Evolutionarily stable preferences

Ingela Alger

Evolutionarily stable preferences*

Ingela ALGER†

December 8, 2022

Abstract

The 50-year old concept of an evolutionarily stable strategy (ESS) 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—a proximate mechanism for behavior—and the distribution of preferences in the population was taken to be fixed and given. 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: individuals inherit their preferences, the preferences determine their strategy choices, which in turn determines evolutionary success. One 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

*I thank Jussi Lehtonen and Xiang-Yi Li for inviting me to contribute to this special issue. I also thank Piret Avila, Pau Juan Bartroli, Péter Bayer, Konrad Dierks, Maria Kleshnina, Laurent Lehmann, Enrico Mattia Salonia, Esteban Muñoz, Jorge Peña, Jörgen Weibull, and two anonymous referees for helpful comments. I 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

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 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 the latter 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 has 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, (s)he 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 prefers to alter his/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, notably McNamara, Gasson, and Houston [4], Taylor and Day [5], Akçay et al. [6], and Hamilton [7, 8].

2 Strategy evolution in biology

2.1 Framework and definition of ESS

Throughout I adopt the following assumptions, which are in line with the standard evolutionary game theory model [9]. First, 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" ([10] p.10, see also the recent book by McNamara and Leimar [11]); this is also in line with standard vocabulary in non-cooperative game theory, see [12]). Second,

I examine a continuum population, in which individuals are randomly matched into pairs to interact, i.e., there is no partner choice. Third, I restrict attention to interactions in which both individuals have access to the same set of strategies, called X, and the material consequences of strategy choices are identical for all individuals in the population. Letting $w(x, y)$ denote the (personal) *fitness* of an individual using strategy x when the other is using strategy y, I will refer to $\Gamma = \langle X, w \rangle$ as the *fitness game*.[1] The question at hand is whether some *resident strategy* x, present in a share $1 - \varepsilon$ of the population, would resist the invasion of some *mutant strategy* y, present in a small share $\varepsilon > 0$ of the population. An evolutionarily stable strategy is then formally defined as follows [9]:

Definition 1. *Consider a population in which individuals are uniformly randomly matched into pairs to interact according to the fitness game* $\Gamma = \langle X, w \rangle$*. 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* $\varepsilon \in (0, \bar{\varepsilon}_y)$*:*

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

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

The population being infinitely large and the interactants being matched in a uniformly random manner, any individual is matched with a resident (who plays x) with probability $1 - \varepsilon$ and with a mutant (who plays y) with probability ε. In (1) the left-hand side is thus 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. 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.

Remark 1. *The present model disregards a number of features that are present in many rich mathematical models in evolutionary biology, in which: the process by which individuals are matched to interact depends on the life-cycle and the population structure; there are individuals of different classes (e.g., men and women, old and young, etc.); there are explicit computations*

[1] Personal fitness measures the individual's reproductive success, i.e., the number of offspring. Note that if the transmission is cultural, then these are not the biological but the cultural offspring

of the expected number of descendants of a single initial mutant; etc. For a recent such general model, see Lehmann and Rousset [13]. The model in Definition 1 is closely related to the special case of theirs, in which there is full dispersal, haploidy, and weak selection, the latter assumption ensuring that the probability of being matched with a mutant does not depend on the mutant strategy itself. In this case the invasion fitness (see their equation (2)) reduces to $w(y,x)/w(x,x)$*, and the condition for* x *to be uninvadable (see their equation (1)) is equivalent to* $w(y,x) \leq w(x,x)$*, a condition which below will be seen to be necessary for* x *to be ESS.*[2] *This bare-bones model presents the advantage of making it easy to derive insights on preference evolution and understand the challenges that its analysis presents.*

While much of the literature focuses on one-shot simultaneous-move games with a finite number of strategies, such as the Prisoner's dilemma or the Hawk-Dove game, the setting in fact encompasses a large number of other kinds interactions, for instance those with an infinite number of pure strategies, and/or where the interactants play in a sequential manner. Several such games are described in detail in Section 1 of the online appendix.[3] While several general results will be presented, I will often refer to the following two examples.

Example 1. *[Fitness game* Γ_1*] This is a (simultaneous-move one-shot) public goods game with strategy set* $X = \mathbb{R}_+$ *and fitness function*

$$w(x,y) = (m + ky)\, x - x^2. \tag{2}$$

The first term is the benefit and the second term the cost from contributing x*. The parameter* $m > 0$ *measures the baseline marginal benefit from the contribution. The parameter* $k \in (-1,1)$ *measures the effect that the other individual's contribution* y *has on the marginal benefit from the contribution: if* $k \in (-1,0)$*,* y *reduces the marginal benefit by* ky*; if* $k \in (0,1)$ *it increases the marginal benefit by* ky*; finally, if* $k = 0$ *it has no effect on the marginal benefit.*

Note that the parameter k, which determines how the other's contribution affects the individual's marginal benefit from contributing, is nothing but the cross-partial derivative of w:

[2] McNamara and Leimar [11] also view $w(y,x)$ as a proxy of invasion fitness.

[3] 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 [12]); due to the complex notation required to rigorously define repeated games and to space restrictions, however, I will not explicitly study these games here.

$\partial^2 w(x,y)/(\partial x \partial y) = k$. The following definition proposes a general classification of fitness games depending on the sign of this cross-partial derivative.[4]

Definition 2. *In a given fitness game $\Gamma = \langle X, w \rangle$, the strategies are:*

1. *strategic substitutes if $\partial^2 w(x,y)/(\partial x \partial y) < 0$ for all $(x,y) \in X^2$;*
2. *strategic complements if $\partial^2 w(x,y)/(\partial x \partial y) > 0$ for all $(x,y) \in X^2$;*
3. *strategically neutral if $\partial^2 w(x,y)/(\partial x \partial y) = 0$ for all $(x,y) \in X^2$.*

Hence, in Example 1 the strategies are strategic substitutes if $k \in (-1, 0)$, strategically neutral if $k = 0$, and strategic complements if $k \in (0, 1)$. Instances of interactions involving strategic substitutes are those where individuals compete over the same resources, such as common-pool resource games, or where individuals benefit from free-riding on each other, such as commonly studied public goods games. By contrast, strategic complementarities are present in interactions where it pays off to coordinate, i.e., where teamwork is valuable. This leads us to Example 2.

Example 2. *[Fitness game Γ_2] This is the (simultaneous-move one-shot) Stag-Hunt game [12], in which each individual has two actions—Stag (S) and Hare (H), and with payoffs as shown in Figure 1. This game captures interactions in which H ("going for the Hare" in the original story) gives a certain payoff of 1 but coordination by both players on S ("going for the Stag") leads to a higher payoff $R > 1$. A (mixed) strategy is the probability of playing S, with H being played with the complementary probability. Hence, the strategy set is $X = [0, 1]$ and the fitness function is*

$$w(x,y) = xyR + 1 - x. \tag{3}$$

While the strategy labels stem from the original story of two hunters, this fitness game can be used to represent a host of interactions where coordination on one strategy enhances the fitness values of both individuals.

[4]Note that fitness games for which $\partial^2 w(x,y)/(\partial x \partial y)$ does not have the same sign for all $(x,y) \in X^2$ fall in neither category.

	S	H
S	R, R	$0, 1$
H	$1, 0$	$1, 1$

Figure 1: The payoff matrix of the simultaneous-move Stag-Hunt game ($R > 1$).

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 fitness earned by residents and that earned by mutants as a function of the share of mutants ε, using what is called the *score function* [14]:

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

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.[5] 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.

As an illustration, in Example 1 the unique ESS is $x^{ESS} = m/(2-k)$. This is found by

[5]Note that this is different from the selection gradient, which is the derivative with respect to the trait value; the slope of the score function instead indicates whether the difference between the average fitness earned by residents and that earned by mutants increases or decreases as the share of mutants ε increases. The selection gradient (for fitness game Γ_1) will be found in equation (6) below.

noting that any ESS x^{ESS} must solve

$$x^{ESS} \in \arg\max_{y \in X} w(y, x^{ESS}). \tag{5}$$

The first-order condition for such a maximum is

$$\frac{\partial w(y, x^{ESS})}{\partial y}_{|y=x^{ESS}} = \big(m + kx^{ESS} - 2y\big)_{|y=x^{ESS}} = 0. \tag{6}$$

This equation has as unique solution $x^{ESS} = m/(2-k)$. Since the second-order derivative with respect to y is strictly negative, it follows that $w(x^{ESS}, x^{ESS}) > w(y, x^{ESS})$ for any $y \neq x^{ESS}$. Result 1 then implies that x^{ESS} is ES against any $y \in X$, $y \neq x^{ESS}$, and Definition 1 in turn implies that x^{ESS} is an ESS. In Example 2, there are two ESS: play S with probability 1, and play H with probability 1.

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., [15]). Such a function is called a *utility function* in economics. Here I will instead refer to it as a *preference function*. 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 personal fitness drives evolutionary success, should we expect evolution to favor the preference function which induces the individual to maximize personal fitness?

Definition 3. *In any given fitness game $\Gamma = \langle X, w \rangle$,* ***a (personal) fitness-maximizer*** *has a preference function that coincides with the personal fitness function, i.e.,*

$$u(x, y) = w(x, y) \quad \textit{for all strategies } x, y \in X. \tag{7}$$

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 is encoded to prefer. 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 common-pool resource game, formalized as fitness game Γ_1 with parameter values $m = 10$ and $k = -1/2$ (see (2)). Having in mind situations such as fishing in a common lake, I will use "extraction effort" or simply "effort" to refer to a strategy. The negative value of k means that an individual's effort has a negative impact on the other's marginal return to effort, the idea being that this effort diminishes the size of the fish population available to the other.

Suppose first that all individuals in the population are fitness-maximizers, i.e., each individual has preferences $u : X^2 \to \mathbb{R}$ defined in (7), namely $u(x, y) = (m + ky)\, x - x^2$, over own and other's efforts, 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* ([12]). In such an interaction, what pair of strategies will be played? While this question has no simple general answer (see, e.g., [16]), it is common in economics to apply the *Nash equilibrium concept*.

Definition 4. *In an interaction between two individuals with preferences $u : X^2 \to \mathbb{R}$ and $v : X^2 \to \mathbb{R}$, respectively, 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:*

$$\begin{cases} x^* \in \arg\max_{x \in X} \; u\,(x, y^*) \\ y^* \in \arg\max_{y \in X} \; v\,(y, x^*)\,. \end{cases} \tag{8}$$

In the example, in an interaction between two fitness-maximizers, (x^*, y^*) is a Nash equilibrium if and only if:

$$\begin{cases} x^* \in \arg\max_{x \in X} \; (10 - y^*/2)\, x - x^2 \\ y^* \in \arg\max_{y \in X} \; (10 - x^*/2)\, y - y^2. \end{cases} \tag{9}$$

Differentiability of the preference function together with an unbounded strategy set implies that (x^*, y^*) must satisfy the following set of first-order conditions for x^* and y^* to be maxima (these

conditions are also sufficient because of the strict concavity of the preference function):

$$\begin{cases} 10 - y^*/2 = 2x^* \\ 10 - x^*/2 = 2y^*. \end{cases} \tag{10}$$

It follows that the unique Nash equilibrium strategy profile between two fitness-maximizers is $(x^*, y^*) = (4, 4)$ (note that this is also the ESS $x^{ESS} = m/(2-k)$). Suppose now that another preference function enters this population. For example suppose that some individuals have the preference function

$$v(x, y) = w(x, y) - \frac{1}{2} w(y, x). \tag{11}$$

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., [17]). 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:

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ (10 - \hat{y}/2)\, x - x^2 \\ \hat{y} \in \arg\max_{y \in X} \ (10 - \hat{x}/2)\, y - y^2 - \frac{1}{2}[(10 - y/2)\, \hat{x} - \hat{x}^2]. \end{cases} \tag{12}$$

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

$$\begin{cases} 10 - \hat{y}/2 = 2\hat{x} \\ 10 - \hat{x}/2 + \hat{x}/4 = 2\hat{y}. \end{cases} \tag{13}$$

It follows that there is a unique Nash equilibrium strategy profile $(\hat{x}, \hat{y}) = (120/31, 140/31) \approx (3.87, 4.52)$. Note that $\hat{y} > x^* > \hat{x}$. By attaching a negative weight to the other's benefit, the spiteful individual extracts more from the common-pool resource than a fitness-maximizer; in turn, compared to when he interacts with another fitness-maximizer and makes extraction

effort $x^* = 4$, here the fitness-maximizer compensates for the higher extraction effort by his opponent by decreasing his extraction effort, to $\hat{x} \approx 3.87$, because the marginal benefit from extracting is reduced when the opponent is extracting more. 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

$$(1-\varepsilon)\cdot w\left(x^*,x^*\right)+\varepsilon\cdot w\left(\hat{x},\hat{y}\right) \tag{14}$$

while the average fitness of the mutant spiteful individuals is

$$(1-\varepsilon)\cdot w\left(\hat{y},\hat{x}\right)+\varepsilon\cdot w\left(\tilde{x},\tilde{x}\right), \tag{15}$$

where $(\tilde{x},\tilde{x}) = (40/9, 40/9) \approx (4.44, 4.44)$ is the unique Nash equilibrium strategy profile in a match between two spiteful individuals. Since $\frac{140\cdot 110}{31^2} = w(\hat{y},\hat{x}) > w\left(x^*,x^*\right) = 16$, it follows 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.

The conclusion that a population of fitness-maximizers would not resist the invasion by some other preference function, here reached in a simple example, has been shown by Heifetz, Shannon, and Spiegel [18] 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 [19]). They show this general result in a model which encompasses any preference function of the form

$$u\left(x,y\right) = w(x,y) + B(x,y,\tau), \tag{16}$$

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 then leads to the question: which preference function, if any, is evolutionarily stable?

Definition 5. *Consider a population in which individuals are uniformly matched into pairs to interact according to the fitness game $\Gamma = \langle X, w\rangle$ under complete information about each other's preferences. Let Θ denote the set of all possible preference functions $u : X^2 \to \mathbb{R}$;*

*any $u \in \Theta$ is such that there exists a unique Nash equilibrium in each matched pair. Then, a preference function $u \in \Theta$ is **evolutionarily stable under complete information (ESC)** against preference function $v \in \Theta$ if there exists $\bar{\varepsilon}_v \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon}_v)$:*

$$(1-\varepsilon) \cdot w\left(x^*, x^*\right) + \varepsilon \cdot w\left(\hat{x}, \hat{y}\right) > (1-\varepsilon) \cdot w\left(\hat{y}, \hat{x}\right) + \varepsilon \cdot w\left(\tilde{x}, \tilde{x}\right), \tag{17}$$

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.[6] It should 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 [20], 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:

$$S_{u,v}(\varepsilon) \;=\; (1-\varepsilon) \cdot \left[w\left(x^*, x^*\right) - w\left(\hat{y}, \hat{x}\right)\right] + \varepsilon \cdot \left[w\left(\hat{x}, \hat{y}\right) - w\left(\tilde{x}, \tilde{x}\right)\right]. \tag{18}$$

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.*

[6] 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.

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 examined the parametric class of preferences originally proposed by Bester and Güth in their seminal paper [21]. In a model with the fitness function given in (2), they examine preference functions of the form

$$u_\alpha(x, y) = w(x, y) + \alpha \cdot w(y, x), \tag{19}$$

where $\alpha \in [0, 1]$ is the evolving trait. Bolle [22] and Possajennikov [23] generalized 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/her strategy choice on the opponent. If $\alpha > 0$, (s)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* [24]. 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 fitness game Γ_1 by Bester and Güth [21], Bolle [22], and Possajennikov [23], is that the evolutionary stability requires α to be of the same sign as k, the parameter that determines whether the strategies are strategic substitutes, strategic complements, or strategically neutral. In order to show how one identifies evolutionarily stable degrees of altruism in interactions under complete information, I turn to the model of Alger and Weibull [17], which generalized the analysis of the same class of preference functions to any fitness game with a differentiable fitness function w such that there exists a unique and differentiable Nash equilibrium in any dyadic interaction; the analysis is further restricted to fitness functions w whereby an individual's strategy has some effect on the other's fitness, i.e.,

such that $\partial w(x,y)/\partial y \neq 0$ for all $(x,y) \in X^2$ (the reason for this will be explained below).

Starting with the behavioral equilibrium in a dyad, let $(x^*(\alpha,\beta), x^*(\beta,\alpha))$ denote the Nash equilibrium strategy profile in a pair where one individual has degree of altruism α and the other has degree of altruism β. Recall that a Nash equilibrium strategy profile is such that neither individual would prefer to deviate to another strategy, given the other's strategy, i.e.:

$$\begin{cases} x^*(\alpha,\beta) \in \arg\max_{x\in X} \; w(x,y) + \alpha \cdot w(y,x) \\ x^*(\beta,\alpha) \in \arg\max_{y\in X} \; w(y,x) + \beta \cdot w(x,y). \end{cases} \tag{20}$$

Since w is differentiable, the pair of first-order conditions is:

$$\begin{cases} w_1(x^*(\alpha,\beta), x^*(\beta,\alpha)) + \alpha \cdot w_2(x^*(\beta,\alpha), x^*(\alpha,\beta)) = 0 \\ w_1(x^*(\beta,\alpha), x^*(\alpha,\beta)) + \beta \cdot w_2(x^*(\alpha,\beta), x^*(\beta,\alpha)) = 0, \end{cases} \tag{21}$$

where the index 1 (resp. 2) indicates the partial derivative with respect to the first (resp. second) argument. Note that for altruism to have an effect on equilibrium strategies, $w_2(\cdot) \neq 0$ is necessary, and this explains why the analysis is restricted to such fitness functions, as announced above.

Turning now to the evolutionary stability analysis, let α denote the degree of altruism in the resident preference function and β that in the mutant preference function. Then, Result 2 implies that for the function with degree of altruism α to be an ESPFC, it is necessary that

$$w(x^*(\alpha,\alpha), x^*(\alpha,\alpha)) \geq w(x^*(\beta,\alpha), x^*(\alpha,\beta)) \quad \forall \beta \in \mathbb{R}, \tag{22}$$

or

$$\alpha \in \arg\max_{\beta\in\mathbb{R}} w(x^*(\beta,\alpha), x^*(\alpha,\beta)). \tag{23}$$

Since w is differentiable, the necessary first-order condition is

$$[w_1(x^*(\beta,\alpha), x^*(\alpha,\beta)) \cdot x_1^*(\beta,\alpha) + w_2(x^*(\beta,\alpha), x^*(\alpha,\beta)) \cdot x_2^*(\alpha,\beta)]_{|\beta=\alpha} = 0, \tag{24}$$

where again an index 1 (resp. 2) indicates the partial derivative with respect to the first (resp. second) argument. Recalling from (21) that $w_1(x^*(\beta,\alpha), x^*(\alpha,\beta)) = -\beta \cdot w_2(x^*(\alpha,\beta), x^*(\beta,\alpha))$,

equation (24) can be rewritten

$$[-\beta \cdot w_2(x^*(\alpha,\beta), x^*(\beta,\alpha)) \cdot x_1^*(\beta,\alpha) + w_2(x^*(\beta,\alpha), x^*(\alpha,\beta)) \cdot x_2^*(\alpha,\beta)]_{|\beta=\alpha} = 0, \quad (25)$$

which reduces to the following simple equation, which in this setting is necessary for the function with degree of altruism α to be an ESPFC [17]:

$$\alpha \cdot x_1^*(\alpha,\alpha) = x_2^*(\alpha,\alpha). \quad (26)$$

In fitness game Γ_1, simple calculations (i.e., solving (21) for w defined in (2)) lead to

$$x^*(\alpha,\beta) = \frac{[2+(1+\alpha)k]m}{4-(1+\alpha)(1+\beta)k^2}. \quad (27)$$

In Section 2 of the online Appendix I reproduce the proof (found for different parameter values in [21, 22, 23]) that the unique evolutionarily stable degree of altruism is

$$\alpha^* = \frac{k}{2-k} \quad (28)$$

for any $k \in (-1,0) \cup (0,1)$, implying that α^* has the same sign as k.

Equation (26) 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 have 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 (26) 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 [17]] Let w be such that $\partial w(x,y)/\partial y \neq 0$ for all $(x,y) \in X^2$. A preference function of the form* (19) *with $\alpha = \alpha^*$ is an ESPFC (within the set of all such functions) only if $x^*(\alpha, alpha)$:*

1. $\alpha^* < 0$ *if the strategies are strategic substitutes;*

2. $\alpha^* > 0$ *if the strategies are strategic complements;*

3. $\alpha^* = 0$ *if the strategies are strategically neutral.*

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. In environments with competition over resources, or where it pays off to free-ride on the other, evolutionary stability requires spiteful preferences. The reason is that spiteful preferences induce a commitment by the individual to engage in higher extraction effort (or to free-ride), and this in turn induces a lower extraction effort (or a higher contribution) by the other; hence they can invade a population consisting of fitness-maximizers or altruistic individuals, as in the introductory example. By contrast, in environments where teamwork is valuable, evolutionary stability requires altruistic preferences. The reason is that altruistic preferences induce a commitment by the individual to engage in higher effort in the team, and this in turn induces a higher effort by the other; for a small enough degree of altruism this enables them to invade a population consisting of fitness-maximizers or spiteful individuals.

The prediction reported in Result 3 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 there is a unique ESS, denoted by x^{ESS}, and that $w(x, x)$ is increasing in x when evaluated around $x = x^{ESS}$. Then in a population where a preference function of the form* (19) *with $\alpha = \alpha^*$ is an ESPFC (within the set of all such functions) 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;*

2. $x^*(\alpha^*, \alpha^*) > x^{ESS}$ *if the strategies are strategic complements;*

3. $x^*(\alpha^*, \alpha^*) = x^{ESS}$ *if the strategies are strategically neutral.*

If the fitness game $\Gamma = \langle X, w \rangle$ is such that $w(x, x)$ is decreasing in x when evaluated at $x = x^{ESS}$, the reverse inequalities hold. Indeed, in fitness game Γ_1, one obtains

$$x^*(\alpha^*, \alpha^*) = \frac{m}{2 - (1 + \alpha^*)k}. \tag{29}$$

Recalling that $x^{ESS} = m/(2-k)$ from (6), one indeed obtains $x^*(\alpha^*, \alpha^*) > x^{ESS}$ if $k \in (0, 1)$, and $x^*(\alpha^*, \alpha^*) < x^{ESS}$ if $k \in (-1, 0)$.

3.2.1 Related models in the biology literature

Following McNamara, Gasson, and Houston [4], 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 opponent 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 [4] and Taylor and Day [5] posit the following rule for an individual playing x in response to the opponent's play of y in the previous round:

$$x = \rho - \lambda \cdot y. \tag{30}$$

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 [4, 5]. 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 (26)) is thus clear. The following remark examines in greater detail the similarities and differences between the seminal models.

Remark 2. *An interesting parallel can be drawn between the response rule in* (30) *and the model analyzed by Bester and Güth [21]. Recalling the fitness function that they posit (see* (2)*), an individual with altruistic preferences chooses a strategy x that maximizes the following expression, where y is the opponent's strategy:*

$$(m - x + ky)\, x + \alpha \cdot (m - y + kx)\, y \tag{31}$$

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

$$m - 2x + ky + \alpha \cdot ky = 0, \tag{32}$$

or

$$x = \tfrac{m}{2} + \tfrac{k(1+\alpha)}{2} y, \tag{33}$$

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 [4] and Taylor and Day [5] (see (30)*) 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 [21] evaluate fitness,*

$$\begin{cases} x^*(\alpha, \alpha') = \frac{m}{2} + \frac{k(1+\alpha)}{2} \cdot x^*(\alpha', \alpha) \\ x^*(\alpha', \alpha) = \frac{m}{2} + \frac{k(1+\alpha')}{2} \cdot x^*(\alpha, \alpha') \end{cases} \tag{34}$$

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 [4] and Taylor and Day [5] evaluate fitness,

$$\begin{cases} x^*((\rho, \lambda), (\rho', \lambda')) = \rho - \lambda \cdot x^*((\rho', \lambda'), (\rho, \lambda)) \\ x^*((\rho', \lambda'), (\rho, \lambda)) = \rho' - \lambda' \cdot x^*((\rho, \lambda), (\rho', \lambda')) \end{cases} \tag{35}$$

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 [21] on the one hand, and McNamara, Gasson, and Houston [4] and Taylor and Day [5] 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 Akçay et al. [6], 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

$$w(x, y) = y^{1/2} - x^2, \tag{36}$$

they consider preference functions of the form

$$u_\beta(x, y) = w(x, y) \cdot [w(y, x)]^\beta, \tag{37}$$

and let $\beta \geq 0$ 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. *[Akçay et al. [6]] 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* (37) *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;*
2. *$x^*(\beta^*, \beta^*) = x^{ESS}$ if the strategies are strategically neutral.*

The qualitative nature of this result is in line with that of a subset of the results found by Alger and Weibull [17] 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 3. *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. [6] (see* (37)*), and defining,*

$$\tilde{u}(x,y) = \ln w(x,y) + \beta \cdot \ln w(y,x), \tag{38}$$

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 [21] (see (19)*).*

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.

3.3 Interactions under incomplete information

In this subsection we will see that fitness-maximizers do prevail when interactions take place 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 [19, 20, 25]. This section summarizes results by closely following the modeling assumptions of Alger and Weibull [25], for a reason that will become clear below. The set Θ of potential preferences is assumed to be less restrictive than above: it is the set of all continuous functions $u : X^2 \to \mathbb{R}$.

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 6. *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*

$$\begin{cases} x^* \in \arg\max_{x \in X} & (1-\varepsilon) \cdot u\left(x, x^*\right) + \varepsilon \cdot u\left(x, y^*\right) \\ y^* \in \arg\max_{y \in X} & (1-\varepsilon) \cdot v\left(y, x^*\right) + \varepsilon \cdot v\left(y, y^*\right). \end{cases} \tag{39}$$

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^*), which realizes with probability $1 - \varepsilon$, and the value that it takes in a match with a mutant (who plays y^*), which realizes with probability ε. Type-homogeneity means that all individuals with the same preference function (or preference type) use the same strategy. For example, consider again fitness game Γ_1, and assume that the resident preference function is of the form (19) with degree of altruism α while the mutant preference function is also of the form (19) but with degree of altruism $\beta \neq \alpha$. It is straightforward to verify that there then exists a unique type-homogenous Bayesian Nash equilibrium strategy profile, which depends on the share of mutants ε as follows:

$$\begin{cases} x^*(\varepsilon) = \frac{m[2+\varepsilon(\alpha-\beta)k]}{2[2-(1+\alpha)k+\varepsilon(\alpha-\beta)k]} \\ \\ y^*(\varepsilon) = \frac{m[2-(1-\varepsilon)(\alpha-\beta)k]}{2[2-(1+\alpha)k+\varepsilon(\alpha-\beta)k]} \end{cases} \tag{40}$$

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

$$W_u\left(x^*(\varepsilon), y^*(\varepsilon), \varepsilon\right) = (1-\varepsilon) \cdot w(x^*(\varepsilon), x^*(\varepsilon)) + \varepsilon \cdot w(x^*(\varepsilon), y^*(\varepsilon)) \tag{41}$$

$$W_v\left(x^*(\varepsilon), y^*(\varepsilon), \varepsilon\right) = (1-\varepsilon) \cdot w(y^*(\varepsilon), x^*(\varepsilon)) + \varepsilon \cdot w\left(y^*(\varepsilon), y^*(\varepsilon)\right) \tag{42}$$

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 7. *[Alger and Weibull [25]] 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 $W_u(x^*(\varepsilon), y^*(\varepsilon), \varepsilon) > W_v(x^*(\varepsilon), y^*(\varepsilon), \varepsilon)$ in all Nash equilibria $(x^*(\varepsilon), y^*(\varepsilon))$ 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 $s = (u, v, \varepsilon) \in \Theta^2 \times (0, 1)$ there exists a unique equilibrium strategy profile. In such a setting the score function is:

$$\begin{aligned} S_{u,v}(\varepsilon) &= (1-\varepsilon) \cdot [w(x^*(\varepsilon), x^*(\varepsilon)) - w(y^*(\varepsilon), x^*(\varepsilon))] \\ &\quad + \varepsilon \cdot [w(x^*(\varepsilon), y^*(\varepsilon)) - w(y^*(\varepsilon), y^*(\varepsilon))]. \end{aligned} \tag{43}$$

Clearly, the score function is not necessarily linear in ε (not even in the simple fitness game Γ_1, as can be readily inferred from the expressions in (40)). What's worse, without further assumptions the score function may even be discontinuous, since the equilibrium strategy profile may vary discontinuously with ε. To see this, recall the Stag-Hunt fitness game Γ_2 with fitness function $w(x, y) = xyR + 1 - x$ (see (3)), and assume that the set of strategies is $\{0, 1/4, 1\}$, i.e., an individual can either choose the pure strategy H with certainty, the pure strategy S with certainty, or to use some randomization device that leads to S with probability $1/4$ and to H with probability $3/4$ (in the hunting story, perhaps the hunter has decided to go for the stag if and only if (s)he sees a snake on the path, an event which happens with probability $1/4$). Consider preference functions of the form

$$u(x, y) = xy(R + K) + 1 - x, \tag{44}$$

where $K \in \mathbb{R}$ can be interpreted, for example, as the joy (if $K > 0$) or the sadness (if $K < 0$) experienced when killing a stag. Suppose that $R = 2$ and that the resident preference function has $K = 2$. In a population consisting entirely of residents, there are three Nash equilibrium strategy profiles: the two pure strategy profiles $(0, 0)$ and $(1, 1)$, as well as the mixed-strategy

equilibrium profile $(1/4, 1/4)$.[7] Suppose now that a share ε of the population strongly dislike killing stags, and has the mutant preference function of the form (44) with $K = -4$. Then a BNE satisfies:

$$\begin{cases} x^* \in \arg\max_{x\in\{0,1/4,1\}} & (1-\varepsilon)\cdot(4xx^* + 1 - x) + \varepsilon\cdot(4xy^* + 1 - x) \\ y^* \in \arg\max_{y\in\{0,1/4,1\}} & (1-\varepsilon)\cdot(-2yx^* + 1 - y) + \varepsilon\cdot(-2yy^* + 1 - y). \end{cases} \tag{45}$$

Clearly, for the mutants the pure strategy H is strictly dominant, meaning that they strictly prefer it to the other two strategies, independent of the opponent's strategy. Hence, in any BNE in this population, the mutants use strategy $y^* = 0$. Hence, the residents choose some strategy such that

$$x^* \in \arg\max_{x\in\{0,1/4,1\}} \quad (1-\varepsilon)\cdot(4xx^* + 1 - x) + \varepsilon\cdot(1 - x). \tag{46}$$

This fixed-point problem admits the solution $x^* = 0$ for any $\varepsilon \in (0, 1)$ and the solution $x^* = 1$ for any $\varepsilon \in (0, 3/4]$, implying that $(x^*, y^*) = (1, 0)$ and $(x^*, y^*) = (0, 0)$ are both BNE for small values of ε. However, for any $\varepsilon > 0$, a resident is no longer indifferent between H and S when all the other residents play $x = 1/4$, since $(1-\varepsilon)4/4 < 1$. Hence $(x^*, y^*) = (1/4, 0)$ is not a BNE, and there is thus a discontinuity at $\varepsilon = 0$: while residents may use the mixed strategy $1/4$ when there are no mutants around, they no longer do so as soon as the mutant preference function at hand is present in the population.

The potential existence of multiple equilibria together with potential discontinuities in the set of BNE introduces a sharp contrast with the linearity in ε of the score functions under strategy evolution (4) and under preference evolution under complete information (18), which implied that analysis of the score function at $\varepsilon = 0$ was sufficient to check evolutionary stability (recall Results 1 and 2). In spite of these challenges, there are conditions that render general analysis possible, even for settings with multiple equilibria. In particular, several authors have shown that the fitness-maximizing preference function is evolutionarily stable.

Prior to stating one such result, there is one additional issue that needs to be addressed, however. Recall that the goal here is to minimally constrain the set of possible preference functions, Θ. In particular, Θ contains several functions that give rise to the same strategy

[7] To see this, note that an individual is indeed indifferent between the two pure strategies S and H, given that the opponent chooses S with probability $1/4$, since $(R + K)/4 = 4/4 = 1$. Such indifference is necessary and sufficient for the individual to be willing to randomize between H and S.

choices (recall Remark 3), and hence that in general there will be no preference function that is ESI against all the other preference functions (since residents must obtain a *strictly* higher average fitness than mutants for it to be evolutionarily stable, see Definition 7). The following definition of *behavioral alikes* addresses this issue.

Definition 8. *Let X_0 be the set of type-homogenous Nash equilibria in a population consisting solely of fitness-maximizers. A preference function f is a* ***behavioral alike*** *to fitness-maximizers if there exists some $x_0 \in X_0$ and some $z \in X$ such that $z \in \arg\max_{x \in X} w(x, x_0)$ and $z \in \arg\max_{x \in X} f(x, x_0)$.*

In words (and somewhat loosely) a behavioral alike to fitness-maximizers is a preference type that would be willing to play a strategy z (perhaps different from x_0) 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. This result is found in [26] (it is a slight variation of the result as stated in the earlier article [25], the difference stemming from a slight difference in how behavioral alikes are defined; the core of the result is not affected, however).

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* (7)*) 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 in ε. 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 [19] 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 [27]).

Why should evolutionary biologists take interest in Result 6? After all, the prediction is not surprising: under incomplete information fitness-maximizers prevail in a panmictic setting. I'd argue that the result is valuable because it solves a dilemma inherent to fitness games with multiple evolutionarily stable strategies. Indeed, for such fitness games analysis under the assumption that selection operates at the level of strategies delivers no clear prediction. By contrast, analysis under the assumption that selection operates at the level of preferences predicts that one particular preference function stands out as being viable from an evolutionary perspective.[8]

4 Strategy and preference evolution in the presence of relatedness

All of the analyses summarized above were derived in the panmictic setting [10], where the probability of being matched with a mutant is the same for residents and mutants, 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 a rare mutant is more likely to be matched with another mutant than a resident is to be matched with a mutant [28, 7, 8]. Relatedness arises in naturally structured populations [29] and is part of the environment of evolutionary adaptation of the human lineage [30]. It has been shown to depend on a number of factors, such as migration rates, and even the strategies present in the population (see, e.g., [13]). However, here I adopt the reduced-form approach originally proposed by Bergstrom [31] for a model of strategy evolution (the term commonly used in the economics literature is assortativity of the matching process rather than relatedness).

Definition 9. *For any given resident preference function u and mutant preference function v,*

[8]Note, however, that for such fitness games the set of Nash equilibria in a population consisting of fitness-maximizers does not necessarily coincide with the set of evolutionarily stable strategies. For example, in the Stag-Hunt game with the (convex) set of strategies $X = [0, 1]$ and parameter $R = 3$ in the fitness function (3), the set of symmetric Nash equilibrium strategies is $\{0, 1/3, 1\}$, while the set of evolutionarily stable strategies is $\{0, 1\}$. This can be understood by recalling Result 1: suppose there is strategy evolution, that the resident strategy is $x = 1/3$ and the mutant strategy is $y = 0$; then $w(x, x) = w(y, x) = 1$, but $w(x, y) = 2/3 < 1 = w(y, y)$, and hence x is not ES against y. However, any ESS is a symmetric Nash equilibrium in a game between two fitness-maximizers (see [9]).

and 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. Assuming that the conditional probability functions are continuous in ε, and that the probability that a mutant is matched with another mutant tends to some number $r \in [0,1]$ as the share of mutants tends to 0, i.e.,

$$\lim_{\varepsilon \to 0} \Pr[v|v,\varepsilon] = r, \tag{47}$$

the number r measures relatedness between interactants.

The analyses summarized above correspond to the special case $\Pr[v|u,\varepsilon] = \Pr[v|v,\varepsilon] = \varepsilon$ for all (u,v,ε), and thus $r = 0$.

Prior to examining preference evolution, it is worth noting that the definition of an evolutionarily stable strategy is readily extended to encompass relatedness (an early analysis can be found in [32]):

Definition 10. *Consider a population in which individuals are randomly matched into pairs to interact according to the fitness game $\Gamma = \langle X, w \rangle$. 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 $\varepsilon \in (0, \bar{\varepsilon}_y)$:*

$$\Pr[x|x,\varepsilon] \cdot w(x,x) + \Pr[y|x,\varepsilon] \cdot w(x,y) > \Pr[x|y,\varepsilon] \cdot w(y,x) + \Pr[y|y,\varepsilon] \cdot w(y,y). \tag{48}$$

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

Hence, in a setting with relatedness $r \in [0,1]$, Result 1 generalizes to (see [25]):

Result 7. *1. If $w(x,x) > w(y,x) + r \cdot [w(y,y) - w(y,x)]$, then x is ES against y.*

2. If $w(x,x) = w(y,x) + r \cdot [w(y,y) - w(y,x)]$, then x is ES against y only if $w(x,y) > w(y,y) + r \cdot [w(y,y) - w(y,x)]$.

3. If $w(x,x) < w(y,x) + r \cdot [w(y,y) - w(y,x)]$, then x is not ES against y.

Remark 4. *Recalling Remark 1 for the panmictic case, the model in Definition 10 can again be seen as the special case of the general model by Lehmann and Rousset [13], in which there is relatedness r, haploidy, and weak selection. In this case the invasion fitness (their equation*

(2)) reduces to $[(1-r)w(y,x)+rw(y,y)]/w(x,x)$*, and the condition for* x *to be uninvadable (their equation (1)) is equivalent to* $(1-r)w(y,x)+rw(y,y)\leq w(x,x)$*, a condition which is necessary for* x *to be ESS.*

Turning now to a summary of the results for preference evolution, the distinction between complete and incomplete information is still called for.

4.1 Interactions under complete information

Starting with interactions under complete information and examining, as above, preference functions of the form (19) whereby an individual attaches some weight $\alpha \in \mathbb{R}$ to the other's individual fitness, the definition of an evolutionarily stable preference function under complete information (see Definition 5) readily generalizes to encompass relatedness by replacing $1-\varepsilon$ and ϵ by the appropriate conditional matching probabilities in (17), to obtain:

$$\Pr[u|u,\varepsilon]\cdot w(x^*,x^*)+\Pr[v|u,\varepsilon]\cdot w(\hat{x},\hat{y})>\Pr[u|v,\varepsilon]\cdot w(\hat{y},\hat{x})+\Pr[v|v,\varepsilon]\cdot w(\tilde{x},\tilde{x})\,. \quad (49)$$

The score function in (18) thus generalizes to:

$$\begin{aligned} S_{u,v}(\varepsilon) &= \Pr[u|u,\varepsilon]\cdot w(x^*,x^*)+\Pr[v|u,\varepsilon]\cdot w(\hat{x},\hat{y}) \\ &- \Pr[u|v,\varepsilon]\cdot w(\hat{y},\hat{x})-\Pr[v|v,\varepsilon]\cdot w(\tilde{x},\tilde{x})\,. \end{aligned} \quad (50)$$

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

Result 8. *[Alger and Weibull [17]] In a population where the matching process entails relatedness* $r\in[0,1]$*, and the conditional probability functions are differentiable in* ε*, a preference function of the form* (19) *with* $\alpha=\alpha^*$ *is an ESPFC only if:*

1. $\alpha^*<r$ *if the strategies are strategic substitutes;*
2. $\alpha^*>r$ *if the strategies are strategic complements;*

3. $\alpha^* = r$ *if the strategies are strategically neutral.*

The complete information setting may be particularly well suited to represent interactions between relatives, especially in view of the fact that relatives often have the opportunity to observe each other's behaviors for many years. Given that the value of α determines how willing individuals are to act generously towards the other, the result suggests that evolution may have led to variation in the degree of intra-family generosity across different regions of the world. Indeed, in our evolutionary past the qualitative nature of the fitness game in any given region may have depended on the local ecological conditions. For example, in Arctic regions whale hunting was common, and such hunting arguably involves strategic complementarities. By contrast, in agricultural societies, food production would arguably have been a game in which production efforts were strategic substitutes. Furthermore, even for a given category of fitness game (i.e., whether strategies are strategic substitutes, complements, or neutral), the specifics of the fitness game is expected to matter. For example, in the production-and-sharing fitness game studied in [33], where the strategies are strategic substitutes, the evolutionarily stable degree of altruism is found to be lower in harsh than in generous environments, for a given relatedness r.

4.2 Interactions under incomplete information

Turning now to interactions under incomplete information, a straightforward generalization of the panmictic setting examined above is possible. Inserting the conditional probabilities into the system of best-response equations (39) in Definition 6 of a Bayesian Nash equilibrium,

$$\begin{cases} x^* \in \arg\max_{x \in X} & \Pr\left[u|u,\varepsilon\right] \cdot u\left(x,x^*\right) + \Pr\left[v|u,\varepsilon\right] \cdot u\left(x,y^*\right) \\ y^* \in \arg\max_{y \in X} & \Pr\left[u|v,\varepsilon\right] \cdot v\left(y,x^*\right) + \Pr\left[v|v,\varepsilon\right] \cdot v\left(y,y^*\right), \end{cases} \tag{51}$$

and into the equilibrium fitnesses of residents and mutants (see (41) and (42)),

$$W_u\left(x^*,y^*,\varepsilon\right) = \Pr\left[u|u,\varepsilon\right] \cdot w(x^*,x^*) + \Pr\left[v|u,\varepsilon\right] \cdot w(x^*,y^*) \tag{52}$$

$$W_v\left(x^*,y^*,\varepsilon\right) = \Pr\left[u|v,\varepsilon\right] \cdot w(y^*,x^*) + \Pr\left[v|v,\varepsilon\right] \cdot w(y^*,y^*), \tag{53}$$

the definition of an evolutionarily stable preference function under incomplete information applies as is (see Definition 7).

As was the case for preference evolution under incomplete information in the panmictic setting, the goal is to impose minimal restrictions on the set of potential preferences. The simple fitness-maximizing preference function (see (7)) is no longer evolutionarily stable, however. Instead, the analyses in [25, 26] reveal that evolution favors *Homo moralis* preferences:

Definition 11. *An individual is a **Homo moralis** if his/her preference function is of the form*

$$u_\kappa(x,y) = (1-\kappa)\cdot w(x,y) + \kappa \cdot w(x,x), \tag{54}$$

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

While it was the mathematical analysis that led to the "discovery" of this preference class, the name *Homo moralis* was inspired by the second term in (54), which can be interpreted as a concern for universalization, reminiscent of Kant's reasoning [34]: 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 [35] for a similar "as if" interpretation, in a model with strategy evolution for interactions between siblings).

As before, it is necessary to precisely define behavioral alikes (recall Definition 8).

Definition 12. *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 f is a **behavioral alike** to such Homo moralis if there exists some $x_r \in X_r$ and some $z \in X$ such that $z \in \arg\max_{x\in X}(1-r)\cdot w(x,x_r) + r\cdot w(x,x)$ and $z \in \arg\max_{x\in X} f(x,x_r)$.*

In words, 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 uses some type-homogenous Bayesian Nash equilibrium strategy in a monomorphic population of such *Homo moralis*, $x_r \in X_r$. Then [26] show the following result (a slight variation can be found in [25], where the main difference is a slight and unimportant difference in the definition of behavioral alikes, already referred to above).

Result 9. *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 Homo moralis preference function with*

degree of morality equal to the coefficient of relatedness, $\kappa = r$ (see (11)*) is ESI against any preference function that is not its behavioral alike.*

A population of *Homo moralis* resists entry by mutants because their preferences make them select a strategy that pre-empts entry by mutants. To see this, note first that the average fitness of vanishingly rare mutants (see (53)), who play some strategy, say z, tends to the following value as ε tends to 0:[9]

$$(1-r)\cdot w(z,x_r)+r\cdot w(z,z), \tag{55}$$

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

$$x_r \in \arg\max_{x\in X}(1-r)\cdot w(x,x_r)+r\cdot w(x,x). \tag{56}$$

A mutant preference type that is not a behavioral alike to *Homo moralis* with degree of morality $\kappa = r$ necessarily plays a strategy which does *not* belong to the set $\arg\max_{x\in X}(1-r)\cdot w(x,x_r)+r\cdot w(x,x)$. Hence, $(1-r)\cdot w(z,x_r)+r\cdot w(z,z) < (1-r)\cdot w(x_r,x_r)+r\cdot w(x_r,x_r) = w(x_r,x_r)$. This in turn implies that the average fitness of vanishingly rare mutants is strictly smaller than the average fitness of residents, which is arbitrarily close to $w(x_r,x_r)$. In sum, a population of *Homo moralis* resists entry by mutants because their preferences make them select a strategy which maximizes the average fitness of vanishingly rare mutants, given that the residents play this strategy.

4.2.1 Related models in the biology literature

Biologists will of course have recognized Hamilton's rule [7, 8] in the results presented in this section (expressed at the relevant level of selection [36], i.e., either at the level of strategies, or at the level of preferences). One may therefore wonder whether analysis of preference evolution brings fundamental new insights to the rich literature that followed in Hamilton's footsteps. I argue that the answer is positive for two reasons.

First, economists propose a rigorous and general analysis of behavioral Nash equilibria that is absent from the biology literature. As shown above, this has made it possible to tackle new questions such as: "What can be said about evolutionarily stable preferences when there are

[9]This observation follows from the upper hemi-continuity that is implied by the topological properties stated in the result (see the discussion following Result 6).

multiple Nash equilibria?", and "What can be said about evolutionarily stable preferences when the set of potential preference functions is the set of all continuous functions?"

Second, and more speculatively, there may be an interesting parallel to be drawn between, on the one hand, the preference evolution literature, and on the other hand the literature that asks whether individuals who use an ESS can be viewed *as if* they are maximizing some goal function (recently reviewed by Lehmann and Rousset [13]; see also [35] for an early such analysis in the economics literature). Specifically, the idea is that this parallel may perhaps shed additional light on the comparison of the gene-centered and the actor-centered view, discussed a length in [13] and in works surveyed therein.

The argument rests on a comparison between the altruistic preference function, defined in (19), and the *Homo moralis* preference function, defined in (54). An individual who has an altruistic preference function evaluates the consequences of his/her strategy on own individual fitness as well as on the interactant's individual fitness. This is reminiscent of the actor-centered view, which aggregates the effects of the individual on the others. An individual who has a *Homo moralis* preference function evaluates the consequences of his/her strategy on own individual fitness as well as on what his/her own individual fitness would be if, hypothetically, the interactant were to use the same strategy instead of the one (s)he is actually using. This is reminiscent of the gene-centered view, which aggregates the effects of others' behaviors on the individual's fitness.

The point is that the literature on preference evolution under incomplete information shows that *Homo moralis* preferences are more robust than altruistic preferences, in the following sense. Recall from Result 9 that *Homo moralis* preferences with degree of morality $\kappa = r$ are evolutionarily stable. It turns out that there are fitness games for which the set of Nash equilibria in a monomorphic population of altruists with degree of altruism $\alpha = r$ coincides with the set of Nash equilibria in a monomorphic population of *Homo moralis* with degree of morality $\kappa = r$. For such games neither preference function is evolutionarily stable against the other, and they are both evolutionarily stable against preference functions which are not behavioral alikes of *Homo moralis*. However, there are also fitness games in which the said set of Nash equilibria do not coincide. In such games, only *Homo moralis* preferences with degree of morality $\kappa = r$ are evolutionarily stable, while altruistic preferences with degree of altruism $\alpha = r$ may fail to be evolutionarily stable. I will now show that the former situation arises in fitness game Γ_1

while the latter case arises in fitness game Γ_2.

In fitness game Γ_1, when interacting with an individual who uses strategy y a *Homo moralis* chooses a strategy that satisfies the necessary first-order condition for x to maximize $u_\kappa(x, y)$:

$$(1-\kappa)(m+ky-2x)+\kappa(m+2kx-2x)=0. \tag{57}$$

This implies that there is a unique any Nash equilibrium strategy in a monomorphic population consisting of *Homo moralis* with degree of morality $\kappa = r$, which is the unique solution to the equation

$$(1-r)(m+kx-2x)+r(m+2kx-2x)=0. \tag{58}$$

In this fitness game an individual with altruistic preferences chooses a strategy that satisfies the necessary first-order condition for x to maximize $u_\alpha(x, y)$

$$m+ky-2x+\alpha ky=0, \tag{59}$$

implying that the unique Nash equilibrium strategy in a monomorphic population consisting of such altruists with degree of altruism $\alpha = r$ solves

$$m+kx-2x+rkx=0. \tag{60}$$

Equations (58) and (60) both yield $x = m/[2-(1+r)k]$: the set of equilibria in both monomorphic populations being the same, both preference functions are evolutionarily stable against preference functions that are not behavioral alikes of *Homo moralis* with degree of morality $\kappa = r$.

Turning now to fitness game Γ_2, a strategy x_α is a symmetric Nash equilibrium in a monomorphic population consisting of altruists with degree of altruism α if and only if an individual would not like to deviate to any other strategy, given that the other plays x_α:

$$(1+\alpha)\left[(x_\alpha)^2 R+1-x_\alpha\right] \geq (1+\alpha)xx_\alpha R+1-x+\alpha(1-x_\alpha) \quad \forall x\in[0,1]. \tag{61}$$

For any value of $\alpha \in [0, 1]$, there are three such equilibrium strategies: $\{0, 1/[(1+\alpha)R], 1\}$. Turning now to *Homo moralis* preferences, a strategy x_κ is a symmetric Nash equilibrium in a

monomorphic population consisting of *Homo moralis* with degree of morality κ if and only if:

$$(x_\kappa)^2 R + 1 - x_\kappa \geq (1-\kappa)(x x_\kappa R + 1 - x) + \kappa\,(x^2 R + 1 - x) \quad \forall x \in [0,1]. \tag{62}$$

Here the set of such equilibrium strategies is $\{0,1\}$ if and only if $\kappa \in [0, 1/R]$, and $\{1\}$ otherwise. In this fitness game, the *Homo moralis* preference function with $\kappa = r$ is evolutionarily stable (as per Result 9). By contrast, the altruistic preference function with $\alpha = r$ is not. To see this, assume that this is the resident preference function and consider a mutant preference function that induces the mutants to choose $y = 1$. Then in one of the Bayesian Nash equilibria (see (51)) residents play the mixed strategy

$$x = \frac{1 - Pr[v|u,\varepsilon](1+\alpha)R}{Pr[u|u,\varepsilon](1+\alpha)R}. \tag{63}$$

Clearly, the mutants obtain a strictly higher average fitness than such resident altruists, when these use this mixed strategy. In this case, the altruistic preferences with $\alpha = r$ are thus not evolutionarily stable, while Homo moralis preferences are (as per Result 9).

5 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* [21], 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 re-

spond to the stimuli and information the individual receives. Hence, individuals are guided by their preferences even in completely novel situations, implying that their behavior may change when, for instance, a hitherto unknown preference function appears in the population. Furthermore, 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 (19) 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 (54) 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. Experimental research shows that these motivations can be distinguished empirically [37, 38].

These observations further suggest three possible future research paths.

First, over the past few decades the behavioral economics literature has proposed and examined a wealth of preference classes to explain observed behaviors in social interactions: altruism [24], warm glow [39], a preference for conformity [40], for reciprocity [41, 42, 43, 44], inequity aversion [45, 46], guilt aversion [47, 48], and image concerns [49, 50]. These preference classes were inspired mostly by research in psychology and sociology. Building the interdisciplinary bridge one step further by evaluating the evolutionary stability properties of these preferences classes would be interesting; note that this is related to the suggestion made by Sober and Wilson [51] that evolutionary viability of psychological motives behind unselfish behaviors. It should be noted in this context that *Homo moralis* preferences [25], studied in this article, 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 [52, 53, 54].

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., [55, 56].

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 [25, 26] has been incorporated by Alger, Weibull, and Lehmann [57] into a standard island model [58], 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 [57] 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 [42, 59, 60, 61, 37]. This function combines material self-interest and a Kantian moral concern à la *Homo moralis* expressed at the material payoff level, and it also has a third component which can be interpreted as altruism/spite towards the opponent, again at the material payoff level. While this function thus differs from the *Homo moralis* function, the *Homo moralis* preference function is still 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 [62], 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 [63]. How would it affect the evolution of preference functions? One possible formalization is provided by Hop-

kins [64], 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 [65] 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., [66], 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. McNamara, C. E. Gasson, and A. I. Houston, "Incorporating rules for responding into evolutionary games," Nature, vol. 401, no. 6751, pp. 368–371, 1999.

[5] 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.

[6] 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.

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

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

[9] J. W. Weibull, Evolutionary Game Theory. Cambridge MA: MIT Press, 1997.

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

[11] J. M. McNamara and O. Leimar, Game Theory in Biology: Concepts and Frontiers. Oxford University Press, USA, 2020.

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

[13] L. Lehmann and F. Rousset, "When do individuals maximize their inclusive fitness?," American Naturalist, vol. 195, no. 4, pp. 717–732, 2020.

[14] I. M. Bomze and B. M. Pötscher, Game Theoretical Foundations of Evolutionary Stability. New York: Springer-Verlag, 1988.

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

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

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

[25] 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.

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

[27] 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.

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

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

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

[31] 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.

[32] A. Grafen, "The hawk-dove game played between relatives," Animal Behaviour, vol. 27, pp. 905–907, 1979.

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

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

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

[36] S. Okasha, "Maynard Smith on the levels of selection question," Biology and Philosophy, vol. 20, no. 5, pp. 989–1010, 2005.

[37] 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.

[38] B. Van Leeuwen and I. Alger, "Estimating social preferences and Kantian morality in strategic interactions," TSE Working Paper 19-1056, 2022.

[39] 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.

[40] B. D. Bernheim, "A theory of conformity," Journal of Political Economy, vol. 102, no. 5, pp. 841–877, 1994.

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

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

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

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

[45] 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.

[46] 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.

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

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

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

[50] 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.

[51] E. Sober and D. S. Wilson, Unto Others: The Evolution and Psychology of Unselfish Behavior. Cambridge MA: Harvard University Press, 1998.

[52] 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.

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

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

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

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

[57] 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.

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

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

[65] 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.

[66] 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.

Online appendix to "Evolutionarily stable preferences"

Ingela ALGER*

December 8, 2022

1 Examples

In this section I provide several examples of games to which the general model in the main text applies.

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" ([1] p.10, see also the recent book by McNamara and Leimar [2]); this is also in line with standard vocabulary in non-cooperative game theory, see [3]).

In one-shot simultaneous-move games—in which the individuals act simultaneously and only once—the set of strategies simply coincides with the set of actions available. In one-shot sequential-move games—in which the individuals are called to act sequentially and they can observe any previous action choices—a strategy is a plan that describes the action to be chosen depending on the order of play and previous action choices. In repeated games, the same simultaneous-move game is played several times, and a strategy is a plan that describes the action to be chosen depending on the action choices in the previous rounds.

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

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

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 ($T > R > P > S$).

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 [3]. 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

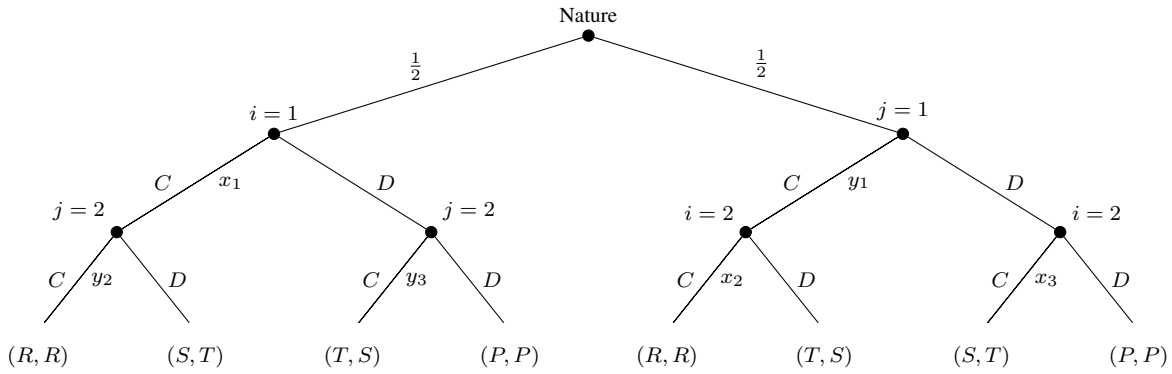

Figure 2: Meta-game protocol for the Sequential Prisoners' Dilemma ($T > R > P > S$)

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]

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 $\partial^2 w(x, y)/(\partial x \partial y) < 0$)
- 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

[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 [2].

– 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:

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

2 Calculating the evolutionarily stable value of α under complete information in fitness game Γ_1

A mutant's equilibrium strategy is

$$x^*(\beta, \alpha) = \frac{m[2 + (1+\beta)k]}{4 - (1+\alpha)(1+\beta)k^2}, \tag{1}$$

and a resident's equilibrium strategy is

$$x^*(\alpha, \beta) = \frac{m[2 + (1+\alpha)k]}{4 - (1+\alpha)(1+\beta)k^2}. \tag{2}$$

Note that an individual's altruism has an effect on its equilibrium strategy only if $k \neq 0$. Hence, the subsequent analysis applies to $k \in (-1, 0) \cup (0, 1)$.

For any $(\alpha, \beta) \in (-1, 1)$, the equilibrium fitness is:

$$\begin{aligned} w(x^*(\alpha,\beta), x^*(\beta,\alpha)) &= \frac{m^2[2 + (1+\alpha)k]}{[4 - (1+\alpha)(1+\beta)k^2]^2} \\ &\quad \cdot\{4 - (1+\alpha)(1+\beta)k^2 + k[2 + (1+\beta)k] - 2 - (1+\alpha)k\} \\ &= \frac{m^2[2 + (1+\alpha)k]}{[4 - (1+\alpha)(1+\beta)k^2]^2} \cdot \{2 - \alpha(1+\beta)k^2 + (1-\alpha)k\} \end{aligned} \tag{3}$$

With the notation used here, Result 2 in the main text is: **Result 2 (bis).**

1. If $w(x^*(\alpha,\alpha), x^*(\alpha,\alpha)) > w(x^*(\beta,\alpha), x^*(\alpha,\beta)$, then u_α is ESC against u_β.

2. If $w(x^*(\alpha,\alpha), x^*(\alpha,\alpha)) = w(x^*(\beta,\alpha), x^*(\alpha,\beta)$, then u_α is ESC against u_β only if

$w(x^*(\alpha,\beta),x^*(\beta,\alpha)) > w(x^*(\beta,\beta),x^*(\beta,\beta))$.

3. If $w(x^*(\alpha,\alpha),x^*(\alpha,\alpha)) < w(x^*(\beta,\alpha),x^*(\alpha,\beta)$, then u_α is not ESC against u_β.

To identify candidates for ESC values of α, I solve the equation (see equation (26) in the main text):

$$\alpha \cdot x_1^*(\alpha,\alpha) = x_2^*(\alpha,\alpha) \tag{4}$$

for α. The partial derivative with respect to the mutant trait β of the mutant's equilibrium strategy,

$$x_1^*(\beta,\alpha) = \frac{mk[4-(1+\alpha)(1+\beta)k^2] + m[2+(1+\beta)k](1+\alpha)k^2}{[4-(1+\alpha)(1+\beta)k^2]^2}, \tag{5}$$

simplifies to

$$x_1^*(\beta,\alpha) = \frac{2mk[2+(1+\alpha)k]}{[4-(1+\alpha)(1+\beta)k^2]^2}. \tag{6}$$

The partial derivative with respect to the mutant trait β of the resident's equilibrium strategy is

$$x_2^*(\alpha,\beta) = \frac{mk^2(1+\alpha)[2+(1+\alpha)k]}{[4-(1+\alpha)(1+\beta)k^2]^2}. \tag{7}$$

Evaluating these partial derivatives for $\beta = \alpha$ gives:

$$x_1^*(\alpha,\alpha) = \frac{2mk[2+(1+\alpha)k]}{[4-(1+\alpha)^2k^2]^2} \tag{8}$$

and

$$x_2^*(\alpha,\alpha) = \frac{mk^2(1+\alpha)[2+(1+\alpha)k]}{[4-(1+\alpha)^2k^2]^2}. \tag{9}$$

Inserting these expressions into equation (4) gives:

$$\alpha \cdot \tfrac{2mk[2+(1+\alpha)k]}{[4-(1+\alpha)^2k^2]^2} = \tfrac{mk^2(1+\alpha)[2+(1+\alpha)k]}{[4-(1+\alpha)^2k^2]^2}, \tag{10}$$

or

$$\alpha \cdot 2mk[2+(1+\alpha)k] = mk^2(1+\alpha)[2+(1+\alpha)k]. \tag{11}$$

For any $k \in (-1,0) \cup (0,1)$ this equation has two solutions, $\alpha = \frac{-2-k}{k}$ and $\alpha = \frac{k}{2-k}$, both of which are thus candidates for an evolutionarily stable α.

I apply the tests in Result 2 (bis) to the first candidate. First, note that for any β

$$w\left(x^*\left(\beta, \frac{-2-k}{k}\right), x^*\left(\frac{-2-k}{k}, \beta\right)\right) = m^2/4, \tag{12}$$

which implies that

$$w\left(x^*\left(\frac{-2-k}{k}, \frac{-2-k}{k}\right), x^*\left(\frac{-2-k}{k}, \frac{-2-k}{k}\right)\right) = w\left(x^*\left(\beta, \frac{-2-k}{k}\right), x^*\left(\beta, \frac{-2-k}{k}\right)\right). \tag{13}$$

One must therefore check whether the strict inequality in the second part of the result is verified. A simple counter-example proves that this inequality is not satisfied for all mutant traits $\beta \neq \frac{-2-k}{k}$. For example, if $\beta = 0$, then

$$w(x^*(\beta,\beta), x^*(\beta,\beta)) \;=\; \frac{m^2}{(2-k)^2}$$

while $w(x^*(\frac{-2-k}{k}, \beta), x^*(\beta, \frac{-2-k}{k})) = 0$, implying that the said condition fails to hold for $\beta = 0$. I conclude that $\alpha = \frac{-2-k}{k}$ does not resist the invasion by the mutant trait $\beta = 0$.

Next, I apply the tests in Result 2 (bis) to the second candidate, $\alpha = \frac{k}{2-k}$. First, note that, upon simplification,

$$w\left(x^*\left(\frac{k}{2-k}, \frac{k}{2-k}\right), x^*\left(\frac{k}{2-k}, \frac{k}{2-k}\right)\right) \;=\; \frac{m^2(2+k)(2-k)}{16(1-k)} \tag{14}$$

and

$$w\left(x^*\left(\beta, \frac{k}{2-k}\right), x^*\left(\frac{k}{2-k}, \beta\right)\right) \;=\; \frac{m^2(2-k)(2+k)[4-k^2(1+\beta)^2]}{4[2(2-k)-(1+\beta)k^2]^2}. \tag{15}$$

Tedious calculations then lead to the following expression for $w(x^*(\frac{k}{2-k}, \frac{k}{2-k}), x^*(\frac{k}{2-k}, \frac{k}{2-k})) - w(x^*(\beta, \frac{k}{2-k}), x^*(\frac{k}{2-k}, \beta))$:

$$\frac{m^2(2+k)(2-k)k^2[(1+\beta)k-2\beta]^2}{16(1-k)[2(2-k)-(1+\beta)k^2]^2}. \tag{16}$$

This expression is strictly positive for any $k \in (-1, 0) \cup (0, 1)$ and $\beta \in (-1, 1)$, $\beta \neq \frac{k}{2-k}$, implying that the first condition of Result 2 (bis) is satisfied. **Q.E.D.**

References

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

[2] J. M. McNamara and O. Leimar, Game Theory in Biology: Concepts and Frontiers. Oxford University Press, USA, 2020.

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