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

Ingela Alger, Laurent Lehmann

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

Ingela Alger* and Laurent Lehmann†

Abstract

We model the evolution of preferences guiding behavior in pairwise interactions in group-structured populations. The model uses long-term evolution theory to examine different interaction scenarios, including conditional preference expression upon recognition of the partner'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 partner also adopted this behavior. We seek 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 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 generally evolves 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) there are generally multiple uninvadable types, including the type whereby an individual is a pure Kantian when interacting with individuals of the same type and applies the Kantian coefficient that is uninvadable under complete information with zero relatedness when interacting with a different typed individual. 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*

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

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

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 models, 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 analysing the genetic evolution of strategies under all sorts of biological scenarios as is illustrated by the literatures on the evolution of fighting, cooperation, and life-histories 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 "evolutionarily stable strategy" since strategies are inherited from parent to offspring and selected among alternatives by way of differential survival and reproduction.

The view that strategies are innate is restrictive, however, as it rules out situations where individuals have capacities to change their own strategy when interacting with their environment. Such processes have been incorporated into evolutionary game theory through several alternative notions, such as the concepts of "culturally stable" and "developmentally stable" strategies (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 the individual's lifespan. In the memorable example detailed by Dawkins (1980), pigs in skinner boxes equilibrate on developmentally stable strategies by action reinforcement in producer-scrounger games, and a large literature has evaluated culturally stable strategies under different sorts of transmission rules (e.g., Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985). 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 and transmission rules induce change in strategies in populations (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 in this literature 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, at the culturally stable or developmentally stable state, 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, given the others' strategies (Mas-Colell et al., 1995). The question of what should be the evolutionarily stable rule for individual strategy selection can thus here be phrased as: 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 for strategic interactions addresses addresses (e.g., Güth, 1995; Ok and Vega-Redondo, 2001; Dekel et al., 2007; Heifetz et al., 2007b,a; Akçay and Van Cleve, 2009; Alger and Weibull, 2010; Akçay and Van Cleve, 2012; Alger and Weibull, 2012, 2013; Wang and Wu, 2023). 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; Wang and Wu, 2023; see Alger and Weibull, 2019; Alger, 2023 for surveys). Focusing on the evolution of preferences gives hope to improve predictions about equilibrium behavior 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) but none empirically fully elucidated (e.g., Jeong et al., 2022).

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 populations. This part can thus 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 the categorical imperative of Nash with that of Kant. Bergstrom (1995) shows 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. Our goal is to analyse 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, but an individual's preferences do 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). It will be seen that the different informational and plasticity assumptions lead to quite different evolutionary outcomes, and that we are not always able to reach general conclusions about convergence stability.

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; Avila and Mullon, 2023; Van Cleve, 2023). 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 ratio) 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 ratio 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 (i.e., $W(\theta, \theta) = 1$ for all $\theta \in \Theta$).

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

[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.

is thus of the resident type θ):

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

Here, $\tilde{w}_i(\tau,\theta)$ 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 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]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, 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, 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 "invasion implies substitution" result was first noted in a special case by Hamilton (1964a) and called "a gift from god" (Hamilton, 1988). See also Eshel et al. (1997) for a different line of argument reaching the same conclusions.

2.3 Behavioral equilibrium

In applications of evolutionary game theory, an individual's type is often taken to be its strategy to be applied in the interaction at hand. Yet many applications require to decouple types from strategies. In order to do this and 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 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. Note that any individual fitness function is subject to the demographic consistency relation $w(y, y, y) = 1$ for all $y \in \mathcal{X}$ and an example thereof is provided in Box 2.

Turning now to the equilibrium strategies, in a population with a mutant type τ and a resident type $\theta \neq \tau$, each group either has zero, one, or two mutants. For groups with two residents (resp. two mutants), we denote by $y_{\mathrm{s}}^*(\theta)$ (resp. $x_{\mathrm{s}}^*(\tau, \theta)$) an equilibrium strategy for each individual, where the subscript "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 $x_{\mathrm{d}}^*(\tau, \theta)$ denote the mutant's equilibrium strategy and $y_{\mathrm{d}}^*(\theta, \tau)$ the resident's equilibrium strategy, where the subscript "d" stands for *different* types. Importantly, throughout we assume that for any type pair $(\theta, \tau) \in \Theta^2$ with $\theta \neq \tau$, there exist unique equilibrium strategies $y_{\mathrm{s}}^*(\theta)$, $x_{\mathrm{s}}^*(\tau, \theta)$, and $(x_{\mathrm{d}}^*(\tau, \theta), y_{\mathrm{d}}^*(\theta, \tau))$. In a population with a mutant $\tau = \theta$ interactions can occur between same and different lineage members having the same type. For this case, we assume that all individuals in all pairs use the same strategy $y_{\mathrm{s}}^*(\theta) = x_{\mathrm{s}}^*(\theta, \theta)$ since they have the same type. This implies that the mappings $\tilde{w}_1$ and $\tilde{w}_2$ used in eq. (2) are well defined[4], as follows:

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

and

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

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 dynamical system drives strategy expression over time, and under some conditions these behavioral dynamics reach an equilibrium. 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

[4]Invasion fitness eq. (2) must be defined for all $\tau \in \Theta$ including $\tau = \theta$. In such a monomorphic population, all individuals use strategy $y_{\mathrm{s}}^*(\theta) = x_{\mathrm{s}}^*(\theta, \theta)$ and thus the demographic consistency relation $W(\theta, \theta) = 1$ will be verified since $w(x_{\mathrm{s}}^*(\theta, \theta), x_{\mathrm{s}}^*(\theta, \theta), y_{\mathrm{s}}^*(\theta)) = w(y_{\mathrm{s}}^*(\theta), y_{\mathrm{s}}^*(\theta), y_{\mathrm{s}}^*(\theta)) = 1$ for all $\theta \in \Theta$.

mappings, $M_{\mathrm{d}} : \Theta^2 \times \mathcal{X}^2 \to \mathbb{R}$ for the mutant type and $R_{\mathrm{d}} : \Theta \times \mathcal{X}^2 \to \mathbb{R}$ 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:

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

The mechanism M_{d} is parametrized by both the mutant and the resident type, while R_{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_{d}^* and y_{d}^* produces the dependence of each strategy on both types, i.e., $x_{\mathrm{d}}^* = x_{\mathrm{d}}^*(\tau, \theta)$ and $y_{\mathrm{d}}^* = y_{\mathrm{d}}^*(\theta, \tau)$.

For the equilibrium strategy used in mutant-mutant interactions, let there be a mapping $M_{\mathrm{s}} : \Theta^2 \times \mathcal{X} \to \mathbb{R}$ which describes the process whereby a mutant interacts with another mutant as a function of the partner's strategy. The strategy in equilibrium is then 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 $y_{\mathrm{s}}^*(\theta)$ is assumed to solve 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. 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 determined by θ and some mutant utility function determined by τ, the strategy $y_{\mathrm{s}}^*(\theta)$ (resp. $x_{\mathrm{s}}^*(\tau, \theta)$)

[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

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 $y_{\rm s}^*(\theta)$ (resp. $x_{\rm s}^*(\tau,\theta)$). Likewise, the strategy $y_{\rm d}^*(\theta,\tau)$ would be the strategy in $\mathcal{X}$ that maximizes the utility of the resident, given that the mutant with whom it interacts uses strategy $x_{\rm d}^*(\tau,\theta)$, while the strategy $x_{\rm d}^*(\tau,\theta)$ would be the strategy in $\mathcal{X}$ maximizing the mutant's utility, given that the resident with whom it interacts uses strategy $y_{\rm d}^*(\theta,\tau)$.

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 who uses strategy x when it neighbour uses strategy y, and given that strategy y^* is played at large in the population, this utility function is defined as

$$u_\kappa(x,y\mid y^*)=(1-\kappa)\,w(x,y,y^*)+\kappa\,w(x,x,y^*), \tag{11}$$

where w is the individual fitness function defined above and $\kappa\in[0,1]$. The first is the individual's realized fitness, given the strategies used. The second term is the fitness that the individual would realize if – hypothetically – the opponent used the same strategy (x) instead of strategy y; since the individual thereby evaluates what would happen if others were to follow the same course of action as itself, the second term can be interpreted as capturing a form of the first formulation of Kant's categorical imperative: "act as if the maxims of your action were to become through your will a universal law of nature" (Kant, 1785). These preferences were 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 the Kantian coefficient and our goal is to 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 restrict attention to settings where w is twice continuously differentiable and the utility function (11) is strictly concave in its first argument for any $\theta\in[0,1]$. 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 $(x,y,z)\in X^3$. We further assume that the sign of this effect is independent of the strategies used, and by convention, we let $\partial w(x,y,z)/\partial y>0$ for all $(x,y,z)\in X^3$, meaning that an increase in the strategy of an individual's partner 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 $(x,y,z)\in X^3$, and we will say that

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.

[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 x and y denote the contributions to the public good. If the interaction at hand is a common pool resource game, then let x and y denote the inverse of contributions to extracting resources.

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's type is the value of their Kantian coefficient taking a value in the interval $[0,1]$, and an individual cannot observe the type of its interaction partner. Still, the individual can have information about the matching distribution in the pairwise interaction, i.e., the probability that the partner belongs 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 their degree of kinship. We assume that individuals hold the belief that the probability of being matched with an individual from the same lineage is given by $r(\tau,\theta)$, which is a correct belief 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. Given these assumptions, an individual can condition its strategy only on the strategy that it expects its partner 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 obtained 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).[7] The behavioral fixed point (13) defines the behavioral mechanisms (8)–(10), which here

[7]Note that our model is different from the one in Alger and Weibull (2013), where a resident faces a positive probability of being matched with a mutant. In our model, the mutant trait appears initially in one single individual, and uninvadability obtains if the lineage created by this initial mutant goes extinct in the population (while invadability obtains if the lineage size created by this initial mutant becomes infinite and thus reaches positive frequency in the infinitely large population). During the time the mutant lineage is around in the population there can thus only be a finite number of mutants, and hence residents face a zero probability of being matched with a mutant in this infinitely large population. See Box 1 or Alger et al. (2020) for a formal explanation.

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):

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

Using eq. (11), these equations become

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

where the second equality of the first line shows that the marginal change in utility can be expressed as the sum of the effect of own behavior on own fitness and the fitness of the partner weighted by the Kantian coefficient θ. The necessary second-order conditions for $(x^*(\tau,\theta),y^*(\theta))$ defined by (14) to be maxima rather than minima are:

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

These inequalities hold strictly by virtue of the assumption that u_θ is strictly concave in its first argument. Hence:

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

This inequality can in turn 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*, which will be seen to play a central role in the evolutionary analysis, is always positive, i.e., $\partial x^*(\tau,\theta)/\partial\tau|_{\tau=\theta}>0$ (see Appendix A for a proof). The intuition is that by making the individual attach a greater value to the effect that the strategy would have on self if it was also adopted by the other individual, an increase in τ makes the individual internalize the positive externality $(\partial w(x,y,y^*)(\partial y)>0)$ that the other's strategy has on self.

3.1.2 Evolutionary equilibrium

Turning now to the analysis of selection on the Kantian coefficient, invasion fitness writes

$$W(\tau,\theta)=[1-r(\tau,\theta)]\,w(x^*(\tau,\theta),y^*(\theta),y^*(\theta))+r(\tau,\theta)\,w(x^*(\tau,\theta),x^*(\tau,\theta),y^*(\theta)), \tag{19}$$

which is differentiable (since individual fitness is differentiable so will be $r(\tau,\theta)$, see Box 1). Then, substituting eq. (19) into $S(\theta)=\partial W(\tau,\theta)/\partial\tau|_{\tau=\theta}$, the selection gradient is

$$S(\theta)=\frac{\partial x^*(\tau,\theta)}{\partial\tau}\left[\frac{\partial w(x,y^*,y^*)}{\partial x}+r(\theta,\theta)\frac{\partial w(x^*,x,y^*)}{\partial x}\right]\Bigg|_{x=x^*=y^*=y^*(\theta)}^{\tau=\theta}, \tag{20}$$

because the term multiplying $\partial r(\tau,\theta)/\partial\tau$ is $w(x^*(\tau,\theta),x^*(\tau,\theta),y^*(\theta))|_{\tau=\theta} - w(x^*(\tau,\theta),y^*(\theta),y^*(\theta))|_{\tau=\theta} = 0$. Selection on the Kantian coefficient thus depends on the (local) mutant behavioral perturbation weighted by the inclusive fitness effect at the strategy level, i.e. the sum of the direct effect, $\partial w(x,y^*,y^*)/\partial x$, and the indirect effect, $\partial w(x^*,x,y^*)/\partial x$, on fitness weighted by neutral relatedness. Moreover, since $\partial w(x,y^*,y^*)/\partial x = -\theta\partial w(y^*,y,y^*)/\partial y$ at $\tau=\theta$ and $x^*=y^*$ (see eq. (15)), we obtain that

$$S(\theta) = [r(\theta,\theta)-\theta]\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,y,y^*)}{\partial y}\bigg|_{\substack{\tau=\theta \\ y=x^*=y^*=y^*(\theta)}}. \tag{21}$$

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 selection gradient shows that the unique singular trait value is

$$\theta^* = r(\theta^*,\theta^*). \tag{22}$$

But is θ^* convergence stable and uninvadable?

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

$$J(\theta^*) = \frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}\bigg|_{\theta=r(\theta^*,\theta^*)} = -\left(1-\frac{\mathrm{d}r(\theta,\theta)}{\mathrm{d}\theta}\right)\left[\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,y,y^*)}{\partial y}\right]_{\substack{\tau=\theta=\theta^* \\ x^*=y=y^*=y^*(\theta)}} \tag{23}$$

is strictly negative. Since the term in the square brackets is strictly positive, we immediately obtain that $\theta^* = r(\theta^*,\theta^*)$ is convergence stable if and only if $\mathrm{d}r(\theta,\theta)/\mathrm{d}\theta < 1$.

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:

$$H(\theta^*) = \frac{\partial^2 W(\tau,\theta)}{\partial\tau^2}\bigg|_{\tau=\theta} = \left[\left(\frac{\partial x^*(\tau,\theta)}{\partial\tau}\right)^2 K\left(r(\theta^*,\theta^*)\right) + 2\frac{\partial r(\tau,\theta)}{\partial\tau}\frac{\partial x^*(\tau,\theta)}{\partial\tau}\frac{\partial w(x^*,y,y^*)}{\partial y}\right]_{\substack{\tau=\theta=\theta^* \\ x^*=y=y^*=y^*(\theta)}}. \tag{24}$$

Since $K\left(r(\theta^*,\theta^*)\right) < 0$, 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 the Kantian coefficient to be evolutionarily stable as follows, where

$$T = -\frac{\frac{\partial x^*(\tau,\theta)}{\partial\tau}K\left(r(\theta^*,\theta^*)\right)}{2\frac{\partial w(x^*,y,y^*)}{\partial y}}. \tag{25}$$

Result 1. *When interactions take place under incomplete 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 trait effects on relatedness are sufficiently weak. More precisely, $\theta^* = r(\theta^*,\theta^*)$ is convergence stable if and only if $\mathrm{d}r(\theta,\theta)/\mathrm{d}\theta < 1$ and it is locally uninvadable if and only if $\partial r(\tau,\theta)/\partial\tau < T$.*

While the condition for uninvadability is consistent with the results of Alger et al. (2020),[8] our analysis reinforces those results by identifying a sufficient condition for the partly Kantian coefficient equal to neutral relatedness to be also convergence stable. When these conditions are satisfied, individuals in a population at the evolutionary equilibrium will thus behave according to Hamilton's (marginal) rule at the strategy level, i.e., their behavioral equilibrium y^* satisfies

$$\left[\frac{\partial w(y,y^*,y^*)}{\partial y}+\tilde{r}(y^*,y^*)\frac{\partial w(y^*,y,y^*)}{\partial y}\right]_{y=y^*}=0, \tag{26}$$

where the value of relatedness may depend endogenously on the strategy.

Interestingly, many biological scenarios do exhibit weak, or even nil, trait effects on relatedness whereby $\tilde{r}(y^*,y^*)=\tilde{r}$ is independent of y^* in eq. (26).[9] First, in family-structured populations, which cover a large class of interactions (e.g., parent-offspring interactions, interactions between 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, which is here again taken to be the partner's Kantian coefficient taking values in the interval $[0,1]$. However, an individual's own utility function cannot be conditioned on that information: this is what we mean by incomplete plasticity. 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 equilibrium strategies y^*_{d} and y^*_{s}, is relevant, as per the behavioral mechanisms (8)–(10). Hence, the equilibrium

[8]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. (24)), 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. (24) becomes $H(\theta^*)=(\partial x^*(\tau,\theta)/\partial\tau)^2\left[K\left(r(\theta^*,\theta^*)\right)+2(\partial\tilde{r}(x,y^*)/\partial x)(\partial w(x^*,y,y^*)/\partial y)\right]$ 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 $\partial^2W(x,y)/\partial x^2\big|_{x=y}$ corresponds to the term in square brackets in $H(\theta^*)$.

[9]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 (Alger and Weibull, 2013).

strategy y_{s}^* used by each resident in an interaction with another resident satisfies

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

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

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

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

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

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 $x_{\mathrm{d}}^*(\theta,\theta) = y_{\mathrm{d}}^*(\theta,\theta)$ owing to the strict concavity of u in its first argument, which implies that to each strategy played by the opponent there exists a unique best response. Hence, if $\tau = \theta$,

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

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 $(x_{\mathrm{d}}^*, y_{\mathrm{d}}^*)$ to be a Nash equilibrium:

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

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

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

where the second equality shows that the marginal change in utility in the resident population can, as under incomplete information, be expressed as the sum of the effect of own behavior on own fitness and the fitness of the partner weighted by the Kantian coefficient θ. Since u_θ is strictly concave in its first argument, the second-order partial derivative of u_θ, evaluated at $y_{\mathrm{s}}^*(\theta) = y_{\mathrm{d}}^*(\theta,\theta)$, is strictly negative:

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

The system of equations in eq. (31) together implicitly define x_{d}^* and y_{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 $\tau = \theta$:

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

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

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 the equilibrium strategy of mutants always changes as a result of a marginal change in the mutant trait value. 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 $w(x_{\rm d}^*(\tau,\tau), x_{\rm d}^*(\tau,\tau), y_{\rm d}^*(\theta,\theta))|_{\tau=\theta} - w(x_{\rm d}^*(\tau,\theta), y_{\rm d}^*(\theta,\tau), y_{\rm d}^*(\theta,\theta))|_{\tau=\theta} = 0$):

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

From eq. (32), we can write $(1-\theta)\partial w(y,x_{\rm d}^*,y_{\rm d}^*)/\partial y + \theta\partial w(y,x_{\rm d}^*,y_{\rm d}^*)/\partial y + \theta\partial w(x_{\rm d}^*,y,y_{\rm d}^*)/\partial y = 0$ at $\tau=\theta$ where $x=y=y_{\rm s}^*=x_{\rm d}^*=y_{\rm d}^*$. Therefore, we can replace $\partial w(x,y_{\rm d}^*,y_{\rm d}^*)/\partial x = \partial w(y,y_{\rm d}^*,y_{\rm d}^*)/\partial y$ by $-\theta\partial w(x_{\rm d}^*,y,y_{\rm d}^*)/\partial y$ in eq. (39), to obtain

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

Comparing eq. (40) to the selection gradient under incomplete information (see eq. (21)), 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 $\theta=1$, then $\rho(\theta)=0$; using eq. (34) in eq. (40), the selection gradient at $\theta=1$ is thus

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

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 for the Kantian coefficient under complete information and incomplete plasticity, and $\tilde{\theta}$ denote a solution to eq.* (42)*. 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 $\tilde{\theta} \in (0, 1)$, in which case $\theta^* = [r(\theta^*, \theta^*) + \rho(\theta^*)]/[1 + \rho(\theta^*) r(\theta^*, \theta^*)]$. 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 and the behavioral equilibrium y^* in a population at an evolutionary equilibrium θ^* now satisfies

$$\left[\frac{\partial w(y, y^*, y^*)}{\partial y} + \left(\frac{\tilde{r}(y^*, y^*) + \rho(\theta^*)}{1 + \rho(\theta^*)\tilde{r}(y^*, y^*)} \right) \frac{\partial w(y^*, y, y^*)}{\partial y} \right]_{y = y^*} = 0, \tag{43}$$

where relatedness will be independent of the strategies (and types) under weak trait effects on relatedness. Whether the Kantian coefficient exceeds or falls short of relatedness depends on whether the fitness function exhibits, respectively, strategic complementarity or substitutability. The reason that strategic complementarity raises the Kantian coefficient above the value of relatedness stems from the fact that a mutation consisting in an increase in the Kantian coefficient then induces a positive correlated response in the strategy expression by its neighbor, implying that the marginal benefit of the Kantian coefficient is larger than under strategic neutrality (compare eq. (21)–eq. (40)). By contrast, under strategic substitutability, a mutation consisting in an increase in the Kantian coefficient has a negative impact on its neighbor's equilibrium strategy, implying that the marginal benefit marginal benefit of the Kantian coefficient is smaller than under strategic neutrality.

Result 2 further shows that 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. Hence, Theorem 1 of this previous work also establishes that whether the exact

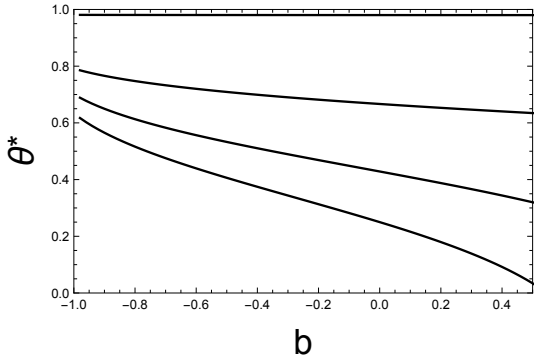

Figure 1: Each curve shows, for the Moran process analyzed in Box 2 with individual fitness (B-i), the singular Kantian coefficient θ^* under complete information and incomplete plasticity, for the linear quadratric fecundity function (44), as a function of parameter b in that function for $a = 0.1$ 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 $m_{\rm b} = 0.01$ whereby $r = (1 - m_{\rm b})/(1 + m_{\rm b}) \approx 0.98$; the second line is for $m_{\rm b} = 0.2$ whereby $r = (1-m_{\rm b})/(1+m_{\rm b}) \approx 0.66$; the third line for $m_{\rm b} = 0.4$ whereby $r = (1-m_{\rm b})/(1+m_{\rm b}) \approx 0.42$; and the last line, where the Kantian coefficient varies over the range $[0, 0.6]$, is for $m_{\rm b} = 0.6$ whereby $r = (1 - m_{\rm b})/(1 + m_{\rm b}) = 0.25$. By computing the Jacobian (A-10) and the Hessian (A-12) coefficients at these values we checked that all these singular Kantian coefficients are indeed both convergence stable and uninvadable.

value of the evolving weight α exceeds or falls short of relatedness depends on whether the fitness function exhibits strategic substitutability, complementarity, or neutrality.[10]

We now examine whether a Kantian coefficient θ^* of Result 2 is convergence stable and uninvadable. Due to the complexity of the expressions for the Jacobian $J(\theta^*)$ and the Hessian $H(\theta^*)$ coefficients at θ^* solving $S(\theta^*) = 0$, presented in Appendix B, 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 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:

$$f(x, y) = 1 + ax - bxy - cx^2 \tag{44}$$

for parameters $a, b, c \in \mathbb{R}$. Then, substituting eq. (44) into individual fitness (B-i), we can evaluate the selection gradient (40), the Jacobian (A-10) and the Hessian (A-12) coefficients. Even for this simple example, eq. (42) is a quartic 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.

[10]Because the behavioral perturbations of other-regarding utility functions, and thus $\rho(\theta)$, typically differ from the ones with the partially Kantian utility function, a singular α will typically differ from the singular Kantian coefficient, however.

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 can not only observe the type of the interaction partner but also condition their preferences on it. Hence, the preferences applied in the interaction become state-specific on the interaction. Specifically, we assume that the type $\theta = (\theta_{\rm d}, \theta_{\rm s})$ of an individual is a two-dimensional quantitative trait ($\theta \in [0,1]^2$) such that $\theta_{\rm s}$ parametrizes an individual's preference (still given by eq. (11)) when individuals in a pair have the same type and $\theta_{\rm d}$ parametrizes an individual's preference when individuals in a pair have different types.

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

$$y_{\rm s}^* \in \arg\max_{y\in\mathcal{X}} u_{\theta_{\rm s}}\left(y, y_{\rm s}^* \mid y_{\rm s}^*\right). \tag{45}$$

The solution thereof defines the equilibrium strategy as a function of only $\theta_{\rm s}$ and we write $y_{\rm s}^*(\theta_{\rm s})$ when this dependence needs to be made explicit. In mutant-mutant interactions both individuals apply the Kantian coefficient $\tau_{\rm s}$ and they both use the equilibrium strategy $x_{\rm s}^*$, satisfying

$$x_{\rm s}^* \in \arg\max_{x\in\mathcal{X}} u_{\tau_{\rm s}}\left(x, x_{\rm s}^* \mid y_{\rm s}^*\right). \tag{46}$$

This defines the equilibrium strategy as a function of both $\tau_{\rm s}$ and $\theta_{\rm s}$ (since $y_{\rm s}^*$ depends on $\theta_{\rm s}$), which we write $x_{\rm s}^*(\tau_{\rm s}, \theta_{\rm s})$. Finally, in mutant-resident pairs where $\tau \neq \theta$, the mutant applies the Kantian coefficient $\tau_{\rm d}$ 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:

$$\begin{cases} y_{\rm d}^* \in \arg\max_{y\in\mathcal{X}}\ u_{\theta_{\rm d}}(y, x_{\rm d}^* \mid y_{\rm s}^*) \\ x_{\rm d}^* \in \arg\max_{x\in\mathcal{X}}\ u_{\tau_{\rm d}}(x, y_{\rm d}^* \mid y_{\rm s}^*), \end{cases} \tag{47}$$

which leads to the dependence of both equilibrium strategies on $\tau_{\rm d}, \theta_{\rm d}$, and $\theta_{\rm s}$ and we write $x_{\rm d}^*(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s})$ and $y_{\rm d}^*(\theta_{\rm d}, \tau_{\rm d}, \theta_{\rm s})$. In mutant-resident pairs with the same type, which occurs if individuals from different lineage interact but are of the same type $\tau = \theta$, then each individual in the pair applies the Kantian coefficient $\theta_{\rm s}$ and therefore expresses strategy $y_{\rm s}^*(\theta_{\rm s})$.

3.3.2 Evolutionary equilibrium

The fixed point eqs (45)–(47) define the behavioral mechanisms (8)–(10) under complete information with complete plasticity and it follows that the invasion fitness can be written as

$$W(\tau, \theta) = [1 - r(\tau, \theta)]\, W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s}) + r(\tau, \theta) W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}), \tag{48}$$

where

$$W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s}) = \begin{cases} w(x_{\rm d}^*(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s}), y_{\rm d}^*(\theta_{\rm d}, \tau_{\rm d}, \theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})) & \text{if } (\tau_{\rm d}, \tau_{\rm s}) \neq (\theta_{\rm d}, \theta_{\rm s}) \\ w(y_{\rm s}^*(\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s})) = 1 & \text{if } (\tau_{\rm d}, \tau_{\rm s}) = (\theta_{\rm d}, \theta_{\rm s}), \end{cases} \tag{49}$$

and

$$W_{\rm s}(\tau_{\rm s},\theta_{\rm s})=w(x_{\rm s}^{*}(\tau_{\rm s},\theta_{\rm s}),x_{\rm s}^{*}(\tau_{\rm s},\theta_{\rm s}),y_{\rm s}^{*}(\theta_{\rm s})). \tag{50}$$

Each component of the mutant trait $\tau=(\tau_{\rm d},\tau_{\rm s})$ thus affects invasion fitness differently. The first trait component, $\tau_{\rm d}$, only affects the part of invasion fitness $W_{\rm d}$ emanating from the interaction of a mutant with a resident. Invasion fitness is not differentiable in trait component $\tau_{\rm d}$ because, as emphasized by eq. (49), there is generally a discrete jump in the equilibrium strategies at $\tau_{\rm d}=\theta_{\rm d}$, from $(x_{\rm d}^{*},y_{\rm d}^{*})$ to $(y_{\rm s}^{*},y_{\rm s}^{*})$ since as long as $\tau\neq\theta$, individuals in mutant-resident pairs play the equilibrium $(x_{\rm d}^{*},y_{\rm d}^{*})$, while at $\tau=\theta$ the pair plays $(y_{\rm s}^{*},y_{\rm s}^{*})$ [see the example below for an illustration]. The second trait component, $\tau_{\rm s}$, not only affects the part of invasion fitness $W_{\rm s}$ emanating from the interaction of a mutant with another mutant, but also $W_{\rm d}$, since if $\tau_{\rm d}=\theta_{\rm d}$ the resident applies the value $\theta_{\rm d}$ when interacting with a mutant if $\tau_{\rm s}\neq\theta_{\rm s}$ but instead the value $\theta_{\rm s}$ if $\tau_{\rm s}=\theta_{\rm s}$. The invasion fitness (48) is thus not differentiable in the mutant type $(\tau_{\rm d},\tau_{\rm s})$. Moreover, as will be shown and explained in detail in an example below, there is typically an infinite number of uninvadable types $\theta\in[0,1]^2$. In spite of these challenges, we have identified two simple sufficient conditions for a type to be uninvadable (but we were unable to conclude on convergence stability).

To state our result, for any $\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}\in[0,1]$ define

$$\begin{cases}\tilde{y}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s})\in\arg\max_{y\in\mathcal{X}}\ u_{\theta_{\rm d}}\left(y,\tilde{x}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})\mid y_{\rm s}^{*}(\theta_{\rm s})\right)\\ \tilde{x}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})\in\arg\max_{x\in\mathcal{X}}\ u_{\tau_{\rm d}}\left(x,\tilde{y}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s})\mid y_{\rm s}^{*}(\theta_{\rm s})\right)\end{cases} \tag{51}$$

and

$$\widetilde{W}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})=w(\tilde{x}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s}),\tilde{y}(\theta_{\rm d},\tau_{\rm d},\theta_{\rm s}),y_{\rm s}^{*}(\theta_{\rm s})). \tag{52}$$

Result 3. *A sufficient condition for $W_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})\leq 1$ for all $\tau_{\rm d}\in[0,1]$ and $W_{\rm s}(\tau_{\rm s},\theta_{\rm s})\leq 1$ for all $\tau_{\rm s}\in[0,1]$ (and thus $W(\tau,\theta)\leq 1$ for all $\tau=(\tau_{\rm d},\tau_{\rm s})\in[0,1]^2$) is that $\theta=(\theta_{\rm d},\theta_{\rm s})=(\theta_{\rm d}^{*},1)$, where $\theta_{\rm d}^{*}\in[0,1]$ is such that $\widetilde{W}(\theta_{\rm d}^{*},\theta_{\rm d}^{*},1)\geq\widetilde{W}(\tau_{\rm d},\theta_{\rm d}^{*},1)$ for all $\tau_{\rm d}\in[0,1]$.*

For a formal proof of this result, see Appendix C. To understand this result, note first that $\theta_{\rm s}=1$ guarantees that a mutant cannot achieve a higher fitness when interacting with another mutant, than a resident does when interacting with another resident. This value of the Kantian coefficient indeed implies that both individuals in the interaction act as "social planners": they choose that strategy which, when chosen by both individuals, maximizes their fitness. Second, note that for $\tau_{\rm d}\neq\theta_{\rm d}$, the system of equations (51) is mathematically equivalent to the system of equations that characterizes the equilibrium strategies between a resident and a mutant (see (47)), and $\widetilde{W}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})$ is equivalent to $W_{\rm d}(\tau_{\rm d},\theta_{\rm d},\theta_{\rm s})$.[11] Hence, the value $\theta_{\rm d}^{*}$ defined in the result is such that if residents apply $\theta_{\rm d}=\theta_{\rm d}^{*}$ when interacting with mutants, the residents get a higher fitness than the mutants do in mutant-resident interactions, for any value of $\tau_{\rm d}\neq\theta_{\rm d}$. Thus, $(\theta_{\rm s},\theta_{\rm d})=(\theta_{\rm d}^{*},1)$ is a "safe bet" as an uninvadable point.

However, and as mentioned before, there are typically many other uninvadable types. To see why, it is useful to draw a parallel between the function $\widetilde{W}$, defined in eq. (52), and the fitness of a mutant under the complete

[11] Note that the definition of the function $\widetilde{W}$ allows us to evaluate W_d under the hypothesis that the resident would apply the Kantian coefficient $\theta_{\rm d}$ even when interacting with another individual with the same type. This is useful for the proof of the result, and for the argument developed in the next paragraph.

information and incomplete plasticity scenario when relatedness is equal to zero, defined in (38). These are mathematically equivalent, except that in eq. (52) individuals in the resident population play strategy $y_{\rm s}^*(\theta_{\rm s})$, where $\theta_{\rm s}$ may differ from $\theta_{\rm d}$. This parallel allows us to realize that the same arguments as for the derivation of Result 2 can be applied here, to conclude that a relevant candidate value for $\theta_{\rm d}^*$ is the one satisfying $\theta_{\rm d}^* = \max\{0, \tilde{\rho}(\theta_{\rm d}^*)\}$ where

$$\tilde{\rho}(\theta_{\rm d}) = \left. \frac{\frac{\partial \tilde{y}(\theta_{\rm d},\tau_{\rm d},1)}{\partial \tau_{\rm d}}}{\frac{\partial \tilde{x}(\tau_{\rm d},\theta_{\rm d},1)}{\partial \tau_{\rm d}}} \right|_{\tau_{\rm d}=\theta_{\rm d}}. \tag{53}$$

From (differentiable) fitness $\widetilde{W}(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s})$ and following the same arguments as those leading up to Result 2, we know that an uninvadable type satisfying $\theta_{\rm d}^* = \tilde{\rho}(\theta_{\rm d}^*)$ cannot be equal to one, i.e., $\theta_{\rm d}^* < 1$. Based on the same argument as used in the previous paragraph, we have $W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}^*, 1) \leq \widetilde{W}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1) < \widetilde{W}(1,1,1) = 1$. The inequality $\widetilde{W}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1) < \widetilde{W}(1,1,1)$ follows from the fact that $\theta_{\rm s} = 1$ implies that both individuals in the interaction act as "social planners" and thus maximize their fitness, and hence interaction partners with $\theta_{\rm d}^* < 1$ do *not* act as social planners, thereby choosing a strategy that does not maximise their fitness (see also Appendix C).

We can thus conclude that $\theta_{\rm d} = \theta_{\rm d}^*$ implies that $W_{\rm d}(\tau, \theta) < 1$ for any $\tau_{\rm d} \neq \theta_{\rm d}^*$. Hence, we can further conclude that $\theta = (\theta_{\rm d}, \theta_{\rm s}) = (\theta_{\rm d}^*, 1)$ implies that $W(\tau, \theta) < 1$ for any $\tau \neq \theta$. It is this discrete jump between the invasion fitness of any mutant $\tau \neq \theta$ and the residents' fitness that opens the door for the multiplicity of uninvadable types. For example, $(\theta_{\rm d}, \theta_{\rm s}) = (\theta_{\rm d}^* - \varepsilon, 1)$ for some sufficiently small $\varepsilon > 0$ would also be uninvadable: with this resident type, a mutant can achieve a higher fitness than a resident in a mutant-resident interaction, but one can always find ε small enough so that this advantage would be too small to overcome the difference $1 - W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}^* - \varepsilon, 1)$. Likewise, $(\theta_{\rm s}, \theta_{\rm d}) = (1 - \varepsilon, \theta_{\rm d}^*)$ for some sufficiently small $\varepsilon > 0$ would also be uninvadable: with this resident type, mutants can achieve a higher fitness when interacting with each other than when two residents interact, but one can always find ε small enough so that this advantage would be too small to overcome the difference $1 - W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}^*, 1 - \varepsilon)$. Note that this example further shows that $\theta_{\rm s} = 1$ is not necessary for θ to be uninvadable.

In order to illustrate Result 3, we work out an example assuming a constant relatedness of $1/2$ and an individual fitness function of the form $w(x_i, x_{-i}, y) = f(x_i, x_{-i})/f(y, y)$, with fecundity $f(x_i, x_{-i}) = 1 + ax_i - bx_i x_{-i} - cx_i^2$ given by the linear-quadratic function (44). This model could be thought of as interactions between siblings in a family-structured semelparous population. Given these assumptions, the invasion fitness (48) is

$$W(\tau, \theta) = \frac{1}{2} W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s}) + \frac{1}{2} W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}), \tag{54}$$

with

$$W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}, \theta_{\rm s}) = \begin{cases} \frac{f(x_{\rm d}^*(\tau_{\rm d},\theta_{\rm d}), y_{\rm d}^*(\theta_{\rm d},\tau_{\rm d}))}{f(y_{\rm s}^*(\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s}))} & \text{if } (\tau_{\rm d}, \tau_{\rm s}) \neq (\theta_{\rm d}, \theta_{\rm s}) \\ 1 & \text{if } (\tau_{\rm d}, \tau_{\rm s}) = (\theta_{\rm d}, \theta_{\rm s}), \end{cases} \tag{55}$$

and

$$W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}) = \frac{f(x_{\rm s}^*(\tau_{\rm s}), x_{\rm s}^*(\tau_{\rm s}))}{f(y_{\rm s}^*(\theta_{\rm s}), y_{\rm s}^*(\theta_{\rm s}))}. \tag{56}$$

In force of eq. (45), the (Nash) equilibrium strategy for a resident pair is

$$y_{\rm s}^*(\theta_{\rm s}) = \frac{a}{2c + b + b\theta_{\rm s}}, \tag{57}$$

while in force of eq. (47), the strategies for a mutant-resident pair are

$$\begin{aligned} x_{\mathrm{d}}^{*}(\tau_{\mathrm{d}},\theta_{\mathrm{d}}) &= \frac{a[2c-b(1-2\theta_{\mathrm{d}}-\tau_{\mathrm{d}})]}{4c^{2}+4bc(\theta_{\mathrm{d}}+\tau_{\mathrm{d}})-b^{2}(1-\theta_{\mathrm{d}}-\tau_{\mathrm{d}}+3\theta_{\mathrm{d}}\tau_{\mathrm{d}})} \\ y_{\mathrm{d}}^{*}(\theta_{\mathrm{d}},\tau_{\mathrm{d}}) &= \frac{a[2c-b(1-2\theta_{\mathrm{d}}-\tau_{\mathrm{d}})]}{4c^{2}+4bc(\theta_{\mathrm{d}}+\tau_{\mathrm{d}})-b^{2}(1-\theta_{\mathrm{d}}-\tau_{\mathrm{d}}+3\theta_{\mathrm{d}}\tau_{\mathrm{d}})}, \end{aligned} \tag{58}$$

and in force of eq. (46), the equilibrium strategy for a mutant pair is

$$x_{\mathrm{s}}^{*}(\tau_{\mathrm{s}}) = \frac{a}{2c+b+b\tau_{\mathrm{s}}}. \tag{59}$$

For this model, we can also write for any $\tau_{\mathrm{d}},\theta_{\mathrm{d}} \in [0,1]$ that

$$\widetilde{W}(\tau_{\mathrm{d}},\theta_{\mathrm{d}},1) = \frac{f(x_{\mathrm{d}}^{*}(\tau_{\mathrm{d}},\theta_{\mathrm{d}}),y_{\mathrm{d}}^{*}(\theta_{\mathrm{d}},\tau_{\mathrm{d}}))}{f(y_{\mathrm{s}}^{*}(1),y_{\mathrm{s}}^{*}(1))}, \tag{60}$$

and the candidate uninvadable Kantian coefficient thus satisfies

$$\theta_{\mathrm{d}}^{*} = \tilde{\rho}(\theta_{\mathrm{d}}^{*}) = \left.\frac{\frac{\partial y_{\mathrm{d}}^{*}(\theta_{\mathrm{d}},\tau_{\mathrm{d}}))}{\partial \tau_{\mathrm{d}}}}{\frac{\partial x_{\mathrm{d}}^{*}(\tau_{\mathrm{d}},\theta_{\mathrm{d}})}{\partial \tau_{\mathrm{d}}}}\right|_{\tau_{\mathrm{d}}=\theta_{\mathrm{d}}^{*}} = -\frac{b(1-\theta_{\mathrm{d}}^{*})}{2(c+b\theta_{\mathrm{d}}^{*})}, \tag{61}$$

which has a relevant root in the interval $[0,1]$ when $b<0$ (which implies that the strategies are strategic complements), given by

$$\theta_{\mathrm{d}}^{*} = \frac{b-2c+\sqrt{4c^{2}-4bc-7b^{2}}}{4b}. \tag{62}$$

It is straightforward to check, for instance by computing $\partial^{2}\widetilde{W}(\tau_{\mathrm{d}},\theta_{\mathrm{d}},1)/\partial\tau_{\mathrm{d}}^{2}$ at $\tau_{\mathrm{d}}=\theta_{\mathrm{d}}=\theta_{\mathrm{d}}^{*}$, that there is a range of parameter values where this Kantian coefficient is uninvadable and this is illustrated in panel A of Fig. 2.

This example illustrates the two features of Result 3 discussed in general terms above. First, as long as $\tau \neq \theta$, individuals in mutant-resident pairs play the equilibrium (58), while at $\tau=\theta$, they play eq. (57). This makes the strategies discontinuous at $\tau_{\mathrm{d}}=\theta_{\mathrm{d}}$, which in turn implies that invasion fitness is not differentiable. Despite this discontinuity embedded in fitness $W_{\mathrm{d}}(\tau_{\mathrm{d}},\theta_{\mathrm{d}},\theta_{\mathrm{s}})$ (eq. (55)), Result 3 shows that we can use the function $\widetilde{W}(\tau_{\mathrm{d}},\theta_{\mathrm{d}},1)$ to straightforwardly calculate $\theta_{\mathrm{d}}=\theta_{\mathrm{d}}^{*}=\max\{0,\tilde{\rho}(\theta_{\mathrm{d}}^{*})\}$, and then use the Hessian to check that $(\theta_{\mathrm{s}},\theta_{\mathrm{d}})=(1,\theta_{\mathrm{d}}^{*})$ is indeed uninvadable. This is illustrated in panel A of Fig. 2. Second, there are multiple uninvadable types; this is illustrated in panels B and C of Fig. 2.

Our formalization of preference evolution under complete information and plasticity makes it difficult to reach any specific conclusion about long term evolution for two reasons. First, because invasion fitness is not differentiable, a different toolkit than the usual multidimensional convergence stability criterion is needed to characterize the attractor points of the evolutionary dynamic (Leimar, 2009). Second, the multiplicity of uninvadable equilibria compounded with the non differentiability makes the focus on monomorphic population questionable and a treatment with polymorphic populations appears required. It would thus be relevant to analyze a full dynamics model of preference evolution with mutation and selection under complete information and plasticity to reach more definite results and also to determine whether the equilibrium $\theta=(\theta_{\mathrm{d}}^{*},1)$ identified in Result 3 plays a special role, for instance whether it is a stochastically stable trait value (*sensu* Foster and Young, 1990).

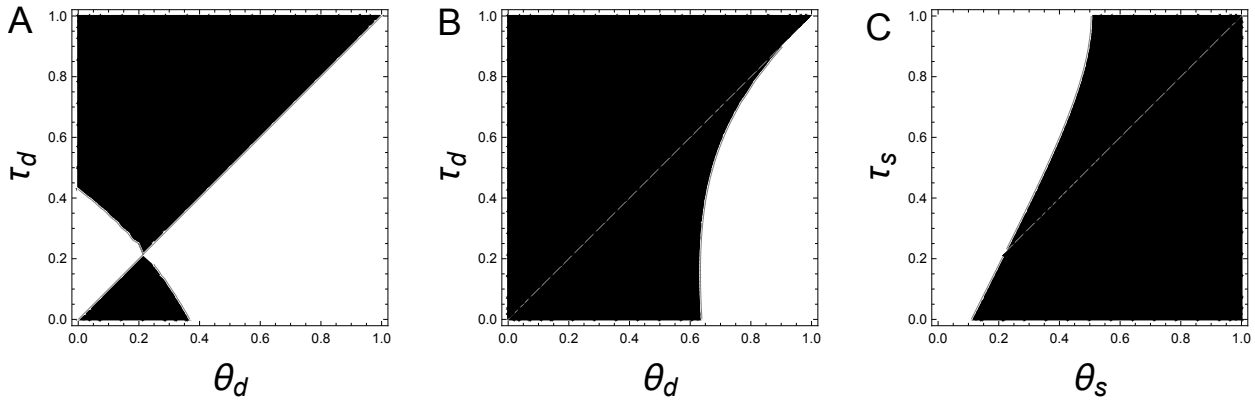

Figure 2: Pairwise invasibility plots. **Panel A** displays the difference $\widetilde{W}(\tau_{\rm d}, \theta_{\rm d}, 1) - \widetilde{W}(\theta_{\rm d}, \theta_{\rm d}, 1)$ for all combinations of $\theta_{\rm d}$ (x-axis) and $\tau_{\rm d}$ (y-axis) values (and thus holding $\theta_{\rm s} = 1$ fixed), for the parameter values $a = 0.1$, $b = -0.5$ and $c = 1$ of the linear quadratic game (44) ($\widetilde{W}(\tau_{\rm d}, \theta_{\rm d}, 1)$ is thus defined by (60). The difference is zero on the diagonal, since $\tau_{\rm d} = \theta_{\rm d}$. The black region depicts all the combinations $(\theta_{\rm d}, \tau_{\rm d})$ such that the difference is negative, while the white region outside the diagonal depicts all the combinations $(\theta_{\rm d}, \tau_{\rm d})$ such that the difference is strictly positive. The candidate given by eq. (62) is $\theta_{\rm d}^* \approx 0.22$, which the graph shows is uninvadable. **Panel B** displays the difference $W(\tau, \theta) - W(\theta, \theta)$ determined by eqs. (54)–(59)of a mutant $\tau = (\tau_{\rm d}, 1)$ in a resident $\theta = (\theta_{\rm d}, 1)$ population for all combinations of resident $\theta_{\rm d}$ (x-axis) and mutant $\tau_{\rm d}$ (y-axis) trait values (and thus holding $\theta_{\rm s} = 1$ and $\tau_{\rm s} = 1$ fixed) for the parameter values $a = 0.1$ $b = -0.5$ and $c = 1$ of the linear quadratic game (44). On the diagonal, the difference equals zero (and invasion fitness equals one) since $\tau_{\rm d} = \theta_{\rm d}$. The black region depicts all combinations $(\theta_{\rm d}, \tau_{\rm d})$ such that the difference is negative, i.e., $W(\tau, \theta) < 1$, so that the mutant $\tau_{\rm d}$ cannot invade, while the white region outside the diagonal depicts all combinations $(\theta_{\rm d}, \tau_{\rm d})$ such that the difference is positive, $W(\tau, \theta) > 1$, so that the mutant can invade. The graph thus shows that all values of $(\theta_{\rm d}, 1)$ with $\theta_{\rm d}$ below approximately 0.6 are uninvadable. There is thus a multiplicity of uninvadable types, which necessarily contains that of Panel A. Note that resident values above approximately $\theta_{\rm d} \approx 0.9$ are invadable by any mutation $\tau_{\rm d} < \theta_{\rm d}$, while resident values between approximatively 0.6 and 0.9 are invadable by mutants with $\tau_{\rm d} = \theta_{\rm d} - \delta$ for $\delta > 0$ large enough. **Panel C** displays $W(\tau, \theta) - W(\theta, \theta)$ determined by eqs. (54)–(59) for a mutant $\tau = (\tau_{\rm d}, \tau_{\rm s})$ in a resident $\theta = (\theta_{\rm d}, \theta_{\rm s})$ population for all combinations of resident $\theta_{\rm s}$ (x-axis) and mutant $\tau_{\rm s}$ (y-axis) trait values holding $\tau_{\rm d} = \theta_{\rm d} = \tilde{\rho}(\theta_{\rm d})$ fixed and given by eq. (61) under the parameter values $a = 0.1$ $b = -0.5$ and $c = 1$ of the linear quadratic game (44). As for Panel B, the black region depicts all combinations $(\theta_{\rm s}, \tau_{\rm s})$ such that the difference is negative, i.e., $W(\tau, \theta) < 1$, so that the mutant $\tau_{\rm s}$ cannot invade, while the white region outside the diagonal depicts all combinations $(\theta_{\rm s}, \tau_{\rm s})$ such that the difference is positive, $W(\tau, \theta) > 1$, so that the mutant can invade. The graph thus shows that all values of $(\theta_{\rm s}, \tilde{\rho}(\theta_{\rm d}^*))$ with $\theta_{\rm s}$ above approximately 0.5 are uninvadable.

4 Discussion

By investigating the evolution of semi-Kantian preferences under different informational and behavioral plasticity assumptions in group-structured populations, we have extended the evolutionary viability analysis of this class of preferences. While we restricted attention to pairwise interactions, and preferences 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 randomly sampled mutant from the same lineage, 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 focused on the more restricted setting where utility functions are parametrized by a single quantitative trait. This allowed us to cover not only uninvadability in a complementary and less abstract way, 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. One 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 the specifics of the individual fitness function so that the Kantian coefficient will depend on life-history parameters such as survival, migration, etc. [recall the example of the fitness function (B-i)]. Gradual evolution thus 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 examined (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. By being observable, a mutant's preference type can thus induce a resident to adopt a different strategy than the one adopted in an interaction with another resident, an effect that is absent under incomplete information. Although we established necessary conditions for a Kantian coefficient value to be uninvadable and illustrated uninvadablility and convergence stability under a linear quadratic game (Fig. (1)), 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 can both observe the opponent's type and also condition its preferences on it. This is akin to a green-beard or secret handshake mechanism (Hamilton, 1964b; Grafen, 1990; Robson, 1990), but at the preference level rather than at 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, individuals are thus no longer committed to respond according to one and the same Kantian coefficient. Residents are therefore less exploitable by mutants and individuals can be regarded as implementing multiple selves (Lester, 2015) since their preferences and motivations are context-dependent. As we showed, this implies that residents can be pure Kantians when interacting with each other, and still be uninvadable: they can 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 then use the strategy which yields the highest 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" preferences, which results in efficient strategy profiles, are stable. It is also reminiscent of results obtained by Wang and Wu (2023) in a model where the search and matching of interacting partners is endogenous but costless: the authors demonstrate the evolutionary stability of a preference for interacting with an individual with the same type as self combined with play of an efficient strategy profiles. We were, however, were not able to characterize convergence stability nor the Kantian coefficient when individuals interact with others with a type different from their own type. Ascertaining this in the non-differentiable setting of complete information with complete plasticity is thus left for future work.

We derived these three results assuming existence of a unique behavioral equilibrium in the resident population and when mutant-resident (or mutant-mutant) interactions occur. Given that it is common that interactions produce a multiplicity of (Nash) equilibria, an important avenue for future research will be to examine the robustness of our results in settings with multiple equilibria. Several questions will need to be addressed. First, one can imagine several alternative formalizations of the interaction between individuals when multiple equilibria are possible. For instance, one could assume that different pairs of individuals in the resident population play different equilibria; such a model

[12]To see that polymorphism is likely under preference evolution 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 pessimism 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.

would likely require tools from the literature on class-structured populations (since an average individual may then play different equilibria and thus be in different contexts or states, the defining feature of class structured populations e.g., Avila and Mullon, 2023). An alternative would be to assume that the same equilibrium is played in all resident-resident interactions; our formalization should then apply to each equilibrium separately (with implications for possibly different relatedness coefficients depending on the equilibrium being played). The same modeling choices will arise when there exist multiple equilibria in mutant-resident pairs. Second, the definitions of uninvadability and convergence stability may need to be adapted. In particular, for uninvadability, would a preference type be deemed uninvadable only if it is uninvadable for all possible formalizations of interactions in the presence of multiple equilibria? A similar question arises for convergence stability. For existing attempts to address some of these questions, see Ok and Vega-Redondo (2001); Dekel et al. (2007); Alger and Weibull (2013, 2016); Alger et al. (2020); Alger and Weibull (2023); Wang and Wu (2023).

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. By contrast to strategy evolution models, which predict behavioral patterns for stationary environments, preference evolution models allow to make predictions about behavioral change in a new environment, and can be tested, using either field data or experimental data. A key prediction of our model is that equilibrium behavior at evolutionary equilibrium should be in accordance with Hamilton's marginal rule expressed at the strategy level only if evolution operated on interactions under incomplete information (recall equation (26)), and that if observed deviations from this rule resulted from interactions taking place under complete information, the deviation from Hamilton's marginal rule should depend on the specifics of the interaction at hand (recall equation (43)). In some experiments under incomplete information, humans do appear to conform to behave according to Hamilton's (marginal) rule at the strategy level (Levy and Lo, 2022).

Economic games can further be used to discriminate between different preferences in controlled laboratory experiments. For example, Fisman et al. (2007) and Bruhin et al. (2019) use dictator games to estimate individuals' preference parameters assuming other-regarding preferences. More recent studies in this literature have been inspired by results in the evolutionary theory of preferences to design economic games that further allow the researchers to discriminate between other-regarding and semi-Kantian preferences (Miettinen et al., 2020), and to estimate the importance of Kantian concerns relative to other-regarding concerns (Van Leeuwen and Alger, 2022). The latter study indicates that many individuals appear to be driven by a combination of self-interest, semi-Kantian concerns, and other-regard. This finding is in line with evolutionary arguments showing that it is important to distinguish between preferences expressed in terms of fitness consequences and preferences expressed in terms of payoff consequences (Lehmann et al., 2015; Alger et al., 2020). The qualitative nature of evolved preferences at the material payoff level indeed 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 populations, but tend to disagree in spatially structured populations owing to the presence of local competition between individuals. Future research could adopt a similar approach to the one we develop here, by analyzing evolution of preferences expressed at the level of material payoffs under a variety of informational assumptions.

In this paper we have sought to bring together modeling tools from evolutionary biology and from economics in order to investigate the evolution of preferences that guide behavior in strategic interactions. We hope that our formalization has illustrated some of the nuances, intricacies, and richness of research endeavours that seek to analyse more completely the evolutionary dynamics of behavioral mechanisms, and that it will inspire future studies into the many open theoretical and empirical research questions.

Declarations

Competing interests : All authors declare that they have no conflicts of interest.

Author's contributions: L.L. conceived the model, I.A. and L.L. derived the results, L.L. wrote the first draft and both authors finalized the manuscript.

Funding : I.A. acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program), and funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC Evolving Economics).

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). It then follows from standard results on multitype branching processes that the lineage of a single τ mutant goes extinct with probability one if, and only if, $W(\tau,\theta)\leq 1$, otherwise the lineage spreads into the population when rare and becomes infinitely large (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 $s \in [0,1]$.

Appendix A Behavioral perturbation under incomplete information

We here show that the behavioral perturbation (18) is always positive, i.e., $\partial x^*(\tau,\theta)/\partial\tau|_{\tau=\theta} > 0$. We proceed in two steps. First, we evaluate the perturbation of a mutant's reaction function and show that this is positive. Second, we show that this latter perturbation is sign equivalent to $\partial x^*(\tau,\theta)/\partial\tau|_{\tau=\theta}$.

For the first step recall from eq. (13) that a specific mutant's expected utility is

$$U(x \mid x^*, y^*) = [1-\tilde{r}(x^*,y^*)]u_\tau\left(x, y^* \mid y^*\right) + \tilde{r}(x^*,y^*)u_\tau\left(x, x^* \mid y^*\right), \tag{A-1}$$

where the notation $\mid x^*, y^*$ emphasises that we are here holding the strategies of the other mutant and resident individuals as given. A mutant's best response to (x^*, y^*) is implicitly defined by the necessary first-order condition:

$$\frac{\partial U(x \mid x^*, y^*)}{\partial x} = [1-\tilde{r}(x^*,y^*)]\frac{\partial u_\tau\left(x, y^* \mid y^*\right)}{\partial x} + \tilde{r}(x^*,y^*)\frac{\partial u_\tau\left(x, x^* \mid y^*\right)}{\partial x} = 0, \tag{A-2}$$

which, using eq. (11), is written

$$\begin{aligned} & [1-\tilde{r}(x^*,y^*)](1-\tau)\frac{\partial w\left(x, y^*, y^*\right)}{\partial x} + [1-\tilde{r}(x^*,y^*)]\tau\frac{\partial w\left(x, x, y^*\right)}{\partial x} \\ + \quad & \tilde{r}(x^*,y^*)(1-\tau)\frac{\partial w\left(x, x^*, y^*\right)}{\partial x} + \tilde{r}(x^*,y^*)\tau\frac{\partial w\left(x, x, y^*\right)}{\partial x} = 0. \end{aligned} \tag{A-3}$$

This equation implicitly defines the specific mutant's optimal strategy $x(\tau, x^*, y^*)$ as a function of τ, x^*, and y^* (which depends on θ). In other words, it defines this mutant's *reaction function*, which specifies its utility-maximizing strategy for each (x^*, y^*) (e.g., Fudenberg and Tirole, 1991, p. 14). We use an index i to denote the partial derivative with respect to the i-th argument of the individual fitness function to rewrite eq. (A-3) as follows:

$$\begin{aligned} & [1-\tilde{r}(x^*,y^*)](1-\tau)w_1\left(x, y^*, y^*\right) + [1-\tilde{r}(x^*,y^*)]\tau[w_1\left(x, x, y^*\right) + w_2\left(x, x, y^*\right)] \\ + \quad & \tilde{r}(x^*,y^*)(1-\tau)w_1\left(x, x^*, y^*\right) + \tilde{r}(x^*,y^*)\tau[w_1\left(x, x, y^*\right) + w_2\left(x, x, y^*\right)] = 0, \end{aligned} \tag{A-4}$$

which simplifies to

$$(1-\tau)\left[[1-\tilde{r}(x^*,y^*)]w_1\left(x, y^*, y^*\right) + \tilde{r}(x^*,y^*)w_1\left(x, x^*, y^*\right)\right] + \tau[w_1\left(x, x, y^*\right) + w_2\left(x, x, y^*\right)] = 0. \tag{A-5}$$

Applying the implicit function theorem holding (x^*, y^*) fixed, we obtain:

$$\frac{\partial x(\tau, x^*, y^*)}{\partial \tau} = -\frac{-[1-\tilde{r}(x^*,y^*)]w_1\left(x, y^*, y^*\right) - \tilde{r}(x^*,y^*)w_1\left(x, x^*, y^*\right) + w_1\left(x, x, y^*\right) + w_2\left(x, x, y^*\right)}{(1-\tau)A + \tau B}, \tag{A-6}$$

where

$$\begin{aligned} A &= [1-\tilde{r}(x^*,y^*)]w_{11}\left(x, y^*, y^*\right) + \tilde{r}(x^*,y^*)w_{11}\left(x, x^*, y^*\right) \\ B &= w_{11}\left(x, x, y^*\right) + 2w_{12}\left(x, x, y^*\right) + w_{22}\left(x, x, y^*\right). \end{aligned} \tag{A-7}$$

Strict concavity of u_τ for all $\tau \in [0,1]$ implies that $A < 0$ and $B < 0$ and so the denominator of eq. (A-6) is strictly negative. At equilibrium, and locally at $\tau = \theta$, we have $x = x^* = y^*$, implying that the numerator reduces to $w_2(x, x, y^*)$, which is strictly positive. Hence, the perturbation $\partial x(\tau, x^*, y^*)/\partial\tau|_{\tau=\theta}$, which measures this mutant's (infinitesimal) change in its reaction function $x(\tau, x^*, y^*(\theta))$ when the value of its Kantian coefficient is infinitesimally increased at $\tau = \theta$, is strictly positive.

We turn now to the second step to show why the result derived in the first step implies that the mutant equilibrium strategy x^* must increase as a result of an (infinitesimal) increase in the value of τ for all mutants. First, note that such an increase in τ has no effect on the resident strategy (see eq. (13)), implying that we can indeed hold y^* fixed, as we did above. Hence, for a given τ and y^*, the equilibrium strategy $x^*(\tau,\theta)$ is a fixed point of the mutant's best-response function (see the second line of eq. (13)) and this is a point on the curve described by the reaction function $x(\tau,x^*,y^*)$ for any given (τ,y^*).. Because $\partial x(\tau,x^*,y^*)/\partial\tau|_{\tau=\theta}$ describes how the response function of one mutant is slightly displaced by the introduction of a small mutant deviation at $\tau=\theta$, and because we proved above that any such deviation is positive for any strategy x^* played by other mutants, the equilibrium strategy $x^*(\tau,\theta)$ must locally vary with the same sign as the reaction function varies, i.e, $\partial x^*(\tau,\theta)/\partial\tau|_{\tau=\theta}\propto\partial x(\tau,x^*,y^*)/\partial\tau|_{\tau=\theta}$.

Appendix B 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

$$J(\theta^*)=\frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\left[(r(\theta^*,\theta^*)-\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2}+(1-\theta^*)(1+r(\theta^*,\theta^*))\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta}+(1-\theta^* r(\theta^*,\theta^*))\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right.$$
$$\left.+\frac{\left[1-r(\theta^*,\theta^*)^2-(1-\theta^{*2})\frac{{\rm d}r(\theta,\theta)}{{\rm d}\theta}\right]}{r(\theta^*,\theta^*)-\theta^*}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right]_{\tau=\theta=\theta^*}. \quad \text{(A-8)}$$

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

$$\theta^*=\frac{r(\theta^*)+\rho(\theta^*)}{1+r(\theta^*,\theta^*)\rho(\theta)^*}, \quad \text{(A-9)}$$

which, on substituting into eq. (A-8), 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

$$J(\theta^*)=-\frac{\partial w(x_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\left[\frac{\left(1-r(\theta^*,\theta^*)^2\right)}{1+r(\theta^*,\theta^*)\rho(\theta^*)}\left(\rho(\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2}-(1-\rho(\theta^*))\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta}-\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right)\right.$$
$$\left.+\frac{\left((1+r(\theta^*,\theta^*)\rho(\theta^*))^2-\frac{{\rm d}r(\theta)}{{\rm d}\theta}(1-\rho(\theta^*)^2)\right)}{1+r(\theta^*,\theta^*)\rho(\theta^*)}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right]_{\tau=\theta=\theta^*}. \quad \text{(A-10)}$$

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

$$H(\theta) = \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2}\left(\frac{\partial w(y,y_{\rm d}^*,y_{\rm d}^*)}{\partial y} + r(\theta,\theta)\frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\right) + 2r(\theta\theta)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta}\left(\frac{\partial w(y,y_{\rm d}^*,y_{\rm d}^*)}{\partial y} + \frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\right)$$
$$+\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\left(r(\theta,\theta)\frac{\partial w(y,y_{\rm d}^*,y_{\rm d}^*)}{\partial y} + \frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\right) + \frac{\partial^2 w(x,y_{\rm d}^*,y_{\rm d}^*)}{\partial x^2}\left[r(\theta,\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + 2r(\theta,\theta)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right.$$
$$\left. + \left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right] + 2\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x\partial y}\left[r(\theta,\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + [1+r(\theta,\theta)]\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau} + r(\theta,\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right]$$
$$+\frac{\partial^2 w(x_{\rm d}^*,y,y_{\rm d}^*)}{\partial y^2}\left[\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2 + 2r(\theta,\theta)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau} + r(\theta,\theta)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\right)^2\right]$$
$$+2\frac{\partial r(\tau,\theta)}{\partial\tau}\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\frac{\partial w(y_{\rm d}^*,y,y_{\rm d}^*)}{\partial y} + \frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\theta}\frac{\partial w(x,y_{\rm d}^*,y_{\rm d}^*)}{\partial x}\right). \quad \text{(A-11)}$$

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-9) yields

$$H(\theta^*) = -\frac{\partial w(x_{\rm d}^*,y,y_{\rm d}^*)}{\partial y}\left[\frac{(1-r(\theta^*,\theta^*))\,(1+r(\theta^*,\theta^*))\left(\rho(\theta^*)\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau^2} - \frac{r(\theta^*,\theta^*)(1-\rho(\theta^*))}{1+r(\theta^*,\theta^*)}\frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\tau\partial\theta} - \frac{\partial^2 x_{\rm d}^*(\tau,\theta)}{\partial\theta^2}\right)}{1+r(\theta^*,\theta^*)\rho(\theta^*)}\right.$$
$$\left. -\frac{2\frac{\partial r(\tau,\theta)}{\partial\tau}(1-\rho(\theta^*)^2)\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}}{1+r(\theta^*,\theta^*)\rho}\right]_{\tau=\theta=\theta^*} + \left(\frac{\partial^2 w(x,y_{\rm d}^*,y_{\rm d}^*)}{\partial x^2}\left[1+\rho(\theta^*)r(\theta^*,\theta^*)(\rho(\theta^*)+2)\right]\right.$$
$$\left. 2\frac{\partial^2 w(x,y,y_{\rm d}^*)}{\partial x\partial y}\left[\rho + r(\theta^*,\theta^*)\left(1+\rho(\theta^*)+\rho(\theta^*)^2\right)\right] + \frac{\partial^2 w(x_{\rm d}^*,y,y_{\rm d}^*)}{\partial y^2}\left[\rho(\theta^*)^2 + 2r(\theta^*,\theta^*)(1+\rho(\theta^*))\right]\right)\left(\frac{\partial x_{\rm d}^*(\tau,\theta)}{\partial\tau}\right)^2. \quad \text{(A-12)}$$

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.

Appendix C Proof of Result 3

We first examine mutant-mutant interactions (the term in eq. (50)) and then turn to mutant-resident interactions (the term in eq. (49)).

Consider some resident type $\theta = (\theta_{\rm d},\theta_{\rm s}) \in [0,1]^2$. For any such type, residents obtain individual fitness $w(y_{\rm s}^*,y_{\rm s}^*,y_{\rm s}^*) = 1$. Visual inspection of the expression in (50) suffices to conclude that, for any value of the resident $\theta_{\rm s}$, the most threatening value $\tau_{\rm s}$ of any mutant type $\tau = (\tau_{\rm d},\tau_{\rm s})$ against the resident is the one that maximizes $W_{\rm s}(\tau_{\rm s},\theta_{\rm s})$, i.e., such that the equilibrium strategy used in a mutant-mutant interaction, $x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s})$, solves $\max_{x\in\mathcal{X}} w(x,x,y_{\rm s}^*)$. Given differentiability of w and $\mathcal{X} = \mathbb{R}$, the strategy $x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s})$ of the most threatening mutant must thus satisfy the following first-order condition:

$$\left[\frac{\partial w(x,x,y_{\rm s}^*)}{\partial x}\right]_{x=x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s})} = 0. \quad \text{(A-13)}$$

Since from eq. (46), $x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s})$ must satisfy the following first-order condition for utility maximization by a mutant,

$$\left[(1-\tau_{\rm s})\frac{\partial w(x,x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s}),y_{\rm s}^*)}{\partial x} + \tau_{\rm s}\frac{\partial w(x,x,y_{\rm s}^*)}{\partial x}\right]_{x=x_{\rm s}^*(\tau_{\rm s},\theta_{\rm s})} = 0, \quad \text{(A-14)}$$

our assumption that $\frac{\partial w(x,y,z)}{\partial y} \neq 0$ for all $(x,y,z) \in \mathcal{X}^3$ implies that $\tau_{\rm s} = 1$ is the unique value of $\tau_{\rm s} \in [0,1]$ which induces mutants to use the strategy that maximizes $w(x,x,y_{\rm s}^*)$. Hence, we can conclude that for any resident type θ, the most threatening mutant type has $\tau_{\rm s} = 1$. Note further that the argument developed above implies that $\theta_{\rm s} = 1$ implies $W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}) = 1$ if $\tau_{\rm s} = 1$ and $W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}) < 1$ if $\tau_{\rm s} < 1$. Hence, $\theta_{\rm s} = 1$ is sufficient for $W_{\rm s}(\tau_{\rm s}, \theta_{\rm s}) \leq 1$ for any $\tau_{\rm s} \in [0,1]$.

We turn now to the term in eq. (49), which measures the individual fitness of a mutant in an interaction with a resident. Suppose that the resident type is $\theta = (\theta_{\rm d}, \theta_{\rm s}) = (\theta_{\rm d}^*, 1)$, where $\theta_{\rm d}^*$ is defined in the statement of the result. By definition of $\theta_{\rm d}^*$, $\widetilde{W}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1) = w(\tilde{y}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1), \tilde{y}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1), y_{\rm s}^*(1)) \geq w(\tilde{x}(\tau_{\rm d}, \theta_{\rm d}^*, 1), \tilde{y}(\theta_{\rm d}^*, \tau_{\rm d}, 1), y_{\rm s}^*(1))$ for all $\tau_{\rm d} \neq \theta_{\rm d}^*$. Since $w(y, y, y_{\rm s}^*(1))$ is strictly concave in y (owing to our assumption that the utility function is strictly concave in its first argument), which achieves its maximum for y satisfying (A-13), it follows that the value of $\kappa \in [0,1]$ that would maximize the function $w(\tilde{y}(\kappa, \kappa, 1), \tilde{y}(\kappa, \kappa, 1), y_{\rm s}^*(1))$ is $\kappa = 1$. Together with the preceding inequality, we can thus conclude that $w(\tilde{x}(\tau_{\rm d}, \theta_{\rm d}^*, 1), \tilde{y}(\theta_{\rm d}^*, \tau_{\rm d}, 1), y_{\rm s}^*(1)) \leq w(\tilde{x}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1), \tilde{y}(\theta_{\rm d}^*, \theta_{\rm d}^*, 1), y_{\rm s}^*(1)) \leq w(\tilde{x}(1,1,1), \tilde{y}(1,1,1), y_{\rm s}^*(1)) = 1$. In other words, for any $\tau_{\rm d} \neq \theta_{\rm d}^*$, $W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}^*, 1) \leq 1$.

The preceding conclusions, namely that $W_{\rm s}(\tau_{\rm s}, 1) \leq 1$ and $W_{\rm d}(\tau_{\rm d}, \theta_{\rm d}^*, 1) \leq 1$, imply that whatever is the value of relatedness $r(\tau, \theta)$, if the resident type is $(\theta_{\rm d}, \theta_{\rm s}) = (\theta_{\rm d}^*, 1)$, then $W(\tau, \theta) \leq 1$ for any $\tau \neq \theta$. **Q.E.D.**

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. 2023. Evolutionarily stable preferences. *Philosophical Transactions of the Royal Society B* 378:20210505.

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. and J. W. Weibull. 2023. Evolution and Kantian morality: A correction and addendum. *Games and Economic Behavior* .

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.

Avila, P. and C. Mullon. 2023. Evolutionary game theory and the adaptive dynamics approach: adaptation where individuals interact. *Philosophical Transactions of the Royal Society B* 378:615255.

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.

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.

Boyd, R. and P. J. Richerson. 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Bruhin, A., E. Fehr, and D. Schunk. 2019. The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association* 17:1025–1069.

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

Cavalli-Sforza, L. and M. W. Feldman. 1981. Cultural Transmission and Evolution. Princeton University Press, NJ.

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. 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.

Fisman, R., S. Kariv, and D. Markovits. 2007. Individual preferences for giving. *American Economic Review* 97:1858–1876.

Foster, D. and H. P. Young. 1990. Stochastic evolutionary game dynamics. *Theoretical Population Biology* 38:219–232.

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. 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.

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

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

Hamilton, W. D. 1964b. The genetical evolution of social behaviour, II. *Journal of Theoretical Biology* 7:17–52.

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.

Jeong, H., A. Taylor, J. R. Floeder, M. Lohmann, S. Mihalas, B. Wu, M. Zhou, D. A. Burke, and V. M. K. Namboodiri. 2022. Mesolimbic dopamine release conveys causal associations. *Science* 378:eabq6740.

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.

Lester, D. 2015. On Multiple Selves. Routledge, London, UK.

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.

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

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 Cleve, J. 2023. Evolutionarily stable strategy analysis and its links to demography and genetics through invasion fitness. *Philosophical transactions of the Royal Society B* 378:20210496.

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.

Wang, Z. and J. Wu. 2023. Preference evolution under stable matching. SSRN Working Paper 4426394.

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.