Evolutionary Models of Preference Formation

Ingela Alger* and Jörgen W. Weibull†

September 17, 2018‡

Abstract

The literature on the evolution of preferences of individuals in strategic interactions is vast and diverse. We organize the discussion around the following question: Supposing that material outcomes drive evolutionary success, under what circumstances does evolution promote *Homo oeconomicus*, defined as material self-interest, and when does it instead lead to other preferences? The literature suggests that *Homo oeconomicus* is favored by evolution only when individuals' preferences are their private information and the population is large and well-mixed so that individuals with rare mutant preferences almost never get to interact with each other. If rare mutants instead interact more often (say, due to local dispersion), evolution instead favors a certain generalization of *Homo oeconomicus* including a Kantian concern. If individuals interact under complete information about preferences, evolution destabilizes *Homo oeconomicus* in virtually all games.

Keywords: Preference evolution, indirect evolution, evolutionary stability, assortativity, altruism, spite, morality.

JEL codes: C73, D01, D03.

*Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, and Institute for Advanced Study in Toulouse. ingela.alger@tse-fr.eu

†Stockholm School of Economics, Institute for Advanced Study in Toulouse, and Toulouse School of Economics. jorgen.weibull@hhs.se

‡This survey was prepared for *Annual Reviews Economics* (doi: 10.1146/annurev-economics-080218-030255). Financial support by ANR-Labex IAST is gratefully acknowledged. J. Weibull also thanks the Knut and Alice Wallenberg Research Foundation and Agence Nationale de la Recherche for funding (Chaire IDEX ANR-11-IDEX-0002-02).

1 Introduction

Economics traditionally takes individuals' motivations—their preferences—as fixed and given. Hence, the predictive power of economics depends to a large extent on the assumptions made regarding these motivations. Since the development of general equilibrium models in the 1950s, consumers are assumed to have fixed and given preferences over the consumption alternatives they face, preferences that do not depend on other consumers' or producers' choices or preferences. Likewise, producers have their expected profits as their sole motivation, and, when economists look inside firms, managers are assumed to only care about their own monetary outcome. Such assumptions separate individual economic agents from each other in a way that allows for clear-cut and powerful analyses of a wide range of economic issues, both under perfect and imperfect competition, and with or without externalities. The same is true in non-cooperative game theory when each player's payoff function traditionally is taken to depend only on the player's own material outcome. However, even slight deviations from standard assumptions may have significant consequences for predictions. For example, individuals who care about fairness may refuse unfair pay increases, competitive individuals may engage in inefficient rat races, morally motivated consumers may prefer expensive "green" products to "brown" products that give them the same consumption utility. And any standard economics model, say an exchange economy under perfect competition, implicitly relies on trust—in other individuals, in institutions, in the issuer of fiat money—which is not part of the model.

By contrast to these models, work by economists in earlier times sometimes included reflections on motivations other than material self interest (e.g., Smith, 1759, Edgeworth, 1881, Veblen, 1899). Moreover, there is now ample empirical evidence in the economics literature—based both on experiments in laboratories and in the field—that people do not behave in line with pure material self-interest. For instance, individuals sometimes refuse unfair offers, or honor trust, even at a cost to themselves, and they sometimes make fair offers at a monetary cost to themselves, and trust strangers, even at a risk of losing money. Such fair offers and trust may be motivated by non-materialistic dispositions such as a concern for fairness, altruism, or morality, for instance, but it may also be driven by pure material self-interest if the decision-maker believes that the other party has such a disposition. In any event, some deviation from the assumption of material self-interest is called for in order to explain observed behaviors. Of course, one may question the external validity of the empirical results. Maybe all or most individuals would "learn" to maximize material self-interest over

time. But this presumes that material self-interest has a long-run survival value over and beyond all forms of other-regarding preferences. Is this presumption generally valid? If not, what other motivational factors may increase the survival value? These are the basic question addressed in the literature on the evolutionary foundations of preferences, a literature we here discuss.

Foundational questions concerning pure self-interest and profit maximization were raised already by Alchian (1950) and Friedman (1953). A few decades later, the literature on the evolutionary foundations of other-regarding preferences took off. Important pioneering contributions were Becker (1976), Hirshleifer (1977, 1978), Frank (1987), Güth and Yaari (1992), to mention a few. Particularly influential for the subsequent formal modelling of preference evolution was the paper by Güth and Yaari. They proposed what was to become called the *indirect evolutionary approach*: "Instead of assuming that individual preferences are exogenously given, we think of an evolutionary process where preferences are determined as evolutionarily stable." (op.cit.) They illustrated this approach by means of a simple two-player game in which players knew each others' preferences, and where reciprocity became evolutionarily stable (see also Frank, 1987). It was in Güth and Kliemt (1998) that the name of the approach was coined, and in this paper the authors showed that trustworthiness, if observable, may be evolutionarily stable. In a sense, these findings confirmed Schelling's (1960) observation about the potential strategic advantage of commitment.

The current literature on preference evolution is vast and methodologically diverse. Instead of trying to cover the whole literature, we extract a few main lessons learned so far. The survey is organized around one question: Does or does not evolution favor pure self-interest as the sole motivation? More precisely, in what context does evolution lead to *Homo oeconomicus*, defined as rational and purely-self interested behavior, and in what contexts does it not? Research results providing affirmative answers to the first question are presented in Sections 3, while research providing affirmative answers to the second question are discussed in Section 4, where we also discuss what other motivations are then favored by evolution. We summarize the answers in these two sections in the form of eleven informally stated observations, major "lessons" that we draw from existing work. However, in order to present and discuss the relevant background research, we first, in Section 2, set up a general model framework encompassing most of the existing models. We conclude in Section 5 by way of discussing shortcomings and lacunas in the existing literature, and by suggesting a few potentially promising directions for future work.

2 Model framework and preliminaries

Much of the literature can be discussed within the following theoretical framework. Consider a population in which individuals are now and then matched into groups of fixed size n to play a symmetric *material game* $\Gamma = \langle n, X, \pi \rangle$, where X is the set of (pure or mixed) strategies, the same for all players, and $\pi(x_i, \boldsymbol{x}_{-i}) \in \mathbb{R}$ is the *material payoff* from using strategy $x_i \in X$ against $\boldsymbol{x}_{-i}$, the strategy profile of all other individuals in the group. The following assumptions will be maintained throughout, unless stated otherwise: (a) the strategy set X is a non-empty and compact set in a normed vector space, (b) the material payoff function $\pi : X^n \to \mathbb{R}$ is continuous and *aggregative* in the sense that the material payoff $\pi(x_i, \boldsymbol{x}_{-i})$ is invariant under permutation of the components of $\boldsymbol{x}_{-i}$, the strategies used by the other players in i's group. In the special case of finite games, we will follow the practice in game theory and let the strategy set X be the unit simplex $\Delta(S)$ of mixed strategies, where S is the finite set of pure strategies available to each player role. The material payoff function π is then linear in the player's own mixed strategy (given the other players' strategies).

Examples of strategic interactions covered by this model include (a) public goods games between identical individuals where the amount of the public good is a symmetric function of their contributions (for example the sum), (b) contests between identical contestants, (c) symmetric coordination games, etc. These games may be one-shot, multi-stage, or repeated. Importantly, while the material game has to be symmetric, it encompasses *ex post* asymmetric interactions as long as all participants are *ex ante* equally likely to be in any given player role. For instance, an ultimatum bargaining game in which the flip of a fair coin determines who is the proposer is *ex post* asymmetric but *ex ante* symmetric. The same is true for an n-player "team-leadership" game in which each player is equally likely to be the team leader, with the others as ordinary team members. Likewise if information, or some other relevant characteristic such as skill, is *ex post* asymmetric but *ex ante* symmetric. In such *ex ante* symmetric, but *ex post* asymmetric interactions, a participant's strategy is conditioned on the player role, information or characteristic given.

There is also a *type space* Θ upon which we will let evolution operate. Each individual in the population is of some such type $\theta \in \Theta$, which may be known or unknown by others. We will restrict attention to population states in which the type distribution μ has finite support, that is, where there are only finitely many types $\theta \in \Theta$ present (with $\mu(\theta) > 0$). More importantly, we will focus on only two extreme cases concerning individuals' information about each others' types: *complete information*, when all individuals in each group know—

and hence can condition their behavior on—the types of all members, that is, their group's *type profile* $\boldsymbol{\theta} = (\theta_1, ..., \theta_n) \in \Theta^n$. In the other extreme case, which we refer to as *incomplete information*, each individual's type is her private information. Types are inherited from one generation to the next, and the question is which types will prevail in the long run, when evolutionary forces are at work. If types are genetically determined, evolution by natural selection is driven by fitness, which essentially is the ability to survive and/or produce viable offspring. If types are culturally determined, it is the ability to produce cultural offspring which determines evolutionary success. In general, fitness is not modeled explicitly in the economics literature, however. Instead, the material payoff function π is taken to represent "fitness" or "evolutionary success".

The indirect evolutionary approach builds on classic evolutionary game theory, in which the type space is the set of (pure or mixed) strategies in the material game, $\Theta = X$, i.e., individuals are "programmed" to play a specific strategy (for a textbook treatment, see Weibull, 1995). We will refer to this case as *strategy evolution*. In the literature on preference evolution, the types are instead *utility functions* $f : X^n \to \mathbb{R}$ over strategy profiles $(x_i, \boldsymbol{x}_{-i}) \in X^n$. Each individual is assumed to strive to maximize the expected value of his or her utility function. In game-theoretic terminology, an individual's utility function is thus his or her (subjective) *payoff function* when matched to play the material-payoff game with other individuals. Focus will here be on the set F of continuous and aggregative utility functions $f : X^n \to \mathbb{R}$. This set contains the material-payoff function π. Obviously, any strictly increasing transformation of any utility function f represents the same preferences over strategy profiles. Hence, by a "type f " we mean the whole (equivalence) class $[f]$ of utility functions in F that represent the same preferences. Individuals with preferences in $[\pi]$ will be called *Homo oeconomicus*.

While individuals of the same type, so defined, have the same set of best replies to every strategy profile, we will also need to keep track of types whose best-reply sets happen to overlap for *some* strategy profiles. Let $X_f \subseteq X$ denote the set of symmetric Nash equilibrium strategies when all players are of the same type $f \in F$:

$$\hat{x} \in \arg\max_{x \in X} \; f\left(x, \hat{\mathbf{x}}^{(n-1)}\right), \tag{1}$$

where $\hat{\mathbf{x}}^{(n-1)}$ is the $(n-1)$-dimensional vector whose components all equal $\hat{x}$. We define the set *behaviorally distinct* types from f as

$$D_f = \left\{ g \in F : \arg\max_{x \in X} \; g\left(x, \hat{\mathbf{x}}^{(n-1)}\right) \cap \arg\max_{x \in X} \; f\left(x, \hat{\mathbf{x}}^{(n-1)}\right) = \varnothing \text{ for all } \hat{x} \in X_f \right\}. \tag{2}$$

In words: in a group where everybody else plays some $\hat{x} \in X_f$, an individual of type g never chooses a strategy that could have been rationally chosen by an individual of type f.

In real-life populations, there is typically a variety of traits present in a population at any point in time. Moreover, the distribution of traits tends to change over time. While some models in the literature are dynamic and allow for multiple types that are simultaneously present in the population, most of the analyses are static and presume that at most two types are simultaneously present. Such analyses often rest on a simple but powerful thought experiment that builds on and extends the notion of evolutionary stability, initially introduced by Maynard Smith and Price (1973).

2.1 Strategy evolution

Classic evolutionary stability analysis considers symmetric and finite two-player games under the hypotheses that *(A)* the population is infinitely large (modelled as a continuum), *(B)* matching is uniformly random, and *(C)* each individual is programmed to play some pure or mixed strategy. The thought experiment behind the definition of evolutionary stability is the following, here generalized to an arbitrary group size n. Suppose that initially there is some strategy, say x, that is used by everyone in the population. This is the *resident* strategy. Suddenly, another strategy, say y, appears in the population. This is the *mutant* strategy. The question is then whether the mutants, who sometimes meet each other and sometimes the residents, earn a higher or lower average material payoff than the residents. Let $\varepsilon \in (0,1)$ be the population share of mutants. The type distribution μ is then binary, with $\mu(x) = 1-\varepsilon$ and $\mu(y) = \varepsilon$. Assumptions *(A)* and *(B)* together imply that the probability that a given resident is matched with precisely m mutants (for $m = 0, 1, ..., n-1$) is binomial,

$$p_m(\varepsilon) = \binom{n-1}{m} \cdot (1-\varepsilon)^{n-m-1} \cdot \varepsilon^m, \tag{3}$$

and that this is also the probability, $q_m(\varepsilon)$, that a given mutant is matched with m *other* mutants. Hence, the average material payoffs to a resident and to a mutant are, respectively,

$$\begin{cases} \bar{\pi}_R(x,y,\varepsilon) = \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot \pi\left(x, \left(\mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right)\right) \\ \bar{\pi}_M(x,y,\varepsilon) = \sum_{m=0}^{n-1} q_m(\varepsilon) \cdot \pi\left(y, \left(\mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right)\right) \end{cases}, \tag{4}$$

where $\mathbf{x}^{(k)}$ is the k-dimensional strategy vector whose components all equal x, $\mathbf{y}^{(k)}$ is the k-dimensional vector whose components all equal y, and m stands for the total number of mutants in the group, and $q_m(\varepsilon) = p_m(\varepsilon)$. Note that $\bar{\pi}_R(x,y,\varepsilon)$ and $\bar{\pi}_M(x,y,\varepsilon)$ are continuous in ε.

The resident strategy is *evolutionarily stable* (Maynard Smith and Price, 1973) if there for every mutant strategy $y \neq x$ exists $\bar{\varepsilon}_y > 0$ such that

$$\bar{\pi}_R(x, y, \varepsilon) > \bar{\pi}_M(x, y, \varepsilon) \quad \forall \varepsilon \in (0, \bar{\varepsilon}_y). \tag{5}$$

Evolutionary stability may be conveniently characterized as follows. Let $\Delta\pi(x, y, \varepsilon) = \bar{\pi}_R(x, y, \varepsilon) - \bar{\pi}_M(x, y, \varepsilon)$. This payoff difference (sometimes called "the score function") is differentiable in ε, and the inequality in (5) holds if and only if either (i) $\Delta\pi(x, y, 0) > 0$, or (ii) $\Delta\pi(x, y, 0) = 0$ and $\partial\Delta\pi(x, y, \varepsilon)/\partial\varepsilon|_{\varepsilon=0} > 0$. We also note that a necessary condition for x to be evolutionarily stable is that

$$\pi\left(x, \mathbf{x}^{(n-1)}\right) \geq \pi\left(y, \mathbf{x}^{(n-1)}\right) \ \forall y \in X. \tag{6}$$

In other words, if a strategy x is evolutionarily stable, then x has to be a symmetric Nash-equilibrium strategy of the game in which all players have payoff function π. A population in which an evolutionarily stable strategy is played is thus behaviorally indistinguishable from a population consisting of *Homo oeconomicus* individuals who freely choose their strategy in order to maximize their own material payoff.

2.2 Preference evolution

The thought experiment underlying the definition of evolutionary stability of preferences is similar. Suppose, thus, that there is some utility function, say $f \in F$, that at some point in history is used by everyone in the population. This is the resident utility function. Suddenly, another utility function, say $g \in F$, appears in the population. This is the mutant utility function. The question is, again, whether the mutants, when rare, earn a higher or lower average material payoff than the residents. However, types are now defined in terms of preferences, and individuals are assumed to optimally adapt their choice of strategy to the situation at hand. The literature on preference evolution has stayed close to standard economic theory by requiring that the two types' average material payoffs should be evaluated when population play is in equilibrium (in terms of individuals' preferences). More precisely, focus is typically on *type homogenous (Bayesian) Nash equilibria*, in which all individuals of the same type use the same strategy. This does not presume that populations always play such equilibria. The requirement is that if the population happens to be in some such equilibrium, at least then it should not be possible for mutants to "invade" the population.[1]

[1] A usual interpretation (see e.g. Sandholm, 2001) in the literature is that behavioral adaptation occurs on a faster time-scale than preference adaptation. The stability condition then is that if behavioral adaptation

Let (f, g, ε) be the *population state* in which the two types f and g are represented in population shares $1-\varepsilon$ and ε, respectively. Of particular interest will be population states in which ε is positive but small, i.e., when the "mutant" type g is rare. Such population states amount to binary type distributions in which $\mu(f) = 1-\varepsilon$ and $\mu(g) = \varepsilon$ for $\varepsilon > 0$ small. For any given population state (f, g, ε), random matching process, and information setting (complete or incomplete information), let $\Pi(f, g, \varepsilon) \subset \mathbb{R}^2$ be the associated set of average equilibrium material-payoff pairs $(\bar{\pi}_R, \bar{\pi}_M) \in \mathbb{R}^2$. In other words, for each $(\bar{\pi}_M, \bar{\pi}_R) \in \Pi(f, g, \varepsilon)$, $\bar{\pi}_R$ is the average material payoff accruing to residents and $\bar{\pi}_M$ is the average material payoff accruing to mutants, in some type homogenous (Bayesian) Nash equilibrium.

We are now in a position to formalize generalized notions of evolutionary stability and instability in the spirit of Maynard Smith and Price (1973). Let $\Theta \subseteq F$. A type $f \in \Theta$ is *evolutionarily stable against a type* $g \in \Theta$ if there is an $\bar{\varepsilon} > 0$ such that $\bar{\pi}_R > \bar{\pi}_M$ for all $(\bar{\pi}_R, \bar{\pi}_M) \in \Pi(f, g, \varepsilon)$ and all $\varepsilon \in (0, \bar{\varepsilon})$. A type $f \in \Theta$ is *evolutionarily stable* in Θ if it is evolutionarily stable against all $g \in \Theta \cap (\sim [f])$. A type $f \in \Theta$ is *evolutionarily unstable* if there is a type $g \in \Theta$ and an $\bar{\varepsilon} > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon})$ there is some $(\bar{\pi}_R, \bar{\pi}_M) \in \Pi(f, g, \varepsilon)$ with $\bar{\pi}_M > \bar{\pi}_R$.

The requirement for stability is stringent, since it requires the residents to strictly materially outperform mutants in all Nash equilibria. By contrast, for a preference type to be unstable it is sufficient that there exists one mutant type that earns a higher material payoff in one Nash equilibrium, whenever the mutant is sufficiently rare. We finally note that, because we require strict outperformance, there are in general utility functions that are neither stable nor unstable.[2]

With this model in hand, we can proceed to identifying conditions under which *Homo oeconomicus* prevails and thereafter conditions under which *Homo oeconomicus* does not prevail.

---

leads towards a Nash equilibrium (in terms of preferences), then it should not be the case that mutants fare materially better in that equilibrium.

[2]We note that this definition of evolutionary stability encompasses also strategy evolution by way of letting $\Theta \subset F$ be the subset of utility functions f for which some strategy $\hat{x} \in X$ is strictly dominant (for example $f(x_i, \boldsymbol{x}_{-i}) \equiv -\|x_i - \hat{x}\|$).

3 Settings in which *Homo oeconomicus* prevails

3.1 Decision problems

The above setting includes decision problems, that is, material-payoff games in which each player's material payoff depends only on his or her own strategy. Formally, $\Gamma = \langle n, X, \pi \rangle$ is a *decision problem* if there exists a continuous function $v : X \to \mathbb{R}$ such that $\pi(x, \boldsymbol{y}) = v(x)$ for all $x \in X$ and $\boldsymbol{y} \in X^{n-1}$. In decision problems, *Homo oeconomicus* is evidently evolutionarily stable against all behaviorally distinct types. Moreover, any type $f \in F$ that is behaviorally distinct from *Homo oeconomicus* is unstable. In sum:

> **Observation 1**: In decision problems, *Homo oeconomicus* is evolutionarily stable against all behaviorally distinct types, and the latter types are evolutionarily unstable.

It should be noted, however, that we here assume that individuals perceive the situation (strategy set and material payoffs) without errors. When such perception errors exist, preferences other than *Homo oeconomicus* may be stable (see, e.g., Rayo and Becker, 2007, and Robson and Samuelson, 2011). The literature on preference evolution that we are concerned with here is not about decision problems, so we now turn to strategic interactions.

3.2 Continuum population, uniform random matching and incomplete information

Assume that *(A)* the population is a continuum, *(B)* player matching is uniformly random, *(C)* the type set is a subset of the set of all continuous utility functions, $\Theta \subseteq F$, and *(D)* each individual's type is his private information. The set of type-homogenous Bayesian Nash Equilibria (BNE) in a population state (f, g, ε), is then the set of strategy pairs $(\hat{x}, \hat{y}) \in X^2$ that satisfy

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot f\left(x, \left(\hat{\mathbf{x}}^{(n-m-1)}\right), \hat{\mathbf{y}}^{(m)}\right)\right) \\ \hat{y} \in \arg\max_{y \in X} \ \sum_{m=0}^{n-1} q_m(\varepsilon) \cdot g\left(y, \left(\hat{\mathbf{x}}^{(n-m-1)}, \hat{\mathbf{y}}^{(m)}\right)\right) \end{cases} . \tag{7}$$

In order to investigate whether $f \in \Theta$ is evolutionarily stable against $g \in \Theta$, one has to evaluate the material payoffs in all solutions $(\hat{x}, \hat{y}) \in X^2$ to this system of fixed-point conditions for all small $\varepsilon > 0$. This may seem a daunting task, since residents may well

vary their equilibrium behavior radically when the population share of mutants changes even slightly. For example, the resident type may play a mixed-strategy equilibrium when no other type is around, that is, when $\varepsilon = 0$, but switch to a pure strategy when ε turns positive. Two approaches have been adopted in the literature to deal with this issue.

Dekel, Ely, and Yilankaya (2007) analyze finite and symmetric two-player material games. They focus on utility functions $f : X^2 \to \mathbb{R}$ of the bilinear form $f(x, y) = \sum_{k,h \in S} x_k u_{kh} y_h$ for some numbers u_{hk}. This defines their type space Θ as a finite-dimensional subspace G of F, those utility functions that meet the expected-utility hypothesis for mixed strategies. Evidently, one may identify each utility function $f \in G$ with its associated subjective payoff matrix $u = (u_{hk})$, and refer to this matrix as its type.[3] Let $\sigma : \Theta \to X$ denote a rule that to every type θ assigns some mixed strategy. Their stability concept differs slightly from the one given above, and applies to any type distribution μ with finite support on $\Theta = G$. They call such a type distribution μ *stable* if there exists a rule σ which is a best response to itself given the type distribution μ, and such that the pair (μ, σ) constitutes what they call a *stable configuration*. The definition of a stable configuration is somewhat technical but essentially requires (i) that all resident types—those in the support of μ—earn the same average material payoff under σ, and (ii) that a small population share ε of mutants cannot destabilize the configuration. Destabilization occurs either if a mutant type materially outperforms the resident types, or if a mutant type induces the residents to switch behavior far from that prescribed by σ. The authors show that when interactions take place under incomplete information, then for a distribution μ to be stable all residents must play some strategy $\hat{x} \in X_\pi$, that is, behave as *Homo oeconomicus* does in symmetric equilibrium in the absence of mutants. They also show that if $(\hat{x}, \hat{x})$ is a strict equilibrium in the material-payoff game for some strategy $\hat{x} \in X_\pi$, then play of $\hat{x}$ is compatible with a stable configuration. In sum, for two-player finite (linear) games:

> **Observation 2**: Under uniform random matching in a continuum population, and unobservable preference types, configuration stability implies symmetric Nash-equilibrium play in the material-payoff game, and symmetric strict equilibrium play in the material-payoff game is configuration stable.

In Alger and Weibull (2013, 2016) we establish sufficient conditions for *Homo oeconomicus*

[3]For any number $m \in \mathbb{N}$ of pure strategies, let $U = \mathbb{R}^{m^2}$. Then each $u \in U$ defines an $m \times m$ payoff matrix where the entries are the row-player's payoffs, and this defines a utility function f_u in the m^2-dimensional subspace G of F, where $f_u(x, y) = \sum_{k,h \in S} x_k u_{kh} y_h \ \forall x, y \in \Delta(S)$.

to be the evolutionary viable in a different model. We consider all continuous and aggregative material games, for which the material payoff function π need not be multi-linear. In those studies, we took the type space to be the whole set of continuous and aggregative utility functions, $\Theta = F$. Utility functions being continuous, the equilibrium correspondence, which maps the mutant share ε to the set of solutions to (7), is then upper hemi-continuous. This means that for small but positive mutant shares $\varepsilon > 0$ each equilibrium is within a neighborhood of *some* resident equilibrium $\hat{x} \in X_f$. Since also the material payoff function is continuous, this in turn implies that to check whether f is evolutionarily stable against g, it is sufficient to evaluate the average material payoffs $\bar{\pi}_R$ and $\bar{\pi}_M$ in the limit as $\varepsilon \to 0$. In other words, it is sufficient to evaluate $\pi\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right)$ and $\pi\left(\hat{y}, \hat{\mathbf{x}}^{(n-1)}\right)$ for each pair $(\hat{x}, \hat{y})$ which satisfies (7) in the limit as $\varepsilon \to 0$, i.e.,

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ f\left(x, \hat{\mathbf{x}}^{(n-1)}\right) \\ \hat{y} \in \arg\max_{y \in X} \ g\left(y, \hat{\mathbf{x}}^{(n-1)}\right). \end{cases} \tag{8}$$

Suppose that the resident type is *Homo oeconomicus*, $f = \pi$. Then, since

$$\hat{x} \in \arg\max_{x \in X} \ \pi\left(x, \hat{\mathbf{x}}^{(n-1)}\right) \tag{9}$$

for all $\hat{x} \in X_f$, there evidently exists no behaviorally distinct mutant type g whose best response $\hat{y}$ would satisfy $\pi\left(\hat{y}, \hat{\mathbf{x}}^{(n-1)}\right) > \pi\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right)$. In other words, *Homo oeconomicus* is evolutionarily stable against all types $g \in D_\pi$. Conversely, if f is behaviorally distinct from *Homo oeconomicus*, i.e., if $f \in D_\pi$, then there does exist some mutant type whose best response $\hat{y}$ is such that $\pi\left(\hat{y}, \hat{\mathbf{x}}^{(n-1)}\right) > \pi\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right)$. In other words, every type $f \in D_\pi$ is evolutionarily unstable. In sum:

> **Observation 3**: Under uniform random matching in a continuum population, and unobservable preference types, *Homo oeconomicus* is evolutionarily stable against all behaviorally distinct types, and such types are evolutionarily unstable.

3.3 A coordination game

To illustrate the above stability concepts, consider a resident population of *Homo oeconomicus* playing the symmetric 2×2 coordination game with material payoff bimatrix (the first number in each entry being the material payoff of the row player):

	a	b
a	$2, 2$	$0, 0$
b	$0, 0$	$1, 1$

(10)

For $f = \pi$, the set of residential equilibrium strategies is $X_f = \{0, 1/3, 1\}$. Consider a mutant type who is "committed" to pure strategy a with, say, utility function $g(x, y) = 4xy + (1-x)(1-y)$, where $x, y \in [0,1]$ is own (respectively, the opponent's) probability of playing pure strategy a.[4] In a sense, this is the most threatening mutant to *Homo oeconomicus,* since g is not behaviorally distinct from f, and, moreover, it always plays according to the best strict equilibrium in terms of material payoffs. The equilibrium condition (7) for the resident *Homo oeconomicus* under incomplete information, when the mutants are in population share ε, is

$$\hat{x} \in \arg\max_{x \in [0,1]} 2\left[(1-\varepsilon)\hat{x} + \varepsilon\right] x + (1-\varepsilon)(1-\hat{x})(1-x). \tag{11}$$

There are three equilibria in population states (f, g, ε) where ε is positive and small. Mutants play a in all of them, while the residents play a in one of them, b in another one, and the mixed strategy $\hat{x}(\varepsilon) = (1-3\varepsilon)/(3-3\varepsilon)$ in the third one. Is *Homo oeconomicus*, $f = \pi$, evolutionarily stable against g? The answer is negative. The reason is that the mutant g earns exactly the same material payoff as f in one of the Nash equilibria in population states (f, g, ε) with ε small, since $\bar{\pi}_f = \bar{\pi}_g = 2$ when $\hat{x} = \hat{y} = 1$. A fortiori, *Homo oeconomicus* is not evolutionary stable in this example. However, *Homo oeconomicus* is not evolutionarily unstable either, since for any mutant type $g \neq \pi$, there exists an $\bar{\varepsilon} > 0$ such that $x = 1$ is an equilibrium strategy for *Homo oeconomicus* in all population states (f, g, ε) with $\varepsilon < \bar{\varepsilon}$. Nonetheless, in force of the general observation above, *Homo oeconomicus* is evolutionarily stable against all mutants $g \in D_\pi$ (and this can be easily verified directly in this example).

We conclude this example by noting that the above-mentioned result by Dekel, Ely and Yilankaya (2007) tells us that in every stable configuration the residents use a strategy in the set $X_f = \{0, 1/3, 1\}$, that there are stable configurations in which they play $x = 0$, and stable configurations in which they play $x = 1$ (since these are strict equilibria).

4 Settings where *Homo oeconomicus* is outperformed

The results summarized above identified the following set of assumptions as being sufficient for *Homo oeconomicus* to be evolutionarily stable: individuals interact under incomplete information, individuals are uniformly randomly matched together to interact, and the population is a continuum. Are these conditions also necessary for *Homo oeconomicus* to be

[4] We note that $g \in G$, with payoff matrix (u_{hk}) with $u_{aa} = 4$ and otherwise $u_{hk} = \pi_{hk}$.

evolutionarily stable? To investigate this question we lift each assumption in turn.

4.1 Complete information

We here study settings where interactions take place under complete information, but keep the assumptions of continuum population and uniform random matching. *Homo oeconomicus* is then evolutionarily unstable. A simple example illustrates why. Suppose that *Homo oeconomicus* is the resident type and that the material game is a two-player prisoner's dilemma,

	c	d
c	$2,2$	$0,4$
d	$4,0$	$1,1$

(12)

Consider a mutant type $g(x,y) = 7xy + 4(1-x)y + (1-x)(1-y)$, where $x, y \in [0,1]$ is own (respectively, the opponent's) probability of cooperating (playing pure strategy c).[5] Under complete information, the unique Nash equilibrium in a game between two residents, and in a game between a mutant and a resident, is for both players to defect (play d). By contrast, in a game between two mutants there are three Nash equilibria: $x = 1$, $x = 1/4$, and $x = 0$. In other words, these preferences allow mutants to cooperate with each other and defect against residents. In any post-entry population in which the mutants play the equilibrium $x = 1$ (or $x = 1/4$) this mutant type strictly outperforms *Homo oeconomicus*, since its average material payoff is then $2-\varepsilon$ (respectively, $(23-7\varepsilon)/16$) while that of *Homo oeconomicus* is 1. Hence, *Homo oeconomicus* is evolutionarily unstable.[6]

Which preferences are stable then, when individuals in a continuum population are uniformly randomly matched to interact under complete information? The literature shows that the answer depends on the set Θ of feasible preferences, and on the material game at hand.

[5] Again, note that $g \in G$, this time with payoff matrix (u_{hk}) with $u_{aa} = 7$ and otherwise $u_{hk} = \pi_{hk}$.

[6] Under strategy evolution in this material game, mutants whose strategy consists, by way of a "secret handshake", in playing c against each other, and d against residents, strictly outperform residents who always play d (Robson, 1990).

4.1.1 Finite two-player games

Dekel, Ely, and Yilankaya (2007) (see Section 3.2 for a description of their setup) call a strategy $x \in X$ *efficient* if $\pi(x,x) \geq \pi(y,y)$ for all $y \in X$. They show that under complete information efficient strategies are the only candidates for configuration stability, and that if an efficient strategy is also its own unique best reply (in terms of material payoffs), then it is configuration stable. In particular, while the efficient strategy profile (a,a) in coordination game (10) is configuration stable, the strategy profile (b,b) is not, and while the efficient profile (c,c) in the prisoners' dilemma (12) is configuration stable, (d,d) is not. The intuition is clear. As shown in the above prisoners' dilemma, the instability of an inefficient strategy is due to the existence of mutants who may behave like residents when they meet residents and do better when meeting each other. However, if a strategy is efficient, and, moreover, is its own unique best reply, then the mutants cannot achieve higher payoff when meeting neither residents nor each other.[7] In sum:

> **Observation 4**: Under uniform random matching in a continuum population, and complete information in finite and symmetric two-player material-payoff games, (material) efficiency is necessary for configuration stability, and efficiency and strict equilibrium (both in material payoffs) are together sufficient for configuration stability.

So far, we have seen how mutants can successfully invade a resident population by behaving differently towards each other than towards residents, while, at the same time, the residents' behavior was unchanged. However, mutants may sometimes successfully enter a population by instead making residents alter their behavior. To see this, suppose that the material game is a hawk-dove game, with material payoffs

	h	d
h	$-1,-1$	$4,0$
d	$0,4$	$2,2$

(13)

Let $x \in [0,1]$ denote the probability that an individual plays h, and $y \in [0,1]$ the probability that his opponent does so. Suppose that *Homo oeconomicus* is the resident type. In a homogenous population, the unique Nash equilibrium in any matched pair is $x = 2/3$. This

[7] See Ockenfels (1993) for an early analysis of preference evolution in the prisoner's dilemma game.

mixed strategy is efficient. Consider now a mutant type, appearing in population share ε, with a utility function $g^h \in G$ for which pure strategy h is strictly dominant.[8] Whether matched with another mutant or with a resident, such a mutant always plays pure strategy h. Under complete information, residents play $x^* = 2/3$ when matched with another resident, but switch to playing d with certainty when matched with a mutant. For small $\varepsilon > 0$, mutants then garner a higher material payoff, $4(1-\varepsilon) - \varepsilon$, than residents, who obtain $2(1-\varepsilon)/3$. Hence, although necessary, efficiency is not sufficient for stability, an observation in line with Proposition 3 in Dekel, Ely, and Yilankaya (2007).

The residential adaptation we saw in this example suggests that maybe there are stable type distributions (in the sense of Dekel, Ely, and Yilankaya, 2007) with more than one type present? Can the two types mentioned above coexist in such population shares that their average material payoffs are equalized, and will this be stable? Indeed, this possibility was pointed out in Banerjee and Weibull (1995). By way of studying populations in which every individual is either a *Homo oeconomicus* or else committed to one particular pure strategy, they showed that a pure *Homo oeconomicus* population is evolutionarily unstable if the material-payoff game has at least one "bully" strategy, a pure strategy that earns a higher payoff against each of its best replies than they earn against it. An example is pure strategy h in the above hawk-dove game. The best reply to h is d, and in such a meeting h earns 4 while d earns zero. Banerjee and Weibull (1995) showed that a certain binary type distribution, with both *Homo oeconomicus* and "hawks" present, is evolutionarily stable when the type space Θ consists of "hawks", represented by a utility function $g^h \in G$ as above, "doves" $g^d \in G$ likewise defined, and *Homo oeconomicus*, π. In the present example, let the type distribution assign probability μ to type h and probability $1-\mu$ to *Homo oeconomicus*. The latter will then play d when matched with a "hawk" and $x^* = 2/3$ when matched with another *Homo oeconomicus*. Hence, the average material payoff to a "hawk" is $4(1-\mu)-\mu$, and the average material payoff to a *Homo oeconomicus* is $2(1-\mu)/3$. The two types earn exactly as much if and only if $\mu = 10/13$. The resulting population behavior, however, is the same as in the unique Nash equilibrium of the material payoff game. These observations hold qualitatively in all Hawk-Dove games. Combining the results of Banerjee and Weibull (1995) with those of Dekel, Ely and Yilankaya (2007), we obtain:

> **Observation 5**: Under complete information in finite and symmetric material-payoff games with a bully strategy, configuration stability may require that *Homo*

[8] For example, let the payoff matrix (u_{hk}) have $u_{hh} = 1$ and otherwise $u_{hk} = \pi_{hk}$.

oeconomicus coexists with a bully type in fixed positive population shares.

In other words: natural selection under complete information may select for a population mix of rational and aggressive individuals. This is due to the, since ancient times, well-known power of commitment (see Schelling, 1960). The commitment power obtained by having preferences that are known by others has also been central to the bulk of the early contributions to the literature on indirect evolution, to which we now turn.

4.1.2 Two-player games with continuum of pure strategies

In Bester and Güth (1998) the pure-strategy set X is taken to be $\mathbb{R}_+$, the non-negative reals, and the material payoff function to be of the non-linear form

$$\pi(x, y) = (b - x)\, x + cxy, \tag{14}$$

for some $b > 0$ and $c \in (-1, 1)$. For c positive (negative), the game exhibits positive (negative) externalities; a higher action by a player increases (decreases) the material payoff to the other player. The authors restrict preferences to a one-dimensional subset of F, by assuming that each individual has a utility function f_λ of the form

$$f_\lambda(x, y) = \lambda \cdot \pi(x, y) + (1 - \lambda) \cdot \pi(y, x), \tag{15}$$

for some $\lambda \geq 1/2$. The type set Θ is thus the subset of utility functions $f_\lambda \in F$ for some $\lambda \geq 1/2$. For $\lambda < 1$, such an individual attaches a positive weight to the other player's material payoff, and is thus *altruistic* (Becker, 1976). By contrast, for $\lambda > 1$, the individual attaches a negative weight to the other player's material payoff and is thus spiteful**.** *Homo oeconomicus* is the intermediate knife-edge case when $\lambda = 1$. We will call $\lambda \geq 1/2$ the type's degree of altruism/spite.

In a match between individuals of arbitrary degrees of altruism/spite $\lambda, \tau \in \Theta$ there is a unique Nash equilibrium $(\hat{x}(\lambda, \tau), \hat{x}(\tau, \lambda)) \in X^2$, and thus a unique pair of equilibrium material payoffs. Bester and Güth (1998) ask which type $\lambda \in \Theta$, if any is evolutionarily stable. This amounts to searching for some value $\lambda \geq 1/2$ such that if u_λ is the resident utility function and u_τ the mutant utility function, for some $\tau \neq \lambda$, the uniquely defined average equilibrium material payoff to a resident, $\bar{\pi}_R$, exceeds that to a mutant, $\bar{\pi}_M$, for all sufficiently small mutant shares $\varepsilon > 0$:

$$\begin{aligned} & (1 - \varepsilon) \cdot \pi[\hat{x}(\lambda, \lambda), \hat{x}(\lambda, \lambda)] + \varepsilon \cdot \pi[\hat{x}(\lambda, \tau), \hat{x}(\tau, \lambda)] \\ > \; & (1 - \varepsilon) \cdot \pi[\hat{x}(\tau, \lambda), \hat{x}(\lambda, \tau)] + \varepsilon \cdot [\hat{x}(\tau, \tau), \hat{x}(\tau, \tau)]. \end{aligned} \tag{16}$$

Both sides being continuous in ε, a sufficient condition for type λ to be evolutionarily stable against type τ is $\pi\left[\hat{x}\left(\lambda,\lambda\right),\hat{x}\left(\lambda,\lambda\right)\right] > \pi\left[\hat{x}\left(\tau,\lambda\right),\hat{x}\left(\lambda,\tau\right)\right]$, while a necessary condition is $\pi\left[\hat{x}\left(\lambda,\lambda\right),\hat{x}\left(\lambda,\lambda\right)\right] \geq \pi\left[\hat{x}\left(\tau,\lambda\right),\hat{x}\left(\lambda,\tau\right)\right]$. Noting that, for a given λ, the right-hand side of these inequalities may be viewed as a function of τ, we see that a necessary condition for type λ to be evolutionarily stable is that λ be a solution to the fixed-point equation

$$\lambda \in \arg\max_{\tau \geq 1/2} \pi\left[\hat{x}\left(\tau,\lambda\right),\hat{x}\left(\lambda,\tau\right)\right]. \tag{17}$$

In the present model, where for a given λ, π is differentiable in τ, this allows the analyst to conveniently identify candidates for evolutionarily stable utility functions by solving the necessary first-order condition for (17) to hold (with subscripts for partial derivatives):

$$\pi_1\left(\hat{x}\left(\tau,\lambda\right),\hat{x}\left(\lambda,\tau\right)\right)\cdot\hat{x}_1\left(\tau,\lambda\right)\big|_{\tau=\lambda} = -\,\pi_2\left(\hat{x}\left(\tau,\lambda\right),\hat{x}\left(\lambda,\tau\right)\right)\cdot\hat{x}_2\left(\lambda,\tau\right)\big|_{\tau=\lambda}. \tag{18}$$

The two sides of this equation measure, for any given resident parameter value λ, the two effects of mutating towards a marginally different parameter value, τ, when the mutation is vanishingly rare. The left-hand side measures the effect on the mutant's own equilibrium strategy and the associated effect on the material payoff. The right-hand side measures the strategic-commitment effect: the mutation causes the opponent (which almost for sure is a resident) to change his equilibrium strategy and this in turn also affects the mutant's material payoff.

With the material payoff function used by Bester and Güth (1998) and given in (14), one obtains $\lambda^* = 1 - c/2$ as the unique solution to (18). By way of considering the second-order condition it is easily verified that this λ^*-value is indeed evolutionarily stable.

Three qualitatively distinct cases arise. First, when $c = 0$, individuals, even if matched into pairs, effectively face a decision problem. Not surprisingly, *Homo oeconomicus* then prevails: $\lambda^* = 1$. Second, if $c > 0$, the strategies are strategic complements in the material payoff function π. Then, $\lambda^* < 1$, which means altruism. The reason is that although such altruism makes an individual contribute more than *Homo oeconomicus* would, which is costly in material terms, this makes also the opponent contribute more, and the net material effect is beneficial. It is important to notice that the reason for why *Homo oeconomicus* is displaced by altruists is *not* that altruists obtain a benefit from interacting with each other. Instead, under complete information altruists may enter a population of *Homo oeconomicus,* by making the latter behave differently than they do when meeting another *Homo oeconomicus*. Under uniform random matching (as we here assume), an altruistic mutant in a population of

resident *Homo oeconomicus* has virtually no chance of being matched with another altruistic mutant.

Third, if $c < 0$, spite obtains; $\lambda^* > 1$**.** This is because when strategies are strategic substitutes, an individual with spiteful preferences contributes less than *Homo oeconomicus* would, and this in turn makes a *Homo oeconomicus* contribute more when matched with a spiteful individual than with another *Homo oeconomicus*.

The astute reader will have noticed that Bester and Güth (1998) in fact reported the value $\lambda^* = 1$ for the evolutionarily stable value of λ in the case $c < 0$. This is because they restricted the set of potential values of λ to the interval $[1/2, 1]$. Bolle (2000) and Possajennikov (2000) contributed by pointing this out, and Possajennikov (2000) derived the evolutionarily stable preferences for the case $c < 0$. Using preferences of the slightly different parametric form

$$f_\alpha(x, y) = \pi(x, y) + \alpha \cdot \pi(y, x), \tag{19}$$

for $\alpha \in \mathbb{R}$, he reports a stable value $\alpha^* = c/2$ which is equivalent to the result reported above (since $\lambda^*/(1-\lambda^*) = c/2$). He further lifts the restriction $c \in (-1, 1)$ imposed by Bester and and Güth (1998), and shows that there exists an evolutionarily stable value of α, namely, $\alpha^* = c/2$, if and only if $c \in [-2, 1] \cup (2, +\infty)$**.**

Remark 1 *If for each possible matched pair there exists a unique Nash equilibrium, the average equilibrium material payoff to a resident and to a mutant is linear in the share of mutants ε (recall (16)). Hence, it is possible to recast the model as a standard two-player evolutionary game, in which the preference traits are strategies, the set Λ of potential preference traits λ is the set of strategies, and the payoff function is $\hat{\pi} : \Lambda^2 \to \mathbb{R}$, where for each pair $(\lambda, \tau) \in \Lambda^2$, $\hat{\pi}(\lambda, \tau)$ is defined as $\pi[\hat{x}(\lambda, \tau), \hat{x}(\tau, \lambda)]$. In this evolutionary game, a "strategy" $\lambda \in \Lambda$ is evolutionarily stable against "strategy" $\tau \in \Lambda$ if either $\hat{\pi}(\lambda, \lambda) > \hat{\pi}(\tau, \lambda)$, or $\hat{\pi}(\lambda, \lambda) = \hat{\pi}(\tau, \lambda)$ and $\hat{\pi}(\lambda, \tau) > \hat{\pi}(\tau, \tau)$, just as in Maynard Smith and Price (1973). Indeed, this is the approach adopted by Bester and Güth (1998), with $\Lambda = [1/2, 1]$*

A more general model, which encompasses the ones discussed above as special cases, was analyzed by Heifetz, Shannon, and Spiegel (2007a,b). They consider a general class of two-player strategic interactions that need not be symmetric. In order to facilitate comparison with other models in this survey, we focus on the symmetric special case of their model, and when the strategy set X is a subset of $\mathbb{R}$. The material payoff function $\pi : X^2 \to \mathbb{R}$ is taken

to be thrice continuously differentiable. They consider utility functions of the form

$$f_\delta(x,y) = \pi(x,y) + v(x,y,\delta), \tag{20}$$

where the function v (also taken to be trice differentiable) represents what the authors call the individual's *disposition*. The parameter δ is taken to be a real number that belongs to some interval $D \subset \mathbb{R}$ containing a neighborhood of 0, and $v(x,y,0) = 0$ for all strategy pairs (x,y). Hence, δ represents the *intensity* of the disposition, with $\delta = 0$ corresponding to *Homo oeconomicus*. Bester and Güth's (1998) model (for π is thrice continuously differentiable) is the special case when $v(x,y,\delta) = \delta \cdot \pi(y,x)$ for $\delta = (1-\lambda)/\lambda$ (and Possajennikov (2000) corresponds to $v(x,y,\delta) = \delta \cdot \pi(y,x)$ for $\delta = \alpha$). Heifetz, Shannon, and Spiegel (2007a,b) proceed by analyzing evolutionary drift in the intensity parameter δ for any given disposition function v. Such an evolutionary analysis thus takes the type space Θ to be the unidimensional manifold $F_v \subset F$ consisting of all functions f_δ of the form (20) for some $\delta \in D$, where *Homo oeconomicus* is represented by f_0.

Using the same notation as above, for any disposition v, and any type pair $f_\delta, f_\tau \in F_v$, let $(\hat{x}(\delta,\tau), \hat{x}(\tau,\delta))$ denote a Nash equilibrium strategy in a pair in which the types are δ and τ. Equilibrium uniqueness and differentiability is imposed as an assumption in Heifetz, Shannon, and Spiegel (2007a).[9]

Given their assumptions, the (total) derivative $d\pi[\hat{x}(\tau,\delta), \hat{x}(\delta,\tau)]/d\tau$, evaluated at $\tau = \delta$ measures the net effect on a individual's (equilibrium) material payoff from a marginal change of his type, away from the resident type δ. Recall from the description of Bester and Güth (1998) that, in a model where the population is a continuum, this derivative also measures the net payoff effect of mutating, at the margin, away from the resident type, evaluated in the limit as the population share of mutants tends to zero. If this marginal effect is strictly positive (negative), it would thus pay off to mutate towards a higher (lower) value in the type space. Hence, a necessary condition for a type δ to be evolutionarily stable is that $d\pi[\hat{x}(\tau,\delta), \hat{x}(\delta,\tau)]/d\tau = 0$ when evaluated at $\tau = \delta$. In a population consisting of *Homo oeconomicus*, i.e., with residents of type $\delta = 0$, this total derivative equals

$$\pi_1[\hat{x}(0,0), \hat{x}(0,0)] \cdot \hat{x}_1(0,0) + \pi_2[\hat{x}(0,0), \hat{x}(0,0)] \cdot \hat{x}_2(0,0). \tag{21}$$

The first term is the effect that the marginal mutation has on the mutant's equilibrium material payoff due to the ensuing change in own strategy. The second term is the effect

[9] In their companion paper (2007b), they instead assume a selection from the Nash equilibrium correspondence that is continuously differentiable at the origin in their type space.

that the marginal mutation has on the mutant's equilibrium material payoff due to the ensuing change in the opponent's strategy. The first term must be nil, because in this setting the first-order condition $\pi_1(\hat{x},\hat{x}) = 0$ must hold for any $\hat{x} \in X_\pi$. Hence, whenever there is a strategic commitment effect of preferences, in the sense that $\hat{x}_2(0,0) \neq 0$, and the individual's material payoff is affected by this (i.e., $\pi_2(\hat{x},\hat{x}) > 0$), then *Homo oeconomicus* is evolutionarily unstable.[10] In sum:

> **Observation 6**: Under complete information, for almost all two-player material-payoff games and for almost any disposition (such as altruism **or** spite), *Homo oeconomicus* is unstable, and instead some non-zero intensity of the disposition will be stable.

Analyzing the stability of preferences within the class of altruistic preferences of the form (19), in Alger and Weibull (2012) we further derived a result which establishes a link between the strategic nature of the material payoff game and the nature of stable preferences. We considered any continuously differentiable material payoff function $\pi : X^2 \to \mathbb{R}$ such that $\pi_2 \neq 0$ and such that for any degrees of altruism $\alpha, \beta \in (-1,1)$ there exists a unique interior and differentiable pair of Nash equilibrium strategies, $(\hat{x}(\alpha,\beta), \hat{x}(\beta,\alpha))$. Using the first-order condition for $(\hat{x}(\alpha,\beta), \hat{x}(\beta,\alpha))$ to be an interior Nash equilibrium strategy, namely,

$$\pi_1[\hat{x}(\beta,\alpha), \hat{x}(\alpha,\beta)] = \alpha \cdot \pi_2[\hat{x}(\alpha,\beta), \hat{x}(\beta,\alpha)], \tag{22}$$

the necessary condition for a degree of altruism $\alpha \in (-1,1)$ to be evolutionarily stable (see (21)) can be written

$$\alpha \cdot \hat{x}_1(\alpha,\alpha) = -\hat{x}_2(\alpha,\alpha). \tag{23}$$

Call the strategies *strategically neutral* if $\pi_{12}(x,y) = 0$ for all strategies x and y, *strategic substitutes* if $\pi_{12}(x,y) < 0$ for all strategies x and y, and *strategic complements* if $\pi_{12}(x,y) > 0$ for all strategies x and y. Then (Alger and Weibull, 2012):

> **Observation 7**: Under uniform random matching in a continuum population with observable preference types, *Homo oeconomicus* is evolutionarily stable if strategies are strategically neutral (in terms of material payoffs), and unstable if strategies are either strategic complements or strategic substitutes. In the case of

[10] For evolutionary stability, they use asymptotic stability in payoff-positive selection dynamics.

strategic complements, altruism emerges, while spite emerges in the case strategic substitutes.

The key point here is that under complete information the specifics of the material game matter for stable preferences. In our evolutionary past, especially in pre-industrial times, the material game may in turn have depended on the environment. To see this, suppose that the material game represents food production in a community. Compare two populations, one in which hunted big game constitutes the main food source, and one in which gathered insects, roots and berries are the main food source. Now, hunting big game typically requires teamwork in which efforts are strategic complements, while efforts in food gathering are strategically neutral or strategic substitutes (they are strategic substitutes, for instance, if food is scarce and the distance an individual has to cover to find food to gather depends on the efforts spent by others in the community on food gathering). The results reported above suggest that evolution by natural selection can sustain a higher degree of altruism in the first population than in the second.

4.1.3 A parallel with contract theory

While models in the literature on preference evolution are typically not cast as descriptions of market interactions, it is clear that strategic commitment may be valuable in markets. For instance, in duopolistic markets under Cournot competition, it can pay off to be aggressive by somehow committing to producing a quantity exceeding the Cournot-Nash equilibrium quantity. This is the point made by Fershtman and Judd (1987). We describe their model here in order to highlight the differences and similarities with preference evolution models.

Consider a duopolistic market in which a firm who produces the quantity x when the competitor produces the quantity y garners profit

$$\pi(x, y) = (b - x - y)\, x - cx, \tag{24}$$

for some $b > c \geq 0$. If the goal of each firm is to maximize its profit, the unique Nash equilibrium quantity is $\hat{x} = (b - c)/3$, and each firm's profit is $\pi(\hat{x}, \hat{x}) = (b - c)^2/9$. Suppose now instead that one firm still seeks to maximize its profit (say, its manager has a contract with a profit bonus), while the other seeks to maximize a weighted sum of profit and sales (say, its manager has a bonus scheme based on both profit and sales)

$$f_\gamma(x, y) = \gamma \cdot [(b - x - y)\, x - cx] + (1 - \gamma) \cdot (b - x - y)\, x. \tag{25}$$

If the weight γ is smaller than one, it is as if the firm's cost was below c. As a result, the firm will then produce more than if it simply sought to maximize profit. In response to this, the competitor, who is assumed to know the goal function f_γ, will produce less. It is easy to verify that the firm whose (manager's) goal function is f_γ obtains a higher equilibrium profit than the firm whose (manager's) goal function is its profit, π.

Based on this observation, Fershtman and Judd (1987) study a two-stage interaction between the two firms to determine which goal function should be expected to emerge as the result of competition between them. In the first stage, the owners of the firms simultaneously choose a weight each in $\mathbb{R}$, a weight to be attached to the profit in the convex combination described in (25). Let γ and μ denote the chosen values. In the second stage, the firm managers simultaneously choose a quantity each to be produced, based on the goal functions chosen by their respective owners in the first stage. Anticipating the equilibrium quantities $\hat{x}(\gamma, \mu)$ and $\hat{x}(\mu, \gamma)$, where

$$\hat{x}(\gamma, \mu) = \frac{b-c}{3} + \frac{c(1+\mu-2\gamma)}{3}, \tag{26}$$

the owner who picks γ anticipates the profit $\Pi(\gamma, \mu) \equiv \pi(\hat{x}(\gamma, \mu), \hat{x}(\mu, \gamma))$, while the owner who picks μ anticipates the profit $\Pi(\mu, \gamma) \equiv \pi(\hat{x}(\mu, \gamma), \hat{x}(\gamma, \mu))$. Hence, it is as if the two owners played a simultaneous-move symmetric game in which the strategies are γ and μ, the common strategy set is $\mathbb{R}$, and the payoff function is $\Pi : \mathbb{R}^2 \to \mathbb{R}$. The unique symmetric equilibrium of this game is

$$\gamma^* = 1 - \frac{b-c}{5c}. \tag{27}$$

In sum, in a duopolistic market in which each firm manager chooses the quantity to be produced, a firm owner should not be expected to request his manager to maximize profit. Instead, if each owner chooses to provide incentives to his managers so that she maximizes a function of the form given in (25), then the firm owners best respond to each other by setting $\gamma = \gamma^*$. In other words, a positive weight is given to sales, and a weight below 1 is given to profit (note that the weight given to profit can even be negative!). This apparently counter-intuitive result apparently hinges on the assumption of complete information about contracts, enabling contracts to act as commitments.

While Fershtman and Judd (1987) do not explicitly model the process by which firms initially appear and perhaps eventually go bankrupt, their result shows that it is necessary to interpret with care Milton Friedman's (1953) claim that "unless the behavior of businessmen in some way or other approximated behavior consistent with the maximization of returns,

it seems unlikely that they would remain in business for long": managers of successful firms may not necessarily seek to maximize profits.

4.2 Assortative random matching

Until now we have only examined settings with uniform random matching, i.e., in which the type distribution that an individual faces in the matching process is independent of his own type. It is hard to think of real societies where uniform random matching would occur. Indeed, natural populations are typically structured into (geographic, cultural, linguistic or socioeconomic) groups, and interactions tend to occur preferentially, though not exclusively, within these groups, for reasons including transportation costs and homophily. This in turn implies that carriers of a rare mutant trait tend to relatively often interact with each other, perhaps even unbeknownst to them. For example, suppose that preferences are genetically transmitted, and suppose that in an initially homogenous population suddenly a new preference type appears in one individual. In the second generation some interactions between carriers of the new trait may occur between siblings and in the third among cousins.

A strand of the evolutionary literature investigates the consequences of such assortative matching on the stability of preference types. The evolution of genetically transmitted traits in structured populations, giving rise to assortativity, was initially formalized in the *island model* (Wright, 1931). And although some work on preference evolution uses the island model (Rogers, 1994, Akçay and van Cleve, 2012, Alger, Weibull, and Lehmann, 2018), in economics assortativity has mostly been modeled as an abstract function that maps the distribution of traits in the population to probabilities governing the matching of interacting individuals. This formalization of assortativity, pioneered by Bergstrom (1995, 2003), is outlined below for n-player interactions (following Alger and Weibull, 2016).[11]

In any population state $s = (f, g, \varepsilon) \in \Theta^2 \times (0, 1)$, the number of mutants—individuals of type g—in a group that is about to play the material-payoff game, is a random variable, T. For any resident drawn at random from the population let $p_m(\varepsilon)$ be the probability that the number of mutants in the resident's group is m, for $m = 0, 1, .., n-1$.[12] Likewise, for any mutant, also drawn at random from the population, let $q_m(\varepsilon)$ be the conditional probability

[11] For a more general formalization of assortative matching rules, with results for strategy evolution, see Jensen and Rigos (2018).

[12] In the special case of uniform random matching, $p_m(\varepsilon)$ is as defined in (3).

that the number of *other* mutants in his or her group is $m = 0, .., n-1$. Let $\boldsymbol{p}(\varepsilon) = (p_0(\varepsilon), ..., p_{n-1}(\varepsilon))$ and $\boldsymbol{q}(\varepsilon) = (q_0(\varepsilon), ..., q_{n-1}(\varepsilon))$ be the so defined probability distributions. We will say that the random matching is uniform if $\boldsymbol{p}(\varepsilon) = \boldsymbol{q}(\varepsilon)$, and assortative if $\boldsymbol{p}(\varepsilon) \neq \boldsymbol{q}(\varepsilon)$.

The literature has examined models in which both $\boldsymbol{p}(\varepsilon)$ and $\boldsymbol{q}(\varepsilon)$ are continuous in the mutant population share, $\varepsilon \in (0, 1)$, and converge to some limit points $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$, respectively, as $\varepsilon \to 0$. These limit points turn out to play a key role in the analysis of stable preferences in a continuum population. In a continuum population, residents virtually never meet mutants when the latter are vanishingly rare, so $\boldsymbol{p}^* = (1, 0, 0, ..., 0)$. Turning now to the limit vector $\boldsymbol{q}^*$, which we call the *assortativity profile* of the matching process, note first that in the special case of uniform random matching, the matching probabilities for mutants are the same as for residents, so then $\boldsymbol{q}^* = \boldsymbol{p}^*$. Under assortative matching, it is useful to study the difference between the probability $\Pr[f|f, \varepsilon]$ for an individual of the resident type f that another, uniformly randomly sampled member of his group also has the resident type, and the probability $\Pr[f|g, \varepsilon]$ of this event for an individual of the mutant type g:

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

This defines the *assortment function* $\phi : (0, 1) \to [-1, 1]$ (the same for all type pairs). Suppose that this function is continuous and that it has a limit value, σ, as the mutant share tends to zero, to be called the *index of assortativity* of the matching process (Bergstrom, 2003). Since $\lim_{\varepsilon \to 0} \Pr[f|f, \varepsilon] = 1$, it is immediate that $\boldsymbol{q}^* = (1 - \sigma, \sigma)$ in the special case of pairwise interactions ($n = 2$). For interactions in larger groups there remains a statistical issue, namely whether or not the types of other members of a mutant's group are statistically dependent of each other or not. An important special case is when a mutant's other group members' types are conditionally independent of each other. This case arises, for example, in groups of siblings when each child's type is an independent random draw from the parents' types, or in groups of students from the same school, when each student's type is independently drawn from the teachers' types. Then $\boldsymbol{q}^*$ is binomial:

$$q_m^* = \binom{n-1}{m} \sigma^m (1 - \sigma)^{n-1-m} \quad \text{for } m = 0, 1, ..., n-1. \tag{28}$$

We now ask whether assortativity affects the stability of preferences, and if so, how.

4.2.1 Incomplete information

In our studies of preference evolution under incomplete information and assortative matching (Alger and Weibull, 2013, 2016), we let the type space be the full set F of continuous and aggregative utility functions. We found that a specific kind of utility function, hitherto unstudied in economics, stands out. We called this class of utility functions *Homo moralis*. An individual is a *Homo moralis* with *morality profile* $\boldsymbol{\kappa} \in \Delta$ if his or her goal function $f \in F$ is of the form

$$f_{\boldsymbol{\kappa}}(x, \boldsymbol{y}) = \mathbb{E}[\pi(x, \mathbf{Y})] \quad \forall (x, \boldsymbol{y}) \in X^n, \tag{29}$$

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

Homo moralis preferences have a Kantian flavor: for any given strategy profile, an individual with such preferences evaluates her or his strategy choice in the light of what the material payoff would be, should some or all other individuals also choose that strategy. At one extreme of the spectrum of *Homo moralis* we find *Homo kantientis*, the variety that has morality profile $\boldsymbol{\kappa} = (0, ..., 0, 1)$, in which case $\mathbb{E}[\pi(x, \mathbf{Y})] = \pi(x, x, ..., x)$. 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. At the opposite extreme we find *Homo oeconomicus*, a *Homo moralis* with morality profile $\boldsymbol{\kappa} = (1, 0, ..., 0)$, in which case $\mathbb{E}[\pi(x, \mathbf{Y})] = \pi(x, \boldsymbol{y})$. The behavior of all other varieties of *Homo moralis* lies between these two extremes. *Homo moralis* with morality profile $\boldsymbol{\kappa} \in \Delta$ behaves as if she followed a probabilistic version of Kant's categorical imperative (Kant, 1785); 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 $\boldsymbol{\kappa}$.[13]

While the function in (29) may be mathematically fairly involved, it is particularly simple if one assumes $\boldsymbol{\kappa} = Bin(n-1, \kappa)$, for some $\kappa \in [0, 1]$. The utility is then the individual's expected material payoff if, hypothetically, each other player would statistically independently switch to his strategy with probability κ. In this one-dimensional case, κ can be referred to as the individual's *degree of morality*. This is particularly clear for pairwise interactions

[13]It should be noted that, for finite games, the utility of an individual with *Homo moralis* preferences is typically non-linear in own mixed strategy. Such preferences do not belong to the set of preferences G considered by Dekel, Ely, and Yilankaya (2007).

(Alger and Weibull, 2013):

$$f_\kappa(x,y) = (1-\kappa)\cdot\pi(x,y) + \kappa\cdot\pi(x,x). \tag{30}$$

Alger and Weibull (2013, 2016) establish:

> **Observation 8**: Under assortative random matching in a continuum population, and unobservable preference types, *Homo moralis* with the assortativity profile of the matching process as its morality profile is evolutionarily stable against all behaviorally distinct types, and the latter are evolutionarily unstable.

In other words, evolutionary stability favors *Homo moralis* preferences of a morality profile that precisely reflects the assortativity of the matching process, i.e., $\boldsymbol{\kappa} = \boldsymbol{q}^*$. In particular, under conditional independence (see (28)), *Homo moralis* with degree of morality $\kappa = \sigma$ is evolutionarily stable.

The intuition for the result is that in a population that consists almost solely of *Homo moralis* with the "right" morality profile, these individuals preempt entry by rare mutants by using some strategy that would maximize the average *material* payoff to a vanishingly rare mutant who would enter this population.[14] To see this, note that by the same (topological) arguments as those used for the special case of uniform random matching (see Observation 3 and the discussion preceding it), it is sufficient to evaluate the average equilibrium material payoffs $\bar{\pi}_R$ and $\bar{\pi}_M$ for each $(\hat{x},\hat{y})$ satisfying

$$\begin{cases} \hat{x} \in \arg\max_{x\in X} \ f\left(x, \hat{\mathbf{x}}^{(n-1)}\right) \\ \hat{y} \in \arg\max_{y\in X} \ \sum_{m=0}^{n-1} q_m^* \cdot g\left(y, \hat{\mathbf{x}}^{(n-1-m)}, \hat{\mathbf{y}}^{(m)}\right), \end{cases} \tag{31}$$

i.e., the system of equations defining (Bayesian) Nash equilibrium in the limit case when $\varepsilon \to 0$ (note that (7) holds for any $\boldsymbol{p}(\varepsilon), \boldsymbol{q}(\varepsilon)$). Importantly, the result does not rely on individuals seeking to behave in a way which materially benefits the group to which they belong; instead, it is driven by individual utility maximizing behavior, and the material payoff benefit that such utility maximization bestows on the individual.

We note that in the special case of two-player constant-sum games, the preferences of *Homo moralis* with any degree of morality $\kappa < 1$ are identical with those of *Homo oeconomicus*; the first utility function is but a positive affine transformation of the second

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

($u_\kappa(x,y) = a\pi(x,y) + b$ for $a > 0$ and $b \in \mathbb{R}$). Such *Homo moralis* are thus behaviorally identical to *Homo oeconomicus*. Not surprisingly, *Homo oeconomicus* prevails in all decision problems, just as under uniform random matching (see Section 3). Indeed, in decision problems material payoffs are unaffected of what others do, and hence, *Homo moralis* preferences are represented by $\mathbb{E}[\pi(x, \mathbf{Y})] = \mathbb{E}[v(x)] = v(x)$, for all strategies x and random strategy profiles Y, irrespective of the morality profile. In sum:

> **Observation 9**: Under assortative random matching in a continuum population, with unobservable preference types, *Homo oeconomicus* is behaviorally identical with every *Homo moralis* except for *Homo kantientis*, in all decision problems, and in all two-player constant-sum games.

Prior to turning to preference evolution under complete information, we briefly discuss a seminal contribution for strategy evolution.

4.2.2 Strategy evolution

In his paper, Bergstrom (1995) focuses on pairwise interactions, and he studies the stability of genetically determined strategies in interactions between siblings. He shows how the necessary condition for a strategy to be evolutionarily stable depends on whether reproduction is asexual or sexual, and in the latter case, on whether the preference trait is autosomal dominant or recessive. Under asexual reproduction, individuals are clones of their single parent, and hence, siblings always have the same preferences. Hence, $\sigma = 1$, any evolutionarily stable strategy must also be a Nash equilibrium in a game where the payoff to each player is f_κ for $\kappa = \sigma = 1$. Under sexual reproduction and an autosomal dominant preference trait, in the limit as the mutant share ε tends to zero, the probability that a sibling of an individual with a mutant strategy is $1/2$, while that probability equals zero for an individual with the resident strategy. Hence, $\sigma = 1/2$, and any evolutionarily stable strategy is what Bergstrom (1995) calls semi-Kantian, i.e., it must be a Nash equilibrium in a game where the payoff to each player is f_κ for $\kappa = \sigma = 1/2$. By contrast, under sexual reproduction and an autosomal recessive preference trait, an evolutionarily stable strategy must be a Nash equilibrium in a game where the payoff to each player involves a mix of selfish, Kantian, and altruistic motives, as follows:

$$w(x,y) = \frac{3}{5} \cdot \pi(x,y) + \frac{1}{5} \cdot \pi(x,x) + \frac{1}{5} \cdot \pi(y,x). \tag{32}$$

A key insight generated by these results is that the transmission process itself matters for which strategies may be evolutionarily stable.[15] This is related to the huge literature in evolutionary biology on the evolution of traits when genetically related individuals interact (Hamilton, 1964, Grafen, 1979, Hines and Maynard Smith, 1979, Rousset, 2004).

4.2.3 Complete information

Returning to preference evolution under assortative matching, models with pairwise interactions under complete information were examined by Alger and Weibull (2010, 2012). Both models focused on the class of altruistic preferences, see (19), and on interactions with a unique and differentiable Nash equilibrium for each pair of preference types $(\alpha, \beta) \in (-1, 1)^2$. Hence, the degree of altruism α is evolutionarily stable if for all $\beta \neq \alpha$, there exists $\bar{\varepsilon}_\beta$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_\beta)$, the average equilibrium material payoff to a resident exceeds that to a mutant:

$$\begin{aligned} & \Pr\left[\alpha|\alpha,\varepsilon\right] \cdot \pi\left[\hat{x}\left(\alpha,\alpha\right),\hat{x}\left(\alpha,\alpha\right)\right] + \Pr\left[\beta|\alpha,\varepsilon\right] \cdot \pi\left[\hat{x}\left(\alpha,\beta\right),\hat{x}\left(\beta,\alpha\right)\right] \\ > \; & \Pr\left[\alpha|\beta,\varepsilon\right] \cdot \pi\left[\hat{x}\left(\beta,\alpha\right),\hat{x}\left(\alpha,\beta\right)\right] + \Pr\left[\beta|\beta,\varepsilon\right] \cdot \pi\left[\hat{x}\left(\beta,\beta\right),\hat{x}\left(\beta,\beta\right)\right]. \end{aligned} \tag{33}$$

Given that the conditional probability functions are continuous (by assumption), it is sufficient to examine this inequality in the limit as ε tends to zero, i.e.:

$$\pi\left[\hat{x}\left(\alpha,\beta\right),\hat{x}\left(\beta,\alpha\right)\right] > \left(1-\sigma\right) \cdot \pi\left[\hat{x}\left(\beta,\alpha\right),\hat{x}\left(\alpha,\beta\right)\right] + \sigma \cdot \pi\left[\hat{x}\left(\beta,\beta\right),\hat{x}\left(\beta,\beta\right)\right]. \tag{34}$$

Noting that the two sides of this inequality are equal when $\beta = \alpha$, the following first-order condition is necessary for $\alpha \in (-1, 1)$ to be evolutionarily stable:

$$\left(1-\sigma\right) \cdot \frac{\partial \pi\left[\hat{x}\left(\beta,\alpha\right),\hat{x}\left(\alpha,\beta\right)\right]}{\partial \beta} + \sigma \cdot \left.\frac{d\pi\left[\hat{x}\left(\beta,\beta\right),\hat{x}\left(\beta,\beta\right)\right]}{d\beta}\right|_{\beta=\alpha} = 0. \tag{35}$$

Using the first-order condition for $\left(\hat{x}\left(\alpha,\beta\right),\hat{x}\left(\beta,\alpha\right)\right)$ to be an interior Nash equilibrium strategy pair (recall (22)), this equation is equivalent to

$$\left(\sigma-\alpha\right) \cdot \hat{x}_1\left(\alpha,\alpha\right) + \left(1-\sigma\alpha\right) \cdot \hat{x}_2\left(\alpha,\alpha\right) = 0. \tag{36}$$

This generalizes the equation (23) from uniform random matching to assortative matching, and it is then straightforward to show:

[15] For an early analysis by economists of evolutionarily stable strategies in the prisoner's dilemma in the presence of assortative matching, see also Bowles and Gintis (1998).

Observation 10: In a continuum population, under assortative matching and observable preferences, evolutionary stability within the class of altruistic preferences requires the degree of altruism to equal the index of assortativity if strategies are strategically neutral (in terms of material payoffs), to exceed the index of assortativity if strategies are strategic complements, and be lower than the index of assortativity if strategies are strategic substitutes.

By contrast to settings with uniform random matching (see Observation 7), then, here a positive degree of altruism may be stable even when strategies are strategic substitutes. Like under uniform random matching, however, the stable degree of altruism depends on the nature of the material game.

Arguably, such dependence on the material game can be explored to investigate how preferences depend on the environment in which the population evolves, in so far as this environment affects material payoffs. In Alger and Weibull (2010) we examine in detail a material game which may well have been relevant in our evolutionary past: grown-up siblings exert effort towards production, whose outcome is uncertain, and, upon observing each other's output each sibling may choose to share some of its output with the other. Properties of the associated evolutionarily stable degree of sibling altruism are derived, and numerical simulations show how it depends on the harshness of the environment, such as the costs of and probabilistic returns to effort. We found that the stable degree of altruism is lower in harsher environments. This result may appear counter-intuitive since risk-sharing has a larger survival value in harsh environments. The result is explained by the fact that altruists are more vulnerable to exploitation by less altruistic siblings in harsher environments. In harsh environments, individuals work harder, so a rare mutant who is slightly less altruistic than the residents is almost certain to be helped by his sibling if his output is low.[16]

4.3 Finite populations

Until now we have focused on infinitely large populations, which, admittedly, is an unrealistic assumption. Accordingly, a strand of the literature has modeled preference evolution in finite populations. The key implication is that, even absent any assortativity in the matching process, the distribution of types that an individual faces in its matches depends on his own

[16]See Alger and Weibull (2008) for a discussion of these results in light of evidence on the strength of family ties in different parts of Europe in pre-industrial times.

type. We begin by illustrating this point with a mini ultimatum bargaining game, analyzed by Huck and Oechssler (1996), and then turn to the main contribution by Ok and Vega-Redondo (2001). Prior to doing this, however, we note that the definition of evolutionary stability can still be applied, by letting the share ε of mutants be the number of mutants divided by total population size.

Individuals in a finite population of finite size N (where N is even), are randomly matched into pairs to play two rounds of the mini ultimatum bargaining game, once in the proposer role and once in the responder role; individuals cannot observe each other's type, and the authors further assume that individuals do not condition play in the second round on the play in the first round. When in the responder role, an individual either accepts or rejects the proposer's offer. If she accepts, the proposer's offer is implemented; otherwise, they both get material payoff of 0. When in the proposer role, a player can choose between the fair split of the endowment, which is set to 2, or an egoistic split, which, if accepted, would give $2-\delta$ to him and $\delta \in (0,1)$ to the responder. Huck and Oechssler (1996) assume that there are two types of players. Type A is *Homo oeconomicus*, while Type B gets subjective utility $\rho \in (\delta, 1)$ from rejecting an unfair offer. Interactions occur under incomplete information. Anticipating that Type A individuals accept all offers, while Type B individuals accept only fair offers, it is optimal for an individual to offer the egoistic split only if he expects to meet a Type B with a sufficiently low probability. Suppose that Type A is the resident type, and that one mutant of Type B appears in the population. For some parameter values, both the residents and the mutant then offer the egoistic split. This results in $\bar{\pi}_B > \bar{\pi}_A$ if the benefit of always getting $2-\delta$ in the proposer role net of the material cost of turning down the egoistic offer, is greater than the benefit of getting $2-\delta$ in the proposer role with probability $(N-2)/(N-1)$ only, which is the case for residents.[17]

Ok and Vega-Redondo (2001) propose a more general model of preference evolution in n-player interactions in finite populations, both under complete and incomplete information. Their goal is to establish results on the stability properties of *Homo oeconomicus* preferences, and they confine their analysis to population states in which there is at least one individual of the mutant type present. Under complete information they generalize the argument illustrated by the hawk-dove game above by showing that in any game where it would pay off materially to be a Stackelberg leader, *Homo oeconomicus* is materially outperformed by a type who is committed to playing the strategy that a Stackelberg leader would play against a

[17] A similar effect was noted by Schaffer (1988) in a model of strategy evolution.

follower with *Homo oeconomicus* preferences. In a match between a *Homo oeconomicus* and this committed type, the latter materially outperforms the former. Hence, if the committed type is a rare mutant in a population with *Homo oeconomicus* as residents, the mutants materially outperform residents (even if they perform worse than residents when matched with each other).[18] Under incomplete information, for settings in which the material payoff function is continuous and strictly concave, and utility functions are in the set F, they show that while *Homo oeconomicus* may be unstable when the population is small, there exists a population size above which *Homo oeconomicus* is evolutionarily stable against behaviorally distinct types. This shows that the corresponding result derived under incomplete information uniform random matching in a continuum population (see Observation 3) is robust with respect to population size.

In sum:

> **Observation 11**: In finite populations, in a large class of settings under incomplete information *Homo oeconomicus* is evolutionarily stable if and only if the population is large enough.

5 Discussion

We conclude by discussing a number of issues that we feel should receive more attention by researchers in this field.

5.1 The preference type space

The results reported above indicate that predictions regarding the evolutionary viability of preferences evidently depend on assumptions regarding the set of potential preferences. In this respect the literature has so far focused on two main approaches. One consists in minimally restricting the set of potential preferences, while the other one restricts attention to a certain parametric class of preferences. In both cases, the domain of preferences is the

[18] Koçkesen, Ok and Sethi (2000a,b) make a similar point with a mutant utility function which consists in maximizing own material payoff relative to that of the opponent(s). They do not, however, conduct an evolutionary analysis.

set of strategy profiles in the material game. Going forward, we see several avenues for future research on this issue.

First, while the literature which restricts attention to a certain parametric class of preferences has tended to focus on altruistic preferences (Becker, 1976), the behavioral economics literature has proposed a menu of parametric classes of preferences. Thus, the evolutionary foundations of warm glow (Andreoni, 1990), inequity aversion (Fehr and Schmidt, 1999), social responsibility (Brekke, Kverndokk, and Nyborg, 2003), or lying costs (Kartik, 2009), among others, remain to be examined.

Second, there is no particular reason for why the domain of preferences should be restricted to the set of strategy profiles in the material game. In view of the behavioral economics literature, two alternative hypotheses seem promising. First, an individual's preferences may depend on the preferences of his opponent. An example of such preferences is reciprocal altruism (Levine, 1998), whereby the weight that an individual attaches to his opponent's material payoff depends on some underlying altruism parameter and the underlying altruism parameter of his opponent. Sethi and Somanathan (2001) adopt a similar preference specification and identify the evolutionary viability of reciprocal altruists compared to *Homo oeconomicus*. But one can imagine many other such preferences. Second, people may have a desire to conform (Bernheim, 1994), a sense of identity (Akerlof and Kranton, 2000), or image concerns (Bénabou and Tirole, 2006). Although some work has been conducted on the evolutionary foundations of social mindedness (Fershtman and Weiss, 1998), more general analyses are called for.

Third, the commonly adopted approach in the economics literature on preference evolution is to interpret the material payoff as fitness.[19] This assumption merits closer examination. Indeed, although it is not outlandish to assume that fitness (loosely speaking, the number of surviving offspring if the preference trait is genetically determined, and the number of cultural "offspring" if the preference trait is acquired through cultural transmission) is monotonic in some material payoff (calories, wealth, physical or mental achievements, cool factor, etc.), it is not clear whether the preferences predicted in models where the material payoff is interpreted as fitness in fact apply at the material-payoff or at the fitness level. This question is key for researchers who want to test the theoretical predictions. Since fitness

[19] A notable exception is Robson (1996), who analyzes the evolutionary foundations of von Neumann-Morgenstern utility functions over material payoffs, when these payoffs determine the number of children. Attention is restricted to decision problems, however.

is a complex notion, which evolutionary biologists have modeled for decades, collaboration between economists and evolutionary biologists on this issue may be fruitful. Some such interdisciplinary efforts have been made in recent years; see Day and Taylor (1998), Akçay et al. (2009), Akçay and van Cleve (2012), Lehmann, Alger, and Weibull (2015), and Alger, Weibull, and Lehmann (2018).

5.2 Observability of preferences

The theory summarized above indicates that the observability of preferences matters significantly for the qualitative nature of stable preferences. First, observability enables mutants to benefit materially either by coordinating on efficient play, or by making residents adopt a different behavior towards them than towards other residents. Second, under complete information evolutionarily stable preferences typically depend on the material game, i.e., on the environment in which the population evolves. While the literature has delivered a rich set of results in the two extreme scenarios of complete and incomplete information, less is known about intermediate scenarios. Heifetz, Shannon, and Spiegel (2007a) consider a scenario in which an exogenously given share of matches interact under complete information, while the remaining share interact under incomplete information. Heifetz, Shannon, and Spiegel (2007b), as well as Frank (1987), analyze a model in which matched individuals each receives a noisy signal of the opponent's preference parameter. Dekel, Ely, and Yilankaya (2007) study the robustness of their results to the assumption that each individual observes his opponent's preferences with some probability $p \in (0, 1)$ (independent of what the opponent observes). Many other intermediate scenarios can be imagined.

The ability of individuals to correctly perceive the preferences of those with whom they interact cannot be dissociated from the issue of mimicry, however. If it pays off materially to be perceived as being of a certain type, would it then not pay off even more to appear to be of this type and in fact maximize own material payoff? Deception is commonplace in the animal and the vegetal kingdom (and other kingdoms of life). There is thus reason to believe that it should also be present in our species. On the other hand, it has often been argued that emotions, such as irrepressible anger, or physical states, such as blushing, are honest signals of preferences (Frank, 1987, 1988, Hirshleifer, 2001). In view of the results reported above, that in some settings the flexible *Homo oeconomicus*, who materially best responds to (its perception of) the social environment can coexist with committed types, it would be interesting to examine heterogeneous population states, with both mimickers and

non-mimickers. Such theoretical investigations should seek to determine, in general settings, conditions which enable such heterogeneous population states to be stable, or prevent them from being so. For recent contributions in this direction, see Mohlin (2012) and Hopkins (2014).

Relatedly, in almost all models individuals do not select their interaction partners. Lifting this assumption is certainly important; for some contributions in this direction, see McNamara et al. (2008) and Izquierdo, Izquierdo, and Vega-Redondo (2010).

5.3 Individuals' environment

The literature on preference evolution in strategic interactions focuses almost exclusively on populations which evolve in a given and fixed environment, represented in the models by the material payoff function, the strategy set, the population size, and the matching process. Hence, no assumptions are needed regarding the ability of individuals to perceive changes in the environment, and *a fortiori* regarding their ability to respond to such changes. Intuitively, a changing environment should favor flexible utility maximizers over committed types. However, intuition alone cannot guide our understanding of whether a changing environment should tilt the balance in favor of *Homo oeconomicus* or other preferences. This is an important agenda for future research.

Another important issue is that the environment itself may partly be a product of preferences. For instance, in a hunter-gather population the marginal cost of gathering food may depend on gathering behaviors of past generations, which in turn may have depended on the time preferences of individuals in these past generations. In spite of the importance of such interdependence between humans and the environment in which they evolve, this topic has hitherto not been received much attention (see, however, Sethi and Somanathan, 1996).

Relatedly, institutions—the "humanly devised constraints that structure political, economic and social interactions" (North, 1991)—are also endogenous. Through their impact on the set of available actions and the associated material payoffs, institutions may be expected to affect preferences. Hence, it is likely that preferences and institutions coevolve, a phenomenon which is not well understood (for recent contributions, see Belloc and Bowles, 2017, and Wu, 2017). In particular, economists are in a good position to develop models that analyze whether and how the market economy has affected the prevalence of certain preferences and traits in humans (Saint-Paul, 2007).

More generally, in view of the importance of markets in the modern world, and of the literature on preference evolution, we would like to highlight the possibility that the traditional stance in economics which consists in treating preferences as primitives may give rise to policy recommendations which fail to achieve long-term goals. Indeed, this would be the case if the implementation of such recommendations led to changes in preference distributions which would deteriorate the situation in the long run.

6 References

Akçay, E., and J. van Cleve (2012): "Behavioral Responses in Structured Populations Pave the Way to Group Optimality," *American Naturalist* 179, 257-269.

Akçay, E., J. van Cleve, M.W. Feldman, and J. 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.

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

Alchian, A. (1950): "Uncertainty, Evolution and Economic Theory," *Journal of Political Economy*, 58, 211–21.

Alger, I., and J. Weibull (2008): "The Fetters of the Sib: Weber Meets Darwin," SSE/EFI Working Paper Series in Economics and Finance No 682.

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 Assortativity," *Econometrica*, 81, 2269-2302.

Alger, I., and J. Weibull (2016): "Evolution and Kantian Morality," *Games and Economic Behavior*, 98, 56-57.

Alger, I., J. Weibull, and L. Lehmann (2018): "Evolution of Preferences in Group-Structured Populations," Toulouse School of Economics WP 18-888.

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

Banerjee, A., and J. Weibull (1995): "Evolutionary Selection and Rational Behavior," in Alan Kirman and Mark Salmon (eds.), *Learning and Rationality in Economics*, Oxford: Basil Blackwell.

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

Belloc, M., and S. Bowles (2017): "Persistence and Change in Culture and Institutions under Autarchy, Trade, and Factor Mobility," *American Economic Journal: Microeconomics*, 9, 245-76.

Bénabou, R., and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652-1678.

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.

Bernheim, B.D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102:841–877.

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.

Bowles, S., and H. Gintis (1998): "The Moral Economy of Communities: Structured Populations and the Evolution of Pro-Social Norms," *Evolution and Human Behavior*, 19, 3-25.

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

Darwin, C. (1859): *The Origin of Species, by Means of Natural Selection.* London: John Murray.

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.

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

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

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

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

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

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

Friedman, M. (1953): *Essays in Positive Economics*. Chicago: University of Chicago Press.

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

Güth, W., and H. Kliemt (1998): "The Indirect Evolutionary Approach: Bridging the Gap Between Rationality and Adaptation," *Rationality and Society*, 10, 377-399.

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

Hopkins, E. (2014): "Competitive Altruism, Mentalizing, and Signaling," *American Economic Journal: Microeconomics*, 6, 272-92.

Hamilton, W.D. (1964): "The Genetical Evolution of Social Behaviour," *Journal of Theoretical Biology*, 7, 1-52.

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.

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

Hirshleifer, J. (1978): "Competition, Cooperation, and Conlict in Economics and Biology", *American Economic Review Papers and Proceedings,* 68, 232-243.

Hirshleifer, J. (2001): "Game-Theoretic Interpretations of Commitment," in Nesse, R. (Ed.) *Evolution and the Capacity for Commitment.* Russell Sage Foundation.

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

Izquierdo, S.S., L.R. Izquierdo, and F. Vega-Redondo (2010): "The Option to Leave: Conditional Dissociation in the Evolution of Cooperation," *Journal of Theoretical Biology*, 267, 76-84.

Jensen, M.K. and A. Rigos (2018): "Evolutionary Games and Matching Rules," *International Journal of Game Theory.* https://doi.org/10.1007/s00182-018-0630-1

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

Kartik, N. (2009): "Strategic Communication with Lying Costs," *Review of Economic Studies*, 76, 1359-1395.

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.

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

Lehmann, L., I. Alger, and J. Weibull (2015): "Does Evolution Lead to Maximizing Behavior?" *Evolution* 69-7, 1858–1873.

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

McNamara, J., Z. Barta, L. Fromhage, and A. Houston (2008): "The Coevolution of Choosi-

ness and Cooperation," *Nature,* 451, 189-192.

North, D. (1991): *Institutions, Institutional Change and Economic Performance.* Cambridge: Cambridge University Press.

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.

Rayo, L. and G.S. Becker (2007): "Evolutionary Efficiency and Happiness," *Journal of Political Economy*, 115, 302-337.

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

Robson, A. (1996): **"**A Biological Basis for Expected and Non-expected Utility," *Journal of Economic Theory*, 68, 397-424.

Robson, A. (2001): "The Biological Basis of Economic Behavior," *Journal of Economic Literature*, 39, 11-33.

Robson, A., and L. Samuelson (2011): "The Evolutionary Optimality of Decision and Experienced Utility," *Theoretical Economics*, 6, 311-339.

Robson, A., and B. Szentes (2014): "A Biological Theory of Social Discounting," *American Economic Review*, 104, 3481-3497.

Rogers, A.R. (1994): "Evolution of Time Preference by Natural Selection," *American Economic Review*, 84, 460-481.

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

Saint-Paul, G. (2007): "On Market Forces and Human Evolution," *Journal of Theoretical Biology*, 247, 397-412.

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

Schaffer, M.E. (1988): "Evolutionarily Stable Strategies for Finite Populations and Variable

Contest Size," *Journal of Theoretical Biology*, 132, 467-478.

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

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

Smith, A. (1759): *The Theory of Moral Sentiments.* Reedited (1976), Oxford: Oxford University Press.

Veblen, T. (1899): *The Theory of the Leisure Class.* Reedited (2009), Oxford: Oxford University Press.

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

Wright, S. (1931): "Evolution in Mendelian Populations," *Genetics*, 16, 97–159.

Wu, J., (2017): "Political Institutions and the Evolution of Character Traits," *Games and Economic Behavior*, 106, 260-276.