Evolutionarily stable strategies, preferences and moral values, in n-player interactions[*]

Ingela Alger[†] and Jörgen W. Weibull[‡]

February 12, 2014. Revised June 23, 2014.

Abstract

We provide a generalized definition of evolutionary stability of heritable types in arbitrarily large symmetric interactions under random matching that may be assortative. We establish stability results when these types are strategies in games, and when they are preferences or moral values in games under incomplete information. We show that a class of moral preferences, with degree of morality equal to the index of assortativity are evolutionarily stable. In particular, selfishness is evolutionarily unstable when there is positive assortativity in the matching process. We establish that evolutionarily stable strategies are the same as those played in equilibrium by rational but partly morally motivated individuals, individuals with evolutionarily stable preferences. We provide simple and operational criteria for evolutionary stability and apply these to canonical examples.

Keywords: Evolutionary stability, assortativity, morality, *homo moralis*, public goods, contests, helping, Cournot competition.

JEL codes: C73, D01, D03.

[*] Support through Knut and Alice Wallenberg Research Foundation and through ANR - Labex IAST is gratefully acknowledged. Ingela Alger also thanks Agence Nationale de la Recherche (ANR) for funding (Chaire d'Excellence). We thank seminar participants at EconomiX, WZB, GREQAM, ETH Zürich, Nottingham, OECD, Séminaire Roy (PSE), IGIER Bocconi, and participants at the Toulouse Economics and Biology Workshop 2014, for helpful comments.

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

[‡] Stockholm School of Economics, KTH Royal Institute of Technology, and Institute for Advanced Study in Toulouse

1 Introduction

Economics provides a rich set of powerful theoretical models of human societies. Since these models usually feature individuals whose motivations—preferences and moral values—are given, their predictive power depends on the accuracy of the assumptions regarding these motivations. However, if the motivations of the members of a society are inherited from past generations, the formation of these motivations may itself be studied theoretically. In particular, one may use evolutionary logic to ask what preferences and moral values have a survival value, and thus ask: what preferences and moral values should humans be expected to have from first principles?

It is of general interest, not the least for economics, to understand if and when selfishness may be favored by evolution, and if not, what kind of preferences are likely to emerge instead. Past research has identified two factors that pull preferences away from pure selfishness. First, as observed already by Schelling (1960), it may be advantageous in strategic interactions to be known or believed to be committed to certain behaviors, or to have preferences or values, even if these commitments or values appear to be at odds with one's material self-interest.[1] The literature on preference evolution confirms this intuition by showing that when interactions occur under incomplete information, selfishness prevails,[2] whereas when interactions occur under complete information this is no longer the case.[3] Secondly, until recently the economics literature has largely disregarded another factor, which has been known and studied in biology for decades, namely that natural selection favors unselfish behaviors between relatives, and more generally, between individuals in structured populations, where most interactions take place within subpopulations.[4]

We propose a general model for the study of the evolutionary foundations of human motivation in strategic interactions in arbitrarily large groups. The model can be applied to

[1] For example, a manager of a firm in Cournot competition, with complete information about managers' contracts, will do better, in terms of equilibrium profits, if the contract rewards both profits and sales, rather than only profits (a literature pioneered by Fershtman and Judd, 1987).

[2] See Ok and Vega-Redondo (2001) and Dekel, Ely and Yilankaya (2007) for analyses of such environments.

[3] See Heifetz, Shannon, and Spiegel (2007a) for a particularly general such result. See also the literature overview in Section 2 for references.

[4] The seminal work on this is Hamilton (1964a,b). For models using game theory, see Grafen (1979, 2006), Hines and Maynard Smith (1979), Bergstrom (1995, 2009), and Day and Taylor (1998).

strategy evolution and to evolution of preferences and/or moral values both under complete and incomplete information, and the random matching may be assortative (as in structured populations). We define evolutionary stability as a property of abstract *types* that can be virtually any characteristic of an individual, such as a behavior pattern or strategy, a goal function, preference, moral value, belief, or cognitive capacity etc. The types may be visible/known or invisible/unknown to others. However, we do not allow that the types directly affect the material payoff consequences of given action profiles. Individuals live in a infinite (continuum) population and are randomly matched in groups of size n, to play a symmetric n-player game with "material" payoffs. Each player's strategy set may be simple, such as in a simultaneous-move game, or very complex, such as in a sequential game with many time periods and information sets. Strategies may be pure or mixed. The game has to be *ex ante* symmetric (in a sense to be defined), but may be *ex post* asymmetric, as long as each player is equally likely to be in any one of the n player roles. The random matching may be assortative, that is, the type distribution of others in the matchings of any given individual may depend on the individual's own type. Individuals of types that result in high payoffs in these random matchings are taken to have a higher survival probability than individuals of types that result in low payoffs. In the model we apply Bergstrom's (2003) algebra of assortative pairwise encounters and show how his *index of assortativity*, a one-dimensional measure of the extent of the assortativity, can be extended to n-player encounters.[5]

We apply the general model to, in turn, evolutionary stability of strategies (Section 3), and evolutionarily stable preferences under incomplete information (Section 4), leaving the study of the evolutionary stability of other types, notably to preferences under complete (or noisy) information, for future research. In Section 5 we illustrate these two applications within several commonly studied games in which an individual's material payoff depends on own strategy and some aggregate measure of the others' strategies. These games are examples of *aggregative games*, and we believe that our results can be fruitfully applied to other such games.[6]

[5] We refer to Bergstrom (2013) and Alger and Weibull (2013) for further discussions of assortativity when $n = 2$.

[6] The notion of aggregative games is, to the best of our knowledge, due to Dubey, Mas-Colell and Shubik (1980). See also Corchón (1996). The key feature is that the payoff to a player depends only on the players' own strategy and some (symmetric) aggregation of others' strategies. For a recent paper on aggregative games, see Acemoglu and Jensen (2013). For work on aggregative games more related to ours, see Haigh and

Two main results are established. First, although we impose minimal restrictions on the set of potential preferences or moral values, our analysis of preference evolution under incomplete information shows that evolution favors a particular class of preferences, namely, a generalization from 2-player games to n-player games of the *homo moralis* preferences defined in Alger and Weibull (2013). For this generalization, one needs to generalize (a) the notion of game symmetry, (b) the notion of assortative matching, and (c) the notion of *homo moralis*. For arbitrary $n \geq 2$, a *homo moralis* evaluates her strategy choice in the light of its effect on her own material welfare as well as on the material welfare that would arise if others were to probabilistically choose the same strategy. We show that generalized *homo moralis* preferences with a degree of morality equal to the index of assortativity are evolutionarily stable. Furthermore, any preferences such that equilibrium behaviors differ from those of *homo moralis* with the right degree of morality are evolutionarily unstable. This generalizes the result in Alger and Weibull (2013) from 2-player to n-player interactions, and the present instability result is somewhat stronger (even for two-player games), since we here allow for multiple equilibria.

Our second main result is that behaviors selected for under strategy evolution are the same as the equilibrium behaviors among *homo moralis* with degree of morality equal to the index of assortativity. This result establishes that evolutionarily stable strategies (under uniform or assortative random matching) need not be interpreted only as resulting when individuals are "programmed" to certain strategies, but can also be interpreted as resulting when individual are rational and free to choose whatever strategy they like, but whose preferences have emerged from natural selection. Together with a first- and second-order characterization result for games in euclidean strategy spaces, with arbitrary index of assortativity, we obtain easy and transparent methods to find the (symmetric) equilibria of n-player games between *homo moralis* with that degree of morality, methods we illustrate in various canonical examples.

2 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

Cannings (1989) and Koçkesen, Ok and Sethi (2000a,b).

means that the probability for an opponent's strategy does not depend on one's own strategy. Broom, Cannings and Vickers (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.[7] 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 in 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. 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.[8] 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[9] or incomplete[10], and whether non-uniform random matching is considered.[11] Few models deal with interactions involving more than two individuals. Like here, the articles in this category focus exclusively on interactions that are symmetric in material payoffs, the payoffs that drive evolution. Unlike us, they restrict attention to uniform random matching. Koçkesen, Ok, and Sethi (2000a,b) show that under complete informa-

[7] Precursors to their work are Haigh and Cannings (1989), Cannings and Whittaker (1995) and Broom, Cannings and Vickers (1996).

[8] See also Frank (1987).

[9] See Robson (1990), Güth and Yaari (1992), Ockenfels (1993), Huck and Oechssler (1996), Ellingsen (1997), Bester and Güth (1998), Fershtman and Weiss (1998), Koçkesen, Ok and Sethi (2000a,b), Bolle (2000), Possajennikov (2000), Sethi and Somanathan (2001), Heifetz, Shannon and Spiegel (2007a,b), Akçay et al. (2009), Alger (2010), and Alger and Weibull (2010, 2012).

[10] See Ok and Vega-Redondo (2001), Dekel, Ely and Yilankaya (2007), and Alger and Weibull (2013).

[11] 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.

tion 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 characterize 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.

3 Model

Consider an infinite (continuum) population of individuals who are randomly matched into groups of $n \geq 2$ individuals to interact according to some game given in normal form $\Gamma = \langle X, \pi, n \rangle$, where X is the set of *strategies* available to each player (individual in the group) and $\pi : X^n \to \mathbb{R}$ is the *material payoff function.* The set X is a non-empty, compact and convex set in topological vector space, and the function π is continuous.[12] The material payoff to any player $i \in \{1, .., n\}$ from using strategy $x_i \in X$ against the strategies $x_j \in X$ $(j \neq i)$ of the others in the group, is denoted $\pi(x_i, \mathbf{x}_{-i})$, where π is *symmetric* in $\mathbf{x}_{-i}$, the strategy profile of all other individuals in the group, in the sense that the payoff $\pi(x_i, \mathbf{x}_{-i})$ is invariant under permutations of the components of $\mathbf{x}_{-i}$. These games may thus be called *aggregative.*[13]

Each individual has some *type* (or *trait*) $\theta \in \Theta$, which may influence his/her choice of strategy, or *behavior* in the material game, where Θ is the set of potential types. Consider a population in which at most two types from Θ are present. For any types θ and τ, and any $\varepsilon \in (0, 1)$, let $s = (\theta, \tau, \varepsilon)$ be the *population state* in which the two types are represented in population shares $1 - \varepsilon$ and ε, respectively. Let $S = \Theta^2 \times (0, 1)$ denote the set of population states. We are particularly interested in states $s = (\theta, \tau, \varepsilon)$ in which ε is small, then calling θ the *resident* type, being predominant in the population, and τ, being rare, the *mutant* type.

In a given population state $s \in S$, the behavioral outcomes, or, more precisely, strategy

[12] All assumptions are not needed for all our claims, but are made at the outset in order to ease the exposition. All results apply, *mutatis mutandis*, also to $n = 1$, in which case Γ is a decision problem.

[13] More precisely: for any $x_i \in X$ and $\mathbf{x}_{-i} \in X^{n-1}$, and any bijection $h : \{2, 3, ..., n\} \to \{2, 3, ..., n\}$: $\pi\left(x_i, x_{h(2)}, x_{h(3)}, ..., x_{h(n)}\right) = \pi(x_i, \mathbf{x}_{-i})$.

profiles used, may, but need not, be uniquely determined. For each population state s, let $V(s) \subset \mathbb{R}^2$ be the set of (average) material-payoff pairs that *can* arise in population state s, where, for any $v = (v_1, v_2) \in V(\theta, \tau, \varepsilon)$, the first component, v_1, is the average material payoff to individuals of type θ, and the second component, v_2, that to individuals of type τ. We assume that $V(s)$ is non-empty and compact for all states $s = (\theta, \tau, \varepsilon)$. Then

$$f(\theta, \tau, \varepsilon) = \min_{(v_1, v_2) \in V(\theta, \tau, \varepsilon)} v_1 - v_2 \tag{1}$$

is well-defined. This is the material payoff difference, in the residents' worst possible outcome as compared with mutants (in terms of material payoffs), across all behavioral outcomes that are possible in state $s = (\theta, \tau, \varepsilon)$. In particular, $f(s) > 0$ if and only if the residents earn a (strictly) higher (average) material payoff than the mutants in all possible outcomes in that state.[14] By definition, $f(\theta, \theta, \varepsilon) = 0$ for all $\theta \in \Theta$ and $\varepsilon \in (0, 1)$.

The following definitions of evolutionary stability and instability are generalizations of the definitions in Alger and Weibull (2013), from $n = 2$ to $n \geq 2$, and from preference evolution under incomplete information to arbitrary types.

Definition 1 *A type θ is **evolutionarily stable against a type** τ if there exists an $\bar{\varepsilon} > 0$ such that $f(\theta, \tau, \varepsilon) > 0$ for all $\varepsilon \in (0, \bar{\varepsilon})$. A type θ is **evolutionarily stable** if it is evolutionarily stable against all types $\tau \neq \theta$. A type θ is **evolutionarily unstable** if there exists a type τ and a sequence $\langle \varepsilon_t \rangle$ from $(0, 1)$ such that $\varepsilon_t \to 0$ and $f(\theta, \tau, \varepsilon_t) < 0$ for all t.*

Clearly, by this definition no type is both evolutionarily stable and unstable, and there may, in general, exist types that are neither stable nor unstable.

Before proceeding, let us briefly consider how these definitions relate to Maynard Smith's and Price's (1973) original definition of an evolutionarily stable (mixed) strategy in a symmetric and finite two-player game under uniform random matching. Suppose thus that X is the unit simplex of mixed strategies in such a game and let $\Theta = X$, that is, let a type be a mixed strategy (as if individuals were "programmed" to strategies). For any population state $s = (x, y, \varepsilon) \in X^2 \times (0, 1)$, the set $V(s)$ of possible material-payoff pairs is then a singleton. Its unique element $v \in V(s)$ has components $v_1 = (1 - \varepsilon)\pi(x, x) + \varepsilon\pi(x, y) = \pi[x, (1 - \varepsilon)x + \varepsilon y]$ and $v_2 = (1 - \varepsilon)\pi(y, x) + \varepsilon\pi(y, y) = \pi[yx, (1 - \varepsilon)x + \varepsilon y]$. In other

[14] The function f is a generalization of the so-called score function in evolutionary game theory, see, e.g., Bomze and Pötscher (1989).

words, v_1 (resp. v_2) is the "post-entry" expected material payoff to strategy x (resp. y). By Definition 1, x is evolutionarily stable against y if $f(x, y, \varepsilon) > 0$ for all $\varepsilon > 0$ sufficiently small, which is equivalent with being evolutionarily stable in the sense of Maynard Smith and Price (1973). Suppose a strategy x is unstable in the sense of Definition 1. Since f is here continuous, there then exists a strategy $y \neq x$ such that $f(x, y, \varepsilon) < 0$ for all $\varepsilon > 0$ sufficiently small, that is, such that this mutant's post-entry expected material payoff exceeds that of the resident strategy x whenever the mutant appears in sufficiently small population shares.

3.1 Matching

The matching process is exogenous. In any population state $s = (\theta, \tau, \varepsilon) \in S$, the number of mutants—individuals of type τ—in a group that is about to play the n-player game π, is a random variable that we will denote T. For any *resident* drawn at random from the population let $p_m(\varepsilon)$ be the conditional probability $\Pr[T = m \mid \theta, s]$ that the total number of mutants in the resident's group is m, for $m = 0, 1, .., n-1$.[15] Likewise, for any mutant, also drawn at random from the population, let $q_m(\varepsilon)$ be the conditional probability $\Pr[T = m \mid \tau, s]$ that the total number of mutants in his or her group is m, for $m = 1, .., n$. We assume that all functions p_m and q_m are continuous, and that each such function has a limit as $\varepsilon \to 0$ (for any given m).

It follows that $p_0(\varepsilon)$ converges to 1 as ε tends to 0 (and hence $\lim_{\varepsilon \to 0} p_m(\varepsilon) = 0$ for all $m > 0$). In other words, residents almost never meet mutants when the latter are vanishingly rare. To formally establish this, we use the *algebra of assortative encounters* developed by Bergstrom (2003) for pairwise interactions. For a given population state $s = (\theta, \tau, \varepsilon)$, let $\Pr[\theta|\theta, \varepsilon]$ denote the conditional probability for an individual of type θ that another, uniformly randomly drawn member of his or her group also is of type θ. Likewise, let $\Pr[\theta|\tau, \varepsilon]$ denote the conditional probability for an individual of type τ that any other uniformly randomly drawn member of his or her group has type θ. Let $\phi(\varepsilon)$ be the difference between the two probabilities:

$$\phi(\varepsilon) = \Pr[\theta|\theta, \varepsilon] - \Pr[\theta|\tau, \varepsilon]. \tag{2}$$

[15] 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.

This defines the *assortment function* $\phi : (0,1) \to [-1,1]$. We assume that, as ε tends to zero, $\phi(\varepsilon)$ converges to some limit $\sigma \in \mathbb{R}$, the *index of assortativity* of the matching process (Bergstrom, 2003). Moreover, by setting $\phi(0) = \sigma$ we extend the domain of ϕ from $(0,1)$ to $[0,1)$.

The following equation is a necessary balancing condition:

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

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 these two individuals are of different types. Equations (2) and (3) together give

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

Now let $\varepsilon \to 0$. Then, from (3), $\Pr[\theta|\theta,\varepsilon] \to 1$, and hence, as claimed, $p_0(\varepsilon) \to 1$. Without loss of generality we may thus uniquely extend the domain of p_m from $(0,1)$ to $[0,1)$, while preserving its continuity, by setting $p_0(1) = 1$.

Turning now to the limit of $q_m(\varepsilon)$ as ε tends to zero (for $m = 1, ..., n$), we first note that in the special case $n = 2$,

$$\lim_{\varepsilon\to 0} q_2(\varepsilon) = \lim_{\varepsilon\to 0} \Pr[\tau|\tau,\varepsilon] = 1 - \lim_{\varepsilon\to 0} \Pr[\theta|\tau,\varepsilon] = 1 - \lim_{\varepsilon\to 0} \phi(\varepsilon) = \sigma.$$

However, for $n > 2$ there remains a statistical issue, namely whether or not, for a given mutant, the types of any two *other* members in her group are statistically dependent or not (in the given population state). Under conditional independence among other group members' types, given the mutant's type, one obtains

$$\begin{aligned} \lim_{\varepsilon\to 0} q_m(\varepsilon) &= \lim_{\varepsilon\to 0} \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} (\Pr[\tau|\tau,\varepsilon])^{m-1} (\Pr[\theta|\tau,\varepsilon])^{n-m} \\ &= \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \sigma^{m-1} (1-\sigma)^{n-m} \end{aligned} \tag{5}$$

for all $m \in \{1, .., n\}$. We will refer to this as *the conditionally independent case.*[16] This generalizes the limit result for group size $n = 2$. In the appendix we present a matching

[16] To be precise, we only require conditional independence in the limit, as expressed in (5).

process with the conditional statistical independence property. We finally note that it follows from (4) that $\sigma \in [0,1]$.[17]

3.2 Homo moralis

In Alger and Weibull (2013), we analyzed a similar model, but for pairwise interactions, and showed that natural selection favors a particular class of preferences that we called *homo moralis*. The utility for a *homo moralis*, with degree of morality $\kappa \in [0,1]$, from playing strategy $x \in X$ against $y \in X$ in a symmetric two-player game with material payoff function $\pi : X^2 \to \mathbb{R}$ is

$$u_\kappa (x,y) = (1-\kappa) \cdot \pi (x,y) + \kappa \cdot \pi (x,x) . \tag{6}$$

Such an individual thus attaches the weight $(1-\kappa)$ to own material payoff and the weight κ to the material payoff that would arise should both players use the same strategy x.

While it is not obvious how one should define such preferences for interactions between more than two individuals, we will see below that natural selection again points in a particular direction. Accordingly, for any player $i \in \{1,..,n\}$, any degree of morality $\kappa \in [0,1]$, and any strategy profile $\mathbf{x} \in X^n$, let $\tilde{\mathbf{x}} : \Omega \to X^n$ be a vector-valued random variable with statistically independent components $\tilde{x}_j$ such that $\Pr[\tilde{x}_j = x_i] = \kappa$ and $\Pr[\tilde{x}_j = x_j] = 1-\kappa$ for all $j \in \{1,..,n\}$. We write utility functions in the same form as the material payoff function, that is, with the player's own strategy as the first argument and the profile of others' strategies as the second argument.

Definition 2 *Player i is a* homo moralis *with degree of morality $\kappa \in [0,1]$ if his or her utility function $u_\kappa : X^n \to \mathbb{R}$ satisfies*

$$u_\kappa (x_i, \mathbf{x}_{-i}) = \mathbb{E}_\kappa [\pi (x_i, \tilde{\mathbf{x}}_{-i}) \mid \mathbf{x}] \quad \forall \mathbf{x} \in X^n. \tag{7}$$

We note that $\tilde{x}_i$ is a degenerate random variable that always takes the constant value x_i, and thus

$$\begin{aligned} u_\kappa (x_i, \mathbf{x}_{-i}) &= (1-\kappa)^{n-1} \cdot \pi (x_i, \mathbf{x}_{-i}) + \kappa^{n-1} \cdot \pi (x_i, x_i, ..., x_i) \\ &\quad + \left[1 - (1-\kappa)^{n-1} - \kappa^{n-1}\right] \cdot \mathbb{E}_\kappa [\pi (\tilde{\mathbf{x}}) \mid \tilde{\mathbf{x}} \neq \mathbf{x}, \tilde{\mathbf{x}} \neq (x_i, x_i, ..., x_i)] . \end{aligned}$$

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

At one extreme, $\kappa = 0$, the individual's goal is to choose a strategy x_i that maximizes her own material payoff, given the strategy profile $\mathbf{x}_{-i}$ for all other participants. At the opposite extreme, $\kappa = 1$, her goal is "to do the right thing" according to Kant's categorical imperative applied to material payoffs, that is, to choose a strategy x_i that would maximize material payoff if all others were to choose that same strategy. We refer to the first case as *homo oeconomicus* and the second as *homo kantientis*. For arbitrary degrees of morality, $0 \leq \kappa \leq 1$, the individual's goal is to maximize her expected material payoff if others were to choose that same strategy with probability κ (and statistically independently of each other).[18]

Since the random variables $\tilde{x}_j$ are statistically independent, the utility $u_\kappa(x_i, \mathbf{x}_{-i})$, for any given strategy profile $\mathbf{x} \in X^n$, is a polynomial function of the degree of morality κ, taking the value $\pi(\mathbf{x})$ at $\kappa = 0$ and the value $\pi(x_i, x_i, .., x_i)$ at $\kappa = 1$.[19] For $n = 2$ and any strategy pair $(x, y) \in X^2$ one obtains the same expression for u_κ as in (6). For $n = 3$ and any strategy triplet $(x, (y, z)) \in X^3$, where x is the player's own strategy, one obtains

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

While the expression in (7) is quite involved, it is easy to determine the set of symmetric Nash equilibrium strategies in a game between n *homo moralis* with the same degree of morality. For any $\kappa \in [0,1]$, let $\beta_\kappa : X \rightrightarrows X$ be defined by

$$\beta_\kappa(x) = \arg\max_{y \in X} u_\kappa\left(y, \mathbf{x}^{(n-1)}\right). \tag{8}$$

The set of symmetric Nash equilibrium strategies is the set $X_\kappa \subseteq X$ of fixed points under β_κ,

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

By symmetry of π the condition $x \in \beta_\kappa(x)$ can be written more explicitly as

$$x \in \arg\max_{y \in X} \sum_{m=1}^{n} \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \kappa^{m-1}(1-\kappa)^{n-m}\, \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right). \tag{10}$$

[18] In his *Grundlegung zür 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." In this vein, *homo moralis* of degree of morality κ can be said to "act according to that maxim whereby you can, at the same time, will that others should do likewise with probability κ." For a discussion of several ethical principles, see Bergstrom (2009).

[19] From a mathematical viewpoint, *homo moralis* defines a homotopy (see e.g. Munkres, 1975), parametrized by $\kappa \in [0,1]$, between selfishness and Kantian morality.

Before applying our general model to preference evolution, in the next subsection we apply it to strategy evolution, the framework used in classic evolutionary game theory. Given the analytical simplicity of strategy evolution, in comparison with preference evolution, a question of interest is whether strategy evolution gives guidance to behavioral predictions under preference evolution.

4 Strategy evolution

Here we adopt the assumption that was used for the original formulation of evolutionary stability (Maynard Smith and Price, 1973), namely, that an individual's type is a strategy that she always uses. Formally, let the set of potential types be $\Theta = X$, the strategy set for the material game $\Gamma = \langle X, \pi, n \rangle$. Thus, in a population where some types $\theta = x$ and $\tau = y$ are present, the unique behavioral outcome is the strategy pair (x, y), where x is played by all individuals of type $\theta = x$ and y is played by all individuals of type $\tau = y$. By symmetry of π, the material payoff to an individual of type x who belongs to a group where a total of $n-m$ individuals have the same, resident, type x, and m individuals have the mutant type y, can be written $\pi\left(x, \mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right)$, where $\mathbf{x}^{(n-m-1)}$ is the $(n-m-1)$-dimensional vector whose components equal x, and $\mathbf{y}^{(m)}$ is the m-dimensional vector whose components equal y. Likewise, the material payoff to an individual of type y who belongs to such a group can be written $\pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right)$. Hence, given a pair of types (x, y), for each ε the average material payoff to a resident is uniquely determined and equal to

$$F(x, y, \varepsilon) = \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot \pi\left(x, \mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right), \tag{11}$$

and likewise for a mutant, whose average material payoff is

$$G(x, y, \varepsilon) = \sum_{m=1}^{n} q_m(\varepsilon) \cdot \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right). \tag{12}$$

Both F and G are continuous by virtue of the assumed continuity of the conditional probabilities for the number of mutants with respect to the population share of mutants.[20]

[20] The functions F and G are generalizations, from uniform to assortative matching, of the functions used by Broom, Cannings and Vickers (1997) in their definition of an evolutionarily stable strategy in symmetric and finite n-player games (here x and y may be mixed strategies in a finite game).

Under strategy evolution, then, the set of (average) material-payoff pairs that can arise in population state s, $V(s) \subset \mathbb{R}^2$, is a singleton for all $s \in S = X^2 \times (0,1)$, and

$$f(x,y,\varepsilon) = F(x,y,\varepsilon) - G(x,y,\varepsilon).$$

Furthermore, for any $x, y \in X$, $f(x,y,\varepsilon)$ converges (to some real number) as ε tends to zero.

By continuity of F and G, f is continuous, and a *necessary* condition for x to be an evolutionarily stable strategy is

$$\lim_{\varepsilon \to 0} f(x,y,\varepsilon) \geq 0 \qquad \forall y \in X. \tag{13}$$

In other words, it is necessary that the residents on average do not earn a lower material payoff than the mutants when the latter are virtually absent from the population. Likewise, a *sufficient* condition for evolutionary stability is that this inequality holds strictly for all strategies $y \neq x$.

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

$$H(y,x) = \lim_{\varepsilon \to 0} G(x,y,\varepsilon). \tag{14}$$

The function value $H(y,x)$ is the average material payoff to a mutant with strategy y in a population where the resident strategy is x and where the population share of mutants is vanishingly small. Since $H(x,x) = \lim_{\varepsilon \to 0} G(x,x,\varepsilon) = \lim_{\varepsilon \to 0} F(x,x,\varepsilon) = \lim_{\varepsilon \to 0} F(x,y,\varepsilon)$, the necessary condition (13) for a strategy x to be evolutionarily stable may be written

$$H(x,x) \geq H(y,x) \qquad \forall y \in X, \tag{15}$$

or, equivalently,

$$x \in \arg\max_{y \in X} H(y,x). \tag{16}$$

This condition says that for a strategy x to be evolutionarily stable, its users have to earn the same average material payoff as the "the most threatening mutants", those with the highest average material payoff that any vanishingly rare mutant can obtain against the resident. In a sense, thus, an evolutionarily stable type *preempts* entry by rare mutants, rather than doing what would be best (in terms of material payoff) for the residents if there were no mutants around.[21]

[21] See also Alger and Weibull (2013) and Robson and Szentes (2014) for a similar observation. Importantly, this logic is very different from that of group selection.

A sufficient condition for a strategy x to be evolutionarily stable is that

$$H(x,x) > H(y,x) \tag{17}$$

for all $y \neq x$. Interestingly, then, irrespective of n, evolutionarily stable types may be interpreted as Nash equilibrium strategies in a derived two-player game, where "nature" plays strategies against each other:

Proposition 1 *Let $\Theta = X$. If x is an evolutionarily stable strategy in a population where individuals are randomly matched to play the symmetric n-player game in material payoffs $\Gamma = \langle X, \pi, n \rangle$, then (x,x) is a Nash equilibrium of the symmetric two-player game in which the strategy set is X and the payoff function is H. If (x,x) is a strict Nash equilibrium of the latter game, then x is an evolutionarily stable strategy in $\Gamma = \langle X, \pi, n \rangle$, while if (x,x) is a not a Nash equilibrium, then x is evolutionarily unstable.*

This proposition allows us to make a first connection between strategy evolution and *homo moralis* preferences. Indeed, while under strategy evolution each individual mechanistically plays a certain strategy—is "programmed" to execute a certain strategy—we will now see that any evolutionarily stable strategy may be viewed as if emerging from individuals' free choice, as if they were striving to maximize a specific utility function.

To see this, consider the conditionally independent case, for which

$$H(y,x) = \sum_{m=1}^{n} \left(\begin{array}{c} n-1 \\ m-1 \end{array} \right) \sigma^{m-1} (1-\sigma)^{n-m} \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right). \tag{18}$$

The expression on the right-hand side is the same as the one in (10) when $\kappa = \sigma$. Combining this observation with Proposition 1 and the fixed-point equation (16), we obtain:

Corollary 1 *Assume conditionally independent random matching. Let $\Theta = X$ (strategy evolution). If x is an evolutionarily stable strategy, then it belongs to X_κ for $\kappa = \sigma$. Every strategy $x \in X_\sigma$ for which $\beta_\sigma(x)$ is a singleton is evolutionarily stable. Every strategy $x \notin X_\sigma$ is evolutionarily unstable.*

This corollary establishes that the behavior induced under strategy evolution and conditionally independent assortative matching is *as if* individuals were equipped with *homo moralis* preferences with degree of morality equal to the index of assortativity. But what if, instead, evolution were to operate at the level of preferences, thereby "delegating" the strategy choice to the individual? In the next section we apply our general model to preference evolution. Compared with strategy evolution, this introduces two main challenges. First, whereas under strategy evolution the set of potential types is identical with the strategy set, under preference evolution there is *a priori* no given set of potential types. Second, whereas under strategy evolution the behavioral outcome is uniquely determined in each population state, under preference evolution this need not be the case.

5 Preference evolution under incomplete information

From now on, we take each type $\theta \in \Theta$ to uniquely determine a continuous and symmetric utility function $u_\theta : X^n \to \mathbb{R}$, a function that its "host" strives to maximize. We focus on the case of incomplete information, where each individual knows only his/her own type. In other words, each individual's utility function is his or her private information. Then an individual's behavior cannot be conditioned on the types of the others with whom (s)he interacts. However, individual behavior may be adapted to the population state at hand (that is, the types present in the population, and their population shares). Arguably, Bayesian Nash equilibrium is a natural criterion to delineate the set $V(s)$ of (average) material-payoff pairs that can arise in a population state s.

More precisely, in any given state $s = (\theta, \tau, \varepsilon) \in \Theta^2 \times (0,1)$, a (type-homogenous Bayesian) Nash equilibrium is a pair of strategies, one for each type, such that each strategy is a best reply for any player of that type in the given population state. In other words, all players of the same type use the same strategy, and each individual player finds his or her strategy optimal, given his or her utility function.

Definition 3 *In any state $s = (\theta, \tau, \varepsilon) \in \Theta^2 \times (0,1)$, a strategy pair $(\hat{x}, \hat{y}) \in X^2$ is a **(type-homogenous Bayesian) Nash Equilibrium** if*

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot u_\theta\left(x, \hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}}^{(n-m-1)}\right) \\ \hat{y} \in \arg\max_{y \in X} \ \sum_{m=1}^{n} q_m(\varepsilon) \cdot u_\tau\left(y, \hat{\mathbf{y}}^{(m-1)}, \hat{\mathbf{x}}^{(n-m)}\right). \end{cases} \tag{19}$$

Let $B^{NE}(s) \subseteq X^2$ denote the set of (type-homogenous Bayesian) Nash equilibria in state $s = (\theta, \tau, \varepsilon)$, that is, all solutions $(\hat{x}, \hat{y})$ of (19). For given types θ and τ, this defines an equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : (0,1) \rightrightarrows X^2$ that maps mutant population shares ε to the associated set of equilibria. As discussed above, under the assumption that all probabilities in (19) are continuous in ε and converge as $\varepsilon \to 0$, the domain of these probabilities was continuously extended to $[0,1)$. This allows us to likewise extend the domain of $B^{NE}(\theta,\tau,\cdot)$. By a slight generalization of the arguments in the proof of Lemma 1 in Alger and Weibull (2013) one obtains:

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

We will henceforth focus on types θ and τ such that $B^{NE}(\theta,\tau,\varepsilon)$ is non-empty for all $\varepsilon \in [0,1)$. This holds, for example, if all functions u_θ are concave in their first argument, the player's own strategy. Given a population state $s = (\theta,\tau,\varepsilon)$ and some Nash equilibrium $(\hat{x},\hat{y}) \in B^{NE}(s)$, the average equilibrium material payoffs to residents and mutants, respectively, equal $F(\hat{x},\hat{y},\varepsilon)$ and $G(\hat{x},\hat{y},\varepsilon)$, where F and G are defined in (11) and (12), respectively.

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

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

and $X_\theta \subseteq X$ its set of fixed points,

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

Given the unrestricted nature of the set of potential types, for any resident type there may be other types such that, if appearing in rare mutants, would give rise to the same behavior as that of the residents. We define the *behavioral clones* to a type θ as those types that, as vanishingly rare mutants among residents of type θ, are behaviorally potentially indistinguishable from residents in the sense that there exists some equilibrium in which they, as rare mutants, behave just as a resident could rationally do. Formally, for any given type $\theta \in \Theta$, this is the subset[22]

$$\tilde{\Theta}(\theta) = \left\{\tau \in \Theta : (x^*, y^*) \in B^{NE}(\theta,\tau,0) \text{ for some } x^* \in X_\theta \text{ and } y^* \in \beta_\theta(x^*)\right\}. \tag{20}$$

[22] This definition labels a slightly wider range of types as "behavioral clones" than according to our definition in Alger and Weibull (2013); $\Theta_\theta \subseteq \tilde{\Theta}(\theta)$. This slight weakening permits a slightly more powerful stability claim.

Examples of such "behavioral alikes" are individuals with utility functions that are positive affine transformations of the utility function of the residents, and also individuals for whom some strategy in X_θ is dominant.[23]

The second statement in the result below will use the following definition (from Alger and Weibull, 2013): the type set Θ is *rich* if for each strategy $x \in X$ there exists some type $\theta \in \Theta$ for which this strategy is strictly dominant. Such a type θ will be said to be *committed* to its strategy x. The following result is a generalization of Theorem 1 in Alger and Weibull (2013) from pairwise interactions ($n = 2$) to interactions with an arbitrary number of participants ($n \geq 2$); furthermore, the instability result now also applies to types for which there are multiple equilibria.

Theorem 1 *Assume conditionally independent assortativity.* Homo moralis *with degree of morality $\kappa = \sigma$ is evolutionarily stable against all types $\tau \notin \tilde{\Theta}(\theta)$. If Θ is rich and $X_\theta \cap X_\sigma = \varnothing$, then θ is evolutionarily unstable.*

Proof: Since π is continuous, and, given $\theta, \tau \in \Theta$, all functions p_m and q_m are continuous in ε by hypothesis, also the two functions $F, G : X^2 \times [0,1) \to \mathbb{R}$ (given $\theta, \tau \in \Theta$) are continuous.

For the first claim, let $\theta = \sigma$ (*homo moralis* of degree of morality σ) and $\tau \notin \tilde{\Theta}(\sigma)$, and suppose that $(x, y) \in B^{NE}(\sigma, \tau, 0)$. Then $x \in X_\sigma$ so $u_\sigma\left(x, \mathbf{x}^{(n-1)}\right) \geq u_\sigma\left(y, \mathbf{x}^{(n-1)}\right)$. Since $\tau \notin \tilde{\Theta}(\sigma)$: $y \notin \beta_\sigma(x)$. Hence, $u_\sigma\left(x, \mathbf{x}^{(n-1)}\right) > u_\sigma\left(y, \mathbf{x}^{(n-1)}\right)$, or, equivalently, $F(x, y, 0) > G(x, y, 0)$. Let $g : X^2 \to \mathbb{R}$ be defined by $g(x, y) = F(x, y, 0) - G(x, y, 0)$. By continuity of F and G, g is continuous. Since $B^{NE}(\sigma, \tau, 0)$ is compact and $g(x, y) > 0$ on $B^{NE}(\sigma, \tau, 0)$, we have $\min_{(x,y) \in B^{NE}(\sigma,\tau,0)} g(x, y) = \delta$ for some $\delta > 0$. Again by continuity of F and G, there exists a neighborhood $U \subseteq X^2 \times [0,1)$ of the compact set $B^{NE}(\sigma, \tau, 0) \times \{0\}$ such that $F(x, y, \varepsilon) - G(x, y, \varepsilon) > \delta/2$ for all $(x, y, \varepsilon) \in U$. Since $B^{NE}(\sigma, \tau, \cdot) : [0,1) \rightrightarrows X^2$ is compact-valued and upper hemi-continuous, there exists an $\bar{\varepsilon} > 0$ such that $B^{NE}(\sigma, \tau, \varepsilon) \times [0, \varepsilon] \subset U$ for all $\varepsilon \in [0, \bar{\varepsilon})$. It follows that $F(x, y, \varepsilon) - G(x, y, \varepsilon) > \delta/2$ for all $\varepsilon \in [0, \bar{\varepsilon})$ and all $(x, y) \in B^{NE}(\sigma, \tau, \varepsilon)$. Setting $V(\sigma, \tau, \varepsilon) = B^{NE}(\sigma, \tau, \varepsilon)$ we thus have $f(\sigma, \tau, \varepsilon) > \delta/2$ for all $\varepsilon \in [0, \bar{\varepsilon})$, establishing the first claim.

For the second claim, let $\theta \in \Theta$ be such that $X_\theta \cap X_\sigma = \varnothing$ and suppose that $x_\theta \in X_\theta$. Then $u_\sigma\left(\hat{x}, \mathbf{x}_\theta^{(n-1)}\right) > u_\sigma\left(x_\theta, \mathbf{x}_\theta^{(n-1)}\right)$ for some $\hat{x} \in X$. If Θ is rich, there exists a type $\tau \in \Theta$

[23] For example, if $x^* \in X_\theta$, let $u(x_i, \mathbf{x}_{-i}) \equiv -(x_i - x^*)^2$.

for which $\hat{x}$ is strictly dominant, so individuals of that type will always play $\hat{x}$. By definition of u_σ,

$$G\left(x_\theta,\hat{x},0\right)=u_\sigma\left(\hat{x},\mathbf{x}_\theta^{(n-1)}\right)>u_\sigma\left(x_\theta,\mathbf{x}_\theta^{(n-1)}\right)=F\left(x_\theta,\hat{x},0\right).$$

Let $\langle x_t,y_t,\varepsilon_t\rangle_{t\in\mathbb{N}}$ be any sequence from $X^2\times(0,1)$ such that $\varepsilon_t\to 0$, $x_t\to x_\theta\in X_\theta$, and $(x_t,y_t)\in B^{NE}\left(\theta,\tau,\varepsilon_t\right)$ for all $t\in\mathbb{N}$. Such a sequence exists by upper hemi-continuity of $B^{NE}\left(\theta,\tau,\cdot\right)$. Then $y_t=\hat{x}$ for all $t\in\mathbb{N}$. Since F and G are continuous, $G\left(x_t,\hat{x},\varepsilon_t\right)>F\left(x_t,\hat{x},\varepsilon_t\right)$ for all t sufficiently large. Let $v^{(t)}=\left(F\left(x_t,\hat{x},\varepsilon_t\right),G\left(x_t,\hat{x},\varepsilon_t\right)\right)$. For $V\left(\theta,\tau,\varepsilon\right)=B^{NE}\left(\theta,\tau,\varepsilon\right)$ we thus have $v^{(t)}\in V\left(\theta,\tau,\varepsilon_t\right)$ and thus $f\left(\theta,\tau,\varepsilon_t\right)<0$ for all t large enough. **Q.E.D.**

This result has implications for a question of particular interest for economists, namely, whether the common assumption of selfishness is justifiable from an evolutionary perspective. To see this, note that a *homo moralis* with degree of morality $\kappa=0$ is selfish, for (s)he cares only about own material welfare. The theorem implies that if there is some assortativity ($\sigma>0$) and if the equilibrium strategy in a n-player group consisting solely of selfish individuals differs from that in a n-player group consisting solely of *homo moralis* with degree of morality $\kappa=\sigma$, the selfishness type would not be evolutionarily stable; it would be vulnerable to small-scale invasions of other, less selfish types. Instead, *homo moralis* with degree of morality equal to the index of assortativity stands out as being favored by evolution. Not only are these preferences (or any other preferences such that residents with such preferences, in a population with no mutants, are behaviorally indistinguishable from homo moralis with $\kappa=\sigma$) evolutionarily stable; preferences that induce behavior that differs from that of *homo moralis* with $\kappa=\sigma$ are evolutionarily unstable, granted the set of potential types is rich.

The following result obtains by combining the previous results and observations.

Corollary 2 *Assume conditionally independent assortativity. In material games $\Gamma=\langle X,\pi,n\rangle$ where* homo moralis *with degree of morality $\kappa=\sigma$ has a unique best reply to each strategy in X_σ, preference evolution under incomplete information induces the same behaviors as strategy evolution.*

This corollary establishes a second connection between strategy evolution and *homo moralis* preferences, a connection that was established for the case $n=2$ in Alger and Weibull (2013): evolutionarily stable strategies may be viewed as emerging from preference evolution when individuals are not programmed to strategies but are rational and play equilibria under incomplete information.

6 Games in euclidean spaces

How does *homo moralis* behave, in particular when compared to *homo oeconomicus*? More precisely, what are the equilibrium strategies among *homo moralis* of the same degree of morality $\kappa \in [0,1]$? We answer this question first for symmetric games in euclidean spaces in general, then in more detail in several canonical such games. In force of Corollaries 1 and 2, under certain regularity conditions it is sufficient to identify the evolutionarily stable strategies.

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 (17) holds for all $y \neq x$, and we will call a strategy $x \in X$ *locally strictly evolutionarily stable* (LSES) if (17) holds for all $y \neq x$ in some neighborhood of x. If, moreover, $\pi : X^n \to \mathbb{R}$ is differentiable, then so is $H : X^2 \to \mathbb{R}$, and standard calculus can be used to find evolutionarily stable strategies. Let $\nabla_y H(y,x)$ be the gradient of H 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:[24]

Proposition 2 *Let $X \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$, and let $x \in int(X)$. If $H : X^2 \to \mathbb{R}$ 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) $\nabla_y H(y,x)_{|y=x} = \mathbf{0}$,

(ii) $(x-y) \cdot \nabla_y H(y,x) > 0$ *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.

[24] 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

Conditions (i) and (ii) in Proposition 2 can be used to obtain remarkably simple and operational and conditions for evolutionarily stable strategies if the strategy set X is one-dimensional ($k = 1$) and π is continuously differentiable. Writing π_h for the partial derivative of π with respect to its h^{th} argument, one obtains:[25]

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*

$$\pi_1(\hat{\mathbf{x}}) + \sigma \cdot (n-1) \cdot \pi_n(\hat{\mathbf{x}}) = 0, \tag{21}$$

where $\hat{\mathbf{x}}$ is the n-dimensional vector whose components all equal $\hat{x}$. If $\hat{x} \in int\,(X)$ does not satisfy (21), then $\hat{x}$ is evolutionarily unstable.

Proof: If π is continuously differentiable, H 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:

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

By symmetry of π, this equation may be written

$$\sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} \left[\pi_1(\mathbf{x}) + (m-1)\,\pi_n(\mathbf{x})\right] = 0, \tag{22}$$

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

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

the expression in (22) simplifies to $\pi_1(\mathbf{x}) + (n-1) \cdot \sigma \cdot \pi_n(\mathbf{x}) = 0$. **Q.E.D.**

Next we study several canonical interactions for which we can use (21) to determine the set of evolutionarily stable strategies. We conclude by briefly considering a game where not all regularity conditions are met.[26]

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

[26] For examples with $n = 2$, see also Sections 4 and 6 in Alger and Weibull (2013).

6.1 Public goods

Consider a material game in which each individual makes a contribution (or exerts an effort) at some personal cost, and where the sum of all contributions give rise to a benefit to all. More specifically, letting $x_i \geq 0$ denote the contribution of individual i, $\mathbf{x}_{-i}$ the vector of others' contributions, and with $X = \mathbb{R}_+$, let

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = B\left(\sum_{j=1}^{n} x_j\right) - C\left(x_i\right)$$

for some continuous (benefit and cost) functions $B, C : \mathbb{R}_+ \to \mathbb{R}_+$ that are twice differentiable on $\mathbb{R}_{++}$ with $B', C' > 0$, $B'' \leq 0$ and $C'' \geq 0$. Under conditionally independent assortativity, the associated function H (see (18)) is concave, implying that (21) is a necessary and sufficient condition for an individual contribution $\hat{x} > 0$ to be evolutionarily stable. The relevant partial derivatives are

$$\pi_1\left(\hat{\mathbf{x}}\right) = B'\left(n\hat{x}\right) - C'\left(\hat{x}\right) \text{ and } \pi_n\left(\hat{\mathbf{x}}\right) = B'\left(n\hat{x}\right),$$

so a contribution $\hat{x} > 0$ is evolutionarily stable if and only if

$$\left[1 + \left(n-1\right)\sigma\right] \cdot B'\left(n\hat{x}\right) = C'\left(\hat{x}\right). \tag{23}$$

This equation has at most one solution, and it has a unique solution $\hat{x} > 0$ if $\left[1 + \left(n-1\right)\sigma\right] \cdot B'\left(0\right) > C'\left(0\right)$, an arguably natural condition in many applications, and which we henceforth assume to be met.[27] Under this condition, the unique evolutionarily stable contribution is increasing in the index of assortativity. For $\sigma = 0$, equation (23) is nothing but the standard formula according to which "own marginal benefit" equals "own marginal cost"; $\hat{x}$ then corresponds to what *homo oeconomicus* would do when playing against other *homo oeconomicus*. At the other extreme, for $\sigma = 1$, the benevolent social planner's solution obtains; then $\hat{x}$ solves $\max_{x \in X}\left[B\left(nx\right) - C\left(x\right)\right]$. For intermediary values of σ, intermediary values of $\hat{x}$ obtain, and this may or may not be decreasing in group size n.

To see this, consider the case when both B and C are power functions; let $B\left(x\right) \equiv x^b$ for some $b \in \left(0,1\right)$ and $C\left(x\right) \equiv x^c$ for some $c \geq 1$. Then the unique evolutionarily stable individual contribution is

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

[27] We also note that this holds true even if B would be linear, granted $C'' > 0$. For although others' contributions are then strategically irrelevant for the individual player, a positive index of assortativity makes the individual willing to contribute more than under uniform random matching.

This contribution is decreasing (increasing) in group size when the index of assortativity is zero (one). See diagram below, showing $\hat{x}$ as a function of n for $\sigma = 0$, 0.25, 0.5, 0.75 and 1 (higher curves for higher σ).

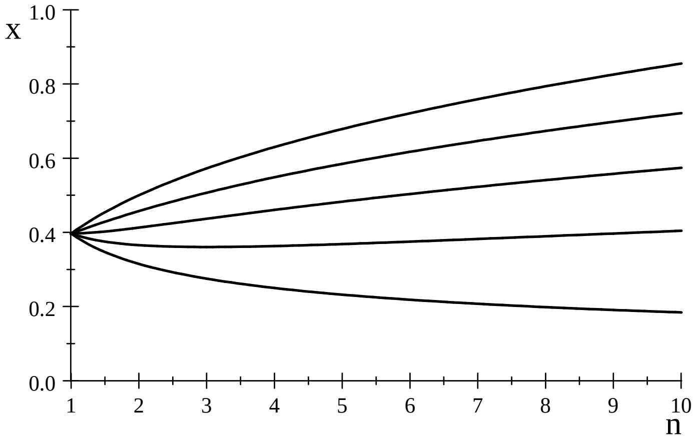

Figure 1: The evolutionarily stable individual contribution in the public-goods game.

The fact that for low σ, the evolutionarily stable individual contribution is decreasing in n is not surprising, since as n increases, the marginal benefit to an individual's material payoff of her contribution decreases as the sum of others' contributions increases. What is more surprising, perhaps, is that the total contribution is increasing in n. To see this, note that whenever C is a power function, (23) can be written as

$$[1 + (n-1)\sigma] \cdot n^{c-1} \cdot B'(n\hat{x}) = C'(n\hat{x}).$$

Given n, the left-hand side is decreasing, and the right-hand side increasing, in the evolutionarily stable aggregate contribution, $n\hat{x}$. The factor before the marginal benefit is increasing in n. The intuition for this is that, beyond the direct, detrimental effect of n on the marginal benefit of a contribution, there is an indirect, beneficial effect, which is related to risk. Indeed, a vanishingly rare mutant faces uncertainty as to the contributions his opponents will make. For $n = 2$, the uncertainty is hefty; a mutant's opponent either makes the same contribution or the resident contribution. As n increases, the mutant's uncertainty becomes less hefty, since then the (empirical) average contribution from other group members is randomly distributed between his own contribution and the resident contribution, with less and less variance as n increases. Hence, adjusting for the direct effect of the higher n on the marginal benefit of making a contribution, the risk associated with mutating decreases as n increases, and hence, a higher total contribution can be sustained.

Remark 1 *The public goods interaction described here is symmetric. However, as noted before, our general model also applies to asymmetric interactions as long as these are* ex ante symmetric*, i.e. such that each individual at the outset is just as likely to be cast in either player role (as, for instance, in a laboratory experiment). To illustrate, suppose that only some individuals are free to give a contribution. More precisely, let $\tilde{A} \subset \{1, ..., n\}$ denote the random set of active players. Suppose further that* ex ante*, each individual faces the same probability $p \in (0,1)$ to get an active player role, that is, to be in the set $\tilde{A}$. A player's strategy, is then a contribution to make if called upon to be active (without being told who else is active). Let x_i denote player i's strategy so defined. We may then write the* ex ante *payoff function of any player i in the symmetric form*

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = p \cdot \mathbb{E}\left[B\left(\sum_{j \in \tilde{A}} x_j\right) \mid i \in \tilde{A}\right] - p \cdot C\left(x_i\right) + (1-p) \cdot \mathbb{E}\left[B\left(\sum_{j \in \tilde{A}} x_j\right) \mid i \notin \tilde{A}\right],$$

where the expectation is taken with respect to the random draw of the subset $\tilde{A}$.

6.2 Team work

Suppose instead that the jointly produced good in the previous example is a private good, split evenly between the members of the group or team. The same analysis applies, with the only difference that the individual benefit be divided by n. One then obtains the following necessary and sufficient condition for the evolutionarily stable individual contribution:

$$\left[\frac{1}{n} + \left(1 - \frac{1}{n}\right)\sigma\right] \cdot B'\left(n\hat{x}\right) = C'\left(\hat{x}\right).$$

Comparing this with the public goods case (equation (23)), we note that the evolutionarily stable individual contribution now is smaller, that it is still increasing in the index of assortativity, and that it is now necessarily decreasing in group size.

6.3 Contests

Many real interactions involve competing for a prize. Examples include competition between job seekers for a vacancy, between firms for a contract, between employees for promotion, etc. Such interactions may be modeled as a contest in which each participant makes a nonnegative effort at some personal cost, and where each participant's effort probabilistically translates to a “result,” and the participant with the “best” result wins the prize. More specifically, let $x_i \geq 0$ be participant i's effort, $\mathbf{x}_{-i}$ the vector of efforts of the others, and let $\tilde{y}_i = x_i + \varepsilon_i$ be participant i’s result (as valued by the "umpire"). With absolutely continuously distributed random terms, ties occur with probability zero. For quadratic costs of effort, the material payoff to participant i is:

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = b \cdot \Pr\left[\tilde{y}_i > \tilde{y}_j \ \forall j \neq i\right] - \frac{1}{2} x_i^2 \tag{24}$$

where $b > 0$ is the value of the prize in question. This defines a continuously and (infinitely) differentiable function on $X^n = \mathbb{R}_+^n$. For Gumbel distributed random terms, the winning probability for each participant i satisfies[28]

$$\Pr\left[\tilde{y}_i > \tilde{y}_j \ \forall j \neq i\right] = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} \quad \forall x \in X^n.$$

From this it is easily verified that a necessary condition (21) for an effort level $\hat{x} > 0$ to be evolutionarily stable boils down to

$$\hat{x} = \frac{1-\sigma}{n} \cdot \left(1 - \frac{1}{n}\right) \cdot b. \tag{25}$$

The evolutionarily stable individual effort is proportional to the value b of the price, linearly decreasing (towards zero) in the index of assortativity, σ, and decreasing in n (recall that $n \geq 2$). Aggregate effort, however, is increasing in n. See diagram, drawn for $b = 2$ and $\sigma = 0.5$.

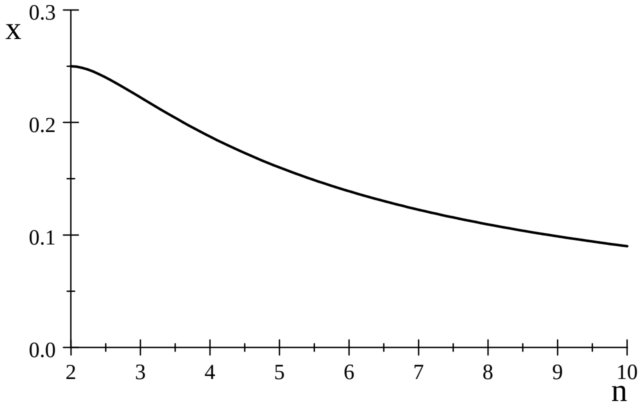

Figure 2: The evolutionarily stable indidividual effort in the contest game.

[28] This is a standard result in random utility theory, see, e.g., Anderson et al. (1992).

6.4 Cournot competition

Consider material payoff functions of the following linear-quadratic form:

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = ax_i - b \cdot \left(\sum_{j \neq i} x_j\right) \cdot x_i - c \cdot x_i^2, \tag{26}$$

for positive a, b and c, where $x_i \geq 0$ is player i's "action". For $b = c$ this is the profit to a firm in Cournot competition among n identical firms facing linear demand with intercept a and slope $-b$, with zero production cost, and with $x_i \geq 0$ being player i's output. Material payoff functions of this form may also represent "common pool" interactions in which the total use by the others affect negatively each individual's benefit from using the common pool.

The material payoff to each player is strictly concave in the player's own strategy, and equation (23) gives

$$\hat{x} = \frac{a}{2c + (1+\sigma)(n-1)b}. \tag{27}$$

The standard result whereby individual (aggregate) output decreases (increases) in n obtains here also, for any $\sigma \in [0,1]$; the more competitors, the less individual output or pool usage but the more aggregate output or usage. Moreover, individual and aggregate output or usage is decreasing in the index of assortativity σ. Interpreted in terms of standard Cournot oligopoly (each firm striving to maximize its profit and $b = c$), the individual output level when there are n firms in the market is then $a/(n+1)$, a result we obtain when $\sigma = 0$. See diagram below, showing the evolutionarily stable output per firm as a function of the number of firms in the market, for $a = b = c = 1$ and $\sigma = 0$, 0.5 and 1 (higher curves for higher σ).

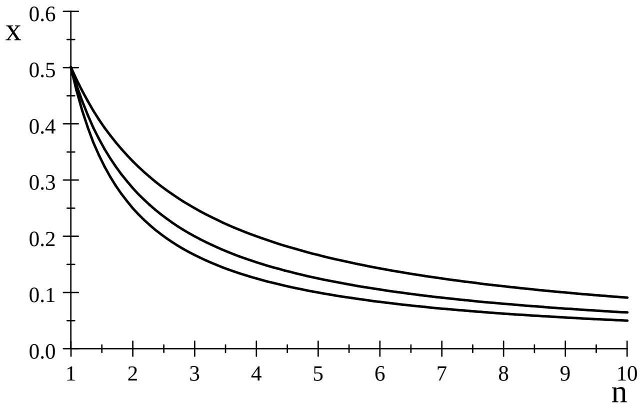

Figure 3: The evolutionarily stable firm output in the Cournot competition game.

By contrast, aggregate output $n\hat{x}$ (or aggregate usage of the common pool) is increasing in the number of firms (users), from $a/(2c)$ towards the limit $a/(b+\sigma b)$ that obtains as the number of firms (users) tends to infinity.

6.5 Helping others

People often help others, also when no reward or reciprocation is expected. To model such behaviors, consider a group of n *ex ante* identical individuals, and suppose that with some exogenous probability $p \in (0,1)$ exactly one individual loses one unit of wealth, with equal probability for all individuals when this happens. The $n-1$ others observe this event, and each of them may then help the unfortunate individual by transferring some personal wealth. These decisions are voluntary and simultaneous. For any individual level of wealth $w \geq 0$, let $v(w)$ be the individual's indirect utility from consumption, where v meets the usual Inada conditions.

We model this as a game where initial wealth is normalized to unity:

$$\pi(x_i, \mathbf{x}_{-i}) = (1-p) \cdot v(1) + p \cdot \left[\left(1 - \frac{1}{n}\right) v(1 - x_i) + \frac{1}{n} v\left(\sum_{j \neq i} x_j\right) \right]$$

Here $x_i \geq 0$ is i's voluntary transfer in case another individual is hit by the wealth loss. Applying equation (23), for an individual transfer $\hat{x} \in (0,1)$ to be evolutionarily stable, it must satisfy

$$v'(1-\hat{x}) = \sigma \cdot v'[(n-1)\hat{x}]. \tag{28}$$

This equation uniquely determines $\hat{x} \in (0,1)$, since the left-hand side is continuously and strictly increasing in $\hat{x}$, from $v'(1)$ towards plus infinity, and the right-hand side is continuously and strictly decreasing in $\hat{x}$, from plus infinity to $v'(n-1)$. It follows immediately from (28) that this transfer is an increasing function of the index of assortativity σ and a decreasing function of group size n. Both effects are intuitively expected; higher assortativity makes helpfulness more worthwhile and more individuals watching the wealth-loss makes free-riding among them the more severe. In the special case when indirect utility is a power function, $v(w) \equiv w^a$ for some $a \in (0,1)$, one obtains

$$\hat{x} = \frac{\sigma^{1/(1-a)}}{n - 1 + \sigma^{1/(1-a)}}.$$

The diagram below shows the evolutionarily stable transfer as a function of group size, for $a = 0.5$ and $\sigma = 0.25$, 0.5 and 1 (higher curves for higher σ). While no transfers are given under uniform random matching ($\sigma = 0$), post-transfer wealth levels are equalized when $\sigma = 1$, so full insurance then holds, while partial insurance obtains for intermediate values of σ.

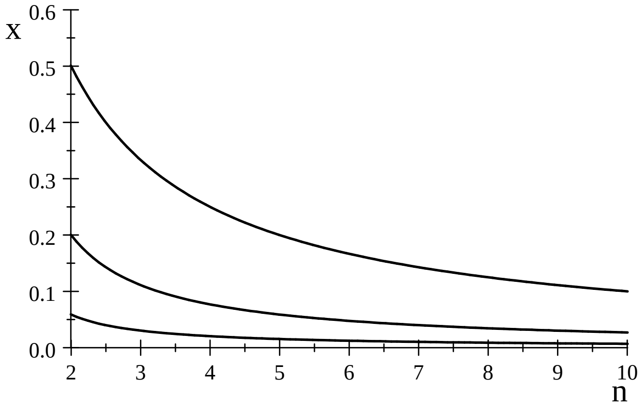

Figure 4: The evolutionarily stable transfer in the helping game.

Furthermore, it is easy to verify that the aggregate transfer, $n\hat{x}$, is increasing in n and converges to $\sigma^{1/(1-a)}$ as $n \to \infty$.

6.6 Coordination

Many strategic interactions involve some element of coordination and thus the possibility of multiple equilibria. In order to clarify whether or not *homo moralis* then has unique or multiple best replies, we carry out this analysis directly in terms of equilibrium play among *homo moralis*. Consider the symmetric two-by-two game with material payoffs

$$\begin{pmatrix} a,a & 0,c \\ c,0 & b,b \end{pmatrix}$$

for $0 < b < a$ and $c < a$. There are three Nash equilibria and two strict equilibria, with payoffs (a, a) and (b, b).[29] Let x be a player's probability for playing the first pure strategy. Then $X = [0, 1]$ and

$$\pi(x, y) = a \cdot xy + b \cdot (1 - x)(1 - y) + c \cdot (1 - x) y$$

[29] The first strict equilibrium thus payoff dominates the second, and the second risk dominates the first iff $b \geq a - c$.

We proceed to identify the *homo moralis* fixed-point set X_κ for all $\kappa \in [0,1]$. For $\kappa = 0$ one immediately obtains

$$X_0 = \left\{0, \frac{b}{a+b-c}, 1\right\}.$$

For $0 < \kappa \leq 1$,

$$u_\kappa(x,y) = (1-\kappa)\cdot(axy + b(1-x)(1-y) + c(1-x)y) + \kappa\cdot\left(ax^2 + b(1-x)^2 + c(1-x)x\right),$$

so

$$\frac{\partial u_\kappa(x,y)}{\partial x} = (1-\kappa)\cdot(ay - b(1-y) - cy) + \kappa\cdot(2ax - 2b(1-x) + c - 2cx)$$

and $\partial^2 u_\kappa(x,y)/\partial x^2 = 2\kappa\cdot(a+b-c) > 0$; the utility function of *homo moralis* of any positive degree of morality is strictly convex in his/her own strategy. Hence, $X_\kappa \subseteq \{0,1\}$ when $0 < \kappa \leq 1$. As is easily verified,

$$u_\kappa(1,1) = a + 4b\kappa + (1-\kappa)c > \kappa b + (1-\kappa)c = u_\kappa(0,1)$$

so $1 \in X_\kappa$ for all $\kappa \in [0,1]$. Likewise, $u_\kappa(0,0) = b$ and $u_\kappa(1,0) = a\kappa$, so that $0 \in X_\kappa$ if and only if $\kappa \leq b/a$. Hence, X_κ has three elements when $\kappa = 0$, two elements when $0 < \kappa \leq b/a$ and one elements when $\kappa > b/a$.

Applying Corollary 1, we conclude that no mixed strategy is evolutionarily stable when $\kappa > 0$, and that $x = 1$ is the unique evolutionarily stable strategy when $\kappa > b/a$. For $\kappa = 0$ it is well-known (see e.g. Weibull, 1995), that both pure strategies are evolutionarily stable while the mixed Nash equilibrium strategy is not. From Corollary 2, we deduce that both pure strategies are evolutionarily stable when $\kappa < b/a$.

In sum: All evolutionarily stable strategies are pure. For $\kappa < b/a$ both pure strategies are evolutionarily stable while for $\kappa > b/a$ only the first pure strategy is evolutionarily stable.

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 motivational types by way of natural selection. The framework is based upon a general definition of an evolutionarily stable type, where an individual's type guides his or her behavior in interactions in groups of any size. The framework may be applied to interactions where others' preferences are known or unknown, and it allows for assortativity in the process by which individuals are matched together to interact. Since our analysis focuses on whether a homogenous population may withstand a small-scale invasion of individuals of a different type, a key factor is the probability with which mutants are matched with other mutants when these are vanishingly rare. In this paper we focus on matching processes in which such assortativity may be conveniently expressed in terms of the probability that another individual with whom a mutant interacts also is a mutant (the index of assortativity; Bergstrom, 2003).

As a benchmark, we first apply the framework to strategy evolution. We then apply the model to preference evolution under incomplete information with few assumptions about the nature of preferences. As in our model with pairwise interactions (Alger and Weibull, 2013) we find that the class of *homo moralis* preferences stands out as a winner in the evolutionary race. Indeed, we find that (a) within the class of utility functions that are continuous in the strategy profile, *homo moralis* preferences are evolutionarily stable, and (b) under quite weak assumptions, any preferences that lead to different behaviors from that of *homo moralis* with the "right" degree of morality are evolutionarily unstable. Furthermore, equilibrium behavior in a homogeneous population consisting of *homo moralis* with this degree of morality is the same as under strategy evolution.

Our model clarifies how group size affects the evolutionarily stable types and the ensuing behaviors. As shown above, group size has no effect on evolutionarily stable types when these are preferences under incomplete information; *homo moralis* preferences with degree of morality equal to the index of assortativity stand out as the winner in the evolutionary race, independent of group size and of the (material) game played. By contrast, group size does affect equilibrium behavior, in groups consisting of identical *homo moralis*, as illustrated in the examples.

Although general, our model relies on a number of simplifying assumptions. Relaxation of these is a task that has to be left for future research. Moreover, we only apply our general definition of evolutionary stability to two cases, strategy evolution and preference evolution under incomplete information. Applications to complete or partially incomplete information are called for, in particular in settings where the random matching is not exogenous, as here, but at least partly endogenous. This is a major analytical challenge, however, opening the door to signalling and mimicry, a very rich, important and exciting research area.

8 Appendix: A class of matching processes

The matching process to be outlined here is a variant of the model of pairwise matching sketched in Section 5.2 of Alger and Weibull (2013). Let n, k and N be integers greater than one, and imagine a finite population of individuals $i \in \{1, 2, ..., N\}$. The population is divided into "islands" of equal size, each island consisting of $k > n$ individuals (and N is some multiple of k). Initially all individuals are of type θ. Suddenly there is an outburst of mutation to τ on one of the islands, and only there. Each individual on that island has probability μ of mutating and individual mutations are statistically independent. Hence, the random number M of mutants is binomially distributed $M \sim Bin\,(k, \mu)$. We note that in this mutation process the same random number M is also the total number of mutants in the population at large, so the population share M/N of mutants is a random variable with expectation $\varepsilon = \mathbb{E}\,[M/N] = \mu k/N$. A group of size n is now formed (to play our game) 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 randomly recruited to form the group, drawn as a random sample without replacement from amongst the k islanders and with equal probability for each islander to be sampled.

Consider an individual i who has been recruited to the group, and assume that the only information we have about her is her type. If the individual is of type τ, it is necessary that $M > 0$ and that she is from the island where the mutation occurred, so the random number of *other* mutants in her group is binomially distributed $Bin\,(n-1, \mu)$. With X_i denoting the type of individual i, and T the total number of mutants in the group, we have, for $m = 1, 2, ..., n$:

$$\Pr\left[T = m \mid X_i = \tau\right] = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \mu^{m-1} \left(1-\mu\right)^{n-m}. \tag{29}$$

If "our" individual i instead is of the resident type θ, then $M = 0$ is possible and she may well be from another island than where the mutation occurred. We thus have

$$\Pr\left[T = m \mid X_i = \theta\right] \leq \frac{k}{N} \cdot \begin{pmatrix} n-1 \\ m \end{pmatrix} \mu^{m} \left(1-\mu\right)^{n-m-1}$$

for all $m > 0$. Moreover, for any two group members i and j:

$$\Pr\left[X_j = \tau \mid X_i = \tau\right] = \mu \quad \text{and} \quad \Pr\left[X_j = \tau \mid X_i = \theta\right] \leq \mu k/N.$$

Keeping μ, n and k constant, we may write $\Pr[\theta|\theta,\varepsilon]$ for $\Pr[X_j = \theta \mid X_i = \theta]$ and $\Pr[\theta|\tau,\varepsilon]$ for $\Pr[X_j = \theta \mid X_i = \tau]$, and these are continuous functions of $\varepsilon = \mu k/N$. In addition, we have $1 - \varepsilon \leq \Pr[\theta|\theta,\varepsilon] \leq 1$ and $\Pr[\theta|\tau,\varepsilon] = 1 - \mu$. Letting $N \to \infty$, we obtain $\varepsilon \to 0$ and $\Pr[\theta|\theta,\varepsilon] \to 1$. Hence, $\lim_{\varepsilon\to 0}\phi(\varepsilon) = \mu$, so $\sigma = \mu$. Moreover, in our general notation: $\lim_{\varepsilon\to 0} p_m(\varepsilon) = 0$, and

$$q_m(\varepsilon) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \sigma^{m-1}(1-\sigma)^{n-m}$$

for all $\varepsilon \in (0,1)$ and $m = 1,..,n$, so (5) follows immediately.

References

Acemoglu, D. and M.K. Jensen (2013) "Aggregate comparative statics," *Games and Economic Behavior*, 81, 27 - 49.

Akçay, Erol, Jeremy Van Cleve, Marcus W. Feldman, and Joan Roughgarden (2009) "A Theory for the Evolution of Other-Regard Integrating Proximate and Ultimate Perspectives," *Proceedings of the National Academy of Sciences,* 106, 19061–19066.

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

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

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

Alger, I. and J. Weibull (2013): "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81:2269-2302.

Anderson, S.P., A. de Palma, and J.-F. Thisse (1992): *Discrete Choice Theory of Product Differentiation.* VCambridge (USA): MIT Press.

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

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

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

Bergstrom, T. (2013): "Measures of Assortativity," *Biological Theory,* 8, 133-141.

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

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

Broom, M., C. Cannings and G.T. Vickers (1996): "Choosing a Nest Site: Contests and Catalysts", *Amer. Nat.* 147, 1108-1114.

Broom, M., C. Cannings and G.T. Vickers (1997): "Multi-Player Matrix Games", *Bulletin of Mathematical Biology* 59, 931-952.

Bukowski, M., and J. Miękisz (2004): "Evolutionary and asymptotic stability in symmetric multi-player games", *International Journal of Game Theory* 33, 41-54.

Cannings, C., and J.C. Whittaker (1995): "The Finite Horizon War of Attrition", *Games and Economic Behavio*r 11, 193-236.

Corchón, L. (1996): *Theories of Imperfectly Competitive Markets.* Berlin: Springer Verlag.

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

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

Dubey, P., A. Mas-Colell, and M. Shubik (1980): "Efficiency Properties of Strategic Market Games", *Journal of Economic Theory* 22, 339-362.

Duffie, D. and Y. Sun (2012): "The Exact Law of Large Numbers for Independent Random Matching", *Journal of Economic Theory* 147, 1105-1139.

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

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

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

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

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

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

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

Haigh, J., and C. Cannings (1989): "The n-Person War of Attrition", *Acta Applic. Math.* 14, 59-74.

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

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

Heifetz, A., C. Shannon, and Y. Spiegel (2007a): "The Dynamic Evolution of Preferences," *Economic Theory*, 32, 251-286.

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

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

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

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

Koçkesen, L., E.A. Ok, and R. Sethi (2000b): "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior* 31, 303-310.

Luenberger, D.G. 1969. *Optimization by Vector Space Methods.* New York: John Wiley & Sons.

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

Munkres, James (1975): *Topology, a First Course.* London: Prentice Hall.

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

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

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

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

Robson, A.J., and B. Szentes (2014): "A Biological Theory of Social Discounting," forthcoming, *American Economic Review.*

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

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

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

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