in X$ if there exists an $\bar{\varepsilon} > 0$ such that $\Pi_R(x, y, \varepsilon) > \Pi_M(x, y, \varepsilon)$ for all $\varepsilon \in (0, \bar{\varepsilon})$, and a strategy $x \in X$ is *evolutionarily unstable* if there exists a strategy $y \in X$ such that for all $\bar{\varepsilon} > 0$ there exists an $\varepsilon \in (0, \bar{\varepsilon})$ such that $\Pi_M(x, y, \varepsilon) > \Pi_R(x, y, \varepsilon)$. A strategy $x \in X$ will be called *evolutionarily stable* if it is evolutionarily stable against all strategies $y \neq x$. This terminology agrees with the standard evolutionary game-theory terminology (see Hines and Maynard Smith, 1979, and Broom et al., 1997). In particular, we obtain the classical definition of Maynard Smith and Price (1973) as the special case when $n = 2$ and the random matching is uniform, $q_0^* = 1$.

A *necessary* condition for $x \in X$ to be evolutionarily stable is

$$\Pi_R(x, y, 0) \geq \Pi_M(x, y, 0) \qquad \forall y \in X. \tag{11}$$

Likewise, a *sufficient* condition for evolutionary stability is that this inequality holds strictly for all strategies $y \neq x$.

Since $\Pi_M(x, x, 0) = \Pi_R(x, x, 0) = \Pi_R(x, y, 0)$, the necessary condition (11) for a strategy x to be evolutionarily stable may be written

$$x \in \arg\max_{y \in X} \Pi_M(x, y, 0). \tag{12}$$

This condition says that for a strategy x to be evolutionarily stable, its users have to earn the same average material payoff as "the most threatening mutants", those with the highest average material payoff that any vanishingly rare mutant can obtain against the resident. As under preference evolution, then, an evolutionarily stable type *preempts* entry by rare mutant types.

Likewise, the sufficient condition for a strategy x to be evolutionarily stable can be written

$$\Pi_M(x, x, 0) > \Pi_M(x, y, 0) \quad \forall y \neq x. \tag{13}$$

Interestingly, then, irrespective of group size n, evolutionarily stable strategies may be interpreted as Nash equilibrium strategies in a derived two-player game, where "nature" plays strategies against each other, and where the payoff from playing strategy y against strategy x is $V(y, x) \equiv \Pi_M(x, y, 0)$. Let $\Gamma = \langle N, X^n, \pi \rangle$ denote the game in material payoffs and $\Gamma_V = \langle \{1, 2\}, X^2, V \rangle$ the derived game.

Proposition 1. *If $x \in X$ is an evolutionarily stable strategy in Γ, then $(x, x) \in X^2$ is a Nash equilibrium of Γ_V. If $(x, x) \in X^2$ is a strict Nash equilibrium of Γ_V, then x is an evolutionarily stable strategy in Γ, while if $(x, x) \in X^2$ is a not a Nash equilibrium of Γ_V, then x is evolutionarily unstable in Γ.*

This proposition allows us to make a connection between strategy evolution and *Homo moralis* preferences. Indeed, while under strategy evolution each individual mechanically plays a certain strategy—is "programmed" to execute a certain strategy—we will now see that any evolutionarily stable strategy may be viewed as emerging from rational individuals' free choice when striving to maximize a specific utility function. To see this, note that thanks to permutation invariance,

$$V(y, x) = \sum_{m=0}^{n-1} q_m^* \cdot \pi\left(y, \left(\mathbf{y}^{(m)}, \mathbf{x}\right)\right), \tag{14}$$

where $\left(\mathbf{y}^{(m)}, \mathbf{x}\right)$, like before, is any vector $\boldsymbol{y} \in X^{n-1}$ with m components equal to $y \in X$ and the other components equal to $x \in X$. Combining this observation with Proposition 1 and the fixed-point equation (12) we obtain:

Corollary 1. *If x is evolutionarily stable, then $x \in X(f^*)$. If $x \in X(f^*)$ and $\beta_{\boldsymbol{f}^*}(x)$ is a singleton, then x is evolutionarily stable. Every strategy $x \notin X(f^*)$ is evolutionarily unstable.*

This corollary establishes that the behavior induced under strategy evolution is *as if* individuals were equipped with *Homo moralis* preferences with a morality profile that exactly matches the assortativity profile. More precisely, in games where *Homo moralis* of morality profile $\boldsymbol{\mu} = \boldsymbol{q}^*$ has a unique best reply to each residential equilibrium strategy, preference evolution under incomplete information induces the same behaviors as strategy evolution.

5. Differentiability

In this section we derive a result that allows to easily characterize the set of evolutionarily stable strategies for the case of differentiable material payoff functions and conditionally independent random matching when the set X is finite-dimensional.

Suppose that X is a non-empty subset of $\mathbb{R}^k$ for some $k \in \mathbb{N}$. We will say that x is *strictly evolutionarily stable* (SES) if (13) holds for all $y \neq x$, and we will call a strategy $x \in X$ *locally strictly evolutionarily stable* (LSES) if (13) holds for all $y \neq x$

in some neighborhood of x. If, moreover, $\pi : X^n \to \mathbb{R}$ is differentiable, then so is $V : X^2 \to \mathbb{R}$, and standard calculus can be used to find evolutionarily stable strategies. Let $\nabla_y V(y,x)$ be the gradient of V with respect to y. We call this the *evolution gradient*; it is the gradient of the (average) material payoff to a mutant strategy y in a population state with residents playing x, and vanishingly few mutants. Writing "$\cdot$" for the inner product and boldface $\mathbf{0}$ for the origin, the following result follows from standard calculus[17]:

Proposition 2. *Let $X \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$, and let $x \in int(X)$. If $V : X^2 \to \mathbb{R}$ is continuously differentiable on a neighborhood of $(x,x) \in X^2$, then condition (i) below is necessary for x to be LSES, and conditions (i) and (ii) are together sufficient for x to be LSES. Furthermore, any strategy x for which condition (i) is violated is evolutionarily unstable.*

(i) $\nabla_y V(y,x)_{|y=x} = \mathbf{0}$,
(ii) $(x-y)\cdot\nabla_y V(y,x) > 0$ *for all* $y \neq x$ *in some neighborhood of x.*

The first condition says that there should be no direction of marginal improvement in material payoff for a rare mutant at the resident type. The second condition ensures that if some nearby rare mutant $y \neq x$ were to arise in a vanishingly small population share, then the mutant's material payoff would be increasing in the direction leading back to the resident type, x.

Conditions (i) and (ii) in Proposition 2 can be used to obtain remarkably simple and operational conditions for evolutionarily stable strategies if the strategy set X is one-dimensional ($k=1$) and π is continuously differentiable. Writing π_j for the partial derivative of π with respect to its jth argument, and $\hat{\mathbf{x}}$ for the n-dimensional vector whose components all equal $\hat{x}$, one obtains[18]:

Proposition 3. *Assume conditionally independent matching with index of assortativity σ, and suppose that π is continuously differentiable on a neighborhood of $\hat{\mathbf{x}} \in X^n$, where $X \subseteq \mathbb{R}$. If $\hat{x} \in int(X)$ is evolutionarily stable, then*

$$\pi_1\left(\hat{\mathbf{x}}\right) + \sigma\cdot(n-1)\cdot\pi_n\left(\hat{\mathbf{x}}\right) = 0. \tag{15}$$

Proof. If π is continuously differentiable, V is continuously differentiable. Hence, if $x \in int(X)$, Proposition 2 holds, and the following condition is necessary for x to be an evolutionarily stable strategy:

$$\nabla V_y(y,x)_{|y=x} = \sum_{m=0}^{n-1}\binom{n-1}{m}\sigma^m(1-\sigma)^{n-m-1}\left[\sum_{j=1}^{m}\pi_j\left(y,\left(\mathbf{y}^{(m)},\mathbf{x}\right)\right)\right]_{|y=x} = 0.$$

Since π is aggregative, this equation may be written

$$\sum_{m=0}^{n-1}\binom{n-1}{m}\sigma^m(1-\sigma)^{n-m-1}\left[\pi_1(\mathbf{x}) + m\cdot\pi_n(\mathbf{x})\right] = 0, \tag{16}$$

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

$$\sum_{m=0}^{n-1}\binom{n-1}{m}\sigma^m(1-\sigma)^{n-m-1}m = (n-1)\cdot\sigma,$$

the expression in (16) simplifies to $\pi_1(\mathbf{x}) + (n-1)\cdot\sigma\cdot\pi_n(\mathbf{x}) = 0$. $\square$

Let f_σ denote the *Homo moralis* goal function with degree of morality σ (that is, $\boldsymbol{\mu} = Bin(n-1,\sigma)$) with associated best-reply correspondence β_σ and set of residential equilibrium strategies X_σ. Together with Corollary 1 and Proposition 2, the preceding proposition implies:

Corollary 2. *Suppose that $X \subset \mathbb{R}$ is an open set, that π is continuously differentiable, and that $\beta_\sigma(x)$ is a singleton for each $x \in X_\sigma$. Then the set X_σ coincides with the set of evolutionarily stable strategies. Moreover, each $x \in X_\sigma$ satisfies (15).*

[17] See, e.g., Theorem 2 in Section 7.4 of Luenberger (1969), which also shows that Proposition 2 in fact holds when the gradient is the Gateaux derivative in general vector spaces.

[18] Symmetry of π implies that $\pi_n\left(\hat{\mathbf{x}}\right) = \pi_j\left(\hat{\mathbf{x}}\right)$ for all $j > 1$.

6. Literature

When introduced by Maynard Smith and Price (1973) the concept of evolutionary stability was defined as a property of *mixed strategies* in *finite and symmetric two-player games* played under *uniform random matching* in an *infinite population*, where uniform random matching means that the probability of meeting a mutant is the same for all individuals in the population. Broom et al. (1997) generalized Maynard Smith's and Price's original definition to *finite and symmetric n-player games*, for $n \geq 2$ arbitrary, while maintaining the assumption of uniform random matching in an infinite population.[19] They noted the combinatorial complexity entailed by this generalization, and reported some new phenomena that can arise when interactions involve more than two parties. Evolutionary stability and asymptotic stability in the replicator dynamic, in the same setting, was further analyzed by Bukowski and Miekisz (2004). Schaffer (1988) extended the definition of Maynard Smith and Price to the case of uniform random matching in *finite populations*, and also considered interactions involving all individuals in the population ("playing the field"). Grafen (1979) and Hines and Maynard Smith (1979) generalized the definition of Maynard Smith and Price from uniform random matching to the kind of *assortative matching* that arises when strategies are genetically inherited and games are played among kin. As mentioned in Section 4, our model generalizes most of the above work within a unified framework.

In a pioneering study, Güth and Yaari (1992) defined evolutionary stability for parametrized *utility functions*, assuming uniform random matching and complete information, that is, every player knows the type of the other player in the match.[20] This approach is often referred to as "indirect evolution." The literature on preference evolution now falls into four broad classes, depending on whether the focus is on interactions where information is complete[21] or incomplete,[22] and whether non-uniform random matching is considered.[23] Few models deal with interactions involving more than two individuals. Like here, the articles in this category focus exclusively on interactions that are symmetric and aggregative in material payoffs, the payoffs that drive evolution. Unlike us, they restrict attention to uniform random matching. Koçkesen et al. (2000a, 2000b) show that under complete information about opponents' preferences, players with a specific kind of interdependent preferences fare better materially than players who seek to maximize their material payoff. Sethi and Somanathan (2001) go one step further and provide sufficient conditions for a population of individuals with the same degree of reciprocity to withstand the invasion of selfish individuals, again in a complete-information framework. By contrast, Ok and Vega-Redondo (2001) analyze the case of incomplete information. They identify sufficient conditions for a population of selfish individuals to withstand the invasion by non-selfish individuals, and for selfish individuals to be able to invade a population of identical non-selfish individuals.

7. Conclusion

To understand human societies it is necessary to understand human motivation. In this paper we build on a large literature in biology and in economics, initiated by Maynard Smith and Price (1973), to propose a theoretical framework within which one may study the evolution of human motivation—preferences—by way of natural selection. The set of potential preferences is taken to be the set of all continuous and aggregative preferences over strategy profiles. Our analysis shows that a particular preference class, that we call *Homo moralis*, comes out as a clear winner in the evolutionary race. An individual with such preferences maximizes a weighted sum of the material payoff that she would obtain in a hypothetical probabilistic scenario in which none, some, or all the individuals with whom she interacts would also use her strategy; the weights represent the probabilities, and we call the hypothetical probability distribution the individual's *morality profile*.

Although quite general, our model relies on a number of simplifying assumptions, such as symmetry. Relaxation of these assumptions, as well as the question of how preferences evolve when interacting individuals can partially or fully observe each other's preferences, is a task that has to be left for future research. Yet another challenge would be to investigate evolutionary neutrality, setwise evolutionary stability and/or evolutionary stability properties of heterogeneous populations.

For the past twenty years or so economists have proposed varieties of pro-social or other-regarding preference in order to explain certain observed behaviors, mostly in laboratory experiments but sometimes in the field, that are at odds with maximization of one's own material payoff. Our research has so far delivered two results of relevance for behavioral economics. First, the result that natural selection selects preferences with a distinct Kantian flavor; it is as if individuals in their strategy choice attach some importance to what would happen if others, hypothetically, would take the same actions as they do. Second, we have the result that the importance that individuals attach to this Kantian morality aspect depends on the assortativity in the matching process, and is independent of the interaction in question. Since historically, assortativity arguably has varied between populations and over time (depending on geography, technology and social structure), this

[19] Precursors to their work are Haigh and Cannings (1989), Cannings and Whittaker (1995) and Broom et al. (1996).

[20] See also Frank (1987).

[21] See Robson (1990), Güth and Yaari (1992), Ockenfels (1993), Huck and Oechssler (1999), Ellingsen (1997), Bester and Güth (1998), Fershtman and Judd (1987), Fershtman and Weiss (1998), Koçkesen et al. (2000a, 2000b), Bolle (2000), Possajennikov (2000), Sethi and Somanathan (2001), Heifetz et al. (2007a, 2007b), Akçay et al. (2009), Alger (2010), and Alger and Weibull (2010, 2012).

[22] See Ok and Vega-Redondo (2001), Dekel et al. (2007), and Alger and Weibull (2013).

[23] In the literature cited in the preceding two footnotes, only Alger (2010), Alger and Weibull (2010, 2012, 2013) allow for non-uniform random matching. Bergstrom (1995, 2003) also allows for such assortative matching, but he restricts attention to strategy rather than preference evolution.

second result suggests that one should expect the importance of morality to differ accordingly. Likewise, our results suggest that if in a population individuals are involved in several different interactions, and assortativity differs between them (e.g., sharing with relatives, and engaging in market interactions with strangers), then one should expect different levels of morality in the different interactions. We hope that our theoretical results, combined with empirical and experimental work, will enhance the understanding of human behavior and motivation.

Appendix A. A class of matching processes

Let n, I and P be positive integers, and imagine a finite population consisting of P individuals. The population is divided into "islands," each island consisting of $I > n$ individuals, and P is some multiple of I. Initially all individuals are of type f. Suddenly a mutation to another type g occurs in one of the islands, and only there. Each individual in that island has probability ρ of mutating, and individual mutations are statistically independent. Hence, the random number M of mutants is binomially distributed $M \sim Bin(I, \rho)$. In this mutation process, the random number M is also the total number of mutants in the population at large, so the population share M/P of mutants is a random variable with expectation $\varepsilon = \rho I/P$. A group of size n is now formed to play a game $\Gamma = \langle N, X^n, \pi \rangle$ (as described in Section 2) as follows, and this is an event that is statistically independent of the above-mentioned mutation. First, one of the islands is selected, with equal probability for each island. Secondly, n individuals from the selected island are recruited to form the group, drawn as a random sample without replacement from amongst the I islanders and with equal probability for each islander to be sampled.

Consider an individual, i, who has been recruited to the group. Let $h_i \in \{f, g\}$ denote the individual's type. If $h_i = g$, it is necessary that $M > 0$ and that the individual is from the island where the mutation occurred, so the random number of *other* mutants in her group is then binomially distributed, $Bin(n-1, \rho)$. With T denoting the total number of mutants in her group, we have, for $m = 0, 1, 2, \ldots, n-1$,

$$q_m = \Pr[T = m+1 \mid h_i = g] = \binom{n-1}{m} \rho^m (1-\rho)^{n-m-1}. \tag{17}$$

If instead $h_i = f$, then $M = 0$ is possible and individual i may well be from another island than where the mutation occurred. We then have

$$p_m = \Pr[T = m \mid h_i = f] \leq \frac{I}{P} \cdot \binom{n-1}{m} \rho^m (1-\rho)^{n-m-1}$$

for $m = 0, 1, 2, \ldots, n-1$.

Likewise, for any *two* group members i and j,

$$\Pr\left[h_j = g \mid h_i = g\right] = \rho \quad \text{and} \quad \Pr\left[h_j = g \mid h_i = f\right] \leq \rho I/P.$$

Keeping ρ, n and I constant, we may write $\Pr[f|f,\varepsilon]$ for $\Pr\left[h_j = f \mid h_i = f\right]$ and $\Pr[f|g,\varepsilon]$ for $\Pr\left[h_j = f \mid h_i = g\right]$. These conditional probabilities are continuous functions of $\varepsilon = \rho I/P$. In addition, we have $1 - \varepsilon \leq \Pr[f|f,\varepsilon] \leq 1$ and $\Pr[f|g,\varepsilon] = 1 - \rho$. Letting $P \to \infty$, we obtain $\varepsilon \to 0$ and $\Pr[f|f,\varepsilon] \to 1$. Hence, $\lim_{\varepsilon \to 0} \phi(\varepsilon) = \rho$, so the index of assortativity is $\sigma = \rho$.

References

Acemoglu, D., Jensen, M.K., 2013. Aggregate comparative statics. Games Econ. Behav. 81, 27–49.
Akçay, E., Van Cleve, J., Feldman, M.W., Roughgarden, J., 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc. Natl. Acad. Sci. 106, 19061–19066.
Alger, I., 2010. Public goods games, altruism, and evolution. J. Public Econ. Theory 12, 789–813.
Alger, I., Weibull, J., 2010. Kinship, incentives and evolution. Amer. Econ. Rev. 100, 1725–1758.
Alger, I., Weibull, J., 2012. A generalization of Hamilton's rule—love others how much? J. Theoret. Biol. 299, 42–54.
Alger, I., Weibull, J., 2013. *Homo moralis*—preference evolution under incomplete information and assortative matching. Econometrica 81, 2269–2302.
Aliprantis, C.D., Border, K.C., 2006. Infinite Dimensional Analysis, 3rd ed. Springer, New York.
Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. Econ. J. 100, 464–477.
Becker, G., 1976. Altruism, egoism, and genetic fitness: economics and sociobiology. J. Econ. Lit. 14, 817–826.
Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Amer. Econ. Rev. 96, 1652–1678.
Bergstrom, T., 1995. On the evolution of altruistic ethical rules for siblings. Amer. Econ. Rev. 85, 58–81.
Bergstrom, T., 2003. The algebra of assortative encounters and the evolution of cooperation. Int. Game Theory Rev. 5, 211–228.
Bergstrom, T., 2009. Ethics, evolution, and games among neighbors. Working paper. UCSB.
Bergstrom, T., 2013. Measures of assortativity. Biol. Theory 8, 133–141.
Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? J. Econ. Behav. Organ. 34, 193–209.
Binmore, K., 1994. Game Theory and The Social Contract, vol. 1: Playing Fair. MIT Press, Cambridge, USA.
Bolle, F., 2000. Is altruism evolutionarily stable? And envy and malevolence? Remarks on Bester and Güth. J. Econ. Behav. Organ. 42, 131–133.
Bramoullé, Y., Rogers, B., 2009. Diversity and popularity in social networks. Discussion papers 1475. Northwestern University, Center for Mathematical Studies in Economics and Management Science.
Brekke, K.A., Kverndokk, S., Nyborg, K., 2003. An economic model of moral motivation. J. Public Econ. 87, 1967–1983.
Broom, M., Cannings, C., Vickers, G.T., 1996. Choosing a nest site: contests and catalysts. Amer. Nat. 147, 1108–1114.
Broom, M., Cannings, C., Vickers, G.T., 1997. Multi-player matrix games. Bull. Math. Biol. 59, 931–952.
Bukowski, M., Miękisz, J., 2004. Evolutionary and asymptotic stability in symmetric multi-player games. Int. J. Game Theory 33, 41–54.
Cannings, C., Whittaker, J.C., 1995. The finite horizon war of attrition. Games Econ. Behav. 11, 193–236.

Corchón, L., 1996. Theories of Imperfectly Competitive Markets. Springer Verlag, Berlin.
Currarini, S., Jackson, M.O., Pin, P., 2009. An economic model of friendship: homophily, minorities and segregation. Econometrica 77, 1003–1045.
Currarini, S., Jackson, M.O., Pin, P., 2010. Identifying the roles of race-based choice and chance in high school friendship network formation. Proc. Natl. Acad. Sci. 107, 4857–4861.
Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Rev. Econ. Stud. 74, 685–704.
Dubey, P., Mas-Colell, A., Shubik, M., 1980. Efficiency properties of strategic market games. J. Econ. Theory 22, 339–362.
Ellingsen, T., 1997. The evolution of bargaining behavior. Quart. J. Econ. 112, 581–602.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.
Fershtman, C., Judd, K., 1987. Equilibrium incentives in oligopoly. Amer. Econ. Rev. 77, 927–940.
Fershtman, C., Weiss, Y., 1998. Social rewards, externalities and stable preferences. J. Public Econ. 70, 53–73.
Frank, R.H., 1987. If *homo economicus* could choose his own utility function, would he want one with a conscience? Amer. Econ. Rev. 77, 593–604.
Grafen, A., 1979. The hawk–dove game played between relatives. Animal Behav. 27, 905–907.
Güth, W., Yaari, M., 1992. An evolutionary approach to explain reciprocal behavior in a simple strategic game. In: Witt, U. (Ed.), Explaining Process and Change—Approaches to Evolutionary Economics. University of Michigan Press, Ann Arbor.
Haigh, J., Cannings, C., 1989. The n-person war of attrition. Acta Appl. Math. 14, 59–74.
Heifetz, A., Shannon, C., Spiegel, Y., 2007a. The dynamic evolution of preferences. Econ. Theory 32, 251–286.
Heifetz, A., Shannon, C., Spiegel, Y., 2007b. What to maximize if you must. J. Econ. Theory 133, 31–57.
Hines, W.G.S., Maynard Smith, J., 1979. Games between relatives. J. Theoret. Biol. 79, 19–30.
Huck, S., Oechssler, J., 1999. The indirect evolutionary approach to explaining fair allocations. Games Econ. Behav. 28, 13–24.
Kant, I., 1785. Grundlegung zur Metaphysik der Sitten. In English: Groundwork of the Metaphysics of Morals. Harper Torch Books, New York, 1964.
Koçkesen, L., Ok, E.A., Sethi, R., 2000a. The strategic advantage of negatively interdependent preferences. J. Econ. Theory 92, 274–299.
Koçkesen, L., Ok, E.A., Sethi, R., 2000b. Evolution of interdependent preferences in aggregative games. Games Econ. Behav. 31, 303–310.
Lehmann, L., Alger, I., Weibull, J., 2015. Does evolution lead to maximizing behavior? Evolution 69, 1858–1873.
Levine, D., 1998. Modelling altruism and spite in experiments. Rev. Econ. Dynam. 1, 593–622.
Luenberger, D.G., 1969. Optimization by Vector Space Methods. John Wiley & Sons, New York.
Maynard Smith, J., Price, G.R., 1973. The logic of animal conflict. Nature 246, 15–18.
McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27, 415–444.
Munkres, James, 1975. Topology, a First Course. Prentice Hall, London.
Ockenfels, P., 1993. Cooperation in Prisoners' dilemma—an evolutionary approach. Europ. J. Polit. Economy 9, 567–579.
Ok, E.A., Vega-Redondo, F., 2001. On the evolution of individualistic preferences: an incomplete information scenario. J. Econ. Theory 97, 231–254.
Possajennikov, A., 2000. On the evolutionary stability of altruistic and spiteful preferences. J. Econ. Behav. Organ. 42, 125–129.
Robson, A.J., 1990. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. J. Theoret. Biol. 144, 379–396.
Robson, A.J., Szentes, B., 2014. A biological theory of social discounting. Amer. Econ. Rev. 104, 3481–3497.
Roemer, J.E., 2010. Kantian equilibrium. Scand. J. Econ. 112, 1–24.
Rousset, F., 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton.
Ruef, M., Aldrich, H.E., Carter, N.M., 2003. The structure of founding teams: homophily, strong ties, and isolation among U.S. entrepreneurs. Amer. Sociol. Rev. 68, 195–222.
Schaffer, M.E., 1988. Evolutionarily stable strategies for finite populations and variable contest size. J. Theoret. Biol. 132, 467–478.
Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. J. Econ. Theory 97, 273–297.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.