WORKING PAPERS

N° 144

Evolutionarily stable preferences

Ingela Alger

Institute for Advanced Study in Toulouse

Evolutionarily stable preferences*

Ingela ALGER†

August 8, 2022

Abstract

The 50-year old definition of an evolutionarily stable strategy provided a key tool for theorists to model ultimate drivers of behavior in social interactions. For decades economists ignored ultimate drivers and used models in which individuals choose strategies based on their preferences. This article summarizes some key findings in the literature on evolutionarily stable preferences, which in the past three decades has proposed models that combine the two approaches: Nature equips individuals with preferences, which determine their strategy choices, which in turn determines evolutionary success. The objective is to highlight complementarities and potential avenues for future collaboration between biologists and economists.

1 Introduction

What drives human behavior in their interactions with others? The premise in evolutionary game theory is that each individual is *programmed* to use a certain strategy. Since the typical life of a human being consists of a large number of different kinds of interactions, Nature should thus have equipped us with automatic play of a certain strategy tailored to each one of them, the *ultimate driver* of the strategies played in a population being natural selection [1]. Such

*I thank Jorge Peña and Maria Kleshnina for helpful and stimulating discussions, and acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics) and IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

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

a worldview is, however, at odds with the idea that we both *understand* the situations we find ourselves in and *choose* how to act. But if this is an accurate description of how strategies are selected, what then guides the strategy choice?

One theory comes from economics, where the overwhelmingly common premise is that each individual is aware of his or preferences over the available strategies, which simply means that if presented with a pair of strategies, say A and B, (s)he can tell whether (s)he prefers A to B, she prefers B to A, or is indifferent between the two strategies. In an interaction with others, the answer may depend on what strategies the others are expected to play. Rational behavior requires that a strategy that is preferred over the others be selected by the individual. A Nash equilibrium strategy profile is such that no interactant wishes to alter his or her strategy given the opponents' strategies. In this approach, the individual's preferences is the *proximate driver* of his/her behavior.

When combining these two strands of thought, the question that follows naturally is: if humans choose strategies in accordance with their preferences, which preferences should we expect evolutionary forces to favor, if any? The literature on preference evolution, initiated by Frank [2] and Güth and Yaari [3], provides some answers to this question. This article summarizes some of the key findings of this literature, found mostly in economics journals, and draws some parallels with related contributions by biologists.

2 Strategy evolution in biology

2.1 Framework and definition of ESS

Throughout I follow John Maynard Smith by defining a " 'strategy' [as] a behavioral phenotype, i.e. it is a specification of what an individual will do in any situation in which it may find itself" ([4] p.10, see also the recent book by McNamara and Leimar [5]); this is also in line with standard vocabulary in non-cooperative game theory, see [6]). To fix ideas, consider first a simultaneous-move one-shot Prisoners' dilemma (PD), in which there are two *actions*—Cooperate (C) and Defect (D), and with payoffs as shown in Figure 1. In this interaction each individual will find itself in only one decision *situation*: a *strategy* can then be formalized as a probability of playing C, with D being played with the complementary probability. In the

simultaneous-move PD a strategy is thus a scalar in the interval $[0, 1]$.

	C	D
C	R, R	S, T
D	T, S	P, P

Figure 1: The payoff matrix of the simultaneous-move Prisoners dilemma.

Consider now instead a Sequential prisoners' dilemma (SPD), played by two individuals, say i and j. Nature first draws the assignment of the individuals to the first-mover and second-mover roles, with equal probability for both assignments. The first-mover then chooses between the two actions C and D, following which the second-mover chooses between the two actions C and D. The *game tree* that represents this interaction is shown in Figure 2. Here an individual's strategy consists of specifying choices in the three situations it may find itself (i.e., at the decision nodes of the game tree, following standard vocabulary associated with sequential games, see [6]. Allowing again for randomization, a strategy is thus a three-dimensional vector in the simplex $[0, 1]^3$. As shown in the game tree, I denote by $x = (x_1, x_2, x_3)$ the strategy of i, where x_1 is the probability that i plays C as a first-mover, x_2 the probability that i plays C as a second-mover following play C by j, and x_3 the probability that i plays C as a second-mover following play D by j; likewise, $y = (y_1, y_2, y_3)$ denotes the strategy used by individual j.

Note that in both the PD and the SPD with role-randomization by Nature, the set of strategies is the same for both individuals: the interval $[0, 1]$ in the PD and the simplex $[0, 1]^3$ in the SPD. We will restrict attention to interactions sharing this feature, and call X the common strategy

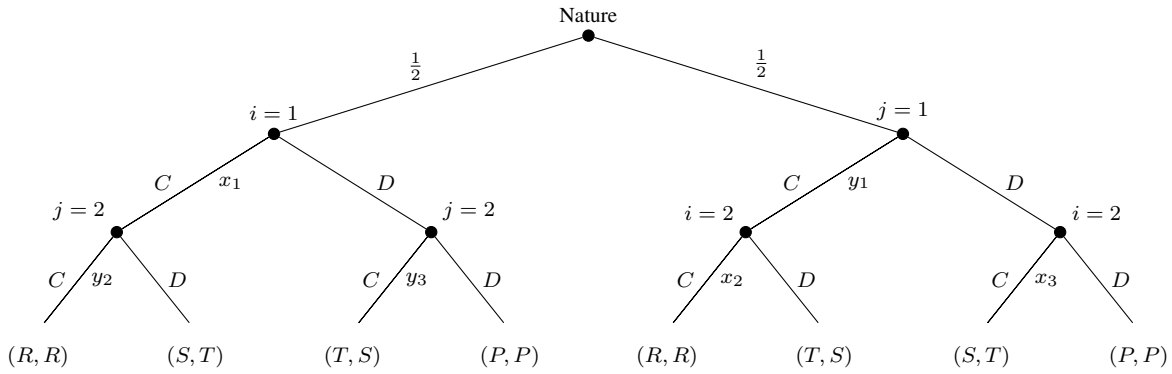

Figure 2: Meta-game protocol for the Sequential Prisoners' Dilemma ($T > R > P > S$)

set. Letting $w(x, y)$ denote the *fitness* of an individual using strategy x when the other is using strategy y, we will refer to $\Gamma = \langle X, w \rangle$ as the *fitness game*.[1]

Using standard notation for the fitness payoffs in the PD (as displayed in the payoff matrix in Figure 1), the (expected) fitness from using strategy x against strategy y is

$$w(x, y) = xyR + (1 - x)yT + x(1 - y)S + (1 - x)(1 - y)P. \tag{1}$$

In the SPD, the (expected) fitness from using strategy $x = (x_1, x_2, x_3)$ against strategy $y = (y_1, y_2, y_3)$ is

$$\begin{aligned} w(x, y) &= \frac{1}{2}[x_1 y_2 R + x_1 (1 - y_2) S + (1 - x_1) y_3 T + (1 - x_1)(1 - y_3) P] \\ &\quad + \frac{1}{2}[y_1 x_2 R + y_1 (1 - x_2) T + (1 - y_1) x_3 S + (1 - y_1)(1 - x_3) P]. \end{aligned} \tag{2}$$

I adopt the standard evolutionary game theory assumption that the population at hand is a continuum population, and that individuals are randomly matched into pairs to interact according to some given fitness game $\Gamma = \langle X, w \rangle$. This setting encompasses a large number of commonly studied games besides the PD and the SPD, for example:

- Simultaneous and one-shot games with a finite number (say, two) of pure strategies, like Hawk-Dove and Coordination (see the payoff matrices in Figure 1). The strategy set is the set of mixed strategies, $X = [0, 1]$.

- The sequential versions of the aforementioned games (following a similar structure as in the Sequential prisoners' dilemma described in detail above).

- Simultaneous and one-shot linear public goods games: $X = [0, E]$ and $w(x, y) = V(x + y) + E - x$, for some endowment $E > 0$ and multiplication factor $V \in (1/2, 1)$

- Simultaneous and one-shot non-linear public goods games where strategies are *strategic substitutes*: $X = \mathbb{R}_+$ and $w(x, y) = (x + y)^\tau - x^2$, for some $\tau \in (0, 1)$ (strategies are strategic substitutes because $\partial^2 w(x, y)/(\partial x \partial y) < 0$)

[1] To simplify the exposition, I here refer to $w(x, y)$ as the individual's fitness, although it should instead be thought of some proxy of invasion fitness, like in [5].

- Simultaneous and one-shot non-linear public goods games where strategies are *strategic complements*: $X = \mathbb{R}_+$ and $w(x,y) = (xy)^{\mu} - x^2$, for some $\mu \in (0,1)$ (strategies are strategic complements because $\partial^2 w(x,y)/(\partial x \partial y) > 0$)

- Simultaneous and one-shot common pool resource games: $X = \mathbb{R}_+$ and $w(x,y) = (a - x - y)\,x - cx$, for some $a > c \geq 0$.

- Helping games:

  - Nature draws the initial wealth distribution: with probability $1/2$, player 1's initial wealth is m^H and 2's is $m^L \leq m^H$, and with probability $1/2$ the players' wealths are reversed

  - the wealthier individual may transfer any amount of his/her wealth to the other

  - let $h : \mathbb{R}_+ \to \mathbb{R}$ measure the material utility from net wealth $m \in \mathbb{R}_+$, where $h' > 0$ and $h'' \leq 0$

  - letting x be player 1's transfer when rich and y 2's transfer when rich, with $x, y \in X = \left[0, m^H\right]$, then the (expected) material payoff is:

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

Note that the framework even applies to both finitely and infinitely repeated games, in which a strategy specifies which action to undertake as a function of the *history of play* (see [6]); due to the complex notation required to rigorously define repeated games and to space restrictions, however, I will not explicitly study these games here.

Letting x denote the resident strategy and ε the share of the population that uses some mutant strategy y, an evolutionarily stable strategy is then formally defined as follows [7]:

Definition 1. *A strategy $x \in X$ is **evolutionarily stable (ES)** against strategy $y \in X$, $y \neq x$, if there exists $\bar{\varepsilon}_y \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_y)$:*

$$(1-\varepsilon)\cdot w(x,x) + \varepsilon \cdot w(x,y) > (1-\varepsilon)\cdot w(y,x) + \varepsilon \cdot w(y,y). \tag{3}$$

*And x is an **evolutionarily stable strategy (ESS)** if it is evolutionarily stable against all $y \in X$, $y \neq x$.*

In (3) the left-hand side is the average fitness of individuals playing the resident strategy, while the right-hand side is the average fitness of individuals playing the mutant strategy, given the share ε of mutants in the population. The population being infinitely large and the interactants being matched in a uniformly random manner, any individual is matched with a resident with probability $1-\varepsilon$ and with a mutant with probability ε. In words, then, an ESS is a strategy which, once it has become prevalent in a population, earns a higher average fitness than any rare mutant strategy.

2.2 An "as if" interpretation of ESS

As a first step towards analysis of preference evolution, it is worth noting that a population in which an ESS is played can be viewed as being populated by individuals who seek to maximize own fitness.

To this end—and also to facilitate description of the analytical challenges that preference evolution sometimes entails—it proves useful to express the difference between the average fitnesses earned by residents and mutants as a function of the share of mutants ε, using what is called the *score function* [8]:

$$S_{x,y}(\varepsilon) \;=\; (1-\varepsilon)\cdot[w(x,x)-w(y,x)]+\varepsilon\cdot[w(x,y)-w(y,y)]. \tag{4}$$

This function being linear in ε, $S_{x,y}(0)\geq 0$ is a necessary condition for x to be ES against y, while $S_{x,y}(0)>0$ is a sufficient condition. Moreover, if $S_{x,y}(0)=0$ then the slope of the score function must be strictly positive for x to be ES against y. This leads to the following result, and also simple test for whether a strategy is evolutionarily stable:

Result 1. *1. If $w(x,x)>w(y,x)$, then x is ES against y.*

2. If $w(x,x)=w(y,x)$, then x is ES against y only if $w(x,y)>w(y,y)$.

3. If $w(x,x)<w(y,x)$, then x is not ES against y.

Since x is ESS only if $w(x,x) \geq w(y,x)$ for all $y \neq x$, an individual in a population where the ESS x is played by everyone can be interpreted *as if* (s)he were choosing the strategy which maximizes his or her fitness, given that any individual (s)he interacts with uses strategy x. This observation brings us to the main question: what if, instead of equipping us with automatic play of a strategy tailored to each possible interaction, Nature has equipped humans with (a) the ability to understand the situations they find themselves in, and (b) some *preferences* that guide their strategy choice in any given interactions? Which preferences should we then expect evolutionary forces to favor, if any?

3 Preference evolution

3.1 In economics, preferences guide behavior

In their analyses of human behavior, economists typically rely on the premise that in any given situation each individual chooses the option that (s)he prefers among all the options that are feasible for him/her. Choosing an option other than a preferred one is deemed irrational. This simple idea is captured by positing that each individual is able to rank all the feasible options. This ranking is then formalized as a preference ordering, which for any pair of feasible options A and B tells whether the individual prefers A, B, or is indifferent between A and B. It turns out that under certain conditions, such a preference ordering can be fully described by a function that to each feasible option associates a real number: the number associated with option A is higher (resp. lower) than that associated with option B if and only if the individual prefers A over B (resp. B over A), and the same number to both options if the individual is indifferent between them (see, e.g., [9]). Such a function is called a *utility function* in economics. Here we will instead refer to it as a *preference function*, or simply *preferences*. In any given situation, an individual is expected to choose that option that yields the highest possible value of the function, since this is the option (s)he prefers. Economists do not interpret this utility maximization literally: it is simply a mathematical tool that the researcher uses to describe behavior that amounts to choosing the preferred item from the *feasible set*.

In the context of a fitness game $\Gamma = \langle X, w \rangle$, an individual's feasible set is the set of strategies X. The individual with whom (s)he is matched to interact—the opponent—also chooses

some strategy in the strategy set X. To capture the fact that an individual's ranking over own strategies may depend on what strategy the opponent is expected to use, a preference function is some function $u : X^2 \to \mathbb{R}$ that to each pair of own and opponent's strategy associates a real number. If the individual strictly prefers some strategy profile, say (x, y), over another, say (x', y'), then u gives a strictly higher number to the former, while if the individual is indifferent, then u gives the same number to both strategy profiles.

The question posed at the end of the previous section can now be formally stated as follows: given that fitness drives evolutionary success, should we expect evolution to favor preferences that are a simple reflection of the fitness function?

Definition 2. *In any given fitness game $\Gamma = \langle X, w \rangle$, **a fitness-maximizer** has a preference function that coincides with the fitness function, i.e.,*

$$u(x, y) = w(x, y). \tag{5}$$

For any given strategy that a fitness-maximizer expects the opponent to choose, (s)he chooses a strategy that maximizes own fitness, since this is the strategy that (s)he prefers. Should we expect evolution to lead humans to be such *fitness-maximizers*? The literature has revealed that information plays a key role in this context.

3.2 Interactions under complete information

An example will serve as an introduction. Consider a population in which individuals are matched pairwise to interact according to a fitness game $\Gamma = \langle X, w \rangle$ which is a simultaneous-move and one-shot non-linear public goods game with strategy set $X = \mathbb{R}_+$ and fitness function

$$w(x, y) = (x + y)^{1/2} - x^2, \tag{6}$$

where the first term is the benefit from the sum of own and other's contribution to the public good and the second term is the individual's cost of making contribution of size x. Suppose that all individuals in the population are fitness-maximizers, i.e., each individual has preferences $u : X^2 \to \mathbb{R}$ defined in (5), namely $u(x, y) = (x + y)^{1/2} - x^2$, over own and other's contribution towards the public good, x and y. Suppose further that individuals who are matched to interact

can observe each other's preference function, i.e., they interact under *complete information* ([6]). In such an interaction, what pair of strategies will be played? While this question has no simple general answer (see, e.g., [10]), it is common in economics to apply the *Nash equilibrium concept*. A pair of strategies (x^*, y^*) is a Nash equilibrium if neither individual would like to deviate from their strategy, given the other's strategy. Formally:

$$\begin{cases} x^* \in \arg\max_{x \in X} \ (x + y^*)^{1/2} - x^2 \\ y^* \in \arg\max_{y \in X} \ (x^* + y)^{1/2} - y^2. \end{cases} \tag{7}$$

Differentiability of the preference function together with an unbounded strategy set implies that (x^*, y^*) must satisfy the following set of first-order conditions (these conditions are also sufficient because of the strict concavity of the preference function):

$$\begin{cases} \frac{1}{2}(x^* + y^*)^{-1/2} = 2x^* \\ \frac{1}{2}(x^* + y^*)^{-1/2} = 2y^*. \end{cases} \tag{8}$$

It follows that $x^* = y^* = (1/32)^{1/3}$. Suppose now that another preference function enters this population. For example suppose that some individuals have the preference function

$$v(x, y) = w(x, y) - \frac{1}{2} w(y, x). \tag{9}$$

For any given strategy used by the individual, x, and any given strategy used by the opponent, y, in this function both the individual's own fitness, $w(x, y)$, and the fitness of the opponent, $w(y, x)$, appears. Since the former enters with a positive sign and the latter with a negative sign, this function means that the individual prefers strategy profiles (x, y) that give a higher fitness to itself and a lower fitness to the opponent. In economics such an individual is said to have *spiteful preferences* (see, e.g., [11]). In an interaction between one fitness-maximizer (with preference function u) and one spiteful individual (with preference function v) a pair of strategies $(\hat{x}, \hat{y})$ is a Nash equilibrium if and only if:

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ (x + \hat{y})^{1/2} - x^2 \\ \hat{y} \in \arg\max_{y \in X} \ (\hat{x} + y)^{1/2} - y^2 - \frac{1}{2}[(\hat{x} + y)^{1/2} - \hat{x}^2]. \end{cases} \tag{10}$$

The following first-order conditions are necessary (and also sufficient because of the strict concavity of the preference functions):

$$\begin{cases} \frac{1}{2}\left(\hat{x}+\hat{y}\right)^{-1/2} = 2\hat{x} \\ \frac{1}{4}\left(\hat{x}+\hat{y}\right)^{-1/2} = 2\hat{y}. \end{cases} \tag{11}$$

It follows that $\hat{x} = 3^{1/3} \cdot (1/32)^{1/3}$ and $\hat{y} = \hat{x}/2$. By valuing the benefit of the public good $(x+y)^{1/2}$ less than a fitness-maximizer, the spiteful individual is less willing to contribute than a fitness-maximizer; in turn, compared to when he interacts with another fitness-maximizer, the fitness-maximizer compensates for the ensuing lower contribution by his opponent by increasing his contribution. Indeed, note that $\hat{x} > x^* > \hat{y}$. Calling the preference function u the resident trait, and the spiteful preference function v the mutant trait, and letting ε denote the share of individuals with the mutant trait, it follows that the average fitness of the resident fitness-maximizers is

$$(1-\varepsilon) \cdot w\left(x^*, x^*\right) + \varepsilon \cdot w\left(\hat{x}, \hat{y}\right) \tag{12}$$

while the average fitness of the mutant spiteful individuals is

$$(1-\varepsilon) \cdot w\left(\hat{y}, \hat{x}\right) + \varepsilon \cdot w\left(\tilde{x}, \tilde{x}\right), \tag{13}$$

where $\tilde{x} = (1/128)^{1/3}$ is the contribution that a spiteful individual makes when matched with another spiteful individual. It follows from $\hat{x} > x^* > \hat{y}$ that $w(\hat{y}, \hat{x}) > w\left(x^*, x^*\right)$, and hence that for ε close enough to zero, the mutants obtain a strictly higher average fitness than the residents. Following the same logic as in standard evolutionary game theory, we conclude that a population of fitness-maximizers would not resist the invasion by mutants with the spiteful preference function v.

This conclusion, here reached in a simple example, has been shown by Heifetz, Shannon, and Spiegel [12] to hold for any fitness game $\Gamma = \langle X, w \rangle$ such that w is a thrice differentiable function and X is an open subset of $\mathbb{R}$ (see also Ok and Vega-Redondo [13]). They show this general result in a model which encompasses any preference function of the form

$$u\left(x, y\right) = w(x, y) + B(x, y, \tau), \tag{14}$$

where $\tau \in E \subseteq \mathbb{R}$ is the evolving trait and B is some thrice differentiable function (the fitness-maximizer is the special case with $B(x, y, \tau) = 0$ for all $(x, y) \in X^2$).

This observation begs the question: which preference function, if any, is evolutionarily stable?

Definition 3. *Consider a population in which individuals are matched into pairs to interact according to the fitness game $\Gamma = \langle X, w \rangle$ and under complete information about each other's preference function. Let Θ denote the set of all possible preference functions $u : X^2 \to \mathbb{R}$ such that there exists a unique Nash equilibrium in each matched pair. Then, a preference function u is **evolutionarily stable under complete information (ESC)** against preference function v if there exists $\bar{\varepsilon}_v \in (0, 1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_v)$:*

$$(1 - \varepsilon) \cdot w(x^*, x^*) + \varepsilon \cdot w(\hat{x}, \hat{y}) > (1 - \varepsilon) \cdot w(\hat{y}, \hat{x}) + \varepsilon \cdot w(\tilde{x}, \tilde{x}), \tag{15}$$

where (x^*, x^*) is the unique Nash equilibrium in an interaction between two residents, $(\hat{x}, \hat{y})$ is the unique Nash equilibrium in an interaction between a resident and a mutant, and $(\tilde{x}, \tilde{x})$ is the unique Nash equilibrium in an interaction between two mutants.

*The preference function u is an **evolutionarily stable preference function under complete information (ESPFC)** if it is evolutionarily stable against all preference functions $v \in \Theta$, $v \neq u$.*

In words, an ESPFC is a preference function which, once it has become prevalent in a population, cannot be displaced by any other preference function, the criterion being fitness evaluated at Nash equilibrium. It is important to note that researchers who have adopted this definition do not necessarily believe that individuals in real life do play some Nash equilibrium; however, it is a useful first approach, and the definition can be readily adapted to other equilibrium notions. It should also be remarked that the definition can be generalized to encompass settings where there exist multiple Nash equilibria; however, most of the literature has restricted attention to settings with a unique Nash equilibrium (an exception is the model of Dekel, Ely, and Yilankaya [14], but their analysis is on the other hand restricted to fitness games with finite action sets).

In settings where there exists a unique Nash equilibrium in each matched pair, the score

function is well defined:

$$S_{u,v}(\varepsilon) \;=\; (1-\varepsilon)\cdot[w(x^*,x^*)-w(\hat{y},\hat{x})]+\varepsilon\cdot[w(\hat{x},\hat{y})-w(\tilde{x},\tilde{x})]. \tag{16}$$

Since $S_{u,v}$ is linear in ε, the following result and simple test obtains:

Result 2. *Let (x^*,x^*) be the unique Nash equilibrium in an interaction between two residents, $(\hat{x},\hat{y})$ the unique Nash equilibrium in an interaction between a resident and a mutant, and $(\tilde{x},\tilde{x})$ the unique Nash equilibrium in an interaction between two mutants. Then:*

1. *If $w(x^*,x^*) > w(\hat{y},\hat{x})$, then u is ESC against v.*
2. *If $w(x^*,x^*) = w(\hat{y},\hat{x})$, then u is ESC against v only if $w(\hat{x},\hat{y}) > w(\tilde{x},\tilde{x})$.*
3. *If $w(x^*,x^*) < w(\hat{y},\hat{x})$, then u is not ESC against v.*

A fundamental difference with strategy evolution is that the set of potential preference functions, Θ, is a priori undetermined. Hence, the researcher must make some assumption. Thus far most of the analyses of preference evolution under complete information have adopted the parametric class of preferences originally proposed by Bester and Güth in their seminal paper [15]. In a model with the following fitness function

$$w(x,y) = (m - x + ky)\,x \tag{17}$$

for some $1 > k > -1$ and $m > 0$, they examine preference functions of the form

$$u_\alpha(x,y) = w(x,y) + \alpha\cdot w(y,x), \tag{18}$$

where $\alpha \in [0,1]$ is the evolving trait. Bolle [16] and Possajennikov [17] generalize the original model by extending the range of possible values of α to $\mathbb{R}_+$. Like in the example studied in detail above (which corresponded to the special case $\alpha = -1/2$), a straightforward interpretation is that an individual with such a preference function attaches some weight, α, to the consequences of his strategy choice on the opponent. If $\alpha > 0$, he is willing to reduce own fitness in order to enhance that of the other, i.e., to act in a *pro-social* manner; economists refer to preferences with $\alpha > 0$ as *altruistic preferences* [18]. By contrast an individual with $\alpha < 0$ is willing

to reduce own fitness in order to reduce that of the other, i.e., to act in an *anti-social* manner; economists refer to preferences with $\alpha < 0$ as *spiteful* ones. Finally, fitness-maximizing individuals correspond to the special case $\alpha = 0$. Although this class of preferences thus encompasses altruism, self-interest, and spite, we will simply refer to α as the *degree of altruism*.

A key insight delivered by the analyses of Bester and Güth [15], Bolle [16], and Possajennikov [17], is that the evolutionarily stable value of α depends on the parameters of the fitness function, k and m. Alger and Weibull [11] subsequently generalized the analysis of the same class of preference functions by considering any fitness game with a continuous fitness function w such that there exists a unique (and regular, meaning that it is differentiable) Nash equilibrium in any dyadic interaction. Letting $x^*(\alpha, \alpha)$ denote the equilibrium strategy employed by both individuals in a dyad where both have degree of altruism α, they first show that the following equation is necessary for a preference function of the form (18) with $\alpha = \alpha^*$ to be evolutionarily stable:

$$\alpha^* \cdot x_1^*(\alpha^*, \alpha^*) = x_2^*(\alpha^*, \alpha^*), \tag{19}$$

where the index 1 (resp. 2) indicates the partial derivative with respect to the first (resp. second) argument. This equation shows that the observability of the opponent's preferences drives a wedge between fitness-maximizing preferences and evolutionarily stable preferences. Indeed, the right-hand side represents the effect that an individual's preferences has on the *opponent's* equilibrium strategy, and fitness-maximizing preferences ($\alpha^* = 0$) are evolutionarily stable if and only if this effect is nil ($x_2^*(0, 0) = 0$). More generally, the characterization in (19) unveils a connection between the qualitative nature of the fitness function w and the sign of the evolutionarily stable value of α.

Result 3. *[Alger and Weibull [11]] A preference function of the form* (18) *with $\alpha = \alpha^*$ is an ESPFC only if:*

1. $\alpha^* < 0$ *if the strategies are strategic substitutes (i.e., $\partial^2 w(x, y)(\partial x \partial y) < 0$).*
2. $\alpha^* > 0$ *if the strategies are strategic complements (i.e., $\partial^2 w(x, y)(\partial x \partial y) > 0$).*
3. $\alpha^* = 0$ *if the strategies are strategically neutral (i.e., $\partial^2 w(x, y)(\partial x \partial y) = 0$).*

This result generates a clear prediction for the relationship between preferences on the one hand, and the specifics of the fitness function w on the other hand, where w presumably depends on

the environment in which the population evolves. Typical examples of interactions involving strategic complementarity are those that require teamwork: if heavy enough, trying to pull up a fishing net is useless unless someone else also pulls; rowing the oar on one side of a boat is useless unless the rower on the other side also rows; going for the Stag rather than the Hare (in the famous Stag-Hunt game) pays off only if the other does so as well; etc. Typical examples of interactions involving strategic substitutability are those where individuals compete over the same resources.

The prediction is testable if the researcher can measure whether individuals in the population at hand are willing to reduce own fitness in order to enhance that of the other (in which case $\alpha > 0$), or rather to reduce it (in which case $\alpha < 0$). If such direct measurement is impossible, the following comparison between the strategy that is employed by individuals in the population at hand and the ESS can be used as an indirect test:

Result 4. *Suppose that the fitness game $\Gamma = \langle X, w \rangle$ is such that $w(x, x)$ is increasing in x, and suppose that there is a unique ESS, denoted x^{ESS}. Then in a population where a preference function of the form* (18) *with $\alpha = \alpha^*$ is an ESPFC and in which the (unique) Nash equilibrium strategy $x^*(\alpha^*, \alpha^*)$ is employed:*

1. *$x^*(\alpha^*, \alpha^*) < x^{ESS}$ if the strategies are strategic substitutes (i.e., $\partial^2 w(x, y)(\partial x \partial y) < 0$).*
2. *$x^*(\alpha^*, \alpha^*) > x^{ESS}$ if the strategies are strategic complements (i.e., $\partial^2 w(x, y)(\partial x \partial y) > 0$).*
3. *$x^*(\alpha^*, \alpha^*) = x^{ESS}$ if the strategies are strategically neutral (i.e., $\partial^2 w(x, y)(\partial x \partial y) = 0$).*

3.2.1 Related models in the biology literature

Following McNamara, Gasson, and Houston [19], a series of contributions in biology have examined the evolutionary stability of *negotiation rules*. This literature takes interest in fitness games whereby individuals engage in a series of interaction rounds which eventually lead to a "negotiated outcome". Compared to the standard strategy evolution setting, where each individual is programmed to employ a certain strategy, here each individual is programmed with a response rule which specifies the strategy to play in response to the strategy used by the op-

ponent in the previous round. This alternating process converges to a pair of strategies—the negotiated outcome—which the interactants then employ in the remaining rounds.

Like in the preference evolution literature, there is *a priori* no clear set of possible negotiation rules. McNamara, Gasson, and Houston [19] and Taylor and Day [20] posit the following rule for an individual playing x in response to the opponent's play of y in the previous round:

$$x = \rho - \lambda \cdot y. \tag{20}$$

The evolving trait is the vector (λ, ρ), which represents the slope and the intercept. A population consisting of individuals with the rule $(\lambda, \rho) = (0, x^{ESS})$ would play the ESS x^{ESS}, and this rule is evolutionarily stable. However, there are also rules (λ, ρ) with $\lambda \neq 0$ that are evolutionarily stable [19, 20]. Note that the non-degenerate slope $\lambda \neq 0$ implies that an individual's behavior is swayed by the opponent's behavior: the similarity with the non-nil effect of an individual's preferences on the opponent's behavior in the model on the evolution of altruistic preferences under complete information (i.e., $x_2^*(\cdot, \cdot) \neq 0$ in (19)) is thus clear. The following remark examines in greater detail the similarities and differences between the seminal models.

Remark 1. *An interesting parallel can be drawn between the response rule in* (20) *and the model analyzed by Bester and Güth [15]. Recalling the fitness function that they posit (see* (17)*), an individual with altruistic preferences chooses a strategy x that maximizes the following expression, where y is the opponent's strategy:*

$$(m - x + ky)\, x + \alpha \cdot (m - y + kx)\, y \tag{21}$$

The necessary (and sufficient) first-order condition for this maximization is

$$m - 2x + ky + \alpha \cdot ky = 0, \tag{22}$$

or

$$x = \tfrac{m}{2} + \tfrac{k(1+\alpha)}{2} y, \tag{23}$$

In other words, the best response of an individual with degree of altruism α to the opponent's strategy is equivalent to the response rule examined by McNamara, Gasson, and Houston [19]

and Taylor and Day [20] (see (20)*) for* $\rho = \frac{m}{2}$ *and* $\lambda = -\frac{k(1+\alpha)}{2}$*. Hence, the system of necessary conditions for a Nash equilibrium strategy profile in a dyad with degrees of altruism* (α, α') *at which Bester and Güth [15] evaluate fitness,*

$$\begin{cases} x^*(\alpha, \alpha') = \frac{m}{2} + \frac{k(1+\alpha)}{2} \cdot x^*(\alpha', \alpha) \\ x^*(\alpha', \alpha) = \frac{m}{2} + \frac{k(1+\alpha')}{2} \cdot x^*(\alpha, \alpha') \end{cases} \tag{24}$$

coincides with the system of equations that define the negotiated outcome in a dyad with response rules $(\rho, \lambda), (\rho', \lambda')$ *at which McNamara, Gasson, and Houston [19] and Taylor and Day [20] evaluate fitness,*

$$\begin{cases} x^*((\rho, \lambda), (\rho', \lambda')) = \rho - \lambda \cdot x^*((\rho', \lambda'), (\rho, \lambda)) \\ x^*((\rho', \lambda'), (\rho, \lambda)) = \rho' - \lambda' \cdot x^*((\rho, \lambda), (\rho', \lambda')) \end{cases} \tag{25}$$

if $\rho = \rho' = m/2$*,* $\lambda = k(1+\alpha)/2$*, and* $\lambda' = k(1+\alpha')/2$*. This comparison highlights two differences between Bester and Güth [15] on the one hand, and McNamara, Gasson, and Houston [19] and Taylor and Day [20] on the other hand. First, in the latter both the slope and the intercept of the response rule are evolving traits, while in the former only the slope evolves. Second, they do not use the same fitness function.*

This remark brings us to the contribution by Akcay et al. [21], which builds a nice bridge between the biology literature on the evolution of negotiation rules on the one hand, and the economics literature on preference evolution under complete information on the other hand. In a model with the fitness function

$$w(x, y) = y^{1/2} - x^2, \tag{26}$$

they consider preference functions of the form

$$u\,(x, y) = w(x, y) \cdot [w(y, x)]^{\beta}, \tag{27}$$

and let $\beta > 0$ be the evolving trait. They derive the best response of an individual with such preferences, they determine the conditions under which a negotiation phase would converge to

Nash equilibrium in a complete information game between two individuals with such preferences, and they characterize the evolutionarily stable value of β. They further derive a result in a general model with generic but differentiable fitness and preference functions, such that in each dyad there exists a unique Nash equilibrium. This result can be described as follows:

Result 5. *[Akcay et al. [21]] Suppose that the fitness game $\Gamma = \langle X, w \rangle$ is such that $w(x, x)$ is increasing in x, and suppose that there is a unique ESS, denoted x^{ESS}. Then in a population where a preference function of the form* (27) *with $\beta = \beta^*$ is an ESPFC and in which the (unique) Nash equilibrium strategy $x^*(\beta^*, \beta^*)$ is employed:*

1. *$x^*(\beta^*, \beta^*) > x^{ESS}$ if the strategies are strategic complements (i.e., $\partial^2 w(x, y)(\partial x \partial y) > 0$).*

2. *$x^*(\beta^*, \beta^*) = x^{ESS}$ if the strategies are strategically neutral (i.e., $\partial^2 w(x, y)(\partial x \partial y) = 0$).*

The qualitative nature of this result is in line with that of a subset of the results found by Alger and Weibull [11] in the case of altruistic/spiteful preference functions (see Result 4 above), an observation which would be expected in light of the following remark.

Remark 2. *It is well-known in economics that any preference ranking over items in an individual's choice set that can be described by some preference function u, can equally well be described by any positive monotone transformation of u. Taking the logarithm of the function posited by Akcay et al. [21] (see* (27)*), and defining,*

$$\tilde{u}\,(x, y) = \ln w(x, y) + \beta \cdot \ln w(y, x), \tag{28}$$

it is clear that this class of preference functions is qualitatively similar to the one adopted in the economics literature that built on Bester and Güth [15] (see (18)*).*

It is still an open question whether the qualitative nature of Results 4 and 5 generalizes to other preference function classes. As mentioned earlier, it is a priori not clear which preference classes should be examined by modelers, a question that will be brought up again in the discussion section below.

3.3 Interactions under incomplete information

By contrast to interactions that take place under complete information, in interactions where the individuals cannot observe each other's preference function their behavior cannot be swayed by the opponent's preference function. However, an individual may still adapt behavior to the *distribution* of preference functions present in the population. This is the assumption adopted in the analyses of preference evolution under incomplete information [13, 14, 22]. This section summarizes results by closely following the modeling assumptions of Alger and Weibull [22], for a reason that will become clear below.

Let a population state $s = (u, v, \varepsilon)$ be defined by the resident preference function $u \in \Theta$, the mutant preference function $v \in \Theta$, and the share ε of mutants. Under the same matching protocol as in the standard framework—i.e., that any individual faces a probability ε of being matched with a mutant—the criterion used in the literature is fitness evaluated at type-homogenous Bayesian Nash equilibrium strategy profiles (below these will simply be referred to as equilibrium strategy profiles, or equilibria), defined as follows.

Definition 4. *In any state $s = (u, v, \varepsilon) \in \Theta^2 \times (0, 1)$, a strategy pair $(x^*, y^*) \in X^2$ is a **type-homogenous Bayesian Nash Equilibrium (BNE)** if*

$$\begin{cases} x^* \in \arg\max_{x \in X} & (1-\varepsilon) \cdot u\left(x, x^*\right) + \varepsilon \cdot u\left(x, y^*\right) \\ y^* \in \arg\max_{y \in X} & (1-\varepsilon) \cdot v\left(y, x^*\right) + \varepsilon \cdot v\left(y, y^*\right). \end{cases} \tag{29}$$

The first (resp. second) equation says that a resident (resp. a mutant) chooses a strategy that maximizes the expected value of the preference function u (resp. v), where the expectation is taken over the value that the preference function takes in a match with another resident, who plays x^*, and the value that it takes in a match with a mutant, who plays y^*. Type-homogeneity means that all individuals with the same preference function (or preference type) use the same strategy.

Given some equilibrium strategy profile (x^*, y^*) associated with population state $s = (u, v, \varepsilon)$, define the equilibrium fitnesses of residents and mutants:

$$W_u\left(x^*, y^*, \varepsilon\right) = (1-\varepsilon) \cdot w(x^*, x^*) + \varepsilon \cdot w(x^*, y^*) \tag{30}$$

$$W_v\left(x^*, y^*, \varepsilon\right) = (1-\varepsilon) \cdot w(y^*, x^*) + \varepsilon \cdot w(y^*, y^*) \tag{31}$$

By contrast to the analyses of interactions under complete information, the typical approach for interactions under incomplete information consists in minimally constraining the set of possible preference functions, Θ. In particular it turns out that it is possible to derive general results even for settings in which there are states $s = (u, v, \varepsilon)$ with multiple equilibria.

Definition 5. *[Alger and Weibull [22]] A preference function $u \in \Theta$ is **evolutionarily stable under incomplete information (ESI)** against a function $v \in \Theta$ if there exists an $\bar{\varepsilon} > 0$ such that $W_u\left(x^*, y^*, \varepsilon\right) > W_v\left(x^*, y^*, \varepsilon\right)$ in all Nash equilibria (x^*, y^*) in all states $s = (u, v, \varepsilon)$ with $\varepsilon \in (0, \bar{\varepsilon})$. A preference function u is an **evolutionarily stable preference function under incomplete information (ESPFI)** if it is ESI against all preference functions $v \neq u$ in Θ.*

To illustrate the analytical challenge that this setting presents, focus momentarily on a setting where in each state $s = (u_\theta, u_\tau, \varepsilon) \in \Theta^2 \times (0, 1)$ there exists a unique equilibrium strategy profile. In such a setting the score function is:

$$\begin{aligned} S_{u,v}\left(\varepsilon\right) \;\; = \;\; & (1-\varepsilon) \cdot \left[w\left(x^*(\varepsilon), x^*(\varepsilon)\right) - w\left(y^*(\varepsilon), x^*(\varepsilon)\right)\right] \\ & + \varepsilon \cdot \left[w\left(x^*(\varepsilon), y^*(\varepsilon)\right) - w\left(y^*(\varepsilon), y^*(\varepsilon)\right)\right], \end{aligned} \tag{32}$$

where I have made it explicit that the equilibrium strategy profile may vary with the share ε of mutants. Clearly, the score function is not necessarily linear in ε. In fact, without further assumptions it may even be discontinuous, since the equilibrium strategy profile may vary discontinuously with ε. This contrasts sharply with the linearity in ε of the score functions under strategy evolution (4) and under preference evolution under complete information (16), which implies that analysis of the score function at $\varepsilon = 0$ is sufficient to check evolutionary stability (recall Results 1 and 2).

Nonetheless, there are conditions that render general analysis possible, even for settings with multiple equilibria. In view of Remark 2, there will, however, typically be many preference functions that are behaviorally equivalent. Clearly, the fitness-maximizing preference function cannot be evolutionarily stable against such *behavioral alikes*, defined as follows.

Definition 6. *Let X_0 be the set of type-homogenous Nash equilibria in a population consisting solely of fitness-maximizers. A preference function u' is a **behavioral alike** to fitness-maximizers*

if there exists some $x_0 \in X_0$ *such that* $x \in \arg\max_{x \in X} u'(x, x_0)$ *and* $x \in \arg\max_{x \in X} w(x, x_0)$.

In words (and somewhat loosely) a behavioral alike to fitness-maximizers is a preference type that would be willing to play a strategy that the fitness-maximizer would also be willing to play, given that the opponent plays some $x_0 \in X_0$. The following result identifies sufficient conditions for the fitness-maximizing preference function to be evolutionarily stable. It is a slight variation of the result as stated by Alger and Weibull [22], found in [23] (the difference stems from a slight difference in how behavioral alikes are defined, which does not affect the core of the result).

Result 6. *If the strategy set* X *is compact and convex, and all the preference functions in* Θ *as well as the fitness function* w *are continuous, then the fitness-maximizing preference function (see* (5)*) is ESI against any preference function that is not its behavioral alike.*

The topological properties stated in the result ensure that the correspondence, which to each population state $s = (u, v, \varepsilon) \in \Theta^2 \times (0, 1)$ associates the set of equilibrium strategy profiles, is upper-hemicontinuous. Hence, even if the introduction of an infinitesimal share of mutants sways the equilibrium strategy of the residents away from the equilibrium strategy played in the absence of mutants, the "new" equilibrium strategy is arbitrarily close to some strategy that the fitness-maximizers could have played in the absence of mutants. Continuity of the fitness function then implies that any mutant which is not a behavioral alike to fitness-maximizers obtains a strictly lower equilibrium fitness than the fitness-maximizers.

Ok and Vega-Redondo [13] adopt similar topological properties, and they show that fitness-maximizers are robust to the entry of non-fitness-maximizers even in finite but large enough populations. By contrast, in small populations the entry of mutants makes the resident fitness-maximizers shift their strategy away from any strategy they would have played in the absence of mutants in many fitness games, and the result no longer holds (this is reminiscent of the fact that a strategy that is ES in infinite populations is not necessarily ES in finite populations [24]).

3.4 Bringing relatedness into the picture

All of the analyses summarized above were derived in the standard panmictic setting [4]. Some of them have been extended to encompass relatedness [25, 26, 27], which arises in naturally

structured populations [28] and is part of the environment of evolutionary adaptation of the human lineage [29].

Relatedness is introduced as follows into the abstract evolutionary stability concept [30]. For any given resident preference function u and mutant preference function v in the considered set of preference functions, and a share $\varepsilon \in (0,1)$ of mutants, let $\Pr[v|u,\varepsilon]$ denote the probability that a resident is matched with a mutant, and $\Pr[v|v,\varepsilon]$ the probability that a mutant is matched with another mutant. In a panmictic population $\Pr[v|u,\varepsilon] = \Pr[v|v,\varepsilon] = \varepsilon$, which implies that as the share of mutants tends to 0, the probability that a mutant is matched with another mutant tends to 0 as well. Relatedness means that rare mutants are more likely to be matched with each other, than a resident is to be matched with a mutant. This is formalized by assuming that

$$\lim_{\varepsilon \to 0} \Pr[v|v,\varepsilon] = r \tag{33}$$

for some *relatedness* $r \in [0,1]$. The analyses summarized above correspond to the special case $r = 0$.

Remark 3. *In the models included in this overview the matching probabilities are exogenously given. Put differently, there is no partner choice.*

3.4.1 Interactions under complete information

Starting with interactions under complete information and preference functions of the form (18), Definition 3 readily generalizes to encompass relatedness by replacing (15) by:

$$\Pr[u|u,\varepsilon]\cdot w(x^*,x^*) + \Pr[v|u,\varepsilon]\cdot w(\hat{x},\hat{y}) > \Pr[u|v,\varepsilon]\cdot w(\hat{y},\hat{x}) + \Pr[v|v,\varepsilon]\cdot w(\tilde{x},\tilde{x}), \tag{34}$$

and the score function in (16) generalizes to:

$$\begin{aligned} S_{u,v}(\varepsilon) &= \Pr[u|u,\varepsilon]\cdot w(x^*,x^*) + \Pr[v|u,\varepsilon]\cdot w(\hat{x},\hat{y}) \\ &- \Pr[u|v,\varepsilon]\cdot w(\hat{y},\hat{x}) - \Pr[v|v,\varepsilon]\cdot w(\tilde{x},\tilde{x})]. \end{aligned} \tag{35}$$

Recall that differentiability of this function facilitates analysis, since it is then sufficient to examine the value (and sometimes the derivative) of $S_{u,v}$ at $\varepsilon = 0$ to establish whether u is ES

against v. Such differentiability obtains if the conditional probability functions are differentiable. Positing such differentiability, Result 3 generalizes to:

Result 7. *[Alger and Weibull [11]] In a population where the matching process entails relatedness $r \in [0,1]$, a preference function of the form* (18) *with $\alpha = \alpha^*$ is an ESPFC only if:*

1. *$\alpha^* < r$ if the strategies are strategic substitutes (i.e., $\partial^2 w(x,y)(\partial x \partial y) < 0$).*
2. *$\alpha^* > r$ if the strategies are strategic complements (i.e., $\partial^2 w(x,y)(\partial x \partial y) > 0$).*
3. *$\alpha^* = r$ if the strategies are strategically neutral (i.e., $\partial^2 w(x,y)(\partial x \partial y) = 0$).*

Given that the value of α determines how willing individuals are to act generously in interactions with relatives, this result suggests that evolution may have led to variation in the degree of intra-family generosity across different regions of the world, the ultimate driving force being the qualitative nature of the fitness game. Furthermore, even for a given category of fitness game (i.e., where strategies are strategic substitutes or complements), the specifics of the fitness game may matter. To wit, in the production-and-sharing fitness game studied in [31], where the strategies are strategic substitutes, the evolutionarily stable degree of altruism is found to be lower, the harsher is the environment.

Note that Result 7 implies that equilibrium behavior in a population with an evolutionarily stable degree of altruism would be at odds with the strategy predicted by Hamilton's rule [26, 27] under strategy evolution [32]. However, the result is in line with Hamilton's rule once it is brought up to the selection of preference functions rather than strategies (see the discussion in [11]).

3.4.2 Interactions under incomplete information

Turning now to interactions under incomplete information and relatedness, a straightforward generalization of the panmictic setting examined above is sufficient. Inserting the conditional probabilities into the system of best-response equations (29) in Definition 4,

$$\begin{cases} x^* \in \arg\max_{x \in X} & \Pr[u|u,\varepsilon] \cdot u(x,x^*) + \Pr[v|u,\varepsilon] \cdot u(x,y^*) \\ y^* \in \arg\max_{y \in X} & \Pr[u|v,\varepsilon] \cdot v(y,x^*) + \Pr[v|v,\varepsilon] \cdot v(y,y^*), \end{cases} \tag{36}$$

and into the equilibrium fitnesses of residents and mutants (see (30) and (31)),

$$W_u(x^*, y^*, \varepsilon) = \Pr[u|u, \varepsilon] \cdot w(x^*, x^*) + \Pr[v|u, \varepsilon] \cdot w(x^*, y^*) \tag{37}$$

$$W_v(x^*, y^*, \varepsilon) = \Pr[u|v, \varepsilon] \cdot w(y^*, x^*) + \Pr[v|v, \varepsilon] \cdot w(y^*, y^*), \tag{38}$$

the definition of an evolutionarily stable preference function under incomplete information applies as is (see Definition 5). The simple fitness-maximizing preference function (see (5)) is no longer evolutionarily stable, however. Instead, the analysis in [22] (see also [23] for a generalization to n-player interactions), reveals that evolution favors *Homo moralis* preferences:

Definition 7. *An individual is a **Homo moralis** if her preference function is of the form*

$$u_\kappa(x, y) = (1 - \kappa) \cdot w(x, y) + \kappa \cdot w(x, x), \tag{39}$$

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

While it was the mathematical analysis that led to the "discovery" of this preference class, the choice of the name *Homo moralis* was triggered by the fact that the second term in (39) can be interpreted as a concern for universalization, reminiscent of Kant's reasoning [33]: what would happen (to the individual's fitness) if the individual's strategy was universalized? The first term being the individual's fitness given own and opponent's actual strategies, the *Homo moralis* preference function can be thought of as representing a form of partial Kantian moral concern (see also [32] for a similar "as if" interpretation, in a model with strategy evolution for interactions between siblings).

The following definition and result generalize Definition 6 and Result 6 to encompass relatedness.

Definition 8. *Let X_r be the set of type-homogenous Nash equilibria in a population consisting solely of Homo moralis with degree of morality $\kappa = r$. A preference function u' is a **behavioral alike** to such Homo moralis if there exists some $x_r \in X_r$ such that $y \in \arg\max_{x \in X} u'(x, x_r)$ and $x \in \arg\max_{x \in X}(1 - r) \cdot w(x, x_r) + r \cdot w(x, x)$.*

A behavioral alike to a *Homo moralis* with a degree of morality $\kappa = r$ is a preference type that would be willing to play a strategy that such a *Homo moralis* would also be willing to play,

given that the opponent plays some $x_r \in X_r$. Modulo the slight difference in the definition of behavioral alikes already referred to, the following result was derived by Alger and Weibull [22, 23].

Result 8. *If the strategy set X is compact and convex, all the preference functions in Θ as well as the fitness function w are continuous, and the conditional probability functions $\Pr[u|u,\varepsilon]$ and $\Pr[u|v,\varepsilon]$ are continuous in ε, then the preference function*

$$u_r(x,y) = (1-r)\cdot w(x,y) + r\cdot w(x,x), \tag{40}$$

is ESI against any preference function that is not its behavioral alike, in a population where the matching process entails relatedness $r \in [0,1]$.

A population of *Homo moralis* resists entry by mutants because their preferences make them select a strategy that maximizes the average fitness of vanishingly rare mutants. Indeed, such mutants, who play some strategy, say z, obtain average fitness which, given the topological properties (see the discussion following Result 6), is arbitrarily close to (see (38))

$$(1-r)\cdot w(z,x_r) + r\cdot w(z,z), \tag{41}$$

where x_r is some equilibrium strategy in a monomorphic population consisting of *Homo moralis*:

$$x_r \in \arg\max_{x\in X} u_r(x,x_r), \tag{42}$$

where

$$u_r(x,x_r) = (1-r)\cdot w(x,x_r) + r\cdot w(x,x). \tag{43}$$

A mutant preference type that is not a behavioral alike to *Homo moralis* with degree of morality $\kappa = r$ must play some strategy z which does not belong to the set $\arg\max_{x\in X}(1-r)\cdot w(x,x_r)+ r\cdot w(x,x)$, and hence obtains an average fitness that is strictly smaller than that of residents, which is is arbitrarily close to $w(x_r,x_r)$.

Interestingly, while altruistic preferences (see (18)) with degree of altruism $\alpha = r$ are sometimes behavioral alikes to *Homo moralis* preferences with degree of morality $\kappa = r$, only the latter preference function is evolutionarily stable for the entire set of fitness games defined by

the assumptions stated in Result 8 (see the discussions in [32], [22], and [34]).

4 Discussion

The first contributions to the literature on the evolution of preferences by natural selection extended the concept of evolutionary stability from the level of strategies [1] to the level of preferences guiding the choice of strategy, an approach that is sometimes referred to as *indirect evolution* [15], since evolution then operates on strategies only indirectly, by "delegating" the strategy choice to the individual. Several novel insights were delivered by these contributions, as summarized above. This literature is arguably still in its infancy, and I here discuss some possible future paths.

To begin, some readers may wonder: *is there really a deep difference between strategy evolution and preference evolution?* After all, and as highlighted in this article, it is typically possible to reformulate preference evolution as evolution of response rules. I'd make the case that there is a fundamental difference, however. Strategies are mere descriptions of behavior. Preferences are expressed within individuals as the result of some process which may involve reasoning, emotions, hormones, and/or other neurobiological mechanisms, and which may respond to the stimuli and information the individual receives. Some preference classes present the advantage of lending themselves to psychological interpretation. For example, one possible interpretation of an individual with altruistic preferences of the form (18) with a positive degree of altruism $\alpha > 0$ is that (s)he has emotions that are swayed by the fitness of the person with whom (s)he interacts: the better off is the opponent, the happier (s)he gets. By contrast, an individual *Homo moralis* preferences of the form (39) with a positive degree of morality $\kappa > 0$ would not react to information about the opponent's fitness: (s)he instead evaluates different courses of action by taking into account what own fitness would be if—hypothetically—the course of action was universalized to all the interactants. This observation suggests three possible future research paths.

First, over the past few decades the behavioral economics literature has proposed a wealth of preference classes to explain observed behaviors in social interactions: altruism [18], warm glow [35], a preference for conformity [36], for reciprocity [37, 38, 39, 40], inequity aversion [41, 42], guilt aversion [43, 44], and image concerns [45, 46]. These preference classes were

inspired mostly by research in psychology and sociology. It should be noted that *Homo moralis* preferences [22], examined above, are novel to behavioral economics: the theory of preference evolution may thus contribute to economics through the discovery of hitherto unstudied preference classes, and future analyses may make further similar discoveries.

Second, the theory of preference evolution may unveil ultimate drivers of the aforementioned preference classes (besides altruistic and *Homo moralis* preferences, already extensively studied). A question of particular interest is whether there may be stable polymorphisms—populations in which several preference classes co-exist—and if so, which factors are expected to affect the stable distribution of preferences. Such theories may help explain observed heterogeneity both within and between populations in survey and experimental data [47, 48, 49].

Third, researchers working with models of preference evolution must make assumptions on the set of potential preference functions. In reality, however, the set of potential preference functions available for a given organism may be determined by physiological constraints. An open question is thus whether findings on the neurobiology of our species would help reduce this set. Such an approach has already been used in the theoretical literature on the evolution of preference functions that govern choices in decision situations other than social interactions, see, e.g., [50, 51].

Readers may also ask: *how realistic is the process by which individuals are matched together in preference evolution models that extend the standard evolutionary game theory model?* An important question is thus whether the results found under this assumption are robust to the extension to other matching processes. Two nascent paths can be mentioned in this context.

First, the model of preference under incomplete information found in [22, 23] has been incorporated by Alger, Weibull, and Lehmann [52] into a standard island model [53], in which the population is structured into groups between which there is limited migration. This approach allowed the researchers to distinguish between preference functions defined over fitness on the one hand and preference functions defined over material payoffs on the other hand. Arguably, the preference function defined over material payoffs that is found to be uninvadable in [52] is more relevant for social scientists who seek to estimate the preferences of individuals by way of observing their behavioral responses to trivial material payoff consequences, such as in the experimental economics literature [38, 54, 55, 56, 57]. This function differs from the *Homo moralis* function, whose evolutionary viability was established under the standard evolutionary

game theory matching protocol. However, the *Homo moralis* preference function is uninvadable when defined over fitnesses rather than material payoffs, thus providing one first robustness test. It remains to be seen which preference functions—or distributions over preference functions—would resist the invasion of mutants in models with more sophisticated modeling of the migration decisions, such as in [58], for example.

Second, individuals are typically free to choose with whom they interact. Such active partner choice is known to matter for the evolution of cooperative strategies [59]. How would it affect the evolution of preference functions? One possible formalization is provided by Hopkins [60], in a model with altruistic preference functions where individuals differ in their ability to understand the mental processes of others.

Readers may further ask: *if preferences emanate from mental and neurobiological processes, is it reasonable to assume that one can observe others' preferences?* In the model proposed by Heller and Mohlin [61] this issue—reminiscent of the well-known "mimicry" issue in biology—is addressed by examining the co-evolution of preferences and the ability to deceive others about preferences and intentions. The extensive work on the commitment role that emotions may have played in our evolutionary past, and the concomitant ability to signal (e.g., through anger) and also detect such emotions (see, e.g., [62], may perhaps also inspire formal work on emotions that can be incorporated into the theory of preference evolution.

The definition of an evolutionarily stable strategy [1] provided a key tool for theorists to model ultimate drivers of behavior in social interactions. Adding the idea that Nature delegates the strategy choice to the individuals by way of equipping them with preferences over strategies [2, 3], arguably brings the theory closer to reality. Although the literature has already delivered many insights, most of the work on evolutionarily viable preferences undoubtedly still lays ahead of us. I hope that this article has underlined the fundamental role played by the bridges built between the models of biologists and economists, both in the past and in the future.

References

[1] J. Smith and G. R. Price, "The logic of animal conflict," Nature, vol. 246, no. 5427, pp. 15–18, 1973.

[2] R. H. Frank, "If Homo economicus could choose his own utility function, would he want one with a conscience?," American Economic Review, vol. 77, no. 4, pp. 593–604, 1987.

[3] W. Güth and M. Yaari, Explaining reciprocal behavior in simple strategic games: an evolutionary approach, pp. 22–34. Ann Arbor, MI: University of Michigan Press, 1992.

[4] J. M. Smith, Evolution and the Theory of Games. Cambridge: Cambridge University Press, 1982.

[5] J. M. McNamara and O. Leimar, Game theory in biology: concepts and frontiers. Oxford University Press, USA, 2020.

[6] D. Fudenberg and J. Tirole, Game Theory. Cambridge MA: MIT Press, 1991.

[7] J. W. Weibull, Evolutionary game theory. Cambridge MA: MIT Press, 1997.

[8] I. M. Bomze and B. M. Pötscher, Game theoretical foundations of evolutionary stability. New York: Springer-Verlag, 1988.

[9] A. Mas-Colell, M. D. Whinston, and J. R. Green, Microeconomic Theory. Oxford: Oxford University Press, 1995.

[10] R. Aumann and A. Brandenburger, "Epistemic conditions for nash equilibrium," Econometrica, vol. 63, no. 5, pp. 1161–1180, 1995.

[11] I. Alger and J. W. Weibull, "A generalization of Hamilton's rule—love others how much?," Journal of Theoretical Biology, vol. 299, pp. 42–54, 2012.

[12] A. Heifetz, C. Shannon, and Y. Spiegel, "What to maximize if you must," Journal of Economic Theory, vol. 133, no. 1, pp. 31–57, 2007.

[13] E. A. Ok and F. Vega-Redondo, "On the evolution of individualistic preferences: An incomplete information scenario," Journal of Economic Theory, vol. 97, no. 2, pp. 231–254, 2001.

[14] E. Dekel, J. C. Ely, and O. Yilankaya, "Evolution of preferences," Review of Economic Studies, vol. 74, no. 3, pp. 685–704, 2007.

[15] H. Bester and W. Güth, "Is altruism evolutionarily stable?," Journal of Economic Behavior & Organization, vol. 34, no. 2, pp. 193–209, 1998.

[16] F. Bolle, "Is altruism evolutionarily stable? and envy and malevolence?: Remarks on Bester and Güth," Journal of Economic Behavior Organization, vol. 42, no. 1, pp. 131–133, 2000.

[17] A. Possajennikov, "On the evolutionary stability of altruistic and spiteful preferences," Journal of Economic Behavior Organization, vol. 42, no. 1, pp. 125–129, 2000.

[18] G. S. Becker, "A theory of social interactions," Journal of Political Economy, vol. 82, no. 6, pp. 1063–1093, 1974.

[19] J. M. McNamara, C. E. Gasson, and A. I. Houston, "Incorporating rules for responding into evolutionary games," Nature, vol. 401, no. 6751, pp. 368–371, 1999.

[20] P. D. Taylor and T. Day, "Stability in negotiation games and the emergence of cooperation," Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 271, no. 1540, pp. 669–674, 2004.

[21] E. Akçay, J. Van Cleve, M. W. Feldman, and J. Roughgarden, "A theory for the evolution of other-regard integrating proximate and ultimate perspectives," Proceedings of the National Academy of Sciences, vol. 106, no. 45, pp. 19061–19066, 2009.

[22] I. Alger and J. W. Weibull, "Homo moralis—preference evolution under incomplete information and assortative matching," Econometrica, vol. 81, no. 6, pp. 2269–2302, 2013.

[23] I. Alger and J. W. Weibull, "Evolution and Kantian morality," Games and Economic Behavior, vol. 98, pp. 56–67, 2016.

[24] M. E. Schaffer, "Evolutionarily stable strategies for a finite population and a variable contest size," Journal of Theoretical Biology, vol. 132, no. 4, pp. 469–478, 1988.

[25] S. Wright, "Coefficients of inbreeding and relationship," American Naturalist, vol. 56, pp. 330–338, 1922.

[26] W. Hamilton, "The genetical evolution of social behaviour. i," Journal of Theoretical Biology, vol. 7, no. 1, pp. 1–16, 1964.

[27] W. Hamilton, "The genetical evolution of social behaviour. ii," Journal of Theoretical Biology, vol. 7, no. 1, pp. 17–52, 1964.

[28] F. Rousset, Genetic Structure and Selection in Subdivided Populations. Princeton: Princeton University Press, 2004.

[29] C. P. Van Schaik, The Primate Origin of Human Behavior. Hoboken, NJ: Wiley-Blackwell, 2016.

[30] T. C. Bergstrom, "The algebra of assortative encounters and the evolution of cooperation," International Game Theory Review, vol. 05, no. 03, pp. 211–228, 2003.

[31] I. Alger and J. W. Weibull, "Kinship, incentives, and evolution," American Economic Review, vol. 100, no. 4, pp. 1725–1758, 2010.

[32] T. C. Bergstrom, "On the evolution of altruistic ethical rules for siblings," American Economic Review, vol. 85, no. 1, pp. 58–81, 1995.

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

[34] I. Alger and J. W. Weibull, "Strategic behavior of moralists and altruists," Games, vol. 8, no. 3, 2017.

[35] J. Andreoni, "Impure altruism and donations to public goods: A theory of warm-glow giving," Economic Journal, vol. 100, no. 401, pp. 464–477, 1990.

[36] B. D. Bernheim, "A Theory of Conformity," Journal of Political Economy, vol. 102, no. 5, pp. 841–877, 1994.

[37] M. Rabin, "Incorporating fairness into game theory and economics," The American Economic Review, pp. 1281–1302, 1993.

[38] G. Charness and M. Rabin, "Understanding social preferences with simple tests," Quarterly Journal of Economics, vol. 117, no. 3, pp. 817–869, 2002.

[39] M. Dufwenberg and G. Kirchsteiger, "A theory of sequential reciprocity," Games and Economic behavior, vol. 47, no. 2, pp. 268–298, 2004.

[40] A. Falk and U. Fischbacher, "A theory of reciprocity," Games and Economic Behavior, vol. 54, no. 2, pp. 293–315, 2006.

[41] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and cooperation," Quarterly Journal of Economics, vol. 114, no. 3, pp. 817–868, 1999.

[42] G. E. Bolton and A. Ockenfels, "Erc: A theory of equity, reciprocity, and competition," American Economic Review, vol. 90, no. 1, pp. 166–193, 2000.

[43] G. Charness and M. Dufwenberg, "Promises and partnership," Econometrica, vol. 74, no. 6, pp. 1579–1601, 2006.

[44] P. Battigalli and M. Dufwenberg, "Guilt in games," American Economic Review, vol. 97, no. 2, pp. 170–176, 2007.

[45] R. Bénabou and J. Tirole, "Incentives and prosocial behavior," American Economic Review, vol. 96, no. 5, pp. 1652–1678, 2006.

[46] T. Ellingsen and M. Johannesson, "Pride and prejudice: The human side of incentive theory," American Economic Review, vol. 98, no. 3, pp. 990–1008, 2008.

[47] A. Falk, A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde, "Global evidence on economic preferences," Quarterly Journal of Economics, vol. 133, no. 4, pp. 1645–1692, 2018.

[48] S. Nunnari and M. Pozzi, "Meta-analysis of inequality aversion estimates," mimeo, 2022.

[49] R. Croson and U. Gneezy, "Gender differences in preferences," Journal of Economic Literature, vol. 47, no. 2, pp. 448–474, 2009.

[50] A. Robson and L. Samuelson, "The evolution of decision and experienced utilities," Theoretical Economics, vol. 6, no. 3, pp. 311–339, 2011.

[51] N. Robalino and A. J. Robson, "The biological foundations of economic preferences," Oxford Research Encyclopedia of Economics and Finance, 2019.

[52] I. Alger, J. W. Weibull, and L. Lehmann, "Evolution of preferences in structured populations: genes, guns, and culture," Journal of Economic Theory, vol. 185, p. 104951, 2020.

[53] S. Wright, "Evolution in mendelian populations," Genetics, vol. 16, pp. 97–159, 1931.

[54] B. R. Fisman, S. Kariv, and D. Markovits, "Individual preferences for giving," American Economic Review, vol. 97, no. 5, pp. 1858–1876, 2007.

[55] M. Blanco, D. Engelmann, and H. T. Normann, "A within-subject analysis of other-regarding preferences," Games and Economic Behavior, vol. 72, no. 2, pp. 321–338, 2011.

[56] A. Bruhin, E. Fehr, and D. Schunk, "The many faces of human sociality: Uncovering the distribution and stability of social preferences," Journal of the European Economic Association, vol. 17, no. 4, pp. 1025–1069, 2019.

[57] T. Miettinen, M. Kosfeld, E. Fehr, and J. W. Weibull, "Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions," Journal of Economic Behavior and Organization, vol. 173, pp. 1–25, 2020.

[58] C. Mullon, L. Keller, and L. Lehmann, "Evolutionary stability of jointly evolving traits in subdivided populations," American Naturalist, vol. 188, no. 2, pp. 175–195, 2016.

[59] J. M. McNamara, Z. Barta, L. Fromhage, and A. Houston, "The coevolution of choosiness and cooperation," Nature, vol. 451, no. 7175, pp. 189–192, 2008.

[60] E. Hopkins, "Competitive altruism, mentalizing and signalling," American Economic Journal MIcroeconomics, vol. 6, pp. 272–292, 2014.

[61] Y. Heller and E. Mohlin, "Coevolution of deception and preferences: Darwin and nash meet machiavelli," Games and Economic Behavior, vol. 113, pp. 223–247, 2019.

[62] J. Tooby and L. Cosmides, The evolutionary psychology of the emotions and their relationship to internal regulatory variables, pp. 114–137. The Guilford Press, 2008.