WORKING PAPERS

N° 1405

January 2023

"Evolution of semi-Kantian preferences in two-player assortative interactions with complete and incomplete information and plasticity"

Ingela Alger and Laurent Lehmann

Toulouse School of Economics

Evolution of semi-Kantian preferences in two-player assortative interactions with complete and incomplete information and plasticity

Laurent Lehmann* and Ingela Alger†

Abstract

We develop a model for the evolution of preferences guiding behavior in pairwise interactions in group-structured populations. The model uses the conceptual platform of long-term evolution theory and covers different interaction scenarios, including conditional preference expression upon recognition of interactant's type. We apply the model to the evolution of semi-Kantian preferences at the fitness level, which combine self-interest and a Kantian interest evaluating own behavior in terms of consequences for own fitness if the interactant also adopted this behavior. We look for the convergence stable and uninvadable value of the Kantian coefficient, i.e., the weight attached to the Kantian interest, a quantitative trait varying between zero and one. We consider three scenarios: (a) incomplete information; (b) complete information and incomplete plasticity; and (c) complete information and complete plasticity, where individuals can, not only recognize the type of their interaction partner (complete information), but also conditionally express the Kantian coefficient upon it (complete plasticity). For (a), the Kantian coefficient tends to evolve to equal the coefficient of neutral relatedness between interacting individuals; for (b), it evolves to a value that depends on demographic and interaction assumptions, while for (c) individuals become pure Kantians when interacting with individuals of the same type, while they apply the Kantian coefficient that is uninvadable in a panmictic population under complete information when interacting with individuals with a different type. Overall, our model connects several concepts for analysing the evolution of behavior rules for strategic interactions that have been emphasized in different and sometimes isolated literatures.

Keywords: evolution of semi-Kantian preferences, group-structured populations, fitness, convergence stability, uninvadability, *Homo moralis*

*Department of Ecology and Evolution, University of Lausanne, Switzerland. laurent.lehmann@unil.ch

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

1 Introduction

This paper is about formalizing natural selection on rules guiding individual behavior in strategic interactions, a central question in evolutionary game theory (Maynard Smith and Price, 1973; Dawkins, 1980; Maynard Smith, 1982). By behavior we mean a "strategy", i.e., "a specification of what an individual will do in any situation in which it may find itself" (Maynard Smith, 1982). In the original evolutionary game theory model, each individual is programmed to play a certain strategy regardless of the strategies used by others in the population. One way to think about this is that the strategy is innate, thus a genetically determined trait. This view led to a vast theoretical literature analyzing the genetic evolution of strategies under all sorts of biological scenarios as is illustrated by the vast literature on the evolution of fighting and cooperation in plants and animals (e.g., the books by Maynard Smith, 1982; Bulmer, 1994; Giraldeau and Caraco, 2000; Vincent and Brown, 2005; McNamara and Leimar, 2020). Here, it is the population genetic process alone that determines the "evolutionary stable strategy" since strategies are inherited from parent to offspring and selected among alternatives by way of differential survival and reproduction. However, the view that strategies are innate is restrictive, as it rules out situations where individuals have capacities to change their strategy when interacting with their environment. Such processes have been incorporated into evolutionary game theory through several alternative notions, such as the concepts of the "culturally stable" and the "developmentally stable" strategy (Dawkins, 1980; Maynard Smith, 1982). Here, the behavior of an individual is the outcome of some updating rule(s), typically imitative or experiential, for strategy selection during its lifespan.

This in turn raises the question of what should be the evolutionarily stable rule for individual strategy selection in strategic interactions? While this question was raised early in the history of evolutionary game theory (Harley, 1981; Maynard Smith, 1982), perhaps more controversy than conclusions where initially reached (e.g., Selten and Hammerstein, 1984), and it is only more recently that this question has gained some renewed theoretical attention in evolutionary biology (e.g., Arbilly et al., 2010; Dridi and Lehmann, 2015; Dridi and Akçay, 2018; McNamara and Leimar, 2020). In the meantime, however, economists and mathematical game theorists also produced insights about how various individual choice rules induce change in population behavior (e.g., the books by Sugden, 1986; Weibull, 1997; Fudenberg and Levine, 1998; Hofbauer and Sigmund, 1998; Samuelson, 1998; Young, 1998; Sandholm, 2011). One obstinate result of this area is that updating rules of strategies–whether imitative or experiential–relying on payoff tend to converge to Nash equilibria (Hofbauer and Sigmund, 1998; Fudenberg and Levine, 1998; Cressman and Tao, 2004). Hence, in behavioral equilibrium, it is as if individuals strive to maximize the payoff function at hand and thus as if they are rational decision makers, in the sense that among a set of options they choose the one they prefer (Mas-Colell et al., 1995).

The next natural question from an evolutionary perspective is then: if the evolving trait is the payoff function to be maximized, which payoff function is evolutionarily stable? This is the question that the literature on preference evolution addresses (e.g., Guth, 1995; Ok and Vega-Redondo, 2001; Dekel et al., 2007; Heifetz et al., 2007b,a; Akçay and Van Cleve, 2009; Alger and Weibull, 2010, 2012, 2013). Because information plays a central role in strategic interactions (Fudenberg and Tirole, 1991), the formalizations of preference evolution have covered a variety of informational scenarios (e.g., Ok and Vega-Redondo, 2001; Dekel et al., 2007; see Alger and Weibull, 2019; Alger, 2022 for surveys). Focusing on the evolution of preferences is useful, because payoff-based choice rules can otherwise

come in endless mechanistic forms–some more biologically and cognitively inspired than others (Sutton and Barto, 1998; Russell and Norvig, 2016) and gives hope to yield some general predictions about equilibrium behavior.

The goal of this paper is to contribute to the literature on the evolution of rules guiding individual behavior in two ways and is thus divided in two parts. In the first part, we connect a number of concepts and results to analyze the long-term evolution (*sensu* Eshel, 1996; Eshel et al., 1998) of behavioral mechanisms for equilibrium action in group structured population, whereby this part can be read as a methodological review. In the second part, we push forward within this framework the evolutionary analysis of the class of preferences involving a mix between self-interest and an interest in evaluating own behavior in the light of the consequences for own payoff if others adopted this behavior. This is the class of semi-Kantian preferences, which, in the words of Binmore (1998, p. 191), can be seen as hybrid preferences combining both the categorical imperative of Nash and that of Kant. Bergstrom (1995) show that the evolutionarily stable strategy in interactions between siblings could be interpreted as if individuals had such preferences, an interpretation that should hold more generally when interactions occur between related individuals. Semi-Kantian preferences have then indeed been shown to be evolutionarily stable and uninvadable under various transmission rules when population structure results from limited genetic or cultural mixing among interacting individuals, when interacting individuals cannot observe each other's preferences (Alger and Weibull, 2013, 2016; Alger et al., 2020). However, so far the evolutionary convergence towards semi-Kantian preferences has not been ascertained and their evolution has not been analyzed under different informational assumptions. We provide analyses of convergence stability and uninvadability of semi-Kantian preferences in three different informational scenarios: (a) incomplete information; (b) complete information and incomplete plasticity (interacting individuals can observe each other's preferences, and an individual's preferences to not depend on the other's preferences); and (c) complete information and complete plasticity (interacting individuals can observe each other's preferences, and an individual's preferences can depend on the other's preferences). As will be seen, the different informational and plasticity assumptions lead to quite different evolutionary outcomes.

Our aim is not to obtain the most general conclusions about the open questions we address, but rather to illustrate how demographic and informational features jointly contribute to the understanding of the long-term evolution of preferences in structured populations. As such, we consider only pairwise interactions and restrict attention to the parametric class of semi-Kantian preferences and the evolution of the Kantian coefficient, a quantitative trait varying between zero and one, which represents the weight attached to the Kantian interest.

2 Evolutionary invasion analysis of behavioral mechanisms

2.1 Biological assumptions for pairwise interactions

We consider a population of asexually reproducing individuals that are demographically homogeneous (no effective age, stage or sex structure). The population occupies a habitat with an infinite and constant number of groups (or demes, or spatial subdivisions), each of which is occupied by exactly two individuals and so the population is of constant size. Each individual is characterized by a type belonging to a type space Θ that affects its phenotype—the collection of any relevant morphological, physiological or behavioral measurable feature of the individual. We

consider a demographic process where the population is censused at discrete time steps, between which the following events occur in cyclic order. (a) In each group, the pair of individuals engage in an interaction. Some process (learning, exchange of information, etc) leads to a pair of equilibrium strategies being expressed. The equilibrium strategy pair, which may depend on the individuals' types as well as the types present in the population at large, determines some outcome (for example, the material payoff of each individual). (b) Each individual in each group produces a large number of juveniles according to the outcome of the pairwise interaction and eventually dies subject to some death process,[1] which may also depend on the outcome of the pairwise interaction. (c) Juveniles remain in the natal group with some fixed probability. With complementary probability, assumed to be non-zero, they migrate out of their natal group and survive dispersal with a certain probability that may depend on the outcome of the interaction between the juvenile's parent and its neighbor. (d) In each group, the open reproductive spots vacated by deceased adults are randomly filled up by competing juveniles, who then become adults.

2.2 Invasion and individual fitness

We adopt a standard invasion analysis framework and consider a population that is monomorphic for some resident type $\theta \in \Theta$ in which a mutant type $\tau \in \Theta$ arises (e.g., Fisher, 1930; Eshel and Feldman, 1984; Parker and Maynard Smith, 1990; Metz et al., 1992; Charlesworth, 1994; Ferrière and Gatto, 1995; Eshel, 1996). It then follows from applications of invasion analysis to our demographic process assumptions of section 2.1 (see Box 1) that any mutation $\tau \in \Theta$, which is introduced in a single individual in a monomorphic population with the resident type $\theta \in \Theta$, eventually goes extinct with probability one if and only if the invasion fitness (the geometric growth rate) of the mutant type, denoted $W(\tau, \theta)$, satisfies

$$W(\tau, \theta) \leq 1. \tag{1}$$

Here, the "1" can be interpreted as the growth rate of a resident type in a monomorphic resident population, which, owing to the fact that the population is of constant size can, on average, only replace itself.

Invasion fitness can be represented as the individual fitness of a randomly sampled mutant τ descending from the individual in which the mutation initially appeared, averaged over the cases where the mutant interacts with another member of the same lineage and those where it interacts with an individual from a different lineage (who is thus of the resident type θ):

$$W(\tau, \theta) = [1 - r(\tau, \theta)]\, \tilde{w}_1(\tau, \theta) + r(\tau, \theta)\, \tilde{w}_2(\tau, \theta), \tag{2}$$

where $\tilde{w}_i(\tau, \theta)$ is the individual fitness of a mutant when there are $i \in \{1, 2\}$ mutants in its group and the population is otherwise monomorphic for θ, and $r(\tau, \theta)$ is the *pairwise relatedness* between a τ mutant and its group neighbor (see Box 1 for a derivation of eq. (2)). Pairwise relatedness is the probability that, conditional on an individual being of type τ, the group neighbor belongs to the same ancestral lineage and is thus also of type τ, whereby both individuals are *identical-by-descent* (Malécot, 1969); note that since migration is assumed non-zero, we have $r(\tau, \theta) < 1$. Whether relatedness $r(\tau, \theta)$ depends on both the mutant and the resident type, only on the resident type, or neither, depends on demographic and interaction assumptions. For instance, relatedness is independent of

[1] While we allow for individuals surviving from one demographic time point to the next, the survival probability is assumed independent of age, so that there is no effective age structure in the population.

the types for family-structured populations, in which case it is determined only by the pedigree relatedness, e.g. $r = 1/2$ for full-siblings [as implied by the model of Michod, 1980, which also entails that eq. (2) applies to sexual reproduction in family-structured populations in the absence of inbreeding].

When $W(\tau, \theta)$ is differentiable (which is not always the case), a resident type θ^* is locally convergence stable if and only if the first two following conditions hold, while it is locally uninvadable if the first and the third conditions hold (Eshel, 1983; Taylor, 1989; Christiansen, 1991; Geritz et al., 1998):

$$S(\theta^*) = \left.\frac{\partial W(\tau, \theta)}{\partial \tau}\right|_{\tau=\theta=\theta^*} = 0 \tag{3}$$

$$J(\theta^*) = \left.\frac{\mathrm{d}S(\tau)}{\mathrm{d}\tau}\right|_{\tau=\theta=\theta^*} < 0 \tag{4}$$

$$H(\theta^*) = \left.\frac{\partial^2 W(\tau, \theta)}{\partial \tau^2}\right|_{\tau=\theta=\theta^*} < 0. \tag{5}$$

Here, $S(\theta)$, $J(\theta)$, and $H(\theta)$, stand respectively for the selection gradient, the selection Jacobian, and the selection Hessian, evaluated at the resident type θ. A type satisfying $S(\theta^*) = 0$ will be called a singular type (or a singularity).[2] There is a non-trivial relationship between the static conditions (3)-(5) obtained from invasion fitness and dynamic stability. Namely, for mutants with small effects on the phenotype, i.e. the difference $|\theta - \tau|$ is small, a singular type θ^* satisfying conditions (4)-(5) is a (i) local attractor of the evolutionary dynamics under gradual evolution and (ii) resistant to invasion by small deviations.[3]

2.3 Behavioral equilibrium

In applications of evolutionary game theory, an individual's type is often simply taken to be its strategy to be applied in the interaction at hand (e.g., Maynard Smith and Price, 1973; Bishop and Cannings, 1978). Yet many models decouple types from strategies, and we do so as well. In order to obtain a full description of how individual fitness depends on own type and neighbor's type — a dependence that in eq. (2) was captured through the mappings $\tilde{w}_1 : \Theta^2 \to \mathbb{R}_+$ and $\tilde{w}_2 : \Theta^2 \to \mathbb{R}_+$, without reference to the strategies used by the individuals — we begin by defining individual fitness as a function of the strategies used, and then we introduce notation and assumptions for how the equilibrium strategies depend on the types.

Letting $\mathcal{X}$ denote the set of strategies that each individual has access to when interacting with its neighbor, the

[2]When invasion fitness is differentiable, the quantities $S(\theta)$, $H(\theta)$, and $J(\theta)$ in fact allow for a complete classification of the singularities of the evolutionary dynamics (Geritz et al., 1998). Thus, when $H(\theta^*) > 0$ and $J(\theta^*) < 0$, a singular type θ^* is an evolutionary branching point; namely, an attractor of the evolutionary dynamics that subsequently splits the population into distinct morphs leading to the coexistence of different types in a protected polymorphism. When $H(\theta^*) < 0$ and $J(\theta^*) > 0$ we have a so-called garden of eden state of the evolutionary dynamics, an uninvadable trait value that is unattainable by gradual evolution. Finally, if $H(\theta^*) > 0$ and $J(\theta^*) > 0$ then the singular type θ^* is a an uninvadable repellor.

[3]This follows from the fact that under the full evolutionary dynamic process of quantitative traits (those whose state space is $\Theta \subset \mathbb{R}$ or more generally $\Theta \subset \mathbb{R}^d$), the selection gradient $S(\theta)$ describes the direction of selection on small trait deviations regardless of population genetic states and demographic structures (Rousset and Billiard, 2000, Rousset, 2004, p. 206, Geritz, 2005, Priklopil and Lehmann, 2021). This entails that any mutant invading the population when rare will eventually substitute the resident and recurrent mutations will drive the trait towards the singularity within its neighborhood when condition (3) is satisfied. This result was first noted in a special case by Hamilton (1964) and called "a gift from god" (Hamilton, 1988). See also Eshel et al. (1997) for a different line of argument reaching the same conclusions.

individual fitness function $w : \mathcal{X}^3 \to \mathbb{R}_+$ is defined such that $w(x_i, x_j, y)$ gives the expected number of descendants (including the surviving self) produced over one demographic time period by an adult individual i expressing strategy x_i when matched to a group neighbour j expressing strategy x_j, when individuals in the population at large all use strategy y. An example is provided in Box 2. Note that any individual fitness function is subject to the demographic consistency relation $w(y, y, y) = 1$ for all $y \in \mathcal{X}$.

Turning now to the equilibrium strategies, in a population with a resident type θ and a mutant type τ, each group either has zero, one, or two mutants. For groups with two residents (resp. two mutants), we denote by $y_{\rm s}^*(\theta)$ (resp. $x_{\rm s}^*(\tau, \theta)$) an equilibrium strategy for each individual, where the index s refers to *same* type (note that we rule out equilibria in which two identical individuals use different strategies). For mixed groups, with one resident and one mutant, let $x_{\rm d}^*(\tau, \theta)$ denote the mutant's equilibrium strategy and $y_{\rm d}^*(\theta, \tau)$ the resident's equilibrium strategy, where the index d stands for *different* types. Importantly, throughout we assume that for any type pair $(\theta, \tau) \in \Theta^2$, there exist unique equilibrium strategies $y_{\rm s}^*(\theta)$,$x_{\rm s}^*(\tau, \theta)$, and $(x_{\rm d}^*(\tau, \theta), y_{\rm d}^*(\theta, \tau))$. This implies that the mappings $\tilde{w}_1$ and $\tilde{w}_2$ used in eq. (2) are well defined, as follows:

$$\tilde{w}_1(\tau, \theta) = \begin{cases} w(x_{\rm d}^*(\tau, \theta), y_{\rm d}^*(\theta, \tau), y_{\rm s}^*(\theta)) & \text{if } \tau \neq \theta \\ w(x_{\rm s}^*(\theta, \theta), x_{\rm s}^*(\theta, \theta), y_{\rm s}^*(\theta)) & \text{if } \tau = \theta, \end{cases} \tag{6}$$

and

$$\tilde{w}_2(\tau, \theta) = w(x_{\rm s}^*(\tau, \theta), x_{\rm s}^*(\tau, \theta), y_{\rm s}^*(\theta)). \tag{7}$$

How do the equilibrium strategies arise? In the evolutionary game theory literature, a variety of processes, or mechanisms, of interdependent strategy expression have been examined, including reactive strategies, behavior response rules, learning rules, or developmental rules (e.g., Maynard Smith, 1982; McNamara et al., 1999; Akçay and Van Cleve, 2009; Killingback and Doebeli, 2002; Taylor and Day, 2004; André and Day, 2007; Dridi and Akçay, 2018; McNamara and Leimar, 2020). In each case, a dynamic system drives strategy expression over time, and these behavioral dynamics reach an equilibrium, which determines survival and reproduction. One way to formalize these mechanisms is to posit that the equilibrium strategies solve a fixed-point problem. Thus, for mixed groups, let there be two mappings, $M_{\rm d} : \Theta^2 \times \mathcal{X}^2 \to \mathbb{R}$ for the mutant type and $R_{\rm d} : \Theta \times \mathcal{X}^2 \to \mathbb{R}$ for the resident type, which capture the process at hand, and which are such that an equilibrium pair of strategies satisfies the fixed-point system of equations:

$$\begin{cases} R_{\rm d}(\theta, y_{\rm d}^*, x_{\rm d}^*) = 0 \\ M_{\rm d}(\tau, \theta, x_{\rm d}^*, y_{\rm d}^*) = 0. \end{cases} \tag{8}$$

The mechanism $M_{\rm d}$ is parametrized by both the mutant and the resident type, while $R_{\rm d}$ is parametrized only by the resident type. This is so because when individuals interact their strategy may depend on (i) their own type and strategy, (ii) the strategy of their interaction partner, and (iii) on strategies in the population at large, which depends only on the resident type when the mutant is rare (see also eq. 10 below). Solving for $x_{\rm d}^*$ and $y_{\rm d}^*$ produces the dependence of each strategy on both types, i.e., $x_{\rm d}^* = x_{\rm d}^*(\tau, \theta)$ and $y_{\rm d}^* = y_{\rm d}^*(\theta, \tau)$. For the equilibrium strategy used in mutant-mutant interactions, let there be a mapping $M_{\rm s} : \Theta^2 \times \mathcal{X} \to \mathbb{R}$ which describes the process whereby a mutant interacts with another mutant as a function of the opponent's strategy. Restricting attention to settings

in which both mutants then use the same strategy in equilibrium, this is assumed to satisfy the fixed-point equation

$$M_{\mathrm{s}}(\tau, \theta, x_{\mathrm{s}}^*) = 0. \tag{9}$$

The behavioral mechanism M_{s} is parametrized by both the mutant and the resident type, because the strategy used in the population at large (which depends on the resident type) may affect the strategy used in a mutant-mutant pair. Hence, the solution of eq. (9) implicitly defines the equilibrium strategy as a function of both τ and θ, so that we can write $x_{\mathrm{s}}^*(\tau, \theta)$. Finally, for the equilibrium strategy in resident-resident interactions, the equilibrium strategy $y_{\mathrm{s}}^*(\theta)$ solves the fixed-point equation

$$R_{\mathrm{s}}(\theta, y_{\mathrm{s}}^*) = M_{\mathrm{s}}(\theta, \theta, y_{\mathrm{s}}^*) = 0, \tag{10}$$

where $R_{\mathrm{s}} : \Theta \times \mathcal{X} \to \mathbb{R}$ is the behavioral mechanism characterizing the (same) equilibrium strategy of each individual in a resident pair.[4] By contrast to the equilibrium strategy between two mutants, which depends both on the mutant and the resident type, the equilibrium strategy between two residents depends only on the resident type.

2.4 Nash equilibrium and utility function

Many formalizations of the behavioral fixed points (8)–(10) consist in assuming that strategies equilibrate by being guided by some payoff function and adopting assumptions such that the dynamics lead to a Nash equilibrium according to this payoff function. In such models, the mappings R_{s}, R_{d}, M_{s}, and M_{d} can be thought of as describing the best response functions according to the payoff function. In equilibrium, it is thus as if individuals maximize this payoff function, given the strategy used by the opponent. One class of such models takes the payoff function to be a *utility function*, which represents an individual's preferences.[5] In our setting, given some resident utility function θ and some mutant utility function τ, the strategy $y_{\mathrm{s}}^*(\theta)$ (resp. $x_{\mathrm{s}}^*(\tau, \theta)$) would be the strategy in $\mathcal{X}$ that maximizes the utility of a resident (resp. that of a mutant), given that the resident (resp. mutant) with whom it interacts also uses strategy $y_{\mathrm{s}}^*(\theta)$ (resp. $x_{\mathrm{s}}^*(\tau, \theta)$). Likewise, the strategy $y_{\mathrm{d}}^*(\theta, \tau)$ would be the strategy in $\mathcal{X}$ that maximizes the utility of the resident, given that the mutant with whom it interacts uses strategy $x_{\mathrm{d}}^*(\tau, \theta)$, while the strategy $x_{\mathrm{d}}^*(\tau, \theta)$ would be the strategy in $\mathcal{X}$ maximizing the mutant's utility, given that the resident with whom it interacts uses strategy $y_{\mathrm{d}}^*(\theta, \tau)$.

[4] Invasion fitness eq. (2) must be defined for all $\tau \in \Theta$ including $\tau = \theta$ in which case it describes the growth ratio of the lineage of a single individual with type $\tau = \theta$ in an otherwise monomorphic population with type θ. Since owing to eq. (10) all individuals use strategy $y_{\mathrm{s}}^*(\theta)$ in such a fully monomorphic populations even if interactions occur between same and different lineage members, all individuals are demographically exchangeable, i.e., all their vital rates are the same, so that the demographic consistency relation $W(\theta, \theta) = 1$ is verified.

[5] An individual's utility function is indeed simply a representation of its preferences. For any pair of strategies x and y, an individual's preferences over available strategies tell whether the individual prefers x, y, or is indifferent between the two. Under certain conditions, such a preference ordering can be fully described by a function that associates a real number to each strategy, namely the utility function (see, e.g., Mas-Colell et al., 1995; Binmore, 2011). An individual is assumed to choose a strategy with the highest possible value of the function, since this is the strategy it prefers. Utility maximization is not to be taken literally: it is simply a mathematical tool used to describe behavior that amounts to choosing the preferred item from the strategy set. A pair of strategies then constitutes a Nash equilibrium if each individual uses a strategy which, given the other individual's strategy, is the one it prefers.

In the remainder of this paper, we endorse this approach and rely on results showing that among the set of all continuous utility functions, a utility function representing semi-Kantian preferences emerges as being particularly viable from an evolutionary perspective (Alger and Weibull, 2013; Alger et al., 2020). For some individual i who uses strategy x_i when it neighbour uses strategy x_j, and given that strategy y^* is played at large in the population, this utility function is defined as

$$u_{\kappa_i}(x_i, x_j \mid y^*) = (1 - \kappa_i)\, w(x_i, x_j, y^*) + \kappa_i\, w(x_i, x_i, y^*), \tag{11}$$

where w is the individual fitness function defined above. In the first term we see i's realized fitness, given the strategies used. In the second term we see the fitness that i would have realized if — hypothetically — the opponent used i's strategy (x_i) instead of strategy x_j. Since the latter implies that the individual evaluates what would happen if others were to follow the same course of action as itself, it can be interpreted as capturing a Kantian moral concern (Kant, 1785). These preferences were therefore dubbed *Homo moralis* (Alger and Weibull, 2013), yet they apply regardless of the organism under consideration in our life-cycle assumptions of section 2.1. We will call the parameter κ_i the Kantian coefficient and restrict it to the interval $[0, 1]$. In the remainder of this paper, we treat the Kantian coefficient as being determined by an individual's type and investigate its evolution under three different scenarios: (a) incomplete information, (b) complete information with incomplete plasticity, and (c) complete information with complete plasticity. Each of these scenarios, together with the utility function (11), defines a specific set of behavioral mechanisms (8)–(10), detailed in the next section.

3 Evolution of the Kantian coefficient

For simplicity, we throughout restrict attention to settings where w is twice continuously differentiable and strictly concave. We further take the strategy space $\mathcal{X}$ to be an open and convex subset of $\mathbb{R}$. These assumptions together imply that any equilibrium strategy must satisfy first-order conditions, and this facilitates the analysis. To rule out trivial settings in which an individual's strategy has no impact on the opponent's fitness, we also assume that $\partial w(x, y, z)/\partial y \neq 0$ for all $(x, y) \in X^2$. By convention, we let $\partial w(x, y, z)/\partial y > 0$ for all $(x, y) \in X^2$, meaning that an increase in the strategy of an individual's opponent enhances the individual's fitness.[6] Finally, we assume that $\partial^2 w(x, y, z)/\partial y \partial x$ has the same sign for all $(x, y) \in X^2$, and we will say that the strategies are *strategic complements* if $\partial^2 w(x, y, z)/\partial y \partial x > 0$, *strategic substitutes* if $\partial^2 w(x, y, z)/\partial y \partial x < 0$, and *strategically neutral* if $\partial^2 w(x, y, z)/\partial y \partial x = 0$.

3.1 Incomplete information

3.1.1 Behavioral equilibrium

Under incomplete information, an individual cannot observe the type of its interaction partner. It can still have information about the matching distribution in the pairwise interaction, i.e., the probability that the partner belongs

[6] This entails no loss of generality, and simply depends on how one defines the strategy set. For example, if the interaction at hand is a public goods game, then let y denote the opponent's contribution to the public good. If the interaction at hand is a common pool resource game, then let y denote the inverse of the size of the opponent's extraction.

to the same lineage. One-shot interactions between perfect strangers are examples of this kind of interaction, as are interactions between family members when the only available information is the degree of kinship between interaction partners. We assume that individuals hold the belief that the probability of being matched with an individual from the same lineage is given by $r(\tau,\theta)$.[7] In this setting, an individual's type is some value of the Kantian coefficient, and the type space is the interval $[0,1]$. Given these assumptions, an individual can condition its strategy only on the strategy that it expects its neighbor to use, given the belief on the matching distribution. Since any individual uses the same strategy whether the neighbor has the same or a different type, we simplify the notation by setting $x_{\rm s}^* = x_{\rm d}^* = x^*$ and $y_{\rm s}^* = y_{\rm d}^* = y^*$, where x^* is the equilibrium strategy of mutants and y^* that of residents. A strategy pair (x^*, y^*) is a (Bayesian) Nash equilibrium if (a) y^* is a preferred strategy for a resident, given that other residents use strategy y^*; and (b) x^* is a preferred strategy for a mutant, given that residents use y^* and the other mutants use x^*, and given that the mutant applies the belief that the probability of being matched with another mutant is

$$r(\tau,\theta) = \tilde{r}(x^*(\tau,\theta), y^*(\theta)), \tag{12}$$

where on the right-hand side relatedness is expressed in terms of the equilibrium strategies of mutant and resident individuals. Formally $\tilde{r} : \mathcal{X}^2 \to [0,1]$ so that $\tilde{r}(x,y)$ is the relatedness of a mutant towards a random group member when mutants play strategy x and residents play strategy y [a concrete example thereof is by setting $x_{\rm s}^* = x_{\rm d}^* = x^*$ and $y_{\rm s}^* = y_{\rm d}^* = y^*$ into the right-hand side of eq. (B-k) of Box 2]. Thus, (x^*, y^*) solves the fixed point system

$$\begin{cases} y^* \in \arg\max_{y\in\mathcal{X}} \ u_\theta\left(y, y^* \mid y^*\right) \\ x^* \in \arg\max_{x\in\mathcal{X}} \ [1-\tilde{r}(x^*,y^*)]u_\tau\left(x, y^* \mid y^*\right) + \tilde{r}(x^*,y^*)u_\tau\left(x, x^* \mid y^*\right), \end{cases} \tag{13}$$

which is fully in line with the model in Alger et al. (2020, eq. 1 and eq. 5) and where the utility functions u_θ and u_τ are defined in eq. (11). The first line ensures that the strategy y^* maximizes the expected utility of a resident, given that any individual it will be matched with is a resident, who uses strategy y^*.[8] The second line ensures that the strategy x^* maximizes the expected utility of a mutant, given that any resident it will be matched with uses strategy y^* and any mutant it will be matched with uses strategy x^*. The behavioral fixed point (13) defines the behavioral mechanisms (8)–(10), which here satisfy $M_{\rm d}(\tau,\theta,x^*,y^*) = M_{\rm s}(\tau,\theta,x^*) = 0$ and $R_{\rm d}(\theta,y^*,x^*) = R_{\rm s}(\theta,y^*) = 0$, since an individual's strategy choice cannot be conditioned on the interactant's type.

Under our mathematical assumptions, the (assumed unique) equilibrium pair of strategies satisfies the necessary first-order conditions for the maximization problems in eq. (13):

$$\begin{cases} \left.\frac{\partial u_\theta(y,y^*|y^*)}{\partial y}\right|_{y=y^*} = 0 \\ [1-\tilde{r}(x^*,y^*)]\left.\frac{\partial u_\tau(x,y^*|y^*)}{\partial x}\right|_{x=x^*} + \tilde{r}(x^*,y^*)\left.\frac{\partial u_\tau(x,x^*|y^*)}{\partial x}\right|_{x=x^*} = 0. \end{cases} \tag{14}$$

[7] This belief is correct in the sense that a randomly drawn mutant in the lineage started by the initial mutant, faces the probability $r(\tau,\theta)$ of being matched with another mutant. However, in any given time period this probability in fact depends on the exact population composition in the preceding period.

[8] This is different from the model in Alger and Weibull (2013), where a resident faces a positive probability of being matched with a mutant. In the model we use, the mutant trait appears initially in one single individual, and uninvadability amounts to requiring that the lineage that this initial mutant creates goes extinct in finite time. During the time the mutant lineage is around there is a finite number of mutants, and hence residents face a zero probability of being matched with a mutant in this infinitely large population. See Alger et al. (2020) for a formal explanation.

Using eq. (11), these equations become

$$\begin{cases} \left[(1-\theta)\frac{\partial w(y,y^*,y^*)}{\partial y} + \theta\frac{\partial w(y,y,y^*)}{\partial y}\right]_{y=y^*} = 0 \\ \left[(1-\tau)\left([1-\tilde{r}(x^*,y^*)]\frac{\partial w(x,y^*,y^*)}{\partial x} + \tilde{r}(x^*,y^*)\frac{\partial w(x,x^*,y^*)}{\partial x}\right) + \tau\frac{\partial w(x,x,y^*)}{\partial x}\right]_{x=x^*} = 0. \end{cases} \tag{15}$$

Turning now to the necessary second-order conditions (for $(x^*(\tau,\theta), y^*(\theta))$ defined by (15) to be maxima rather than minima), these are:

$$\begin{cases} \left.\frac{\partial^2 u_\theta(y,y^*|y^*)^2}{\partial y^2}\right|_{y=y^*} \leq 0 \\ [1-\tilde{r}(x^*,y^*)]\left.\frac{\partial^2 u_\tau(x,y^*|y^*)}{\partial x^2}\right|_{x=x^*} + \tilde{r}(x^*,y^*)\left.\frac{\partial^2 u_\tau(x,x^*|y^*)}{\partial x^2}\right|_{x=x^*} \leq 0. \end{cases} \tag{16}$$

Our analysis is restricted to settings in which these inequalities hold, since we assume equilibrium existence.

Because of strict concavity of w, the following inequality holds:

$$K(\theta) = \left.\frac{\partial^2 u_\theta\,(y, y^* \mid y^*)}{\partial y^2}\right|_{y=y^*} = (1-\theta)\frac{\partial^2 w(y,y^*,y^*)}{\partial y^2} + \theta\left.\frac{\partial^2 w(y,y,y^*)}{\partial y^2}\right|_{y=y^*} < 0. \tag{17}$$

This inequality will be used to evaluate how the mutant's equilibrium strategy would change if the mutant trait value changed. To see this, by applying the implicit function theorem, one obtains by totally differentiating the second line of eq. (15) with respect to τ and solving the resulting linear equation for $\partial x^*(\tau,\theta)/\partial\tau$:

$$\left.\frac{\partial x^*(\tau,\theta)}{\partial\tau}\right|_{\tau=\theta} = -\left.\frac{\frac{\partial w(x^*,y,y^*)}{\partial y}}{K(\theta) + r(\theta,\theta)(1-\theta)\frac{\partial^2 w(x,y,y^*)}{\partial x\partial y}}\right|_{\substack{x=y=y^* \\ x^*=y^*}}. \tag{18}$$

This (local) *mutant behavioral perturbation* will be seen to play a central role in the evolutionary analysis. Defining

$$C(\theta) = \left.\frac{\partial^2 w(x,y,y^*)}{\partial x\partial y}\right|_{x=y=y^*=y^*(\theta)} \tag{19}$$

and recalling our assumptions on w (see the first paragraph of this Section), we conclude from this expression that:

(i) $\left.\frac{\partial x^*(\tau,\theta)}{\partial\tau}\right|_{\tau=\theta} > 0$ if $r(\theta,\theta)(1-\theta)C(\theta) < |K(\theta)|$, which is true if w is such that the strategies are strategic substitutes, strategically neutral, or moderately complementary;

(ii) $\left.\frac{\partial x^*(\tau,\theta)}{\partial\tau}\right|_{\tau=\theta} < 0$ if $r(\theta,\theta)(1-\theta)C(\theta) > |K(\theta)|$, which is true if w is such that the strategies are strongly complementary.

3.1.2 Evolutionary equilibrium

Turning now to the analysis of selection on the Kantian coefficient, we obtain that invasion fitness under the present scenario (where recall that $x^*_{\rm s} = x^*_{\rm d} = x^*$ and $y^*_{\rm s} = y^*_{\rm d} = y^*$) writes

$$W(\tau,\theta) = [1-r(\tau,\theta)]\,w(x^*(\tau,\theta), y^*(\theta), y^*(\theta)) + r(\tau,\theta)\,w(x^*(\tau,\theta), x^*(\tau,\theta), y^*(\theta)), \tag{20}$$

which is differentiable. Then, substituting eq. (20) into $S(\theta) = \partial W(\tau,\theta)/\partial\tau|_{\tau=\theta}$, the selection gradient is

$$S(\theta) = \frac{\partial x^*(\tau,\theta)}{\partial\tau}\left[\frac{\partial w(x,y^*,y^*)}{\partial x} + r(\theta,\theta)\frac{\partial w(x^*,x,y^*)}{\partial x}\right]\Bigg|_{x=x^*=y^*=y^*(\theta)}^{\tau=\theta}, \tag{21}$$

because the term multiplying $\partial r(\tau,\theta)/\partial\tau$ is $w(x^*(\tau,\theta),x^*(\tau,\theta),y^*(\theta))|_{\tau=\theta} - w(x^*(\tau,\theta),y^*(\theta),y^*(\theta))|_{\tau=\theta} = 0$. The expression in eq. (21) can further be simplified by noting that the second equation of eq. (15) reduces to $\partial w(x,y^*,y^*)/\partial x = -\theta\partial w(y^*,x,y^*)/\partial x$ when $\tau=\theta$ (and thus $x^*=y^*$). Hence:

$$S(\theta) = [r(\theta,\theta)-\theta]\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,x,y^*)}{\partial x}\bigg|_{\substack{\tau=\theta \\ x=x^*=y^*=y^*(\theta)}}. \tag{22}$$

Since (by assumption) $\partial w(x,y,y^*)/\partial y \neq 0$ for all $(x,y,y^*)\in X^3$, which also implies that $\partial x^*(\tau,\theta)/\partial\tau \neq 0$ (see eq. (18)), this equation shows that the unique singular trait value is

$$\theta^* = r(\theta^*,\theta^*). \tag{23}$$

But is θ^* convergence stable and uninvadable?

Let us first consider convergence stability, by determining whether the Jacobian

$$J(\theta^*) = \frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}\bigg|_{\theta=r} = -\left(1-\frac{\mathrm{d}r(\theta,\theta)}{\mathrm{d}\theta}\right)\left[\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,y,y^*)}{\partial y}\right]_{\substack{\tau=\theta=\theta^* \\ x^*=y=y^*=y^*(\theta)}} \tag{24}$$

is strictly negative. Since $\partial w(x^*,y,y^*)/\partial y > 0$, we immediately obtain that $\theta^* = r(\theta^*,\theta^*)$ is convergence stable if and only if

$$\left(1-\frac{\mathrm{d}r(\theta,\theta)}{\mathrm{d}\theta}\right)\frac{\partial x^*(\tau,\theta)}{\partial\tau}\bigg|_{\substack{\tau=\theta=\theta^* \\ x^*=y=y^*=y^*(\theta)}} > 0. \tag{25}$$

Using results derived above, we further conclude that $\theta^* = r(\theta^*)$ is convergence stable if and only if either $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta < 1$ and $r(\theta,\theta)[1-r(\theta,\theta)]C(\theta) < |K(\theta)|$, or $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta > 1$ and $r(\theta,\theta)[1-r(\theta,\theta)]C(\theta) > |K(\theta)|$.

What about local uninvadability? To ascertain this, we examine whether the Hessian is strictly negative. Given that $\partial w(x,y^*,y^*)/\partial x = -\theta\partial w(y^*,x,y^*)/\partial x$ when $\tau=\theta$ (as noted already above), we obtain:

$$H(\theta^*) = \frac{\partial^2 W(\tau,\theta)}{\partial\tau^2}\bigg|_{\tau=\theta} = \left[\left(\frac{\partial x^*(\tau,\theta)}{\partial\tau}\right)^2 K\left(r(\theta^*)\right) + 2\frac{\partial r(\tau,\theta)}{\partial\tau}\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,y,y^*)}{\partial y}\right]_{\substack{\tau=\theta=\theta^* \\ x^*=y=y^*=y^*(\theta)}}. \tag{26}$$

Since $K\left(r(\theta^*)\right) < 0$, the first term is strictly negative. Hence, a sufficient condition for $\theta^* = r(\theta^*,\theta^*)$ to be (locally) uninvadable is that the local perturbation of relatedness, $\partial r(\tau,\theta)/\partial\tau$, be nil. The relatedness perturbation can be different from zero, however (for example, see eq. (B-j) for the expression of $r(\theta,\tau)$ for a Moran process), and its sign typically depends on demographic and interaction assumptions in non-trivial ways. Moreover, it does not involve second-order derivatives of individual fitness (Mullon et al., 2016), and thus does not vary systematically according to the strategic substitutability or complementarity of the strategies. Hence, in settings where behavior affects relatedness $\partial r(\tau,\theta)/\partial\tau \neq 0$, it is challenging to identify general conditions that would guarantee that $J(\theta^*) < 0$ and $H(\theta^*) < 0$. Yet, it is known that in certain settings (summarized below) both $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta$ and $\partial r(\tau,\theta)/\partial\tau$ are negligible. We refer to this as *weak trait effects on relatedness*. We can then summarize sufficient conditions for $\theta = r(\theta,\theta)$ to be evolutionarily stable as follows.

Result 1. *When interactions take place under incompete information, the Kantian coefficient equal to the neutral relatedness, $\theta^* = r(\theta^*,\theta^*)$, is the unique singular trait value. It is both an evolutionary attractor (convergence stable) and locally uninvadable if:*

(i) either the strategies in the interaction are strategic substitutes, strategically neutral, or moderately complementary, i.e., $r(\theta,\theta)[1-r(\theta,\theta)]C(\theta) < |K(\theta)|$*, and the trait effects on relatedness are sufficiently weak, i.e.,* $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta < 1$ *and* $\partial r(\tau,\theta)/\partial\tau \approx 0$*;*

(ii) or the strategies in the interaction are strongly complementary, i.e., $r(\theta,\theta)[1-r(\theta,\theta)C(\theta) > |K(\theta)|$*, and the effect of the resident equilibrium strategy on relatedness is strong, i.e.,* $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta > 1$*, but the relatedness perturbation is weak, i.e.,* $\partial r(\tau,\theta)/\partial\tau \approx 0$*.*

While the conditions for uninvadability are consistent with the results of Alger et al. (2020),[9] our analysis reinforces those results by identifying sufficient conditions for the partly Kantian preference $\theta = r(\theta,\theta)$ to be also convergence stable.

Interestingly, many biological scenarios do exhibit weak, or even nil, trait effects on relatedness.[10] First, in family-structured populations, which cover a large class of interactions (e.g., parent-offspring interactions, interactions beween sibling or cousins, etc...), relatedness is independent of the types ($\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta = \partial r(\tau,\theta)/\partial\tau = 0$). Second, relatedness is also independent of the types in spatially-structured populations when selection is weak in the sense that the strategies in the interaction affect fitness only marginally (see, e.g., Alger et al., 2020). Such independence can extend to cases where effects are not so marginal because when the migration probability is exogenous, both $\mathrm{d}r(\theta,\theta)/d\theta$ and $\partial r(\tau,\theta)/\partial\tau$ tend to be negligible for several games (Wakano and Lehmann, 2014; Mullon et al., 2016). Finally, for certain demographic processes, like the Moran process when behavior affects only reproduction, one has $\mathrm{d}r(\theta,\theta)/\,\mathrm{d}\theta = 0$ and $\partial r(\tau,\theta)/\partial\tau = 0$ (Mullon et al., 2016), but, as implied by eq. (B-k) of Box 2, the relatedness perturbation is non-zero in the Moran process when behavior affects survival.

3.2 Complete information with incomplete plasticity

3.2.1 Behavioral equilibrium

Under complete information individuals have information about the type of their interaction partner, but an individual's preferences cannot be conditioned on that information: this is what we mean by incomplete plasticity. An individual's type is thus again some value of the Kantian coefficient, and the type space is the interval $[0,1]$. Because individuals can observe the type composition of their group, whenever the mutant type differs from the resident type, the distinction between a mutant's equilibrium strategies x^*_{d} and x^*_{s}, as well as between a resident's

[9] This may not be immediately apparent, for the results in Alger et al. (2020) (see Propositions 1 and 2) state as a necessary and sufficient condition for a utility function to be uninvadable, that the equilibrium strategy in a population where all individuals have this utility function be an uninvadable strategy (i.e., uninvadable in a setting where the set of traits is the set of strategies, a setting that Alger et al. (2020) call strategy evolution). One can check that the condition for $\theta^* = r(\theta^*,\theta^*)$ to be uninvadable in our setting, i.e., $H(\theta^*) < 0$ (see eq. (26)), coincides with the condition for the equilibrium strategy $y^*(\theta^*)$ to be an uninvadable strategy under strategy evolution. This is so because owing to eq. (12), $\partial r(\tau,\theta)/\partial\tau = (\partial\tilde{r}(x,y^*)/\partial x)(\partial x^*(\tau,\theta)/\partial\tau)$, whereby eq. (26) becomes $H(\theta^*) = (\partial x^*(\tau,\theta)/\partial\tau)^2\left[K\left(r(\theta^*)\right) + 2(\partial\tilde{r}(x,y^*)/\partial x)(\partial w(x^*,y,y^*)/\partial y)\right]$ at $\tau=\theta$; and under strategy evolution the invasion fitness of mutant type x in a monomorphic resident population y is $W(x,y) = [1-\tilde{r}(x,y)]\,w(x,y,y) + \tilde{r}(x,y)\,w(x,x,y)$ and $\partial^2 W(x,y)/\partial x^2\big|_{x=y}$ corresponds to the term in square brackets in $H(\theta^*)$.

[10] Such independence was originally assumed in evolutionary game theory models with assortative interactions (Hines and Maynard Smith, 1978; Maynard Smith, 1982) and later used in preference evolution models (Bergstrom, 1995; Alger and Weibull, 2013).

equilibrium strategies $y_{\rm d}^*$ and $y_{\rm s}^*$, is relevant, as per the behavioral mechanisms (8)–(10). Hence, the equilibrium strategy $y_{\rm s}^*$ used by each resident in an interaction with another resident satisfies

$$y_{\rm s}^* \in \arg\max_{y\in\mathcal{X}} u_\theta\left(y, y_{\rm s}^* \mid y_{\rm s}^*\right), \tag{27}$$

the equilibrium strategy $x_{\rm s}^*$ used by each mutant in an interaction with another mutant satisfies

$$x_{\rm s}^* \in \arg\max_{x\in\mathcal{X}} u_\tau\left(x, x_{\rm s}^* \mid y_{\rm s}^*\right), \tag{28}$$

and the equilibrium pair of strategies $(x_{\rm d}^*, y_{\rm d}^*)$ used by a mutant and a resident, respectively, in a mutant-resident interaction solves the fixed point system

$$\begin{cases} y_{\rm d}^* \in \arg\max_{y\in\mathcal{X}}\ u_\theta\left(y, x_{\rm d}^* \mid y_{\rm s}^*\right) \\ x_{\rm d}^* \in \arg\max_{x\in\mathcal{X}}\ u_\tau\left(x, y_{\rm d}^* \mid y_{\rm s}^*\right). \end{cases} \tag{29}$$

The fixed point equations (27)–(29) define the behavioral mechanisms (8)–(10) under complete information with incomplete plasticity. Note that if $\tau = \theta$ in eq. (29), then $x_{\rm d}^*(\theta,\theta) = y_{\rm d}^*(\theta,\theta)$ owing to the strict concavity of w, which implies that to each strategy played by the opponent there exists a unique best response. Hence, if $\tau = \theta$,

$$x_{\rm d}^*(\theta,\theta) = y_{\rm d}^*(\theta,\theta) = x_{\rm s}^*(\theta) = y_{\rm s}^*(\theta). \tag{30}$$

As in the incomplete information scenario, in the evolutionary analysis we use the expressions that capture how the equilibrium strategies are modified by marginal changes in the mutant trait and the resident trait. To obtain these behavioral perturbations, we first write the necessary first-order conditions for $(x_{\rm d}^*, y_{\rm d}^*)$ to be a Nash equilibrium:

$$\begin{cases} \left.\frac{\partial u_\tau(x, y_{\rm d}^* | y_{\rm s}^*)}{\partial x}\right|_{x=x_{\rm d}^*} = \left[(1-\tau)\frac{\partial w(x, y_{\rm d}^*, y_{\rm s}^*)}{\partial x} + \tau\frac{\partial w(x,x,y_{\rm s}^*)}{\partial x}\right]_{x=x_{\rm d}^*} = 0 \\ \left.\frac{\partial u_\theta(y, x_{\rm d}^* | y_{\rm s}^*)}{\partial y}\right|_{y=y_{\rm d}^*} = \left[(1-\theta)\frac{\partial w(y, x_{\rm d}^*, y_{\rm s}^*)}{\partial y} + \theta\frac{\partial w(y,y,y_{\rm s}^*)}{\partial y}\right]_{y=y_{\rm d}^*} = 0. \end{cases} \tag{31}$$

Therein, the monomorphic resident behavioral equilibrium $y_{\rm s}^*$ solves

$$\left.\frac{\partial u_\theta\left(y, y_{\rm s}^* \mid y_{\rm s}^*\right)}{\partial y}\right|_{y=y_{\rm s}^*} = \left[(1-\theta)\frac{\partial w(y, y_{\rm s}^*, y_{\rm s}^*)}{\partial y} + \theta\frac{\partial w(y,y,y_{\rm s}^*)}{\partial y}\right]_{y=y_{\rm s}^*} = 0. \tag{32}$$

Since u_θ is strictly concave, the second-order partial derivative of u_θ, evaluated at $y_{\rm s}^*(\theta) = y_{\rm d}^*(\theta,\theta)$, is strictly negative:

$$\tilde{K}(\theta) = \left.\frac{\partial^2 u_\theta\left(y, y_{\rm s}^* \mid y_{\rm s}^*\right)}{\partial y^2}\right|_{y=y_{\rm s}^*} = \left[(1-\theta)\frac{\partial^2 w(y, y_{\rm s}^*, y_{\rm s}^*)}{\partial y^2} + \theta\frac{\partial^2 w(y,y,y_{\rm s}^*)}{\partial y^2}\right]_{y=y_{\rm s}^*=y_{\rm d}^*(\theta,\theta)} < 0. \tag{33}$$

The system of equations in eq. (31) together implicitly define $x_{\rm d}^*$ and $y_{\rm d}^*$ as functions of τ and θ. Applying the implicit function theorem, we obtain the following expressions for the behavioral perturbation of the equilibrium strategy of a mutant and of a resident with respect to the mutant trait value, evaluated locally at $\tau = \theta$:

$$\left.\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial \tau}\right|_{\tau=\theta} = -\left.\frac{\frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\tilde{K}(\theta)}{\left(\tilde{K}(\theta) + (1-\theta)\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}\right)\left(\tilde{K}(\theta) - (1-\theta)\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}\right)}\right|_{x=y=y_{\rm d}^*=y_{\rm d}^*(\theta,\theta)} \tag{34}$$

$$\left.\frac{\partial y_{\rm d}^*(\theta,\tau)}{\partial \tau}\right|_{\tau=\theta} = \left.\frac{(1-\theta)\frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}}{\left(\tilde{K}(\theta) + (1-\theta)\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}\right)\left(\tilde{K}(\theta) - (1-\theta)\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}\right)}\right|_{x=y=y_{\rm d}^*=y_{\rm d}^*(\theta,\theta)}. \tag{35}$$

In the evolutionary analysis, it is the ratio of the resident's to the mutant's behavioral perturbation that will matter:

$$\rho(\theta) = \left. \frac{\frac{\partial y_{\rm d}^*(\theta,\tau)}{\partial \tau}}{\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial \tau}} \right|_{\tau=\theta}. \tag{36}$$

This is well defined, since the assumption $\partial w(y_{\rm d}^*, y, y_{\rm d}^*)/\partial y \neq 0$ implies $\partial x_{\rm d}^*(\tau,\theta)/\partial \tau \neq 0$, which means that a mutant always alters its equilibrium strategy if the mutant trait value were to change. Because $\rho(\theta)$ measures the extent to which an individual's neighbour's strategy varies with own strategy variation, we follow previous terminology and refer to $\rho(\theta)$ as the *response coefficient* (Akçay and Van Cleve, 2012). Inserting eq. (34) and eq. (35) into eq. (36), we obtain

$$\rho(\theta) = \left. -\frac{(1-\theta)\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x \partial y}}{\tilde{K}(\theta)} \right|_{x=y=y_{\rm d}^*(\theta,\theta)}, \tag{37}$$

implying that the response coefficient has the same sign as $\partial^2 w(x, y, y_{\rm d}^*)/\partial x \partial y$, and this will play a role in the analysis of selection on the Kantian coefficient, to which we now turn.

3.2.2 Evolutionary equilibrium

To begin, note that eq. (30) implies that we can write invasion fitness (2) as follows:

$$W(\tau,\theta) = [1 - r(\tau,\theta)]\, w(x_{\rm d}^*(\tau,\theta), y_{\rm d}^*(\theta,\tau), y_{\rm d}^*(\theta,\theta)) + r(\tau,\theta)\, w(x_{\rm d}^*(\tau,\tau), x_{\rm d}^*(\tau,\tau), y_{\rm d}^*(\theta,\theta)), \tag{38}$$

which is differentiable. Substituting this into the selection gradient $S(\theta) = \partial W(\tau,\theta)/\partial \tau|_{\tau=\theta}$, and simplifying yields (since the term multiplying $\partial r(\tau,\theta)/\partial \tau$ is $w(x_{\rm d}^*(\tau,\tau), x_{\rm d}^*(\tau,\tau), y_{\rm d}^*(\theta,\theta))|_{\tau=\theta} - w(x_{\rm d}^*(\tau,\theta), y_{\rm d}^*(\theta,\tau), y_{\rm d}^*(\theta,\theta))|_{\tau=\theta} = 0$):

$$S(\theta) = \left[\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial \tau} + r(\theta,\theta)\frac{\partial y_{\rm d}^*(\theta,\tau)}{\partial \tau} \right) \frac{\partial w(x, y_{\rm d}^*, y_{\rm d}^*)}{\partial x} + \left(r(\theta,\theta)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial \tau} + \frac{\partial y_{\rm d}^*(\theta,\tau)}{\partial \tau} \right) \frac{\partial w(x_{\rm d}^*, y, y_{\rm d}^*)}{\partial y} \right]_{\substack{\tau=\theta \\ x=y=x_{\rm d}^*=y_{\rm d}^* \\ y_{\rm d}^*=y_{\rm d}^*(\theta,\theta)}}. \tag{39}$$

Using eq. (32), we can replace $\partial w(x, y_{\rm d}^*, y_{\rm d}^*)/\partial x$ by $-\theta \partial w(x_{\rm d}^*, y, y_{\rm d}^*)/\partial y$, to obtain

$$S(\theta) = \frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial \tau} \frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y} \left([r(\theta,\theta) - \theta] + [1 - \theta r(\theta,\theta)]\, \rho(\theta) \right) \Bigg|_{\substack{\tau=\theta \\ y=y_{\rm d}^*=y_{\rm d}^*(\theta,\theta)}}. \tag{40}$$

Comparing eq. (40) to the selection gradient under incomplete information (see eq. (22)), we see that if the response coefficient is nil, i.e., if $\rho(\theta) = 0$, the two selection gradients are identical, and $\theta = r(\theta,\theta)$ is then the unique singularity. This is not surprising since under incomplete information changes in the mutant trait value has no effect on the resident's equilibrium strategy. We further observe that when $\theta = 1$, $\rho(\theta) = 0$ and using eq. (34) in eq. (40), the selection gradient at $\theta = 1$ is

$$S(1) = \frac{(1 - r(1,1))}{\tilde{K}(1)} \left(\frac{\partial^2 w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y^2} \right)^2 \Bigg|_{y=y_{\rm d}^*=y_{\rm d}^*(\theta,\theta)} < 0. \tag{41}$$

Since $\tilde{K}(1) < 0$ and $r(\theta,\theta) < 1$ for all $\theta \in [0,1]$, we obtain $S(1) < 0$, which implies that $\theta = 1$ is always counter-selected, and can neither be convergence stable nor uninvadable. By contrast, nothing allows to rule out that

$\theta = 0$ could be convergence stable and/or uninvadable. More generally, since $\partial w(x_{\rm d}^*, y, y_{\rm d}^*)/\partial y \neq 0$ (by assumption), eq. (40) implies that $S(\theta) = 0$ if and only if

$$\theta = r(\theta, \theta) + [1 - \theta r(\theta, \theta)]\rho(\theta). \tag{42}$$

Let $\tilde{\theta}$ denote a solution to this equation. Since $r(\theta, \theta) < 1$ for all $\theta \in [0, 1]$, so that $1 - \theta r(\theta, \theta) > 0$ for any $\theta \in [0, 1]$, it follows immediately from eq. (42) that $\tilde{\theta} = r(\tilde{\theta})$ if $\rho(\tilde{\theta}) = 0$, $\tilde{\theta} > r(\tilde{\theta})$ if $\rho(\tilde{\theta}) > 0$, and $\tilde{\theta} < r(\tilde{\theta})$ if $\rho(\tilde{\theta}) < 0$. Recalling that the sign of $\rho(\theta)$ depends on the sign of $\partial^2 w(x, y, y_{\rm d}^*)/\partial x \partial y$ (see eq. (37)), and that we restrict the Kantian coefficient to take values between 0 and 1, the following result obtains.

Result 2. *Let θ^* denote a singularity under complete information and incomplete plasticity. Then:*

(i) $\theta^* = 0$ if $\tilde{\theta} \leq 0$, which requires w to be such that strategies are strategic substitutes or strategically neutral;

(ii) $\theta^* = \tilde{\theta}$ if $\tilde{\theta} \in (0, 1)$ and if this is the case, then $\theta^* = [r(\theta^*, \theta^*) + \rho(\theta^*)]/[1 + \rho(\theta^*) r(\theta^*, \theta^*)]$. In particular, $\theta^* = r(\theta^*, \theta^*)$ if w is such that strategies are strategically neutral, while $\theta^* < r(\theta^*, \theta^*)$ (resp. $\theta^* > r(\theta^*, \theta^*)$) if w is such that strategies are strategic substitutes (resp. complements), and $\theta^* = \rho(\theta^*)$ if $r(\theta^*, \theta^*) = 0$.

By contrast to the incomplete information setting where the Kantian coefficient must coincide with the coefficient of relatedness, here it can be either larger or smaller. Moreover, a singular Kantian coefficient can in principle take any value in the range $[0, 1)$ depending on demographic and behavioral parameters. Interestingly, eq. (42) along with eq. (36) is identical to the corresponding equation in the model of Alger and Weibull (2012) (see their eq. (29)), wherein they examine the class of other-regarding utility functions whereby an individual may attach some evolving weight $\alpha \in (-1, 1)$ to the interactant's individual fitness.[11] In spite of this difference, Theorem 1 of this previous work also establishes that whether the exact value of the evolving weight α exceeds or falls short of relatedness depends on whether the fitness function exhibits strategic substitutability, complementarity, or neutrality.

We now examine whether type θ^* of Result 2 is convergence stable and uninvadable. Due to the complexity of the expressions for the Jacobian $J(\theta^*)$ and the Hessian $H(\theta^*)$ at θ^* solving $S(\theta^*) = 0$, presented in Appendix A, we were unable to reach generic answers to these questions, and further assumptions may be needed to reach more definite results. However, we verify that convergence stability and uninvadability can obtain, by resorting to an illustrating example. Consider a Moran demographic process (i.e. individual fitness takes the form of eq. (B-i)) with constant death rate μ and juvenile survival probability s, and that individual face an a pairwise interaction such that their expected fecundity (number of offspring produced at stage (b) of the life cycle of section 2.1) is linear-quadratic in the two players' actions:

$$f(x, y) = 1 + ax - bxy - cx^2 \tag{43}$$

for parameters $a, b, c \in \mathbb{R}$. Then, substituting eq. (43) into individual fitness (B-i), we can evaluate the selection gradient (40), the Jacobian (A-3) and the Hessian (A-5) coefficients. Even for this simple example, eq. (42) is a quadratic function that cannot be solved explicitly and so we analyse the selection gradient numerically. Fig. (1) displays how for fixed but different values of the backward migration rate $m_{\rm b}$ (eq. B-l), θ^* varies when b is varied. Fig. (1) shows that by depending on b, the Kantian coefficient takes a value above or below that of relatedness.

[11] Because the behavioral perturbations of other-regarding utility functions typically differ from the ones with the partially Kantian utility function, a singular α will typically differ from the singular Kantian coefficient, however.

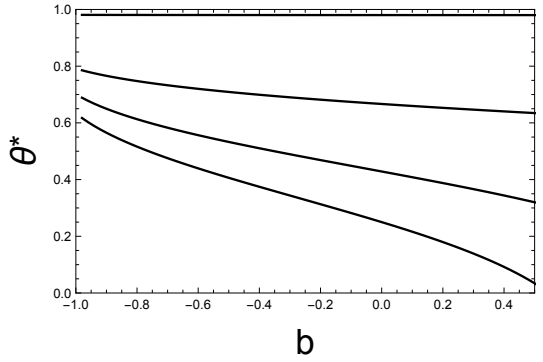

Figure 1: Each curve shows, for the Moran process analyzed in Box 2 with individual fitness (B-i), the singular Kantian coefficient θ^* under complete information and incomplete plasticity, for the linear quadratric fecundity function (43), as a function of parameter b in that function for $a = 0.1$ and $c = 1$. Each of the four lines corresponds to a different value of the "backward migration probability", which depends on the exogenously given migration probability m (see eq. (B-l) and the description following it). Starting from the top, the first line, where the Kantian coefficient remains essentially constant at $\theta^* = 0.98$ is for $m_\mathrm{b} = 0.01$ whereby $r = (1 - m_\mathrm{b})/(1 + m_\mathrm{b}) \approx 0.98$; the second line is for $m_\mathrm{b} = 0.2$ whereby $r = (1 - m_\mathrm{b})/(1 + m_\mathrm{b}) \approx 0.66$; the third line for $m_\mathrm{b} = 0.4$ whereby $r = (1 - m_\mathrm{b})/(1 + m_\mathrm{b}) \approx 0.42$; and the last line, where the Kantian coefficient varies over the range $[0, 0.6]$, is for $m_\mathrm{b} = 0.6$ whereby $r = (1 - m_\mathrm{b})/(1 + m_\mathrm{b}) = 0.25$. By computing the Jacobian (A-3) and the Hessian (A-5) coefficients at these values we checked that all these singular Kantian coefficients are indeed both convergence stable and uninvadable.

3.3 Complete information and plasticity

3.3.1 Behavioral equilibrium

The defining assumption of our complete information with complete plasticity scenario is that individuals are assumed to be able to both express different strategies conditionally on interacting with individuals having different types and to also use different preferences. Hence, the preferences applied in the interaction become state-specific on the interaction. Specifically, we assume that the type $\theta = (\theta_\mathrm{d}, \theta_\mathrm{s})$ of an individual is a two-dimensional quantitative trait ($\theta \in [0,1]^2$) such that θ_s parametrizes an individual's preference (still given by eq. (11)) when individuals in a pair have the same type and θ_d parametrizes an individual's preference when individuals in a pair have different types.

In terms of the equilibrium strategies, consider some resident type $\theta = (\theta_\mathrm{d}, \theta_\mathrm{s})$ and some mutant type $\tau = (\tau_\mathrm{d}, \tau_\mathrm{s})$ that is different from θ (either because $\theta_\mathrm{d} \neq \tau_\mathrm{d}$, or because $\theta_\mathrm{s} \neq \tau_\mathrm{s}$, or because both $\theta_\mathrm{d} \neq \tau_\mathrm{d}$ and $\theta_\mathrm{s} \neq \tau_\mathrm{s}$). Then, a resident applies the Kantian coefficient θ_s when interacting with another resident, in which case they both play the equilibrium strategy y_s^*, satisfying

$$y_\mathrm{s}^* \in \arg\max_{y \in \mathcal{X}} u_{\theta_\mathrm{s}}\left(y, y_\mathrm{s}^* \mid y_\mathrm{s}^*\right). \tag{44}$$

In mutant-mutant interactions both individuals apply the Kantian coefficient τ_s and they both use the equilibrium strategy x_s^*, satisfying

$$x_\mathrm{s}^* \in \arg\max_{x \in \mathcal{X}} u_{\tau_\mathrm{s}}\left(x, x_\mathrm{s}^* \mid y_\mathrm{s}^*\right). \tag{45}$$

Finally, in mutant-resident pairs, the mutant applies the Kantian coefficient τ_d while the resident applies the Kantian

coefficient $\theta_{\rm d}$, and the Nash equilibrium strategies $x_{\rm d}^*$ and $y_{\rm d}^*$ are best responses to each other according to these preferences:

$$\begin{cases} x_{\rm d}^* \in \arg\max_{x\in\mathcal{X}} \; u_{\tau_{\rm d}}(x, y_{\rm d}^* \mid y_{\rm s}^*) \\ y_{\rm d}^* \in \arg\max_{y\in\mathcal{X}} \; u_{\theta_{\rm d}}(y, x_{\rm d}^* \mid y_{\rm s}^*)\,. \end{cases} \tag{46}$$

These fixed point eqs (44)–(46) define the behavioral mechanisms (8)–(10) under complete information with complete plasticity.

3.3.2 Evolutionary equilibrium

We begin by noticing that in the complete information complete plasticity scenario, invasion fitness can be written

$$W(\tau,\theta) = [1-r(\tau,\theta)]\, w(x_{\rm d}^*(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y_{\rm d}^*(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})) + \, r(\tau,\theta) w(x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s}), x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})), \tag{47}$$

where the inclusion of $\theta_{\rm s}$ among the arguments of the equilibrium strategies emphasizes their dependence on the preference type that residents apply in resident-resident interactions (the utility functions in eqs (45)–(46) depend on $y_{\rm s}^*(\theta_{\rm s})$). This invasion fitness is not necessarily differentiable in τ because there may be a discrete jump in the equilibrium strategies at $\tau=\theta$, from $(x_{\rm d}^*, y_{\rm d}^*)$ to $(x_{\rm s}^*, y_{\rm s}^*) = (y_{\rm s}^*, y_{\rm s}^*)$. This precludes an application of eqs. (3)-(5) to each component of type $\tau = (\tau_{\rm d}, \tau_{\rm s})$. In spite of this challenge, we can show the existence of an uninvadable type, as stated in the following result, but we were unable to conclude on convergence stability.

Result 3. *The type $(\theta_{\rm s}, \theta_{\rm d}) = (1, \max\{\theta^*, 0\})$, where θ^* solves the equation $\theta^* = \rho(\theta^*)$, is uninvadable.*

Proof: Consider some resident type $\theta = (\theta_{\rm d}, \theta_{\rm s}) \in [0,1]^2$. A necessary and sufficient condition for it to be uninvadable is that invasion fitness evaluated at the mutant type $\tau = (\tau_{\rm d}, \tau_{\rm s})$ that maximizes its value, is smaller than 1. Define the following separate terms, which both appear in the invasion fitness expression in eq. (47):

$$W_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}) = w(x_{\rm d}^*(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y_{\rm d}^*(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})) \tag{48}$$

and

$$W_{\rm s}(\tau_{\rm s},\theta_{\rm s}) = w(x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s}), x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})). \tag{49}$$

Starting with the latter, consider the mutant type τ with $\tau_{\rm s} = 1$. Then $x_{\rm s}^*(\tau_{\rm s}|\theta_{\rm s})$ solves (see eq. (45))

$$\max_{x\in\mathcal{X}} w(x, x, y_{\rm s}^*). \tag{50}$$

Recall the necessary first-order condition for the equilibrium strategy in resident-resident interactions (see eq. (44)):

$$\left[(1-\theta_{\rm s})\frac{\partial w(y, y_{\rm s}^*, y_{\rm s}^*)}{\partial y} + \theta_{\rm s}\frac{\partial w(y, y, y_{\rm s}^*)}{\partial y}\right]_{y=y_{\rm s}^*} = 0. \tag{51}$$

By strict concavity of w, eq. (51) implies that $W_{\rm s}(\tau_{\rm s},\theta_{\rm s}) > w(y_{\rm s}^*, y_{\rm s}^*, y_{\rm s}^*)$ for any $\theta_{\rm s} \in [0,1)$, while $W_{\rm s}(\tau_{\rm s},\theta_{\rm s}) = w(y_{\rm s}^*, y_{\rm s}^*, y_{\rm s}^*)$ for $\theta_{\rm s} = 1$.

Turning now to the term in eq. (48), it is identical to invasion fitness in the complete information incomplete plasticity scenario with relatedness equal to 0 (see eq. (38)), except that here $y_{\rm s}^*$ does not depend on $\theta_{\rm d}$. Inserting

$r(\tau,\theta) = 0$ into eq. (42), and following the same reasoning as for Result 2, we immediately obtain that $\theta_{\rm d} = \max\{\theta^*, 0\}$, where θ^* solves the equation $\theta^* = \rho(\theta^*)$, is the value of $\theta_{\rm d}$ for which there exists no $\tau_{\rm d}$ such that the individual fitness of a mutant in a mutant-resident interaction, $w(x^*_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y^*_{\rm d}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), y^*_{\rm s}(\theta_{\rm s}))$, exceeds the individual fitness of a resident in a mutant-resident interaction, $w(y^*_{\rm d}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), x^*_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y^*_{\rm s}(\theta_{\rm s}))$. In fact, given $\theta_{\rm d} = \max\{\theta^*, 0\}$, strict concavity of w implies that the mutant obtains a strictly smaller fitness than the resident in a mutant-resident interaction unless $\tau_{\rm d} = \theta_{\rm d}$. In sum:

(i) if $\theta_{\rm s} = 1$, then $W_{\rm s}(\tau_{\rm s},\theta_{\rm s}) \leq W_{\rm s}(\theta_{\rm s},\theta_{\rm s}) = 1$ for any $\tau_{\rm s} \in [0,1]$, and the inequality is strict if and only if $\tau_{\rm s} \neq 1$;

(ii) if $(\theta_{\rm s},\theta_{\rm d}) = (1, \max\{\theta^*, 0\})$, where θ^* solves the equation $\theta^* = \rho(\theta^*)$, then $W_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})$ is strictly smaller than $w(y^*_{\rm d}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), x^*_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y^*_{\rm s}(\theta_{\rm s}))$ for any $\tau_{\rm d} \neq \theta_{\rm d}$; strict concavity of w further implies that $w(y^*_{\rm s}, y^*_{\rm s}, y^*_{\rm s}) = 1$ is strictly larger than $w(y^*_{\rm d}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}), x^*_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}), y^*_{\rm s}(\theta_{\rm s}))$.

Together with the fact that invasion fitness is a convex combination of $W_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})$ and $W_{\rm s}(\tau_{\rm s},\theta_{\rm s})$, these conclusions imply that whatever is the value of $r(\tau,\theta)$, if the resident type is $(\theta_{\rm s},\theta_{\rm d}) = (1, \max\{\theta^*, 0\})$, then $W(\tau,\theta) < 1$ for any $\tau \neq \theta$. **Q.E.D.**

4 Discussion

By investigating the evolution of semi-Kantian preferences under different informational and behavioral plasticity assumptions, we have extended the evolutionary viability analysis of this class of preferences. While we restricted attention to pairwise interactions in group-structured populations, and preferences being characterized by a single evolving quantitative trait, the Kantian coefficient, our model weaves together different threads of the literature and shows how long-term evolution concepts can be used to analyze preferences under gradual evolution. We obtained three main results on the convergence stability and uninvadability of the value of the Kantian coefficient.

First, when interacting individuals have no information about each other's Kantian coefficient and mutants hold beliefs about the probability of being matched with another mutant that are consistent with the average such probability for the lineage created by the initial mutant, we confirm that an uninvadable Kantian coefficient must equal the coefficient of relatedness (Alger and Weibull, 2013; Alger et al., 2020). But instead of considering the set of possible utility functions to be the set of all continuous functions as done in this previous work, we focus on the more restricted setting where utility functions are parametrized by a single quantitative trait. This allows us to cover not only uninvadability in a complementary and less abstract way to that in the previous work but also to cover convergence stability. In Result 1 we show that the Kantian coefficient equal to the coefficient of relatedness is both convergence stable and uninvadable when trait effects on relatedness are sufficiently weak. Thus, we characterize conditions where gradual evolution drives preferences to induce individuals to behave according to Hamilton's (marginal) rule at the strategy level. A relevant avenue for future research for preference evolution under incomplete information is to consider more realistic demographic scenarios of class structured population (e.g., by sex, age, or stage).

Second, when interacting individuals can observe each other's type, but each individual has the same preferences regardless of the other's type, we showed that an uninvadable value of the Kantian coefficient can exceed, fall short of,

or equal the coefficient of relatedness. Moreover, we showed that the sign of the discrepancy is determined by whether an individual's equilibrium strategy is correlated positively, negatively, or not at all with the opponent's equilibrium strategy. This response coefficient in turn depends on an easily distinguishable property of the individual fitness function. Thus, gradual evolution now drives preferences to induce individuals to behave according to a context-specific Kantian coefficient, which combines both the relatedness and the response coefficients. This result is fully in line with previous models under complete information and incomplete plasticity, which have all considered other parametric classes of preferences than the one we examine (e.g. Bester and Güth, 1998; Bolle, 2000; Possajennikov, 2000; Heifetz et al., 2007b,a; Akçay and Van Cleve, 2009 for models without relatedness, and Alger, 2010; Alger and Weibull, 2010, 2012; Akçay and Van Cleve, 2012 for models with relatedness. The dependence of an uninvadable value of the Kantian coefficient on the response coefficient stems from the commitment to a particular behavioral response that an individual's preferences induces (when its preferences cannot be conditioned on the opponent's preference type, like in the complete plasticity scenario). By being observable, a mutant's preference type can thus induce a change in the resident's strategy, compared to when the resident interacts with another resident, an effect that is absent under incomplete information. Since the magnitude of this effect depends on the specifics of the fitness function, an uninvadable Kantian coefficient thus depends on this as well. Although we established clear necessary conditions for a Kantian coefficient value to be uninvadable, we did not succeed in identifying simple general sufficient conditions, neither for uninvadability nor for convergence stability. In particular, we cannot rule out *evolutionary branching points* (which obtain when a singular Kantian coefficient value is convergence stable but not uninvadable; recall footnote 3, and see Geritz et al., 1998 for a general discussion and McNamara and Leimar, 2020 for typical evolutionary game theory applications). An avenue for future research on preference evolution under complete information is thus to analyze conditions leading to polymorphism in preferences.[12]

Finally, we considered the case of complete information with complete plasticity where individuals both can observe the opponent's type and also condition its preferences on it. This is akin to a green-beard or secret handshake mechanism (Dawkins, 1976; Grafen, 1990; Robson, 1990), but at the preference level rather than the strategy level as in most previous work. Since an individual's Kantian coefficient may depend on the type of the interaction partner, a type is now a two-dimensional quantitative trait. Compared to the complete information incomplete plasticity scenario, residents are thus no longer committed to respond according to one and the same Kantian coefficient, and are therefore less exploitable by mutants. As we showed, this implies that residents can be pure Kantians when interacting with each other, and still be uninvadable: they prevent entry by mutants by using the Kantian coefficient equal to the response coefficient when interacting with individuals with a different type than theirs. In such a population, when interacting with each other residents use the strategy which yields the highest

[12]As an example, consider Proposition 1 of Heifetz et al. (2007a), which establishes conditions for the evolutionary viability of pessimism or optimism for a particular individual fitness function. Using our notation, these preferences entail the utility function $u_{\theta_i}(x_i, x_{-i} \mid y^*) = w(x_i, x_{-i}, y^*) + \theta_i x_i$, where θ_i is the evolving quantitative trait that can be taken to describe optimism when $\theta_i > 0$ and pessiminism when $\theta_i < 0$. It is straightforward to check that the singularity in Proposition 1 of Heifetz et al. (2007a) is both convergence stable and uninvadable, as it should under the measure dynamics they consider (e.g.,Cressman and Hofbauer, 2005). However, it is also straightforward to find parameter values of the fitness function they use Heifetz et al., 2007a, eq. 8 such that the singularity is convergence stable and invadable, and thus conducive to an adaptive polymorphism in dispositions, e.g., the coexistence of optimists and pessimists.

possible individual fitness. In other words, they use the strategy that yields an efficient outcome. This is reminiscent of a result by Dekel et al. (2007), who showed that a class of "coordination" preference, which results in efficient strategy profiles, are stable. We did, however, not characterize convergence stability under complete information with complete plasticity, because invasion fitness is not differentiable in this case, and hence a different toolkit than the usual multidimensional convergence stability criterion would be required (Leimar, 2009). Ascertaining convergence stability in this setting is thus left for future work.

Our three results show within the same model how different information and behavioral flexibility assumptions lead to different values of the Kantian coefficient, and thus to different equilibrium strategies. At the qualitative level, this range of outcomes under preference evolution are similar to those observed previously under strategy evolution models. Models of preference evolution, however, remain distinctively useful as their aim is to examine how ultimate imperatives translate into proximate ones and provide behavioral predictions that can be tested across different games in different lab experiments. Lab experiments, however, typically focus on behavior being incentivized at the payoff level and this can be distinct from incentives at the fitness level. Indeed, the qualitative nature of evolved preferences at the material payoff can differ from that at the fitness level and this difference depends on demographic properties under genetic evolution and transmission rules under cultural evolution (Alger et al., 2020). Payoff and fitness incentives tend to agree in panmictic and family structured population but tend to disagree in spatially structured populations owing to the presence of local competition between individuals (a feature occurring in the example of Box 2). In some experiments under incomplete information, humans do appear to conform to behave according to Hamilton's (marginal) rule at the action level (Levy and Lo, 2022) and to be driven by semi-Kantian concerns combined with other-regard (Van Leeuwen and Alger, 2022), as predicted by Alger et al. (2020) for preference evolution at the level of material payoffs. Understanding whether such results vindicate individuals being endowed with such preferences, how behavior in the laboratory varies with informational assumptions, and analysing more completely the evolutionary dynamics of behavioral mechanisms remain open research questions. We hope that our formalization has illustrated some of the nuances, intricacies, and richness of such research endeavours.

Appendix A Jacobian and Hessian under complete information with incomplete plasticity

Differentiating eq. (40) with respect to θ and evaluating at θ^* satisfying $S(\theta^*)=0$ yields

$$J(\theta^*)=\frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\left[(r(\theta^*)-\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2}+(1-\theta^*)(1+r(\theta^*))\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta}+(1-\theta^* r(\theta^*))\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right.$$
$$\left.+\frac{\left[1-r(\theta^*)^2-(1-\theta^{*2})\frac{{\rm d}r(\theta)}{{\rm d}\theta}\right]}{r(\theta^*)-\theta^*}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right]_{\tau=\theta=\theta^*}. \quad \text{(A-1)}$$

Using eq. (40) at $S(\theta^*)=0$ and using eq. (36) we can express the singular trait value implicitly as

$$\theta^*=\frac{r(\theta^*)+\rho(\theta^*)}{1+r(\theta^*)\rho(\theta)^*}, \quad \text{(A-2)}$$

which, on substituting into eq. (A-1), using $\rho(\theta) = \left(\partial x_{\rm d}^*(\tau,\theta)/\partial\theta\right)/\left(\partial x_{\rm d}^*(\tau,\theta)/\partial\tau\right)$ and simplifying produces

$$J(\theta^*) = -\frac{\partial w(x_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\left[\frac{\left(1-r(\theta^*)^2\right)}{1+r(\theta^*)\rho(\theta^*)}\left(\rho(\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2} - (1-\rho(\theta^*))\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta} - \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right) + \frac{\left((1+r(\theta^*)\rho(\theta^*))^2 - \frac{{\rm d}r(\theta)}{{\rm d}\theta}(1-\rho(\theta^*)^2)\right)}{1+r(\theta^*)\rho(\theta^*)}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right]_{\tau=\theta=\theta^*}. \quad \text{(A-3)}$$

The second order behavioral perturbations $\partial^2 x_{\rm d}^*(\tau,\theta)/\partial\tau^2, \partial^2 x_{\rm d}^*(\tau,\theta)/(\partial\tau\partial\theta), \partial^2 x_{\rm d}^*(\tau,\theta)/\partial\theta^2$ appearing in this Jacobian can be computed by using implicit differentiation in eq. (31). The resulting expressions are complicated and lengthy and we were unable to infer some general information from these expressions, although they can be handled easily with a symbolic manipulation system such as Mathematica (Wolfram Research, 2016). A Mathematica notebook with all algebraic computations of the paper is available on request.

Now using invasion fitness (38) to evaluate $H(\theta) = \partial^2 W(\tau,\theta)/\partial\tau^2\big|_{\tau=\theta}$, we find that

$$\begin{aligned} H(\theta) = {} & \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2}\left(\frac{\partial w(y, y_{\rm d}^*, y_{\rm d}^*)}{\partial y} + r(\theta)\frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\right) + 2r(\theta)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta}\left(\frac{\partial w(y, y_{\rm d}^*, y_{\rm d}^*)}{\partial y} + \frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\right) \\ & + \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\left(r(\theta)\frac{\partial w(y, y_{\rm d}^*, y_{\rm d}^*)}{\partial y} + \frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\right) + \frac{\partial^2 w(x, y_{\rm d}^*, y_{\rm d}^*)}{\partial x^2}\left[r(\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + 2r(\theta)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right. \\ & \left. + \left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right] + 2\frac{\partial^2 w(x, y, y_{\rm d}^*)}{\partial x\partial y}\left[r(\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + [1+r(\theta)]\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau} + r(\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right] \\ & + \frac{\partial^2 w(x_{\rm d}^*, y, y_{\rm d}^*)}{\partial y^2}\left[\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + 2r(\theta)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau} + r(\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right] \\ & + 2\frac{\partial r(\tau,\theta)}{\partial\tau}\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\frac{\partial w(y_{\rm d}^*, y, y_{\rm d}^*)}{\partial y} + \frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial w(x, y_{\rm d}^*, y_{\rm d}^*)}{\partial x}\right). \end{aligned} \quad \text{(A-4)}$$

Substituting into this expression $\partial w(x, y_{\rm s}^*, y_{\rm s}^*)/\partial x = -\theta\partial w(x, y, y_{\rm s}^*)/\partial y$ and eq. (A-2) yields

$$\begin{aligned} H(\theta^*) = {} & -\frac{\partial w(x_{\rm d}^*, y, y_{\rm d}^*)}{\partial y}\left[\frac{(1-r(\theta^*))\,(1+r(\theta^*))\left(\rho(\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2} - \frac{r(\theta^*)(1-\rho(\theta^*))}{1+r(\theta^*)}\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta} - \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right)}{1+r(\theta^*)\rho(\theta^*)}\right. \\ & \left. - \frac{2\frac{\partial r(\tau,\theta)}{\partial\tau}(1-\rho(\theta^*)^2)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}}{1+r(\theta^*)\rho}\right]_{\tau=\theta=\theta^*} + \left(\frac{\partial^2 w(x, y_{\rm d}^*, y_{\rm d}^*)}{\partial x^2}\left[1+\rho(\theta^*)r(\theta^*)(\rho(\theta^*)+2)\right]\right. \\ & \left. 2\frac{\partial^2 w(x, y, y_{\rm d}^*)}{\partial x\partial y}\left[\rho + r(\theta^*)\left(1+\rho(\theta^*)+\rho(\theta^*)^2\right)\right] + \frac{\partial^2 w(x_{\rm d}^*, y, y_{\rm d}^*)}{\partial y^2}\left[\rho(\theta^*)^2 + 2r(\theta^*)(1+\rho(\theta^*))\right]\right)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2. \end{aligned} \quad \text{(A-5)}$$

A key distinction between the Jacobian $J(\theta^*)$ and the Hessian $H(\theta^*)$ is that the sign of the Jacobian does not depend directly on fitness derivatives, while the Hessian does. Both expressions remain complicated and we did not manage to obtain general information from them. Hence, they need to be evaluated on a case by case basis.

References

Akçay, E. and J. Van Cleve. 2009. A theory for the evolution of other-regarding motivations integrating proximate and ultimate perspectives. *Proceedings of the National Academy of Sciences of the United States of America* 106:19061–19066.

Akçay, E. and J. Van Cleve. 2012. Behavioral responses in structured populations pave the way to group optimality. *American Naturalist* 179:257–269.

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

Alger, I. 2022. Evolutionarily Stable Preferences. IAST Working Paper 22-144.

Alger, I. and J. W. Weibull. 2010. Kinship, Incentives, and Evolution. *American Economic Review* 100:1725–1758.

Alger, I. and J. W. Weibull. 2012. A generalization of Hamilton's rule—love others how much? *Journal of Theoretical Biology* 299:42–54.

Alger, I. and J. W. Weibull. 2013. Homo Moralis: preference evolution under incomplete information and assortative matching. *Econometrica* 6.

Alger, I. and J. W. Weibull. 2016. Evolution and Kantian morality. *Games and Economic Behavior* 98:56–67.

Alger, I. and J. W. Weibull. 2019. Evolutionary models of preference formation. *Annual Review of Economics* 11:329–354.

Alger, I., J. W. Weibull, and L. Lehmann. 2020. Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory* 185:1–45.

André, J. B. and T. Day. 2007. Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. *Journal of Theoretical Biology* 247:11–22.

Arbilly, M., U. Motro, M. W. Feldman, and A. Lotem. 2010. Co-evolution of learning complexity and social foraging strategies. *J Theor Biol* 267:573–81.

Bergstrom, T. 1995. On the evolution of altruistic ethical rules for siblings. *American Economic Review* 85:58–81.

Bester, H. and W. Güth. 1998. Is altruism evolutionarily stable? *Journal of Economic Behavior and Organization* 34:193–209.

Binmore, K. 1998. Just Playing: Game Theory and the Social Contract 2. MIT Press, Cambridge, MA.

Binmore, K. 2011. Rational decisions. Princeton University Press, Princeton, NJ.

Bishop, D. and C. Cannings. 1978. A generalized war of attrition. *Journal of Theoretical Biology* 70:85—124.

Bolle, F. 2000. Is altruism evolutionarily stable? And envy and malevolence? Remarks on Bester and Güth. *Journal of Economic Behavior and Organization* 42:131–133.

Bulmer, M. G. 1994. Theoretical Evolutionary Ecology. Sinauer Associates, Massachusetts.

Charlesworth, B. 1994. Evolution in Age-Structured Populations. Cambridge University Press, Cambridge, 2th edn.

Christiansen, F. B. 1991. On conditions for evolutionary stability for a continuously varying character. *American Naturalist* 138:37–50.

Cressman, R. and J. Hofbauer. 2005. Measure dynamics on a one-dimensional continuous trait space: theoretical foundations for adaptive dynamics. *Theoretical Population Biology* 67:47–59.

Cressman, R. and Y. Tao. 2004. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 11:10810–10817.

Dawkins, R. 1976. The Selfish Gene. Oxford University Press, Oxford.

Dawkins, R. 1980. Good strategy or evolutionarily stable strategy? In Barlow, G. W. (ed.), *Sociobiology: Beyond Nature/Nurture?* Westview Press, Boulder, Colorado.

Dekel, E., J. Ely, and O. Yilankaya. 2007. Evolution of preferences. *Review of Economic Studies* 74:685–704.

Dridi, S. and E. Akçay. 2018. Learning to cooperate: The evolution of social rewards in repeated interactions. *American Naturalist* 191:58–73.

Dridi, S. and L. Lehmann. 2015. A model for the evolution of reinforcement learning in fluctuating games. *Animal Behaviour* 104:1–28.

Eshel, I. 1983. Evolutionary and continuous stability. *Journal ot Theoretical Biology* 103:99–111.

Eshel, I. 1996. On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. *Journal of Mathematical Biology* 34:485–510.

Eshel, I., M. Feldman, and A. Bergman. 1998. Long-term evolution, short-term evolution, and population genetic theory. *Journal of Theoretical Biology* 191:391–396.

Eshel, I. and M. W. Feldman. 1984. Initial increase of new mutants and some continuity properties of ESS in two-locus systems. *The American Naturalist* 124:631–640.

Eshel, I., U. Motro, and E. Sansone. 1997. Continuous stability and evolutionary convergence. *Journal of Theoretical Biology* 074:222–232.

Ferrière, R. and M. Gatto. 1995. Lyapunov exponents and the mathematics of invasion in oscillatory or chaotic populations. *Theoretical Population Biology* 48:126–171.

Fisher, R. A. 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Fisher, R. A. 1941. Average excess and average effect of a gene substitution. *Annals of Human Genetics* 11:53–63.

Fudenberg, D. and D. K. Levine. 1998. Theory of Learning in Games. MIT Press, Cambridge, MA.

Fudenberg, D. and J. Tirole. 1991. Game Theory. MIT Press, Massachusetts.

Geritz, S. 2005. Resident-invader dynamics and the coexistence of similar strategies. *Journal of Mathematical Biology* 50:67–82.

Geritz, S. A. H., E. Kisdi, G. Meszéna, and J. A. J. Metz. 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology* 12:35–57.

Giraldeau, L. and T. Caraco. 2000. Social Foraging Theory. Princeton University Press, Princeton, NJ.

Grafen, A. 1990. Do animals really recognize kin? *Animal Behaviour* 39:42–54.

Guth, W. 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory* 24:323–344.

Hamilton, W. D. 1964. The genetical evolution of social behaviour, 1. *Journal of Theoretical Biology* 7:1–16.

Hamilton, W. D. 1970. Selfish and spiteful behavior in an evolutionary model. *Nature* 228:1218–1220.

Hamilton, W. D. 1988. This week's citation classic. *Current Contents* 40:16.

Harley, C. B. 1981. Learning the evolutionary stable strategy. *Journal of Theoretical Biology* 89:611–633.

Harris, T. E. 1963. The Theory of Branching Processes. Springer, Berlin.

Heifetz, A., C. Shannon, and Y. Spiegel. 2007a. The dynamic evolution of preferences. *Economic Theory* 32:251–286.

Heifetz, A., C. Shannon, and Y. Spiegel. 2007b. What to maximize if you must. *Journal of Economic Theory* 133:31 – 57.

Hines, W. G. S. and J. Maynard Smith. 1978. Games between relatives. *Journal of Theoretical Biology* 79:19–30.

Hofbauer, J. and K. Sigmund. 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge.

Iosifescu, M. 2007. Finite Markov Processes and Their Applications. Dover, New York.

Kant, I. 1785. Grundlegung zur Metaphysik der Sitten. Hartknoch, Riga.

Karlin, S. and H. M. Taylor. 1975. A First Course in Stochastic Processes. Academic Press, San Diego.

Killingback, T. and M. Doebeli. 2002. The continuous prisoner's dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *American Naturalist* 160:421–438.

Lehmann, L., I. Alger, and J. W. Weibull. 2015. Does evolution lead to maximizing behavior? *Evolution* 69:1858–1873.

Lehmann, L. and F. Rousset. 2020. When do individuals maximize their inclusive fitness? *The American Naturalist* 195:717–732.

Leimar, O. 2009. Multidimensional convergence stability. *Evolutionary Ecology Research* 11:191–208.

Levy, M. and A. W. Lo. 2022. Hamilton's rule in economic decision-making. *Proceedings of the National Academy of Sciences of the United States of America* 119:1–5.

Malécot, G. 1969. The Mathematics of Heredity. W. H. Freeman and Company, San Francisco.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. Microeconomic Theory. Oxford Unversity Press, Oxford.

Maynard Smith, J. 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge.

Maynard Smith, J. and G. R. Price. 1973. The logic of animal conflict. *Nature* 246:15–18.

McNamara, J. M., C. E. Gasson, and A. I. Houston. 1999. Incorporating rules for responding into evolutionary games. *Nature* 401:368–71.

McNamara, J. M. and O. Leimar. 2020. Game Theory in Biology. Oxford University Press.

Metz, J. A. J., R. M. Nisbet, and S. A. H. Geritz. 1992. How should we define fitness for general ecological scenarios? *Trends in Ecology and Evolution* 7:198–202.

Michod, R. 1980. Evolution of interactions in family-structured populations: mixed mating models. *Genetics* 96:275–96.

Moran, P. A. P. 1962. The Statistical Processes of Evolutionary Theory. Clarendon Press, Oxford.

Mullon, C., L. Keller, and L. Lehmann. 2016. Evolutionary stability of jointly evolving traits in subdivided populations. *American Naturalist* 188:175–195.

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

Parker, G. A. and J. Maynard Smith. 1990. Optimality theory in evolutionary biology. *Science* 349:27–33.

Possajennikov, A. 2000. On the evolutionary stability of altruistic and spiteful preferences. *Journal of Economic Behavior and Organization* 42:125–129.

Priklopil, T. and L. Lehmann. 2021. Metacommunities, fitness and gradual evolution. *Theoretical Population Biology* 142:12–35.

Robson, A. J. 1990. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Evolutionary Biology* 144:379–396.

Rousset, F. 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton, NJ.

Rousset, F. 2015. Regression, least squares, and the general version of inclusive fitness. *Evolution* 69:2963–2970.

Rousset, F. and S. Billiard. 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. *Journal of Evolutionary Biology* 13:814–825.

Russell, S. and P. Norvig. 2016. Artificial Intelligence: a Modern Approach. Pearson, Edinburgh, UK.

Samuelson, L. 1998. Evolutionary Games and Equilibrium Selection. MIT Press, Cambridge, MA.

Sandholm, W. H. 2011. Population Games and Evolutionary Dynamics. MIT Press, Cambridge, MA.

Selten, R. and P. Hammerstein. 1984. Gaps in Harley's argument on evolutionarily stable learning rules and in the logic of "tit for tat". *Behavioral and Brain Sciences* 7:115–116.

Sugden, R. 1986. The Economics of Rights, Cooperation and Welfare. Palgrave Macmillan, New York.

Sutton, R. S. and A. G. Barto. 1998. Reinforcement Learning. MIT Press, Cambridge, MA.

Taylor, P. D. 1989. Evolutionary stability in one-parameter models under weak selection. *Theoretical Population Biology* 36:125–143.

Taylor, P. D. and T. Day. 2004. Stability in negotiation games and the emergence of cooperation. *Proceedings of the Royal Society of London Series B-Biological Sciences* 271:669–674.

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

Vincent, T. L. and J. S. Brown. 2005. Evolutionary Game Theory, Natural Selection, and Darwinian Dynamics. Cambridge University Press, Cambridge.

Wakano, J. Y. and L. Lehmann. 2014. Evolutionary branching in deme-structured populations. *Journal of Theoretical Biology* 351:83–95.

Weibull, J. W. 1997. Evolutionary Game Theory. MIT Press, Cambridge, MA.

Wolfram Research, I. 2016. Mathematica. Wolfram Research, Inc., Champaign, Illinois.

Young, H. P. 1998. Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, Princeton, NJ.

Box 1. Invasion fitness as eigenvalue. The invasion fitness of a type is its geometric growth ratio when rare in a resident population (Fisher, 1930; Eshel and Feldman, 1984; Metz et al., 1992; Ferrière and Gatto, 1995). When the resident population is monomorphic for θ, the invasion fitness $W(\tau,\theta)$ of mutant τ under our demographic assumptions (section (2.1)) is obtained as the leading eigenvalue of the matrix

$$\mathbf{A}(\tau,\theta)=\begin{pmatrix}a_{11}(\tau,\theta) & a_{12}(\tau,\theta)\\ a_{21}(\tau,\theta) & a_{22}(\tau,\theta)\end{pmatrix}, \tag{B-a}$$

where a_{ij} stands for the expected number of groups with $i\in\{1,2\}$ mutants that over one demographic time period descend (either through local change or through migration) from a focal group with $i\in\{1,2\}$ mutants, when the population is otherwise monomorphic for θ. Matrix $\mathbf{A}(\tau,\theta)$ is assumed to be regular (irreducible and aperiodic, Iosifescu, 2007, p. 123) and it then follows from standard results on multi-type branching processes that the lineage of a single τ mutant goes extinct with probability one if, and only if, $W(\tau,\theta)\leq 1$, while otherwise it spreads into the population when rare (Harris, 1963; Karlin and Taylor, 1975). By definition of invasion fitness, $W(\tau,\theta)\mathbf{u}(\tau,\theta)=\mathbf{A}(\tau,\theta)\mathbf{u}(\tau,\theta)$, where $\mathbf{u}(\tau,\theta)=(u_1(\tau,\theta),u_2(\tau,\theta))$ is the only non-negative right eigenvector of $\mathbf{A}(\tau,\theta)$, where, by normalization, $u_1(\tau,\theta)+u_2(\tau,\theta)=1$. The eigenvector $\mathbf{u}(\tau,\theta)$ can be interpreted as the quasi-stationary distribution of mutant group types as it is invariant to multiplication by $\mathbf{A}(\tau,\theta)$, whereby $u_i(\tau,\theta)$ is the frequency of groups with $i\in\{1,2\}$ mutants among groups with at least one mutant. Following previous developments (Mullon et al., 2016), we can left multiply $W(\tau,\theta)\mathbf{u}(\tau,\theta)=\mathbf{A}(\tau,\theta)\mathbf{u}(\tau,\theta)$ by the vector $(1,2)$. Rearranging terms, this produces

$$W(\tau,\theta)=[1-r(\tau,\theta)]\,\tilde{w}_1(\tau,\theta)+r(\tau,\theta)\tilde{w}_2(\tau,\theta), \tag{B-b}$$

where

$$\tilde{w}_1(\tau,\theta)=a_{11}(\tau,\theta)+2a_{21}(\tau,\theta) \tag{B-c}$$

$$\tilde{w}_2(\tau,\theta)=a_{12}(\tau,\theta)/2+a_{22}(\tau,\theta) \tag{B-d}$$

$$r(\tau,\theta)=\frac{2u_2(\tau,\theta)}{u_1(\tau,\theta)+2u_2(\tau,\theta)}. \tag{B-e}$$

Here, $\tilde{w}_i(\tau,\theta)$ is the expected total number of individuals produced (including the surviving self) by a single τ individual over one demographic time step when there are $j\in\{1,2\}$ τ individuals in its group and the population is otherwise monomorphic for θ; and $r(\tau,\theta)$ is the probability that, for any given descendant of the initial mutant, the neighbor of that mutant is also a mutant. An explicit example of these invasion fitness components is given in Box 2.

Eq (B-b) is a recipient-centered representation of the mutant's geometric growth ratio since it is expressed as the average of the expected fitness of a type τ individual, who is necessarily the recipient of the traits of others. An actor-centered representation of the growth ratio, which focuses on the consequence on others of an individual expressing the mutant instead of the resident trait value can also be obtained (Hamilton, 1970; Rousset, 2015). Such an actor-centered representation of invasion fitness can be reached by rearranging the components of eq. (B-b) (Lehmann and Rousset, 2020). Indeed, owing to the fact that $\tilde{w}_2(\theta,\theta)=1$, we have the equality

$$W(\tau,\theta)=1-c(\tau,\theta)+r(\tau,\theta)b(\tau,\theta), \tag{B-f}$$

where

$$-c(\tau,\theta)=\frac{1}{1+r(\tau,\theta)}\left(\tilde{w}_1(\tau,\theta)-\tilde{w}_2(\theta,\theta)\right)+\frac{r(\tau,\theta)}{1+r(\tau,\theta)}\left(\tilde{w}_2(\tau,\theta)-\tilde{w}_1(\theta,\tau)\right)$$

$$b(\tau,\theta)=\frac{1}{1+r(\tau,\theta)}\left(\tilde{w}_1(\theta,\tau)-\tilde{w}_2(\theta,\theta)\right)+\frac{r(\tau,\theta)}{1+r(\tau,\theta)}\left(\tilde{w}_2(\tau,\theta)-\tilde{w}_1(\tau,\theta)\right).$$

Here, $-c(\tau,\theta)$ is the *average effect* (sensu Fisher, 1941) on the number of mutant gene copies produced by a single individual when expressing a copy of the mutant instead of the resident allele. The average thus being over the two possible contexts in which an individual expressing τ instead of θ can be: interacting with a neighbor that carries or not the mutant. The actor-centered perspective of eq. (B-f) is then born out from the fact that $b(\tau,\theta)$ is the average effect on the expected number of offspring produced by an individual's neighbour, which stemming from the actor switching to expressing a copy of the mutant instead of the resident allele.

Box 2. Moran process example. We illustrate the invasion fitness components described in Box 1 by considering a process where exactly one individual dies in each group during a demographic time step (i.e., an instance of a Moran process, Moran, 1962). For this case, the entries of matrix (B-a) are

$$a_{11} = 1 - b_1 - d_1 + e_1, \qquad a_{21} = b_1, \qquad a_{12} = d_2 + e_2, \qquad a_{22} = 1 - b_2 - d_2, \tag{B-g}$$

with b_i and d_i standing, respectively, for the probability that there is a mutant descendant and mutant death, and e_i is the expected number of succesful emigrant mutants, in a group with i mutants. These variables are given by

$$\begin{aligned} b_k(\tau,\theta) &= \frac{(2-k)\mu_k(\theta)}{k\mu_k(\tau)+(2-k)\mu_k(\theta)}\left[\frac{(1-m)kf_k(\tau)}{(1-m)\left[kf_k(\tau)+(2-k)f_k(\theta)\right]+m2f_0(\theta)s_0(\theta)}\right] \\ d_k(\tau,\theta) &= \left[1-\frac{(2-k)\mu_k(\theta)}{k\mu_k(\tau)+(2-k)\mu_k(\theta)}\right]\left[1-\frac{(1-m)kf_k(\tau)}{(1-m)\left[kf_k(\tau)+(2-k)f_k(\theta)\right]+m2f_0(\theta)s_0(\theta)}\right] \\ e_k(\tau,\theta) &= \frac{1}{2}\frac{mkf_k(\tau)s_k(\tau)}{(1-m)f_0(\theta)+mf_0(\theta)s_0(\theta)} \end{aligned} \tag{B-h}$$

where $f_k(\theta')$, $\mu_k(\theta')$, $s_k(\theta')$ are, respectively, the fecundity, death-factor, juveniles' survival probability during migration, of a single type $\theta' \in \{\tau,\theta\}$ adult individual when there are exactly k mutants in its group (see Lehmann et al., 2015; Mullon et al., 2016 for more details on the derivation and the case where there are more than 2 individuals per group). On setting $f_1(\tau) = f(x_{\rm d}^*, y_{\rm d}^*)$, $f_2(\tau) = f(x_{\rm s}^*, x_{\rm s}^*)$, $f_0(\theta) = f(y_{\rm s}^*, y_{\rm s}^*)$, $f_1(\theta) = f(y_{\rm d}^*, x_{\rm d}^*)$, $\mu_1(\tau) = \mu(x_{\rm d}^*, y_{\rm d}^*)$, $\mu_2(\tau) = \mu(x_{\rm s}^*, x_{\rm s}^*)$, $\mu_0(\theta) = \mu(y_{\rm s}^*, y_{\rm s}^*)$, $\mu_1(\theta) = \mu(y_{\rm d}^*, x_{\rm d}^*)$, $s_1(\tau) = s(x_{\rm d}^*, y_{\rm d}^*)$, $s_2(\tau) = s(x_{\rm s}^*, x_{\rm s}^*)$, and $s_0(\theta) = s(y_{\rm s}^*, y_{\rm s}^*)$, where x refers to mutant and y to resident strategies [recall eqs. (6)–(7)] and $f : \mathcal{X}^2 \to \mathbb{R}_+$, $\mu : \mathcal{X}^2 \to \mathbb{R}_+$, and $\mu : \mathcal{X}^2 \to \mathbb{R}_+$, then algebraic rearrangements show that the fitness function $w : \mathcal{X}^3 \to \mathbb{R}_+$ in eqs. (6)–(7) for the Moran process is defined as

$$\begin{aligned} w(x_i, x_{-i}, y) &= 1 - \frac{\mu(x_i, x_{-i})}{\mu(x_i, x_{-i}) + \mu(x_{-i}, x_i)} \\ &+ \frac{1}{2}\left[\frac{(1-m)f(x_i, x_{-i})}{(1-m)\left[f(x_i, x_{-i}, y) + f(x_{-i}, x_i)\right] + m\, f(y,y)s(y,y)} + \frac{mf(x_i, x_{-i})s(x_i, x_{-i})}{(1-m)f(y,y) + mf(y,y)s(y,y)}\right] \end{aligned} \tag{B-i}$$

(see Box 1 of Lehmann et al., 2015 for a biological interpretation of each term).

Even for this Moran process, the expression for relatedness eq. (B-e) is complicated, but its computation can be alleviated by using an invasion fitness proxy. An invasion fitness proxy is by definition any fitness measure $P(\tau,\theta)$ that is sign equivalent to $W(\tau,\theta)$ such that the evolutionary invasion analysis can be carried out from this measure (i.e. $P(\tau,\theta) \leq 1 \iff W(\tau,\theta) \leq 1$). An invasion fitness proxy for $W(\tau,\theta)$ can be obtained by keeping the functional form eq. (B-b) but relatedness, instead of being given by the complicated expression eq. (B-e), is given by

$$r(\tau,\theta) = \frac{2b_1(\tau,\theta)}{2b_1(\tau,\theta) + d_2(\tau,\theta)}, \tag{B-j}$$

which can be readily evaluated using eq. (B-h). Conceptually, this simplification obtains by substituting $u_i \to t_i$ in eq. (B-e), where t_i is the sojourn time with $i \in \{1,2\}$ mutants of the mutant lineage in a single group where $t_1 = 1/d_1$ and $t_2 = b_1/(d_1 d_2)$ (see Lehmann et al., 2015; Mullon et al., 2016 for more details). Substituting eq. (B-h) into eq. (B-j) and using the expression for the vital rates in terms of strategies and assuming, for simplicity that fecundity f is independent of the types, one can then check that relatedness can be written as

$$r(\tau,\theta) = \frac{(1-m)\mu(y_{\rm d}^*(\theta,\tau), x_{\rm d}^*(\tau,\theta))}{(1-m)\mu(y_{\rm d}^*(\theta,\tau), x_{\rm d}^*(\tau,\theta)) + m\left[\mu(x_{\rm d}^*(\tau,\theta), y_{\rm d}^*(\theta,\tau)) + \mu(y_{\rm d}^*(\theta,\tau), x_{\rm d}^*(\tau,\theta))\right] s(y_{\rm s}^*(\theta), y_{\rm s}^*(\theta))}, \tag{B-k}$$

where we made explicit all functional dependencies. Further, in a monomorphic population relatedness boils down to

$$r(\theta,\theta) = \frac{1 - m_{\rm b}(\theta)}{1 + m_{\rm b}(\theta)} = \frac{1-m}{1 - m\left[1 - 2s(y_{\rm s}^*(\theta), y_{\rm s}^*(\theta))\right]}, \tag{B-l}$$

where $m_{\rm b}(\theta) = (1-m)/[1 - m + ms(y_{\rm s}^*(\theta), y_{\rm s}^*(\theta)]$ is the backward migration probability, i.e., the probability that an individual randomly sampled in a patch is of philopatric origin. Eq. (B-l) displays two generic features about relatedness. First, it is a monotonic decreasing function of dispersal and of juvenile survival. Second, relatedness can depend endogeneously on the interactions, because the spatial structure is an outcome of survival and reproduction which are themselves functions of interactions between individuals. If survival $s(y_{\rm s}^*(\theta), y_{\rm s}^*(\theta))$ were independent of strategies, then neutral relatedness would be independent of the types and reduce to $r = (1-m)/(1-m(1-2s))$, as it should (Mullon et al., 2016), for parameters $m \in (0,1]$ and $s \in [0,1]$.