Evolution of preferences in structured populations: genes, guns, and culture

Ingela Alger*, Jörgen W. Weibull†, Laurent Lehmann‡

October 24, 2019§

Abstract

During human evolution, individuals interacted mostly within small groups that were connected by limited migration and sometimes by conflicts. Which preferences, if any, will prevail in such scenarios? Building on population biology models of spatially structured populations, and assuming individuals' preferences to be their private information, we characterize those preferences that, once established, cannot be displaced by alternative preferences. We represent such uninvadable preferences in terms of fitness and in terms of material payoffs. At the fitness level, individuals can be regarded to act as if driven by a mix of self-interest and a Kantian motive that evaluates own behavior in the light of the consequences for own fitness if others adopted this behavior. This Kantian motive is borne out from (genetic or cultural) kin selection. At the material-payoff level, individuals act as if driven in part by self-interest and a Kantian motive (in terms of material payoffs), but also in part by other-regarding preferences towards other group members. This latter motive is borne out of group resource constraints and the risk of conflict with other groups. We show how group size,

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

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

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

§All authors together conceived the model, I.A. and L.L. derived the main results with input from J.W., and I.A. wrote the manuscript with input from all authors. We thank Lee Dinetan, Gustav Karreskog, Jonathan Newton, Jorge Peña, Peter Wikman, and seminar audiences at Paris School of Economics, University of Cambridge, University of Gothenburg, Université Catholique de Louvain, as well as participants at the conference "Neuroeconomics and the Biological Basis of Preferences and Strategic Behavior" at Simon Fraser University, the Toulouse Economics and Biology Workshop, and the Learning, Evolution, and Games 2019 conference at Bar-Ilan University for helpful comments. Support by Knut and Alice Wallenberg Research Foundation is gratefully acknowledged. I.A. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics). I.A. and J.W. acknowledge IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program), and Chaire d'Excellence ANR-12-CHEX-0012-01 for I.A., and Chaire IDEX ANR-11-IDEX-0002-02 for J.W.

the migration rate, the risk of group conflicts, and cultural loyalty shape the relative strengths of these motives.

Keywords: strategic interactions, preference evolution, evolution by natural selection, cultural transmission, pro- and anti-sociality, Kantian moral concerns
JEL codes: A12, A13, B52, C73, D01, D91.

1 Introduction

Preferences are fundamental to economic theory.[1] If preferences are transmitted across generations and if they affect the expected survival and reproduction—the fitness—of their bearer: which preferences are likely to be favored by evolution and which preferences are likely to disappear? Analysis of the long-term evolution of preference distributions can help understand the proximate drivers and motivation of human behavior in social and economic interactions (Hirshleifer, 1977, Bergstrom, 1996, Binmore 1998, Robson, 2001, Newton, 2018, Alger and Weibull, 2019). Here we build on previous work on strategy evolution in structured populations (Lehmann, Alger, and Weibull, 2015) by studying preference evolution in such populations.

For more than a million years, our ancestors most likely lived in groups of hunter-gatherers (probably ranging from 5 to 150 grown-ups), extending beyond the nuclear family (Grueter, Chapais, and Zinner, 2012, Malone, Fuentes, and White, 2012, van Schaik, 2016, Layton, O'Hara, and Bilsborough, 2012). This population structure, whose defining features are small group size and limited migration between groups (i.e., not all individuals migrate), is thus part of the environment of evolutionary adaptation of the human lineage (e.g., van Schaik, 2016). Analysis of the long-term evolution of preferences should thus take such population structure into account. We here do exactly that, and we ask how such structural features as group size, migration rates between groups, and the risk of conflicts between groups, determine the qualitative nature of the preferences that evolution favors. Combining the economics paradigm of utility-maximizing behavior with methods from population genetics, we obtain predictions about the nature of individuals' preferences and motivations in the canonical model of evolution in structured populations, the so-called island model of migration originally due to Wright (1931, 1943). The model allows us to examine both

[1] Throughout this paper we use concepts and terminology that are standard in economics, and model behavior as a choice of action (or stream of actions) from a set of feasible actions, where this choice is guided by a striving to maximize some goal (utility) function. The utility function together with the information and the constraints imposed by the environment are thus what biologists would call the proximate causes driving behavior. Furthermore, by contrast to the evolutionary biology literature where the terms "altruism" and "spite" are used to refer to the fitness consequences of a behavior on the actor and others, in economics they are used to describe the proximate causes behind behaviors. Thus, in economics, an individual who has a utility function which puts a positive weight on another individual's material well-being is altruistic; and an individual who has a utility function which puts a negative weight on another individual's material well-being is spiteful. For further discussion of the meaning of these terms in different academic disciplines, we refer to West, Griffin, and Gardner (2007) and Bshary and Bergmüller (2008).

genetic and cultural transmission of preferences in such structured populations.

The island model is a textbook evolutionary biology model (see, e.g., Cavalli-Sforza and Bodmer, 1971, Frank 1998, Rousset 2004, Hartl and Clark, 2007), which formally captures in a tractable and stylized way the fact that all natural populations, human or otherwise, are structured into small groups (or bands, or villages; patches for plants) connected to each other by limited migration (or dispersal). Limited migration causes limited genetic and/or cultural mixing in the population, and this results in several individuals from the same group possibly having a recent common ancestor. For example, suppose that a genetically transmitted new trait suddenly appears in one individual. In the next generation, multiple carriers of the new trait may coexist in the same group. Hence, the immediate descendants of the initial mutant are more likely to interact with each other than are individuals sampled at random from the whole population. Such assortative matching, induced by limited migration, even when the mutant trait is rare in the population at large, tends to favor mutant behaviors that promote the survival and/or reproductive success of others in their group. The reason is that such behavior is more likely to benefit other mutants than it would be if all offspring always migrated and matching therefore would be uniformly random (Hamilton, 1964, 1971, Grafen, 1985, Frank, 1998, Rousset 2004). This is the so-called mechanism of *kin selection* in evolutionary biology (Maynard Smith, 1964).[2] In the biology literature, assortative matching between pairs of individuals is usually quantified by the *coefficient of relatedness*–which indicates the likelihood that interacting individuals share a common ancestor–a quantity that depends on such features of the population structure as group size and migration rates.

By the same token, however, individuals who share a local common ancestor are also more likely to expose each other to fitness externalities, than are randomly selected individuals from the overall population. Indeed, through the local interactions, which occur in islands of finite size, related individuals may harm or enhance each other's fitnesses (think of young siblings fighting over candy, or individuals teaming up to fight off a common enemy), and such externalities have an impact on selected traits (Hamilton, 1971, Schaffer, 1988, Frank, 1998, Rousset 2004). As assortative matching and local fitness externalities can, in general, not be separated, their joint effects need to be taken into account in order to understand the evolutionary success of traits under limited dispersal, a question that has received much attention in the evolutionary biology literature (see e.g. Hamilton, 1967, and Taylor, 1992,a,b, for pioneering and paradigmatic examples, and Frank, 1998, and Rousset, 2004, for general theoretical treatments).

While clearly relevant for the understanding the evolution of traits relevant in human social interactions, the evolutionary biology literature is yet of limited direct value for economists, because in the bulk of these analyses: (a) the focus is on the evolution of strategies, not preferences, (b) predictions are derived at the level of basic fitness components, such as reproduction and survival, and not at the level of the material payoffs obtained in strategic

[2] A necessary and sufficient condition for kin selection to take place is that an evolving genetic (or cultural) trait tends to more strongly affect the survival and/or reproduction of individuals who are genetically (or culturally) related to the actor than under uniform random matching (Michod, 1982, p.20). This is true whether or not relatives recognize each other. Our analysis will be based on the assumption that they cannot recognize each other.

interactions, (c) transmission is genetic instead of cultural, while cultural evolution is also relevant for the understanding of human behavior. Our model enriches the analysis in all of these dimensions.

We propose a framework in line with that of economists and game theorists, and model the following thought experiment that takes place in a large population over an infinite sequence of demographic time periods. The population is structured into a large number of groups or islands of equal size. Within each group or island, individuals engage in a strategic interaction in which all individuals' strategy choices may affect the material payoffs to all participants. The strategic interaction is modelled as a game in material payoffs, and the game may be arbitrarily complex and may take place over many stages within each demographic time period. By material payoff we mean a one-dimensional summary measure, like income (or calories). The expected material payoffs, realized in a demographic time period, in turn determine the fitness of each individual in the population in that demographic time period. An individual's fitness is defined as the expected number of individuals in the following time period who have acquired their trait from him (or her). If transmission is genetic, an individual's fitness is the number of his surviving offspring and the individual himself if he survives. If transmission is cultural, an individual's fitness is the number of individuals in the next time period who acquired their cultural trait from this individual. Offspring may migrate to other groups or islands, or stay in their natal group or island. Many different transmission scenarios are covered by this model framework. For instance, generations may or may not be overlapping, islands may wage wars against each other, traits may be transmitted culturally from parent to child or by imitation of materially successful individuals, etc.

In all our scenarios, genetic and cultural, the population is initially homogenous; all individuals are *ex ante* identical. Suddenly, a different, mutant heritable trait spontaneously appears in exactly one individual. The original, resident trait is uninvadable if there exists no mutant trait, such that the initial mutant produces enough descendants for its trait to be maintained in the population in the long run.

To study preference evolution, we let the heritable traits be continuous utility functions, defined over all strategy profiles that are possible in the material game that represent the interaction on each island. Together with the individual's (probabilistic) belief about other group members' strategy choices, an individual's utility function guides his or her choice of strategy in the local interaction. We evaluate a utility function's fitness consequences for its carriers in terms of the expected material payoffs that result in all (Bayesian) Nash equilibria under incomplete information, that is, when each individual's utility function is his or her private information, but individuals' beliefs about each others' strategies are consistent with some (Bayesian) Nash equilibrium. We ask if there exist utility functions that are uninvadable in the sense that any mutant utility function does worse, in terms of its carriers' expected material payoffs, than the residents, in *all* equilibria. Thus bridging the gap between economics and biology, we obtain links between preferences, material incentives, and population structure (including migration and potential group conflicts). The following four main results emerge from our analysis.

First, we obtain a necessary and sufficient condition for a utility function to be unin-

vadable. This characterization says that a utility function is uninvadable if and only if all strategies used in any Nash equilibrium among individuals with this utility function are, when viewed as heritable strategies, uninvadable by other strategies.

Second, we identify a class of utility functions that, for any given game in material payoffs, contains an uninvadable utility function. Each utility function in this class can be interpreted as a mix of self-interest and a Kantian concern, both expressed at the fitness level. Specifically, the Kantian concern, driven by kin selection, consists in evaluating one's behavior in the light of what one's own fitness would be if others in one's group were to behave in the same way. This concern vanishes under unlimited migration (that is, when all offspring always migrate) and when groups are very large.

Third, when material payoffs only have marginal effects on fitnesses (a property which arguably holds for many human interactions), uninvadable preferences generically involve a mix of self-interest, a Kantian concern, and also a concern for neighbors, all concerns being expressed at the level of material payoffs. The weight given to the Kantian motive is then proportional to the coefficient of relatedness, but it also depends on fitness externalities between neighbors. The weight on other group members' material payoffs may be negative ("spite") or positive ("altruism"), and it depends on the *coefficient of fitness interdependence*, which measures the effect on own fitness that an individual obtains relative to his neighbors by diminishing or enhancing their material payoffs.

Finally, we provide sufficient conditions for the uninvadability of preferences of a particularly simple form, namely, a convex combination of own material payoff and the own material payoff that would arise should all others choose the same behavior. Under these specific conditions, the weight given to the second, Kantian, component is determined by the *coefficient of scaled relatedness*, a coefficient that combines the (standard) coefficient of relatedness with the coefficient of fitness interdependence. This weight allows to determine whether, on balance, equilibrium behaviors are pro-or anti-social, in the sense that equilibrium material payoffs are higher or lower than under selfishness. We show that an increased risk for group conflicts make preferences less anti-social, and, at a critical level of the risk of group conflict, preferences are neither anti- nor pro-social, while at higher risk levels, preferences turn pro-social. Hence, at this intermediate risk of conflict, preferences have only a self-interested and a Kantian component, while at lower (higher) risks, a third component appears, a component that expresses envy or spite if the risk is low, and empathy or altruism if the risk is high. We also show that cultural transmission of preferences may trigger anti-sociality because of local competition for proselytes.

Compared to the existing economics literature on preference evolution in social interactions (see Alger and Weibull, 2019, for a recent survey), our model makes two key innovations.[3] First, it explicitly analyzes the effects of population structure and limited dispersal

[3]In the economics literature it has been shown that the following conditions are sufficient (and necessary except in knife-edge settings) for populations of self-interested individuals to resist invasion by non-self-interested individuals: *(i)* the population is very large and homogeneous (no subdivision by sex, age, size, etc.) and reproduction is clonal, *(ii)* interacting individuals do not know each other's preferences but have statistically correct beliefs, and *(iii)* interactions are uniformly random in the population, in the sense that each encounter is just as likely (see Ok and Vega-Redondo, 2001, Dekel, Ely, and Yilankaya, 2007).

upon behavior and preferences. While Alger and Weibull (2013, 2016) investigated the evolutionary stability of preferences under incomplete information, they did so in an abstract model of assortative matching which did not explicitly account for the demographics and population dynamics.[4] They found that preferences expressing a certain combination of self-interest and a Kantian concern are evolutionarily stable, and that preferences that are behaviorally distinct from these are evolutionarily unstable. They also showed how the weight given to the Kantian concern depends on the assortativity in group formation. While assortativity in those models is treated as an abstract primitive, it here arises explicitly and endogenously from the population structure; group size, rates of survival, migration, and conflicts together determine the probability that rare mutants get to interact with each other—i.e., relatedness. The present model thus contributes to this strand of literature by explicitly modeling the population structure and how it gives rise to assortativity.

Second, it establishes a clear distinction between preferences at the fitness level and preferences at the material payoff level. In the existing economics literature on preference evolution, these are taken to coincide. The model makes it clear that, when preferences are expressed at the level of material payoffs, relatedness must go hand in hand with local fitness interdependence, a force which does not appear in Alger and Weibull (2013, 2016). We here also show how relatedness and fitness interdependence can be formally traced back to group size and limited migration. While we already made this distinction in Lehmann, Alger, and Weibull (2015), we then did not analyze preference evolution. Instead, we asked under what conditions, if any, evolving strategies can be interpreted as chosen by rational individuals endowed with specific utility functions (we examined three candidate utility functions, two of which are described above). The value added of the present paper is that we here analyze preference evolution, rather than strategy evolution, in group-structured populations. In addition, we (a) examine other utility functions than those used to establish the "as if" results in Lehmann, Alger, and Weibull (2015), (b) obtain new results concerning fitness interdependence and scaled relatedness, and (c) analyze a wider class of evolutionary scenarios.

Apart from our previous work, the most closely related work is by Akçay and van Cleve (2012). They investigated the evolutionary stability of preferences parameterized by scalar traits (in the vein of Heifetz, Shannon and Spiegel, 2007a-b). In addition to focusing on complete rather than incomplete information, their model differs from ours in two broad respects. First, since they focus only on the effects of traits on reproduction under genetic transmission, they do not obtain results for preferences over strategy profiles or material

[4]By contrast to the present model, assortativity was there modeled as an abstract function that maps the distribution of traits in the population to probabilities governing the matching of interacting individuals. This formalization of assortativity was pioneered in economics by Bergstrom (1995, 2003), who focused on strategy evolution; see also Bowles and Gintis (1998), as well as Alger and Weibull (2010, 2012) for analyses of preference evolution under complete information. This formalization of assortativity, which implicitly assumes marginal effects of traits on fitness, goes back to Hamilton (1971); Michod and Hamilton (1980) discuss how different formalizations of assortativity are equivalent to each other. It should further be noted that Rogers (1994) studied the evolution of time preference in an age-structured population; a setting that allows for kin selection but not kin competition. Finally, alternative models of endogenous assortativity have been proposed by Nax and Rigos (2016), Newton (2017), and Wu (2017, 2019).

payoffs distinct from fecundity. Second, they focus only on necessary first-order conditions. These conditions express how many offspring an individual is willing to forgo, at the margin, in order to marginally increase the number of offspring of other group members.

The paper is organized as follows. Section 2 describes the model and provides a characterization of an uninvadable trait. Section 3 presents the analysis. In Section 4 we illustrate our results in three canonical evolutionary scenarios, including genetic and cultural evolution, as well as potential "wars" between groups. Section 5 concludes. Mathematical proofs are provided in an appendix.

2 Model

This section presents the building blocks of the analysis—the population structure and individuals' life-cycles—and defines the evolutionary stability criterion that will be used. It also describes what is novel compared to the existing literature.

2.1 Population structure

Consider a countably infinite population, divided into infinitely many identical *islands* (groups, locations, or villages), each of constant size n. Evolution takes place perpetually and stochastically over time, and time is divided into *demographic time periods*. Individuals are called children or offspring in their first demographic period, the period in which they are born, and grown-ups or *adults* in all other periods of their life. No age distinction is made between adults. Each demographic time period consists of two phases:

- In *Phase 1*, the n adults in each island engage in a social or economic interaction with each other, the same on all islands and at all times. The strategies used determine each individual's *material payoff*, which we take to be income (expressed, for example, in money or in calories). An individual's strategy choice in the interaction is assumed to be determined by her *preferences* and her beliefs about the strategies used by her island neighbors. Preferences are inherited in childhood from exactly one adult, the individual's genetic or cultural parent, and are fixed throughout her life.

- In *Phase 2*, the realized material payoffs determine each adult's survival, and, in case there are exogenous random shocks to entire groups (e.g., warfare, environmental catastrophies), the adult's entire group's survival. Individual and group survival probabilities are assumed to be independent of age. The realized material payoffs also determine each adult's fecundity, which is its number of *offspring* (where offspring are biological if preferences are coded for genetically, and cultural if they are transmitted by way of a cultural process). Following reproduction, offspring, and only offspring, may migrate from their native island to other islands (and this migration takes place in the period they are born). The migration probability m is the same for all offspring at all times, and is strictly positive. Moreover, migration is blind in the sense that

any migrant picks a destination in a uniformly random fashion.[5] After migration and competition among the offspring for securing a place on an island, there are exactly n adults in each island (group). Offspring who did not secure a place on an island die.

Phase 1 and 2 taken together determine an adult's *individual fitness*. This is the expected number of her *immediate descendants*, defined as those adults in the next demographic time period who have inherited their preferences from her. These immediate descendants consist of those of the individual's (genetic or cultural) offspring who survived, and thus became adults in the next demographic time period, as well as the individual herself if she survived into the next period. We next describe in more detail how the interactions and the ensuing individual fitness are formalized in the subsequent analysis.

2.2 The interaction

2.2.1 The material game

The material game is formalized as a symmetric non-cooperative normal-form n-player game in which each player has access to the same set of strategies (which may be pure or mixed), X, a non-empty compact set in some normed vector space. The expected material payoff[6] accruing to any (adult) individual $i \in \{1, 2, ..., n\}$ on a given island depends on her own strategy, $x_i \in X$, and on the vector $\boldsymbol{x}_{-i} \in X^{n-1}$ of strategies used by the others on i's island, her *neighbors*. The material payoff function $\pi : X^n \to \mathbb{R}$ is continuous, and $\pi(x_i, \boldsymbol{x}_{-i})$ is invariant under permutation of the components of the vector $\boldsymbol{x}_{-i} \in X^{n-1}$.[7] Such permutation invariance holds if, for example, strategies are real numbers, and an individual's material payoff depends on her own strategy and either the sum, product, maximum or minimum of her island neighbors' strategies. The material game may be a simple simultaneous-move game or a multi-stage game in which individuals interact over many stages within the demographic time period.

2.2.2 The subjective game

Every (adult) individual in the population has (personal) preferences over strategy profiles, preferences that can be represented by some continuous utility function.[8] More precisely,

[5] Technically, we study the limit of uniform random dispersal among finitely many islands as the number of islands tend to infinity.

[6] For simplicity, we will henceforth use "material payoff" to refer to "expected material payoff".

[7] More precisely, for any $x_i \in X$ and $\boldsymbol{x}_{-i} \in X^{n-1}$, and any bijection $h : \{2, 3, ..., n\} \to \{2, 3, ..., n\}$: $\pi\left(x_i, x_{h(2)}, x_{h(3)}, ..., x_{h(n)}\right) = \pi(x_i, \boldsymbol{x}_{-i})$.

[8] Continuity is inessential for much of the analysis. However, it is important for some existence results, and for our stability analysis since we there invoke Berge's maximum theorem. (A form of upper semi-continuity would be sufficient for these results, but such a generalization does not seem to be of primary interest here.)

every individual i has a complete and transitive preference ordering $\succeq_i$ over strategy profiles $(x_i, \boldsymbol{x}_{-i}) \in X^n$, such that there exists a continuous function $u_i : X^n \to \mathbb{R}$ satisfying $u_i(x_i, \boldsymbol{x}_{-i}) \geq u_i(y_i, \boldsymbol{y}_{-i})$ if and only if $(x_i, \boldsymbol{x}_{-i}) \succeq_i (y_i, \boldsymbol{y}_{-i})$. Letting $\mathcal{F}$ be the set of continuous functions $f : X^n \to \mathbb{R}$, each individual has preferences that admit representation by some function $u \in \Theta \subseteq \mathcal{F}$, where Θ is the subset of *preference types*, or simply *types*. Each individual of any given type $u \in \Theta$ chooses her strategy so as to maximize the expected value of her utility function under her probabilistic beliefs about her island neighbors' strategy choices.

In order to carry out our evolutionary analysis of preferences, we need to evaluate the fitness consequences of preferences, and we will do so when individuals' strategy choices constitute Nash equilibria in the subjective game. (Which is not to say that we assume or believe that interactions are always in equilibrium. We use Nash equilibrium as a systematic reference point.)

2.3 Individual fitness

An individual's fitness may depend on (a) own material payoff, (b) the material payoffs to the individual's island neighbors, and (c) the material payoffs in the population at large. Dependence on own material payoff is self-evident. Dependence on neighbors' material payoffs arises as soon as neighbors' survival and number of offspring influences the competition that one's own offspring meet when competing for succession of deceased adults on the native island. Own and others' material payoffs may also affect one's island's success probability in wars with other islands. Dependence on material payoffs in other islands has two sources; migration and potential wars between islands. First, an individual's offspring face competition with offspring from other islands, both when competing for succession of deceased adults on her native island and on other islands. Second one's island's success probability in wars may depend on those islands' material payoffs.

The individual (or direct) fitness function $w : \mathbb{R}^{n+1} \to \mathbb{R}$ is assumed to be continuously differentiable. We write $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ for i's fitness, where $\pi_i \in \mathbb{R}$ is own material payoff, $\boldsymbol{\pi}_{-i} \in \mathbb{R}^{n-1}$, is the vector of her neighbors' material payoffs, and $\pi^* \in \mathbb{R}$ is the average material payoff in the population at large.[9] We also assume that $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ is invariant under permutation of the components of the vector $\boldsymbol{\pi}_{-i}$. Owing to the assumption of constant group size, average fitness in the population is always equal to 1. The subsequent analysis further presumes that an individual's fitness is strictly increasing in her own material payoff, strictly decreasing in the average material payoff in the population at large, and that it may be decreasing or increasing, or non-monotonic in other group members' material payoffs, but never increase more from a neighbor's increase in material payoff than from the same increase in own material payoff. Formally:

[M] *(i)* $\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)/\partial \pi_i > 0$, *(ii)* $\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)/\partial \pi_j \leq \partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)/\partial \pi_i$ *for all*

[9] Individual fitness is thus assumed to be expressible in terms of expected material payoff only, and we therefore neglect effects of variance in payoff (see Appendix 6.1 and 6.2 for a justification).

$j \neq i$, *(iii)* $\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)/\partial \pi^* < 0$.

Denote by $\mathcal{P} = \langle n, X, \pi, w, \Theta \rangle$ a population with (countably) infinitely many islands of size n, strategy set X, material payoff function π, fitness function w, and type set $\Theta \subseteq \mathcal{F}$. In each demographic time period t there is some type distribution $\mu_t \in \Delta(\Theta)$ in the population at large. The focus of the analysis is on the dynamics of this type distribution. Analysis will be restricted to type distributions with at most two types present in the population at any given point in time.

Our model adds two novelties to the existing island model literature. First, an individual's strategy choice is guided by her preferences and beliefs about her island neighbors' strategy choices. Second, we distinguish material payoffs here interpreted as income from both vital rates (e.g., survival and fecundity) and individual fitness (see concrete examples of fitness functions in section 5). Having this three-fold distinction is a novelty for the literature on preference evolution in economics, in which fitness is equated with material payoff, and for the evolutionary biology literature, in which payoff tends to be equated to fecundity.

2.4 Uninvadability in structured populations

Consider a population $\mathcal{P}$ which prior to some date $t = 0$ is *homogenous* with some *resident* type u. Can this population be invaded by some *mutant* type $v \neq u$ that appears at time $t = 0$ in a single adult individual? By "invasion" is meant long-run survival of the mutant type, or, more precisely, that it does not go extinct within finite time. In our model all adults have a positive probability of dying in each period of their lives, and offspring migrate with positive probability, so for a mutant type $v \neq u$ to be able to invade the population, it must spread beyond the island where the first mutant appeared. The initial mutant's descendants may, by way of migration, colonize new islands that were before inhabited exclusively by the resident type u. Such colonization, as well as survival and fecundity, depends on all individuals' material payoffs, which in turn depend on the strategy profiles used in the population. The analysis of a residential type's invadability or uninvadability is thus an analysis of (non-linear) stochastic population processes involving both (global) demography and (local) strategic interactions.

In order to be able to apply results in the biology literature for stochastic evolution in structured populations to preference evolution, we impose the following *homogeneity assumption* concerning individual's equilibrium behavior in the subjective game:

> [**H**] On all islands with the same number of mutants, and irrespective of calendar time, the same Nash equilibrium is played, and all residents use the same strategy (say, $x \in X$), and all mutants use the same strategy (say, $y \in X$).

As a consequence, on any given island and in any given demographic time period: all residents obtain the same fitness (which in general depends on their strategy, their island

neighbors' strategies, and on strategy profiles in the population at large), and all mutants obtain the same fitness (which likewise depends on their strategy, their island neighbors' strategies, and on strategy profiles in the population at large). Under these conditions, it is possible to obtain results for the long-run survival, or extinction, of mutants. We proceed in steps towards a definition and characterization of uninvadability.

First, in a population where all individuals are of the same type u, all individuals use the same strategy, to be called the *resident strategy*, and this strategy $\tilde{x}$ has to satisfy

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

where $\tilde{\mathbf{x}}^{(n-1)}$ is the $(n-1)$-dimensional vector whose components all equal $\tilde{x}$. We write X_u for the set of strategies $\tilde{x}$ that satisfy (1).[10] Condition (1) follows from the homogeneity assumption **[H]** and the Nash equilibrium requirement that every individual chooses a strategy that is optimal, given her preferences.

Second, consider a population $\mathcal{P}$ initially populated by some resident type u and in which some strategy in X_u is played by everyone. Let some mutant type v appear in exactly one individual at time $t = 0$. Under assumption **[H]**, and for any selection of Nash equilibria, one equilibrium for all islands with $k = 0, 1, ..., n$ mutants, respectively, this defines a probability distribution over fitness levels in all future demographic time periods. We define the resident type u to be *uninvadable by* v if, for every Nash equilibrium selection, the mutant type v goes extinct in finite time with probability one.[11] A type $u \in \Theta$ is *uninvadable* in Θ if it is uninvadable by all mutant types $v \in \Theta$.

The notion of *lineage fitness* plays a key role in our characterization of uninvadability. An individual's *lineage* consists of all of the individual's descendants, that is, her immediate descendants (her offspring and also herself if she survives), the immediate descendants of her immediate descendants, etc. *ad infinitum*. The individual's *local lineage* is the subset of her lineage members who live, as adults, in the island where she herself became an adult. Our assumption that the migration rate is positive and constant implies that the random time T of first extinction of any individual's *local lineage* is finite with probability one, and that in time periods before T local lineage members may produce emigrants settling on other islands.[12]

Any selection of Nash equilibrium (in the case of complete information, one equilibrium for each number of mutants in a group) defines a Markov chain that induces a unique invariant probability distribution over possible mutant local lineage size realizations (including the

[10] By continuity of the utility function and compactness of the set X, the set of maximands in (1) is non-empty and compact (by Weierstrass' maximum theorem). Moreover, by Berge's maximum theorem, the set of maximands also define an upper hemi-continuous correspondence. By Kakutani's fixed-point theorem, the set X_u is therefore non-empty (and compact) if the function u is also quasi-concave in its first argument (the player's own strategy).

[11] Extinction is defined as the event that no individual in the population is of the mutant type.

[12] Even if locally extinct, members of the individual's lineage may still live on other islands. Moreover, some lineage members may later move to the mutant's native island. However, the probability that this event occurs in finite time is zero.

realization of the random extinction time T), and this occurs irrespectively of whether an island of residents is colonized by a single or several successful mutant emigrants. This probability distribution in turn can be taken to determine the *lineage fitness* of the mutant type v given this equilibrium selection, defined as the average w fitness of a mutant, the average being taken over (a) all possible local lineage size realizations (each one before the associated random period T) and (b) over all possible initial conditions of a local lineage (single or multiple simultaneous emigrant mutants). As long as the mutant is rare in the population, the number of mutants is finite, so the average material payoff earned by individuals of the resident type u in those periods is simply $\pi^* = \pi(x, x, ..., x)$, where $x \in X_u$ is the resident strategy in the equilibrium in question.[13] For any given selection of Nash equilibrium, under assumption **[H]** the lineage fitness of a mutant type $v \in \Theta$ in an otherwise homogeneous population in which all individuals are of type $u \in \Theta$, can be written in the form

$$W(v,u) = \sum_{k=0}^{n-1} p_k(v,u) \cdot w(\pi(v|k), \langle \pi(v|k), \pi(u|k) \rangle, \pi^*), \tag{2}$$

where for each $k = 0, ..., n-1$, $p_k(v,u)$ is the probability for a mutant uniformly drawn from a local lineage, that $k = 0, 1, ..., n-1$ of her neighbors are from this lineage, $\pi(v|k)$ is the material payoff to the mutant at hand, and $\langle \pi(v|k), \pi(u|k) \rangle \in \mathbb{R}^{n-1}$ is the vector of material payoffs to the mutant's $n-1$ island neighbors (among whom k have the mutant trait v, and the other $n-1-k$ have the resident trait u). Hence, the lineage fitness of a mutant is the average individual fitness of a representative carrier of the mutant trait. Note that if there are multiple Nash equilibria, there may be several matching probability distributions $\boldsymbol{p}(v,u) = (p_0(v,u), ..., p_{n-1}(v,u))$, one for each selection of Nash equilibrium. Note further that the lineage fitness of the mutant type is well-defined if the mutant type happens to be identical with the resident type; then all individuals in the population have the same lineage fitness, namely $W(u,u) = 1$ (since population size is constant).

A positive probability weight $p_k(v,u)$ in the definition of W for some $k > 0$ means that descendants of the initial mutant face a positive probability of being matched with each other. The overall level of such assortative matching can be usefully quantified by the *coefficient of pairwise relatedness*, defined as

$$r(v,u) = \sum_{k=0}^{n-1} \frac{k}{n-1} \cdot p_k(v,u). \tag{3}$$

This coefficient measures, for any descendant of the initial mutant of type v, the average share of island neighbors who are also descendants of the initial mutant. When migration is complete ($m = 1$, see Section 2.1) or when groups are infinitely large ($n \to \infty$), no two group members can be traced back to an initial common ancestor, and thus $p_k(v,u) = 0$ for all $k > 0$, and hence $r(v,u) = 0$. But since real-life groups are of finite size, and owing to the cost of dispersal, essentially all natural populations display positive relatedness between

[13] Note that because the analysis focuses on the fitness of rare mutants in an otherwise homogenous population, our assumption that the fitness function depends on the average material payoff in the population at large, and not on the distribution of the material payoffs therein, is innocuous.

group members, i.e., $r(v,u) > 0$. This in turn implies that $p_k(v,u) > 0$ for at least some $k > 0$.[14]

We denote by $\mathcal{W}(v,u)$ the set of lineage fitness levels induced by all Nash equilibria compatible with types v and u in a given population $\mathcal{P}$. The (potentially empty) set $\mathcal{W}(v,u) \subseteq \mathbb{R}$ is compact. Extending the characterization in Lehmann et al. (2016) from types with unique lineage fitness values to types with sets of potential lineage fitness values,[15] uninvadability can be succinctly characterized as follows: A type $u \in \Theta$ with $\mathcal{W}(v,u) \neq \varnothing$ is *uninvadable* if and only if

$$\max \mathcal{W}(v,u) \leq 1 \quad \forall v \in \Theta. \tag{4}$$

For each mutant type $v \in \Theta$, this characterization compares the highest possible average lineage fitness of a single initial v-mutant, $\max \mathcal{W}(v,u)$, with the lineage fitness of any resident individual, $W(u,u) = 1$. An uninvadable type u thus preempts entry into the population in the sense of obtaining (weakly) higher average lineage fitness that any mutant type can ever obtain.

2.5 Nash equilibrium

In order to apply our characterization of uninvadability to preference evolution we need to get a handle on the set of Nash equilibria, which in turn depends on the informational assumptions about the strategic interactions on the islands. We know of three settings that are compatible with homogeneity assumption [**H**], and that admit analysis. In the first setting, each type in the type space $\Theta \subset \mathcal{F}$ has exactly one strategy that it will always use. This is the easiest case, and it can be referred to as "strategy evolution." In the second setting, all types are permitted, $\Theta = \mathcal{F}$, and interactions take place under (maximally) incomplete information, i.e., each individual's type is his or her private information. In the third setting, interactions take place under complete information, i.e., every individual knows the types of all individuals on her island. Under the homogeneity assumption, each of these settings is amenable to analysis. While one could argue that individuals are likely more knowledgeable about the type distribution in their own island than in the overall population, we nonetheless adopt the incomplete information assumption here, and leave analysis of complete information for future research. The reason is that the incomplete information setting is not only known to provide benchmark results to which results derived under complete information assumptions can be fruitfully compared (for a recent survey, see Alger and Weibull, 2019), but is also likely to be the default case under genetic transmission since information about genotype is generally incomplete (e.g., Frank 1998, chapter 6).

---

[14] This would be true even if migration probabilities were endogenous, as long as migration entails some cost (for literature with endogenous dispersal decisions, see, e.g., Clobert et al., 2001, Frank, 1998, and Rousset, 2004, and Hartl and Clark, 2007). The model by Newton (2017) can be interpreted as having costless migration.

[15] In Lehmann, Alger, and Weibull (2015) we proved this result for scenarios where new islands can be colonized only by singleton mutants. Lehmann et al. (2016, eqs. (14)-(16)) extended that result to allow for scenarios in which multiple offspring from the same group can reproduce in the same non-natal island.

It remains to define the set of Nash equilibria that will be used to calculate any mutant's lineage fitness under incomplete information. We assume that individuals' probabilistic beliefs about the type distribution among their neighbors are statistically correct. In particular, every individual of the resident type u (correctly) believes that all other individuals on his or her island are (with probability one) of her type, and every mutant (correctly) believes that the types of his or her island neighbors are drawn according to the mutant lineage's probability distribution $\boldsymbol{p}(v,u) = (p_0(v,u), ..., p_{n-1}(v,u))$. For any given resident strategy $\tilde{x} \in X_u$, all mutants are, by homogeneity assumption **[H]**, assumed to use one and the same strategy, say $\tilde{y}$ that, moreover, is a best response for them, with their utility function v, and given the matching probability distribution that they face:

$$\tilde{y} \in \arg\max_{y \in X} \sum_{k=0}^{n-1} p_k(\tilde{y}, \tilde{x}) \cdot v\left(y, \tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{x}}^{(n-k-1)}\right). \tag{5}$$

Here $\tilde{\mathbf{y}}^{(k)}$ is the strategy vector whose k components all are $\tilde{y} \in X$, and $\tilde{\mathbf{x}}^{(n-k-1)}$ the strategy vector whose $n-k-1$ components all are $\tilde{x}$, and the matching probabilities are from now on and throughout written directly as a function of the equilibrium strategies played. Given the resident and mutant types, $u, v \in \Theta$, a strategy pair $(\tilde{x}, \tilde{y}) \in X^2$ is a (type-homogenous) *Nash equilibrium* if $\tilde{x} \in X_u$ and $\tilde{y}$ satisfies (5). Let $B_{\mathrm{NE}}(u,v) \subseteq X^2$ denote the set of such Nash equilibria. Any such Nash equilibrium defines all remaining material payoffs $\pi(v|k)$ and $\pi(u|k)$ in (2).[16]

3 Analysis

It turns out that it is useful, as a first step, to examine preference types which induce commitment to some particular strategy. To be more specific, let $\Theta \subset \mathcal{F}$ consist of all utility functions $u : X^n \to \mathbb{R}$ of the form $u(x_i, \boldsymbol{x}_{-i}) \equiv \|x_i - x\|^2$ for some $x \in X$. All individuals with types in this set Θ each have a unique dominant strategy, and we will identify types by their dominant strategy; $\Theta = X$. Under such *strategy evolution*, for a resident type x and a mutant type y, the set $\mathcal{W}(y,x)$ is a singleton, and $\max \mathcal{W}(y,x) = W(y,x)$, where

$$W(y,x) = \sum_{k=0}^{n-1} p_k(y,x) \cdot \tilde{w}\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right). \tag{6}$$

Here $\tilde{w} = w \circ \pi$ is the composite function which gives the fitness of any individual i who plays strategy $x_i \in X$ when the others on his or her island play $\boldsymbol{x}_{-i} \in X^{n-1}$, while some strategy $x^* \in X$ is played by all individuals on all other islands:

$$\tilde{w}(x_i, \boldsymbol{x}_{-i}, x^*) = w\left(\pi(x_i, \boldsymbol{x}_{-i}), (\pi(x_j, \boldsymbol{x}_{-j}))_{j \neq i}, \pi(x^*, ..., x^*)\right). \tag{7}$$

[16] The reader may worry about model robustness at this point. For if the total population is large but finite, then the probability is not zero, but small and positive, that there will at least one mutant in a given resident's island. However, for sufficiently large populations (with fixed island size n), the probability that a mutant is present in a resident's island is so small that, by upper hemi-continuity of the best-reply correspondence of any $u \in \mathcal{F}$, the set of Nash equilibrium strategies for the residents when mutants are very rare in their islands, can be kept within arbitrarily small distance from the set X_u.

The population size being constant over time, we note that $\tilde{w}\left(x, \mathbf{x}^{(n-1)}, x\right) = 1$ for all $x \in X$. A necessary and sufficient condition for a strategy $x \in X$ to be uninvadable under strategy evolution is readily obtained by applying condition (4), resulting in:

$$\sum_{k=0}^{n-1} p_k\left(y, x\right) \cdot \tilde{w}\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right) \leq 1 \quad \forall y \in X. \tag{8}$$

Equivalently, an uninvadable strategy x can be seen as preempting entry into the population by earning the maximal lineage fitness that can be obtained in a population where the resident strategy is x; that is

$$x \in \arg\max_{y \in X} \sum_{k=0}^{n-1} p_k\left(y, x\right) \cdot \tilde{w}\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right). \tag{9}$$

In other words, a strategy is uninvadable if and only if it is a best reply to itself in terms of lineage fitness. We denote by $\hat{X}\left(\mathcal{P}\right)$ the (potentially empty) set of uninvadable strategies in population $\mathcal{P} = \langle n, X, \pi, w, \Theta \rangle$.

As a second step we use these observations to write the condition for a type $u \in \mathcal{F}$ to be uninvadable, (4), in an operational form:

$$\sum_{k=0}^{n-1} p_k\left(y, x\right) \cdot \tilde{w}\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right) \leq 1 \quad \forall v \in \mathcal{F} \text{ and } \forall\left(x, y\right) \in B_{\mathrm{NE}}\left(u, v\right). \tag{10}$$

This immediately leads to our first result.

Proposition 1 *In a population $\mathcal{P}$, a utility function u is uninvadable in $\mathcal{F}$ if and only if $X_u \subseteq \hat{X}\left(\mathcal{P}\right)$.*

In other words, for a utility function to be uninvadable, it must induce resident Nash equilibrium strategies that are all uninvadable. However, a utility function does not need to give rise to all uninvadable strategies in resident Nash equilibrium; any strategy in $\hat{X}\left(\mathcal{P}\right)$ that would not belong to X_u would simply not be played by residents, and would thus not be subject to potential invasion by mutants.[17]

The expression on the left-hand side of (10), however, shows that characterization of uninvadable preferences involves a major challenge, because the matching probabilities may depend both on the resident and mutant strategies, in all time periods when mutants are around.[18] In the second part of the analysis below, we analyze a model in which independence

[17] Note that the proposition implies that strategy-committed types that are uninvadable by other strategy-committed types, are uninvadable by all preference types. Indeed, if $\hat{X} = \{\hat{x}\}$, then $u \in \mathcal{F}$ is uninvadable if and only if $X_u = \{\hat{x}\}$. Moreover, this is true even if the residents would have preferences that do not entail commitment to a particular strategy, as long as these preferences induce them to play the uninvadable strategy in residential Nash equilibrium.

[18] Obtaining exact expressions for the matching probabilities is typically hard. However, their values can be approximated (see Appendix 6.7 for an approximation method).

of the matching probabilities on the strategies played arises endogenously. This model will allow us to fully characterize the set of uninvadable preferences at the level of material payoffs. Prior to that, however, we report results on uninvadable preferences at the level of fitnesses.

3.1 Utility and fitness

In spite of the challenge posed by the dependence of the matching probabilities on the strategies played by residents and mutants, we show that one particular class of utility functions stands out, in the sense that there always exists a utility function in this class for which some resident strategy is uninvadable.

3.1.1 The general case

For any given strategy $x^* \in X$, let the utility function $u_{x^*} : X^n \to \mathbb{R}$ be defined by

$$u_{x^*}(x_i, \boldsymbol{x}_{-i}) = \mathbb{E}_{\boldsymbol{p}(x_i, x^*)}\left[\tilde{w}(x_i, \tilde{\boldsymbol{z}}_{-i}, x^*) \mid (x_i, \boldsymbol{x}_{-i})\right] \quad \forall (x_i, \boldsymbol{x}_{-i}) \in X^n, \tag{11}$$

where $\boldsymbol{p}(x_i, x^*) = (p_0(x_i, x^*), p_1(x_i, x^*), ..., p_{n-1}(x_i, x^*))$ is the vector of matching probabilities that would be induced in population $\mathcal{P}$ if residents played x^* and mutants played x_i. Here $\tilde{\boldsymbol{z}}_{-i}$ is a *random strategy-profile* such that with probability $p_k(x_i, x^*)$ (for each $k = 0, 1, .., n-1$) exactly k of the $n-1$ components in $\boldsymbol{x}_{-i}$ are replaced by x_i, with equal probability for each such subset of k replaced components, while the remaining components in $\boldsymbol{x}_{-i}$ keep their original value. Then:

Proposition 2 *Any utility function $u_{\hat{x}}$ of the form (11) such that $X_{u_{\hat{x}}} = \{\hat{x}\}$ is uninvadable in $\mathcal{F}$. Moreover, each uninvadable strategy $\hat{x}$ is also a resident strategy under the utility function $u_{\hat{x}}$.*

This proposition identifies a sufficient condition for a utility function of the form (11) to be uninvadable. The condition is that the utility function has a unique resident strategy. Moreover, if the population structure admits multiple uninvadable strategies, then there are multiple utility functions of the form (11) that may be uninvadable, one for each $\hat{x} \in \hat{X}(\mathcal{P})$. Interestingly, then, different utility functions may arise in different populations with the same population structure. The reason is that the residential strategy affects the matching probabilities (which further explains why this result differs sharply from models in which the assortativity in the matching process is exogenous). There may of course be other uninvadable utility functions than those of the form (11) (see below). Nonetheless, Proposition 2 has a powerful implication: any uninvadable utility function must give rise to a resident strategy that is also a resident strategy under $u_{\hat{x}}$ for some $\hat{x} \in \hat{X}(\mathcal{P})$.

An individual with the utility function u_{x^*} can be seen as following a probabilistic version of Kant's categorical imperative (Kant, 1785) at the fitness level; she evaluates the strategies at her disposal in the light of what would happen to her own fitness in the hypothetical

scenario in which others would probabilistically use her strategy, according to the probability distribution $\boldsymbol{p}(x_i, x^*)$.[19] For illustrative purposes, we state u_{x^*} explicitly for $n = 2$ (then calling own strategy x_i and the opponent's strategy x_j) and $n = 3$ (then calling the opponents' strategies x_j and x_k):

$$u_{x^*}(x_i, x_j) = p_0(x_i, x^*) \cdot \tilde{w}(x_i, x_j, x^*) + p_1(x_i, x^*) \cdot \tilde{w}(x_i, x_i, x^*) \tag{12}$$

$$\begin{aligned} u_{x^*}(x_i, x_j, x_k) = {} & p_0(x_i, x^*) \cdot \tilde{w}(x_i, x_j, x_k, x^*) + \frac{p_1(x_i, x^*)}{2} \cdot \tilde{w}(x_i, x_i, x_k, x^*) \\ & + \frac{p_1(x_i, x^*)}{2} \cdot \tilde{w}(x_i, x_j, x_i, x^*) + p_2(x_i, x^*) \cdot \tilde{w}(x_i, x_i, x_i, x^*). \end{aligned} \tag{13}$$

Note that the weights $\boldsymbol{p}(x_i, x^*)$ in the u_{x^*} utility function depend on the individual's own strategy x_i in the present, whereas in the lineage fitness the matching probabilities depend on the strategy played by mutants individuals living over several (and past) demographic time periods. This highlights the difference between lineage fitness, which is an objective measure, and utility, which is subjective. The dependence of the weights $\boldsymbol{p}(x_i, x^*)$ on own strategy x_i, however, questions the operational relevance of u_{x^*} as an analytically and conceptually useful utility function. As such, we now turn to study cases where the matching probabilities in the lineage fitness no longer depend on the mutants' strategy; this will enable us to turn to utility functions with weights that do not depend on the individual's strategy.

3.1.2 The differentiable case

Suppose that the following differentiability assumption holds:[20]

[D] *(i)* $X = \mathbb{R}$, *(ii)* $\pi : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable, and (iii)* $p_k : X^2 \to [0, 1]$ is differentiable for all $k \in \{0, 1, ..., n-1\}$.

In the next proposition, which states a necessary condition for a strategy $\hat{x}$ to be uninvadable, $r(\hat{x}, \hat{x})$ is the coefficient of pairwise relatedness (see (3)) in a population where all individuals play $\hat{x}$, and a subscript i on $\tilde{w}$ denotes the partial derivative with respect to the i-th argument.

Proposition 3 *If [D] holds and* $\hat{x} \in \hat{X}(\mathcal{P})$*, then*

$$\tilde{w}_1\big(\hat{x}, \hat{\mathbf{x}}^{(n-1)}, \hat{x}\big) + r(\hat{x}, \hat{x}) \sum_{j=2}^{n} \tilde{w}_j\big(\hat{x}, \hat{\mathbf{x}}^{(n-1)}, \hat{x}\big) \quad = \quad 0. \tag{14}$$

---

[19] These preferences are reminiscent of *homo moralis* preferences (Alger and Weibull, 2013, 2016). However, there are two important distinctions. First, here the utility function is defined for a certain reference strategy. Second, the weights attached to the different terms in the utility function depend on the strategy used by the individual at hand.

[20] The uni-dimensionality assumption is inessential. All analysis can be carried out in terms of gradients, but with little gain in terms of qualitative insight. For brevity and clarity, we therefore stay with the unidimensional case.

The first term is the marginal fitness benefit of the individual's own strategy, while the second term is the sum of the marginal fitness benefits conferred by others, weighted by the coefficient of pairwise relatedness. This equation is nothing but the marginal version of Hamilton's rule (Hamilton, 1964, 1970, Franck, 1998), which provides the necessary first-order condition for an (interior) strategy to be uninvadable (see equation (3) in Taylor and Frank, 1996, or equation (7.5) in Rousset, 2004).[21] Such first-order conditions are standard in the biology literature, but for the sake of completeness we provide a proof in the appendix.

Consider the utility function defined by

$$\tilde{u}_{\hat{x}}\left(x_i, \boldsymbol{x}_{-i}\right) = \left[1 - r\left(\hat{x}, \hat{x}\right)\right] \cdot \tilde{w}\left(x_i, \boldsymbol{x}_{-i}, \hat{x}\right) \; + \; r\left(\hat{x}, \hat{x}\right) \cdot \tilde{w}\left(x_i, \mathbf{x}_i^{(n-1)}, \hat{x}\right), \tag{15}$$

where $\hat{x} \in \hat{X}\left(\mathcal{P}\right)$, and $\mathbf{x}_i^{(n-1)} \in X^{n-1}$ is the strategy vector whose $n-1$ components all equal x_i. Clearly, Propositions 1 and 3 together imply that if $\hat{x}$ is the unique resident strategy under $\tilde{u}_{\hat{x}}$, then this utility function is uninvadable.

An individual equipped with the utility function in (15) evaluates her strategy, x_i, both in terms of how it affects her own fitness, given the neighbors' strategies and the strategy played in the population at large, reflected in the first term, and how her strategy x_i would affect her fitness should her neighbors, hypothetically, also use it, reflected in the second term. This is reminiscent of *homo moralis* preferences (see, in particular, Proposition 3 of Alger and Weibull, 2016), although an important difference is that here the utility function in (15) is defined for a certain reference strategy.

3.2 Utility and material payoffs

We now turn to an approach in which the matching probabilities still depend on the transmission process but are independent of the strategies used. This approach, in biology called *weak selection* (see, e.g., Nagylaki, 1992, 1993), assumes that fitness effects from the interaction in question are small. Arguably, this approach is highly relevant for the social sciences, since it generates predictions regarding those preferences that guide behaviors in minor everyday interactions, those with only small effects on lifetime fitness.

3.2.1 Weak selection

Formally: for each $x \in X$ and $\boldsymbol{y} \in X^{n-1}$ let an individual's material payoff be a convex combination of two terms,

$$\bar{\pi}^{(\delta)}\left(x, \boldsymbol{y}\right) = \left(1 - \delta\right) \cdot \pi_0 + \delta \cdot \pi\left(x, \boldsymbol{y}\right), \tag{16}$$

[21] First-order conditions like equation (14) apply more generally to traits if lineage fitness and individual fitness are differentiable in trait values. The aforementioned evolutionary dynamics literature focuses on the evolution of phenotypes—the composite of an organism's characteristics—thus subsuming virtually any heritable trait and can be applied to essentially any demographic scenario (see Rousset, 2004, for general results).

where π_0 is baseline material payoff, assumed identical for all individuals, and $\delta \in (0,1)$ is the share of the material payoff that emanates from the present material game interaction. This factor δ is the *intensity of selection*. Thus, for $\delta \in (0,1)$ fixed, the fitness of individual i is now of the form

$$w\left(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*\right) = w\left(\left(1-\delta\right)\pi_0 + \delta\pi_i, \left(\left(1-\delta\right)\pi_0 + \delta\pi_j\right)_{j\neq i}, \left(1-\delta\right)\pi_0 + \delta\pi^*\right). \tag{17}$$

Weak selection amounts to considering the limit as δ tends towards 0.[22] Importantly, under weak selection, the matching probabilities, while still depending on the transmission process, do not depend on the strategies x and y (for any population $\mathcal{P} = \langle n, X, \pi, w, \Theta\rangle$ satisfying assumption [**M**]). The probability for a randomly drawn descendant of an ancestor, be it a resident or mutant, to coexist in its island with k other descendants of the same ancestor is then solely determined by the vital rates in a population in which everybody uses the same strategy x, no matter which. In biology this is referred to as the *neutral process*. This in turn has profound implications for the ability of a mutant trait to invade, since it means that the strategy played by residents matters only insofar as it affects the local success of mutants.

Let $\boldsymbol{p}^0 = \left(p_0^0, p_1^0, ..., p_{n-1}^0\right)$ denote the vector of matching probabilities induced by the neutral process. Proposition 2 still holds under weak selection: individuals playing some uninvadable strategy $\hat{x} \in \hat{X}\left(\mathcal{P}\right)$ may be viewed as if they were striving to maximize the utility function $u_{\hat{x}}$, with the matching profile now given by $\boldsymbol{p}^0$. This utility function is a sum of individual fitnesses. However, as is shown in the next proposition, under weak selection there is also an uninvadable utility function which describes preferences at the level of material payoffs, and which does not depend on any reference strategy. Let $v^0 : X^n \to \mathbb{R}$ be defined by

$$v^0\left(x_i, \boldsymbol{x}_{-i}\right) = \mathbb{E}_{\boldsymbol{p}^0}\left[\pi\left(x_i, \tilde{\boldsymbol{z}}_{-i}\right) - \lambda_0 \cdot \sum_{j\neq i}\pi\left(\tilde{z}_j,\ \tilde{\boldsymbol{z}}_{-j}\right) \mid \left(x_i, \boldsymbol{x}_{-i}\right)\right] \quad \forall\left(x_i, \boldsymbol{x}_{-i}\right) \in X^n, \tag{18}$$

where $\tilde{\boldsymbol{z}}_{-i}$ is defined in the same way as in (11), and

$$\lambda_0 = \lim_{\delta\to 0}\lambda\left(x\right) \tag{19}$$

is the *coefficient of fitness interdependence* under weak selection, where

$$\lambda\left(x\right) = -\left(\sum_{j\neq i}\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j}\right) / \left(\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i}\right), \tag{20}$$

evaluated when all individuals in the population use the same strategy $x \in X$. Hence, λ_0 measures the marginal effect of neighbors' material payoffs on own fitness, relative to the marginal effect of own material payoff on own fitness, in a population in which all individuals play the same strategy, and in the limit as δ tends to zero. A positive coefficient λ_0 can be

[22] This formalization of weak selection corresponds to what Wild and Traulsen (2007) call w-weak selection.

interpreted as there being competition for local resources: an increase in the material payoffs to neighbors then reduces an individual's fitness. A negative coefficient λ_0 means that there is a positive externality at the level of material payoffs between neighbors: an increase in the material payoffs to neighbors then increases an individual's fitness.

Our next result establishes that selection favors the so-defined utility function, which was used in the context of strategy evolution in Lehmann, Alger, and Weibull (2015), and rejects other utility functions unless they induce an identical best reply to some resident equilibrium for v^0:

Proposition 4 *The utility function v^0 is uninvadable in $\Theta = \mathcal{F}$ under weak selection. A utility function $u \in \Theta$ is invadable under weak selection if $\exists \tilde{x} \in X_u$ such that $\tilde{x} \notin X_{v^0}$. Moreover, $1 - n \leq \lambda_0 \leq 1$.*

An individual with the utility function v^0 is but that of the familiar *Homo oeconomicus* if $\lambda_0 = 0$ and $p_0^0 = 1$. By contrast, if $\lambda_0 \neq 0$ and $p_0^0 < 1$, the individual evaluates any strategy profile $(x_i, \boldsymbol{x}_{-i})$ by pondering his expected *material payoff advantage* over his neighbors, $\pi(x_i, \tilde{\boldsymbol{z}}_{-i}) - \lambda_0 \cdot \sum_{j \neq i} \pi(\tilde{z}_j, \tilde{\boldsymbol{z}}_{-j})$, if all, some, or none of the others in her island would use the same strategy as herself (drawn randomly according to $\boldsymbol{p}^0$), instead of playing their strategies, given by $\boldsymbol{x}_{-i}$.

To illustrate how the v^0 goal function is related to preferences studied in behavioral and experimental economics, we briefly consider the two-player case. By writing the utility function as

$$\begin{aligned} v^0(x_i, x_j) &= (1 - \lambda_0)\left(1 - p_1^0\right)\pi(x_i, x_j) + (1 - \lambda_0)\, p_1^0 \pi(x_i, x_i) \\ &+ \lambda_0\left(1 - p_1^0\right)\left[\pi(x_i, x_j) - \pi(x_j, x_i)\right], \end{aligned} \tag{21}$$

it can be interpreted as the sum of three terms, where the first represents "pure self-interest" (own material payoff), the second a Kantian concern (what is the "right thing to do if others in the population act like me"), and the third a "comparison with the Joneses" (the difference between own material payoff and that of the neighbor). Note also that a positive weight $\lambda_0 > 0$ expresses a form of *envy* or *spite*; if instead $\lambda_0 < 0$, then it is as if individuals care positively, or *altruistically*, about their neighbors' material payoffs.

Remark 1 *Part of the economics literature on the evolutionary stability of strategies and preferences relies on models in which rare mutants may have a positive probability of being matched with each other, even in the limit as the share of mutants tends to zero (Bergstrom, 2003, Alger and Weibull, 2013, 2016). These limit matching probabilities are taken to be independent of the strategies being played. Hence, they may be interpreted as the vector of matching probabilities $\boldsymbol{p}^0$ in the neutral process.*

3.2.2 The differentiable case

We finally turn to return to the general model, i.e., selection need not be weak, and consider settings where **[D]** holds. In a population in which all individuals play x, let $\kappa(x)$ denote

the coefficient of scaled relatedness, defined as

$$\kappa(x) = \frac{r(x,x) - \frac{1}{n-1}\lambda(x)\left[1 + (n-2)\, r(x,x)\right]}{1 - \lambda(x)\, r(x,x)}. \tag{22}$$

Then we obtain a result that (unlike Proposition 3) is new to the evolutionary biology literature:[23]

Proposition 5 *If [D] holds and* $\hat{x} \in \hat{X}(\mathcal{P})$*, then*

$$\left[1 - \kappa(\hat{x})\right] \cdot \pi_1\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) \; + \; \kappa(\hat{x}) \cdot \sum_{j=1}^{n} \pi_j\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) \quad = \quad 0. \tag{23}$$

Like $r(\hat{x},\hat{x})$, the coefficient $\kappa(\hat{x})$ can be interpreted as a marginal substitution rate: it gives the number of units of own material payoff that any given individual is willing to forgo to increase the material payoff of each neighbor by one unit. Absent any fitness interdependence, i.e., if $\lambda(\hat{x}) = 0$, $\kappa(\hat{x})$ would simply equal relatedness $r(\hat{x},\hat{x})$. To see exactly how $\kappa(\hat{x})$ accounts for fitness interdependence, consider first the case when there is but one neighbor, that is $n = 2$. A payoff transfer to this neighbor increases competition from the neighbor at rate $\lambda(\hat{x})$ (since $\lambda(\hat{x})$ measures the relative increase in competition in the neighborhood of an individual when its payoff is varied, see (20)). The fitness benefit to the donor from giving the transfer to the neighbor is thus reduced by $\lambda(\hat{x})$, so that the numerator in (22) becomes $r(\hat{x},\hat{x}) - \lambda(\hat{x})$. Moreover, a transfer of resources to the neighbor alleviates the competition that the neighbor experiences, and the neighbor is related to the donor according to coefficient $r(\hat{x},\hat{x})$. Hence, the cost of the transfer is reduced by $\lambda(\hat{x})\, r(\hat{x},\hat{x})$, which explains the denominator in (22).

When there are multiple neighbors, $n > 2$, a transfer given to one neighbor enhances the competition by $\lambda(\hat{x})/(n-1)$, but also for the $(n-2)$ other neighbors, each of which is related to the donor according to coefficient $r(\hat{x},\hat{x})$. Therefore, the fitness benefit of the transfer to the donor is reduced by $\lambda(\hat{x})/(n-1)$ times the term in square brackets in the numerator; which explains the numerator of $\kappa(\hat{x})$. In the denominator, the cost of the transfer is still reduced by $\lambda(\hat{x})\, r(\hat{x},\hat{x})$, which is the expected alleviation of competition that

[23] To be more explicit about this statement, we note that first-order conditions similar to the one in (23) appear elsewhere in the evolutionary biology literature, but then under the form $f_1\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) + \tilde{\kappa}(\hat{x}) \cdot \sum_{j\neq 1}^{n} f_j\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) = 0$, where f is the fecundity of an individual (see Lehmann and Rousset, 2010, Akçay and van Cleve, 2012, Van Cleve, 2015, Dos-Santos and Peña, 2017). We would obtain the exact same expression if in our model fitness depended solely on fecundity, since then derivatives of fecundity with respect to material payoffs would cancel from first-order conditions (to see this, set survival to zero in the fitness function in (26) in Section 4). Our model generalizes previous models, since Proposition 5 applies regardless of whether fitness depends only on fecundity, or also on individual and/or group survival (see Section 4 for examples of fitness functions), and it makes explicit the role of the coefficient of fitness interdependence. Further, it demonstrates that even if $r(\hat{x},\hat{x}) = 0$, the substitution rate $\kappa(\hat{x})$ can be substantial depending on the scenario (see, in particular, the examples in Section 4.2). As such, our results unify and extend previous ones of the evolutionary biology literature.

the transfer induces for the individual's neighbors (recall that $\lambda(\hat{x})$ accounts for all neighbors through the term $(n-1)$).

In view of the necessary first-order condition (23), it may be of interest to consider the utility functions $v_{\hat{x}} \in \mathcal{F}$ defined by

$$v_{\hat{x}}(x_i, \boldsymbol{x}_{-i}) = [1 - \kappa(\hat{x})] \cdot \pi(x_i, \boldsymbol{x}_{-i}) \; + \; \kappa(\hat{x}) \cdot \pi\left(x_i, \mathbf{x}_{-i}^{(n-1)}\right), \tag{24}$$

where $\hat{x} \in \hat{X}(\mathcal{P})$. Since (23) implies that $\hat{x}$ satisfies the necessary first-order condition for an interior symmetric Nash equilibrium of the n-player game in which all players have utility function $v_{\hat{x}}$, Proposition 1 implies that $v_{\hat{x}}$ is an uninvadable utility function if $\hat{x}$ is the unique resident strategy under $v_{\hat{x}}$.

In sum, in a population in which all individuals play some interior uninvadable strategy $\hat{x}$, these individuals may (under some conditions) be perceived as having a Kantian concern at the fitness level as well as at the material payoff level. Importantly, the strength of the Kantian (or other-regarding) concern at the fitness level, measured by $r(\hat{x}, \hat{x})$, typically differs from the strength of the Kantian (or other-regarding) concern at the material payoff level, measured by $\kappa(\hat{x})$, as shown next:

Proposition 6 *Suppose that **[D]** holds and that $v_{\hat{x}}$ is uninvadable. The weight $\kappa(\hat{x})$ attached to the neighbors' material payoffs in the function $v_{\hat{x}}$ lies in the interval $[-1, 1]$. Furthermore, $\kappa(\hat{x}) > r(\hat{x}, \hat{x})$ if and only if $\lambda(\hat{x}) < 0$.*

We note that a necessary and sufficient condition for $\lambda(\hat{x})$ to be negative is that in a population where everybody plays $\hat{x}$, $\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)/\partial \pi_j$ is positive (this partial derivative being the same for all $j \neq i$). We finally note that under weak selection (22) becomes:

$$\kappa_0 = \frac{r_0 - \frac{1}{n-1}\lambda_0 [1 + (n-2) r_0]}{1 - \lambda_0 r_0}, \tag{25}$$

where λ_0 is defined in (19) and $r_0 = \lim_{\delta \to 0} r(x, x)$.

4 Three canonical scenarios

We have reported general theoretical results on how fitness consequences of material payoffs may be expected to affect preferences over material payoff outcomes. In this section we apply these general results by examining three canonical evolutionary scenarios. For each scenario we calculate the associated coefficients of relatedness r, fitness interdependence λ, and scaled relatedness κ (all the calculations can be found in the appendix). Once these coefficients have been identified, equations (18), (15) and (24) provide closed-form representations of uninvadable utility functions, expressed in terms of the material payoff function that represents the strategic interaction at hand. We note that these coefficients are independent of the material payoff function in question, so the obtained utility representations

can be carried over from one material game to any other material game. Also, for all these scenarios the approximate explicit expression for the matching probabilities can be applied (see equation (78) in the Appendix), so the preferences can be fully evaluated in terms of the aforementioned coefficients.

4.1 Scenario A: Genes

If types are genetically determined, a possible fitness function is:

$$\begin{aligned} w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right) &= s\left(\pi_i\right) + m \cdot \left[1 - s\left(\pi^*\right)\right] n \cdot \frac{f\left(\pi_i\right)}{n f\left(\pi^*\right)} \\ &+ (1-m) \cdot \left(n - \sum_{j=1}^{n} s\left(\pi_j\right)\right) \cdot \frac{f\left(\pi_i\right)}{(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nm f\left(\pi^*\right)}, \end{aligned} \tag{26}$$

where $s\left(\pi_i\right) \in [0,1]$ is the probability that i survives to the next demographic time period, $f\left(\pi_i\right) > 0$ is i's expected number of offspring (who will have inherited i's type), and $0 < m \leq 1$ is the probability for each offspring to migrate to another island. These vital events—survival, reproduction and migration—are assumed to be statistically independent. In each island the deceased adults, if any, are replaced by (uniformly) randomly drawn aspiring offspring, native and immigrant. The fortunate ones settle and become adults while the unfortunate ones die. The third term is thus the expected number of i's offspring who manage to secure a "breeding spot" on the natal island. It is the product of three factors: (a) the probability for not migrating, $(1-m)$; (b) the number of available spots on the island; and, for each available spot, (c) the competition for the spot, among native and migrating offspring from other islands, where $f\left(\pi^*\right)$ is he fecundity in the population at large. The second term is the expected number of i's offspring who migrate and manage to secure a breeding spot on another island: each offspring who migrates to another island competes against $nf\left(\pi^*\right)$ other individuals for the n available spots.

Remark 2 *For a more detailed derivation of an equation like (26) from the random variables that underlie survival and reproduction, see Lehmann and Balloux (2007). In particular, since the total number of islands is infinite, the probability is zero for the event that more than one of i's offspring happen to migrate to the same island. Moreover, when the expected number of offspring is large, as we here assume, then the event that there are fewer aspiring offspring than there are available slots in an island is negligible, and the ratios between the expected numbers of offspring, in (26), equal the expectations of the ratios of the underlying random numbers of offspring.*

Considering the case where the survival probability is constant, $s\left(\pi_i\right) = s_0$, the coefficient of relatedness equals

$$r\left(x,x\right) = \frac{\left(1-m\right)^2 + s_0\left(1-m^2\right)}{n - \left(n-1\right)\left(1-m\right)^2 + s_0\left[1 + \left(n-1\right)m^2\right]}, \tag{27}$$

and the coefficient of fitness interdependence equals

$$\lambda(x) = \frac{(n-1)(1-m)^2}{n-(1-m)^2}. \tag{28}$$

Both coefficients turn out to be independent of the reference strategy x. Hence, the utility functions u_x and v_x, defined in equations (15) and (24), are independent of what strategy x is used in the population at large and can, in this evolutionary scenario, be explicitly parametrized in terms of the migration rate m, group size n, and survival probability s_0. Both $r(x,x)$ and $\lambda(x)$ are strictly positive for all n, all $m \in (0,1)$, and all $s_0 \in (0,1)$. By contrast, if $m = 1$, the probability of interacting with an individual from the same lineage is nil, $r(x,x) = 0$, and, moreover, there is no fitness benefit from out-competing neighbors materially, $\lambda(x) = 0$. Moreover, both coefficients are decreasing in m, and $r(x,x)$ is increasing in s_0. Substituting (27) and (28) into (22), we obtain:

$$\kappa(x) = \frac{2(1-m)s_0}{2(1-m)s_0 + n[2-m(1-s_0)]}, \tag{29}$$

which is strictly positive for any $m \in (0,1)$ and $s_0 > 0$, but nil for $m = 1$ and for $s_0 = 0$. In other words, in this evolutionary scenario, when $s_0 = 0$ but $m \in (0,1)$, any uninvadable utility function must be as if individuals are pro-social at the level of fitnesses ($r(x,x) > 0$), but are purely selfish at the level of material payoffs ($\kappa(x) = 0$). Furthermore, a positive survival probability $s_0 > 0$ induces pro-sociality ($\kappa(x) > 0$). However, note that $\kappa(x)$ is decreasing in island size n and in migration rate m. In fact, it vanishes as n becomes infinitely large. Figure 1 shows how $\kappa(x)$ depends on m when $s_0 = 1/n$, for $n = 2$ (black solid) and $n = 10$ (black dashed), and when $s_0 = 0.8$ for $n = 2$ (blue) and $n = 10$ (blue dashed), as well as $s_0 = 0$ (pink).

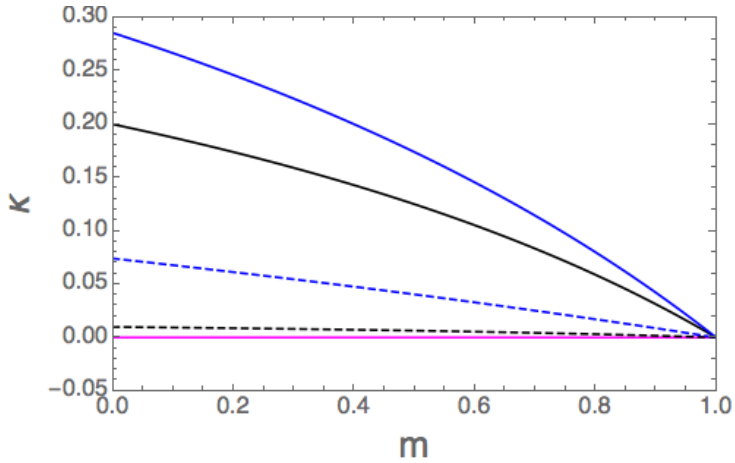

Figure 1: The value of $\kappa(x)$ as a function of the migration rate m.

Remark 3 *In the biology literature, the island model has become a work-horse model to analyze conditions favoring pro- and anti-sociality at the level of survival and reproduction in spatially structured populations. This literature has a well-known result known as Taylor's*

cancellation result, a useful yardstick for understanding how changes in the transmission scenario can tip the balance either in the direction of pro-sociality or anti-sociality. Our result that $\kappa(x) = 0$ *for* $s_0 = 0$ *is in line with this result. When* $s_0 = 0$*, (14) boils down to* $n^2 m(2-m) f'(\pi(\hat{x})) = 0$*, which implies that* $f'(\pi(\hat{x})) = 0$*: in spite of a positive relatedness, uninvadability requires simple maximization of own fecundity. This holds true even if fecundity depends directly on the underlying trait, the standard assumption in the biology literature (see also Footnote 23), without being a function of some material payoff. It is this observation which is known as Taylor's cancellation result, noticed initially in agent-based simulations by Wilson, Pollock, and Dugatkin (1992), proven formally by Taylor (1992a) for the island model, and then shown to hold for arbitrary migration patterns between groups (e.g., Taylor, 1992b, Rousset, 2004, and Ohtsuki, 2012). To see that it is in line with our result that* $\kappa(x) = 0$ *for* $s_0 = 0$*, note that since* f *is strictly increasing in* π*,* $f'(\pi(\hat{x})) = 0$ *in turn implies that* $\hat{x}$ *maximizes* π*, i.e.,* $\pi'(\hat{x}) = 0$*. Finally, we note that the same expression as that in the right hand side of eq. (29) was first obtained by Taylor and Irwin (2000, their eq. A.10), as a marginal cost to benefit ratio at the level of fecundity (see also Akçay and van Cleve, 2012). There is by now an extensive theoretical literature seeking to delineate how the assumptions pertaining to demography, life-history, the environment, and the modes of transmission, tip the balance in favor of pro- or anti-sociality at the survival or fecundity level (see, e.g., Eshel, 1972, Aoki, 1982, Gardner and West, 2006, Johnstone and Cant, 2008, Lehmann, Foster, and Feldman, 2008, Lion and Gandon, 2010, Bao and Wild, 2012, and Micheletti, Ruxton, and Gardner, 2017, for a some representative case studies, and Lehmann and Rousset, 2010, for a review).*

4.2 Scenario B: Guns

Take the biological scenario A with non-overlapping generations (set $s(\pi) = 0$ for all π), and augment it by introducing wars between groups. Following play of the material game in a demographic time period, but before reproduction, death of the adults, and migration by the offspring, islands are randomly engaged in pairwise wars, under exogenous uniform random matching. In each war, one island wins and the other loses. All individuals in the losing island thus die before they reproduce; the winning island takes over all reproductive resources of the other island and thus doubles its members' fecundity. Technically, the double-sized pool of offspring of the winning island will split in two halves, one for each of the two islands, that they will treat as their "home" island. Let $0 \leq \rho \leq 1$ denote the probability that any given island is drawn into war, the *war risk*, and let $g(\boldsymbol{\pi}, \pi^*)$ denote the conditional probability that an island with material payoff profile $\boldsymbol{\pi} \in \mathbb{R}^n$ wins a war when the average payoff in the rest of the population is π^*, conditional on being drawn into war. Here g is assumed to be increasing and permutation invariant with respect to the material payoffs earned by the inhabitants of the island in question. In other words, for π^* fixed, g

has the properties of standard welfare functions. In this scenario the fitness function is

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right) = \left[(1-\rho) + 2\rho g\left(\boldsymbol{\pi}, \pi^*\right)\right] \cdot \left[m \cdot \frac{f\left(\pi_i\right)}{f\left(\pi^*\right)} + (1-m)\, n \cdot \frac{f\left(\pi_i\right)}{(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)}\right]. \tag{30}$$

The difference with the baseline scenario is the first factor, which contains two terms: the probability that the individual's island will not go to war $(1-\rho)$, and the probability that the island will go to war and win times two $(2\rho g\left(\boldsymbol{\pi}, \pi^*\right))$, where the factor two comes from the assumption that a winning island doubles its fecundity and spreads its offspring uniformly over the two islands it now possesses. To see why the second factor is the same as the right-hand side of (26), note that migrants who arrive at any island, irrespective of whether this island has been involved in war or not, come with probability $1-\rho$ from an island that was not in war, and (recalling that the average probability of winning a war is $1/2$) with probability $\rho/2$ from an island that won a war. Moreover, victorious islands send out twice as many migrants as islands that did not go to war. Hence, the expected number of migrants who compete for the breeding spots in any given island is $m\left(1-\rho+2\rho/2\right)\cdot f\left(\pi^*\right) = mf\left(\pi^*\right)$, the same as in the absence of wars.

The coefficient of relatedness turns out to coincide with that in the preceding scenario (for $s_0 = 0$). This is because the only event in which a randomly drawn individual can belong to the same local lineage as a randomly drawn neighbor, is when both belong to an island which did not lose a war, and both stayed in their natal island. Since the risk of losing a war applies to the whole island, while the migration probability applies to the individual, only the latter matters for relatedness. The coefficient of fitness interdependence equals

$$\lambda\left(x\right) = -\frac{(n-1)\left[2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) - \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)}\right]}{2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) - \left[n-(1-m)^2\right]\frac{f'(\pi^*)}{nf(\pi^*)}}, \tag{31}$$

where $\boldsymbol{\pi}^*$ is the n-dimensional vector whose components all equal π^* and g_n denotes the partial derivative of g with respect to the n-th argument (since g is evaluated in a homogenous population here, and since g is invariant under permutation of the n first arguments, g_n simply captures the marginal effect of an increase in the material payoff of any island member on the probability of winning a war). While the expression is involved, it can readily be seen (by considering a scenario in which $f'\left(\pi^*\right) = 0$, for instance) that the effect of material payoffs on the strength in wars can make $\lambda\left(x\right)$ negative, while in the scenario without wars studied above, it is always positive. In other words, conflicts between groups reduces spite, and may even reduce it so much that it turns into altruism, i.e., a positive weight is attached to the neighbors' material payoffs. Indeed, by substituting (27) (for $s_0 = 0$) and (31) into (22), we obtain:

$$\kappa\left(x\right) = \frac{\rho}{\rho + \frac{(2-m)m}{2g_n(\boldsymbol{\pi}^*, \pi^*)}\frac{f'(\pi^*)}{f(\pi^*)}}, \tag{32}$$

which is increasing in the marginal effect g_n on the probability of winning wars.

We next turn to weak selection in order to obtain more explicit results on the effects of wars on fitness interdependence and scaled relatedness. Recalling the notation under weak selection (see (16)), let each individual's fecundity be exponentially increasing in the individual's material payoff,

$$f(\bar{\pi}_i) = f_0 \cdot \exp\left((1-\delta_f)\cdot\pi_0 + \delta_f \cdot \pi_i\right), \tag{33}$$

where $f_0 > 0$ is baseline fecundity and $\delta_f > 0$ represents the intensity of selection with respect to fecundity. Furthermore, assume that the probability of winning a war depends on the two islands' aggregate material payoffs according to

$$g(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\pi}}^*) = \frac{\exp\left(V(\bar{\boldsymbol{\pi}})\right)}{\exp\left(V(\bar{\boldsymbol{\pi}})\right) + \exp\left(V(\bar{\boldsymbol{\pi}}^*)\right)}, \tag{34}$$

where $\bar{\boldsymbol{\pi}} = (1-\delta_\nu)\boldsymbol{\pi}_0 + \delta_v \boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}^* = (1-\delta_\nu)\boldsymbol{\pi}_0 + \delta_v \boldsymbol{\pi}^*$ (where $\boldsymbol{\pi}_0$ is the n-dimensional vector whose components all equal π_0) and $V : \mathbb{R}^n \to \mathbb{R}$ is a strictly increasing symmetric function (like any standard welfare function). Its values $V(\bar{\boldsymbol{\pi}})$ and $V(\bar{\boldsymbol{\pi}}^*)$ represent the "strengths" of the two islands. This is a logistic version of the Tullock contest function (Tullock, 1980), see Skaperdas (1996). It spans a continuum of cases, from all islands having the same chance to win any war, if the intensity of selection with respect to wars be nil, to the case in which the materially wealthiest island is almost sure to win any war (is the intensity of selection is infinitely large). Letting $\delta_f = \sigma_f \cdot \delta$ in equation (33) and $\delta_v = \sigma_v \cdot \delta$, for non-negative parameters $\sigma_f \geq 0$, $\sigma_v \geq 0$, and $\delta > 0$, we can let both sensitivity parameters tend to zero at proportional rates by focusing on the limit as $\delta \to 0$. Below, however, we let $\sigma_v = \sigma_f$, and thus write δ for δ_v.

Many scenarios can be imagined, of which we consider two. First, if an island's strength is proportional to its total material payoff, i.e., if $V(\bar{\boldsymbol{\pi}}) = (1-\delta)\, n\pi_0 + \delta \sum_{i=1}^{n} \pi_i$, then fitness interdependence takes the following form (see the appendix):

$$\lambda_0 = \frac{(n-1)(1-m)^2 - \rho(n-1)n/2}{n - (1-m)^2 + \rho n/2}. \tag{35}$$

This changes sign when the risk of war is $\rho^* = 2(1-m)^2/n$; it is positive at lower risks of war and negative at higher risk levels for war. Since in the baseline scenario with non-overlapping generations uninvadability under weak selection requires individuals to be selfish on balance (see Section 3.2), the reduction in fitness interdependence that the war risk entails, leads to pro-sociality on balance; indeed, for any $\rho > 0$ we obtain $\kappa_0 > 0$:

$$\kappa_0 = \frac{\rho}{\rho + 2m(2-m)}. \tag{36}$$

Moreover, the threat of war ($\rho > 0$) nourishes pro-sociality: κ_0 is increasing in the risk of war, ρ, and is independent of group size, n.[24] Figure 2 shows κ_0 as a function of the migration

[24] In the early evolutionary biology literature, which considered traits affecting environmentally induced group extinction (e.g., Eshel, 1972, Aoki, 1982), pro-sociality at the fecundity level (the equivalent of $\tilde{\kappa}$ referred to in Footnote 23) is usually a decreasing function of n (see also Lehmann and Rousset, 2010).

rate m, for war risk $\rho = 0$ (the pink curve), $\rho = 0.4$ (the orange curve), and $\rho = 0.8$ (the blue curve).[25]

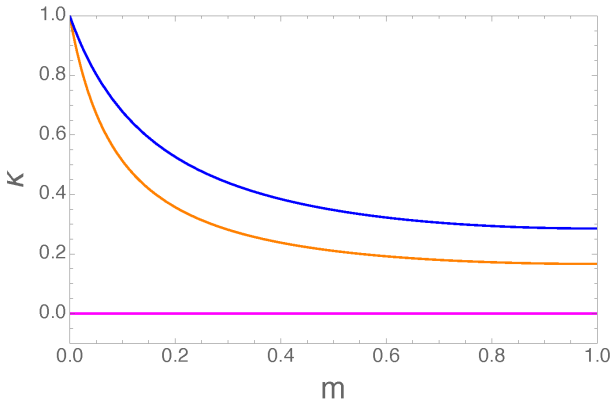

Figure 2: The value of κ_0 as a function of the migration rate m.

Second, while it is arguably a natural benchmark case to assume that the probability of winning a war depends on the group's total material payoff, sometimes the success or failure in conflicts depends on the strongest or the weakest member of one's group.[26] A general case, that allows for intermediate cases between dependence on the group's total material payoff and its minimal payoff, is obtained by using a CES-functional form. Let

$$V(\bar{\boldsymbol{\pi}}) = \left[(1-\delta)\cdot n\pi_0^c + \delta\cdot\sum\nolimits_{i=1}^{n}\pi_i^c\right]^{1/c} \tag{37}$$

for some $c \neq 0$. For $c = 1$ we obtain the previous case, and as $c \to -\infty$, $V(\bar{\boldsymbol{\pi}}) \to \min_i\{(1-\delta)\cdot\pi_0 + \delta\cdot\pi_i\}$ (Leontieff production function). Hence, when c is negative and large in absolute terms, an increase in the poorest group member's material payoff will increase the winning probability, and hence have a positive effect on others' fitness. This suggests a Rawlsian, rather than a Benthamite concern for other group members' material well-being. Individuals with medium or high material payoffs may then behave as if they had a particular concern for individuals with low payoff.

4.3 Scenario C: Culture

Suppose now that types are carried over from one generation to the next by cultural transmission. In every demographic time period, each adult dies and is replaced by exactly one

[25] The analytical models of Bowles (2006, 2009) for the evolution of "parochial altruism" are also close to our scenario with wars; in particular, the expected number of groups $[1-\rho+2\rho\nu(\boldsymbol{\pi},\bar{\pi}^*)]$ to which a focal group has access for reproduction after warfare also appears in Bowles's formalization. However, since in his model there are no explicit assumptions that allow to close the lifecycle, it is impossible to derive the explicit values of λ_0, r_0, and κ_0 for his model.

[26] A host of other hypotheses about group strength could be explored, see, e.g., Konrad (2014) and the references therein.

child, who searches for a type to emulate, from its deceased (single) parent, another adult in its island, or an adult in another island. With probability $s(\pi_i) \in [0,1]$, the loyalty of i's child, the (unique) child of individual i, emulates its parent's type. With probability $1-m$ a non-loyal child searches for a type to emulate among the (now dead) grown-ups in its natal island (including its own parent). With the complementary probability, $m > 0$, such a child draws a sample of n grown-ups from the population at large, and emulates the type of one of them. The probability that an adult on any island is chosen as role model, when compared to others in her island (by a non-loyal child), depends on her type's attractiveness relative to the attractiveness of the other grown-ups' types in her island. Likewise, the probability that a child who searches outside its native island will pick a certain island, when looking for a "role model", is assumed to be proportional to the island's relative attractiveness in the world at large. Fitness $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ is then the expected number of children who emulate their type from an individual with material payoff π_i when the other island members earn the material payoff vector $\boldsymbol{\pi}_{-i}$, and individuals in all other islands earn material payoff π^*:

$$w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) = s(\pi_i) + \; m \cdot [1 - s(\pi^*)] \cdot \frac{f(\pi_i)}{f(\pi^*)} + \; (1-m) \cdot \left(n - \sum_{j=1}^{n} s(\pi_j) \right) \cdot \frac{f(\pi_i)}{\sum_{j=1}^{n} f(\pi_j)}, \tag{38}$$

where, for any individual j in i's island, $f(\pi_j) > 0$ is the attractiveness of the type used by j. The first term in (38) is the probability that i's child loyally emulates its parent's type, without comparison with other adults' types.[27] The second term concerns the event that children from other islands emulate their type from one of the parents on i's island. Written more explicitly, this term can be spelled out as

$$mn\,[1 - s(\pi^*)] \cdot \frac{\sum_{j=1}^{n} f(\pi_j)}{n \cdot f(\pi^*)} \cdot \frac{f(\pi_i)}{\sum_{j=1}^{n} f(\pi_j)}, \tag{39}$$

where the first factor is the expected number of children who search outside their native islands, the second factor is the probability for each such child to decide for i's island, and the third is the conditional probability that it will then choose i as role model. The third term concerns the event that some or all the children in i's island emulate their type from one among the parents on the island. The product of the first two factors in this term is the expected number of such children and the third factor is the probability, for each such

[27] In the economics literature on cultural transmission of traits, a commonly used model is that of Bisin and Verdier (2001). Like in our model, in Bisin and Verdier (2001) each grown-up has exactly one child, and each child inherits its parent's trait with some probability, and otherwise it inherits the trait of another grown-up in the population. By contrast to our model, the population is not structured into islands, and there is no strategic interaction between individuals. Furthermore, in their model a parent cares about whether her child has the same trait as her, but not about whether the child inherited this trait from the parent or from someone else. Denoting by q_θ the population share of individuals with trait θ in the population, and by $s(q_\theta)$ the probability that a child inherits its trait vertically from its parent, their equation (1) says that the unique child of a parent with trait θ acquires trait θ with probability $s(q_\theta) + [1 - s(q_\theta)] \cdot q_\theta$. Thus, in the their model it is only the frequency of the trait that determines the transmission probability; in our model the attractiveness of a trait also plays a role.

child, that it will choose to imitate individual i. Note that, comparing this scenario to the biological scenario with overlapping generations, loyalty plays a similar role to survival, and attractiveness to fecundity. Moreover, the *cultural import propensity* m plays a similar role to migration. (These observations motivated the notation.)

In this scenario,

$$r(x,x) = \frac{(1-m)^2 + s(\pi(\mathbf{x}))(1-m^2)}{n-(n-1)(1-m)^2 + s(\pi(\mathbf{x}))[1+(n-1)m^2]}, \tag{40}$$

where $\mathbf{x} = (x, ..., x) \in X^n$, and

$$\lambda(x) = \frac{(n-1)(1-m)}{n-1+m}, \tag{41}$$

which leads to

$$\kappa(x) = -\frac{(1-m)[1-s(\pi(\mathbf{x}))]}{2n-[m(n-1)+1][1-s(\pi(\mathbf{x}))]}. \tag{42}$$

Comparison with the biological scenario with overlapping generations reveals that the coefficients of relatedness are the same, but that for any $m < 1$ the coefficient of fitness interdependence is larger under cultural transmission. The enhanced competitiveness is strong enough to lead to anti-sociality, since $\kappa(x) < 0$ obtains if and only if $(1-m)[1-s(\pi(\mathbf{x}))] < (2-m[1-s(\pi(\mathbf{x}))]) \cdot n$, an inequality which holds for all parameter values.[28] In this example, cultural transmission thus leads to anti-sociality, and anti-sociality is stronger at low values of m. This is because a low cultural import rate enhances fitness interdependence. Note that although genetic and cultural transmission here lead to opposite predictions regarding sociality, one qualitative similarity that appears is that like survival under genetic transmission, loyalty under cultural transmission has a positive effect on sociality, $\kappa(x)$. We also note that the negative pro-sociality vanishes as groups tend to become infinitely large: $\kappa(x) \to 0$ as $n \to \infty$.

To illustrate this, Figure 3 shows that $\kappa(x)$ is strictly negative for all $m < 1$, for different loyalty rates and different island sizes: for $s_0 = 0$ and $n = 2$ (the pink curve), $s_0 = 0.4$ and $n = 2$ (the orange curve), $s_0 = 0.8$ and $n = 2$ (the blue curve), $s_0 = 0$ and $n = 10$ (the pink dashed curve), $s_0 = 0.4$ and $n = 10$ (the orange dashed curve), $s_0 = 0.8$ and $n = 10$ (the blue dashed curve).

[28] In evolutionary biology, the same expression as the right hand side of (42) was obtained for the case of no cultural loyalty as a marginal fecundity cost to benefit threshold ratio under which the mutant is favored in a public good game (eq. 26 of Lehmann, Foster, and Feldman, 2008).

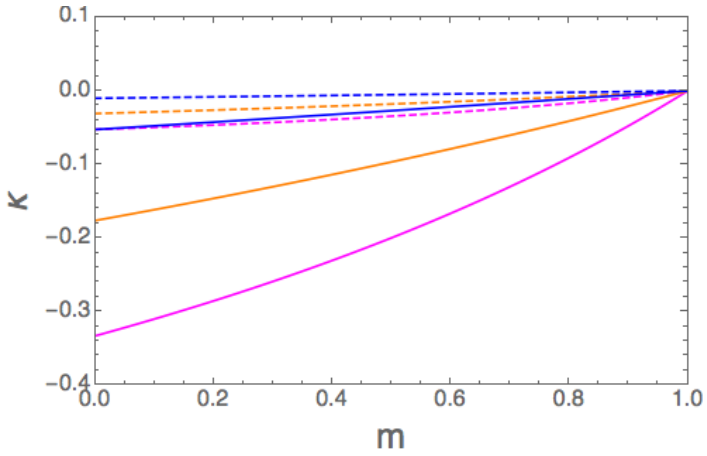

Figure 3: The value of $\kappa(x)$ as a function of the cultural import rate m, for different degrees of background loyalty of offspring towards parents.

5 Conclusion

By combining non-cooperative game theory and evolutionary biology, we have derived several novel insights on the evolutionary viability of preferences in social interactions. In particular, our model enables analysis of how the tendency of individuals to interact in fairly small groups, between which there is limited migration, and between which there may be conflicts, affects such preferences. A key strength of the model is that it makes a distinction between material payoffs, which typically is the level at which data analysis by economists is conducted, and individual fitness. Our results clearly show that the qualitative nature of evolutionarily viable preferences is typically different at the material payoff than at the individual fitness level. Furthermore, our results provide an evolutionary justification for preferences as drivers of choice, by connecting stability at the strategy level with equilibrium behavior under certain preferences. Our results thus address a criticism of the literature on preference evolution, according to which it conflates revealed preferences with preferences that drive choice, see, e.g. Newton (2018).

The cognitive assumption we make is that individuals understand what interaction is at hand, but they need not know the material payoffs to others or the preferences of others. Moreover, our formalization allows for the possibility that in fact there are (finitely) many interactions going on simultaneously, or that are randomly selected, and even that each interaction involves only a subset of the inhabitants in an island. What is required is symmetry in the sense that all individuals face the same probabilities of being involved in any one of the interactions and that the interaction at hand is aggregative and symmetric.

However, if individuals also understand the mapping from strategies to material payoffs, a remarkable result emerges from our analysis. Under weak selection the nature of the derived preferences are independent of the nature of the strategic interaction within islands. This is because the matching profiles then depend only on the population structure, without any reference to material payoffs. Hence, the utility function v^0 (see (18)) would remain

uninvadable even if the mapping from material payoffs to fitness and/or the mapping from strategies to material payoffs were to change over time, as long as these changes do not affect the matching probabilities, and as long as individuals understand the mapping from strategies to material payoffs and adjust the material payoff terms in v^0 accordingly.[29] Such robustness, however, presumes that Nash equilibrium play under the adjusted v^0 would be reached. Furthermore, the utility function at the fitness level would generally not remain uninvadable. Given that the aforementioned mappings have certainly changed over the course of human history, future research should lift the assumption of time-invariant mappings.

While our model is general in the sense that we allow for essentially any type of interactions within groups, it also has several limitations. Perhaps the strongest is that we only analyze type-homogeneous play and homogeneous populations subject to a single mutant. More realistic models, with heterogeneous individuals, heterogeneous islands and resident populations with multiple types are called for. Our hope is that the model proposed here can be fruitfully used to this end.

6 Appendix

6.1 Fitness and randomness

We here give details about how we justify the expression of individual direct fitness $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ in our model. First, we note that, generically in the infinite island model, lineage fitness of a mutant trait τ in a resident population with trait θ is defined as

$$W(\tau, \theta) = \sum_{k=0}^{n-1} p_k(\tau, \theta) \cdot \bar{w}(\tau, \theta, k), \tag{43}$$

where $\bar{w}(\tau, \theta, k)$ is the expected number of settled offspring in the next demographic time period that descend from a given adult mutant with trait τ in a group with exactly k other mutants, and thus $n-k-1$ individuals with trait θ. More formally, $\bar{w} : \Theta^2 \times \mathbb{N}_0 \to \mathbb{R}_+$ is defined as the expectation of the random number $W_{\tau,\theta,k} \in \mathbb{N}_0$ of settled offspring descending from the given mutant (including herself through survival), conditional on the event that in the parental demographic time period her island is in *state* $s = (\tau, \theta, k)$, that is, with k other mutants (with trait τ) and the other $n-k-1$ individuals with trait θ. The stochasticity in the random variable $W_{\tau,\theta,k}$ is due to *within-generation variability*.

Let $\boldsymbol{\Pi} = (\Pi_1, \Pi_2, ..., \Pi_n, \Pi^*)$ be the *random payoff vector* on an island, where Π_i is the random material payoff obtained by individual $i = 1, ..., n$ at the end of phase one of a demographic time period, and Π^* is the random payoff earned by a representative individual in an island where all individuals have trait θ. According to our decomposition

[29] To be more specific, and using Scenario A to illustrate this point, if the mappings f and s in (26) change, while the migration probability m as well as the function w remain unchanged, then the weights attached to the components in v^0 remain unchanged.

of a demographic time period into two phases (see section 2.1):

$$\bar{w}(\tau,\theta,k)=\mathbb{E}\left[W_{\tau,\theta}\mid k\right]=\mathbb{E}_1\left[\mathbb{E}_2\left[W_{\tau,\theta,k}\mid\mathbf{\Pi}\right]\mid k\right], \tag{44}$$

where $\mathbb{E}_1$ is the expectation over all stochastic events occurring during phase 1 of the demographic time period (potential randomness in the actions taken by individuals, and hence in payoffs obtained), while $\mathbb{E}_2$ is the expectation over all stochastic events occurring during phase 2 of the demographic time period (randomness in reproduction, survival, and/or sampling among competing offspring).

We note that three sources of within-generation variability can be distinguished in our model: (i) within-island *trait variability* (randomness in the number of other mutants), (ii) within-island *interaction and payoff variability* (for given number of mutants, randomness in the payoff vector), (iii) within-individual variability (for given number of mutants and payoffs, randomness in survival and number of surviving offspring). Hence, equation (43) can be viewed as a three-level iterated expectation:

$$W(\tau,\theta)=\mathbb{E}_0\left[\mathbb{E}_1\left[\mathbb{E}_2\left[W_{\tau,\theta,k}\mid\mathbf{\Pi}\right]\mid k\right]\right]. \tag{45}$$

This is the grand expectation of the random number $W_{\tau,\theta}$ of settled offspring descending from a mutant randomly sampled from the local lineage of the initial mutant, sampled during the random time interval until the first extinction of the local lineage, a time interval that is finite with probability one.

We are now in a position to introduce the continuously differentiable *individual fitness function* $w:\mathbb{R}^{n+1}\rightarrow\mathbb{R}$ that maps realized material payoff vectors to the expected number of offspring, conditional to the island state $s=(\tau,\theta,k)$:

$$w(\Pi_\tau,\langle\Pi_\tau,\Pi_\theta\rangle_k,\Pi^*)=\mathbb{E}_2\left[W_\tau\mid\mathbf{\Pi},\tau,\theta,k\right], \tag{46}$$

where $\langle\Pi_\tau,\Pi_\theta\rangle_k$ is the random vector of the island neighbors' payoffs, when k neighbors (of the given mutant) are mutants and the others are residents. We note that in a homogeneous population, that is, where all individuals carry the same trait, irrespective what that trait is, and all individuals use the same strategy, the random payoffs are identically and independently distributed, and hence, for $\tau=\theta$: $w(\Pi_\tau,\langle\Pi_\tau,\Pi_\theta\rangle_k,\Pi^*)=w(\tilde{\Pi},\left\langle\tilde{\Pi},\tilde{\Pi}\right\rangle,\tilde{\Pi})=1$, due to the constancy of the population in our model. Hence,

$$\bar{w}(\tau,\theta,k)=\mathbb{E}_1\left[w(\Pi_\tau,\langle\Pi_\tau,\Pi_\theta\rangle_k,\Pi^*)\mid\tau,\theta,k\right]. \tag{47}$$

6.2 Functions and randomness

So far, we imposed no restrictions on the effect of within-generation uncertainty. A key assumption we make in the analysis in the main text is that

$$\mathbb{E}_1\left[w(\Pi_\tau,\langle\Pi_\tau,\Pi_\theta\rangle_k,\Pi^*)\mid\tau,\theta,k\right]=w\left(\pi(\tau|k),\langle\pi(\tau|k),\pi(\theta|k)\rangle,\pi^*(\theta)\right), \tag{48}$$

where $\pi(\tau|k)=\mathbb{E}_1\left[\Pi_\tau\mid\tau,\theta,k\right]$, $\pi(\theta|k)=\mathbb{E}_1\left[\Pi_\theta\mid\tau,\theta,k\right]$, $\pi^*=\mathbb{E}_1\left[\Pi_\theta\mid\theta,\theta,k\right]$. Hence, we replace the expectation of a function by the function of the expectation for uncertainty in

phase 1 (type (ii) uncertainty above), which is a substantial assumption, except when all functions are affine.

When the game under consideration is but one source for individuals' fitness and there is variance in payoff, equation (48) is less restrictive under weak selection than may first be thought. To see this, suppose that the total random payoff to an individual i (who may have trait θ or τ) is the convex combination of two random variables, one exogenous random background payoff (from other interactions, say), Π_b, and the random payoff $\tilde{\Pi}_i$ from the material game under consideration in our model:

$$\Pi_i = (1-\delta)\Pi_b + \delta\tilde{\Pi}_i \tag{49}$$

where $\delta \in (0,1)$ is small. Then, by way of a Taylor expansion with respect to δ at $\delta = 0$, and using the zero-sum property of effects on individuals' fitnesses, we get (with subscripts on the function w denoting partial derivatives):

$$w(\Pi_i, \boldsymbol{\Pi}_{-i}, \Pi^*) = 1 + \delta \cdot w_1(\boldsymbol{\Pi}_b) \cdot \tilde{\Pi}_i + \delta \cdot \sum_{j=2}^{n} w_j(\boldsymbol{\Pi}_b) \cdot \tilde{\Pi}_j + \delta \cdot w_{n+1}(\boldsymbol{\Pi}_b) \cdot \tilde{\Pi}^* + \mathcal{O}(\delta^2). \tag{50}$$

Suppose further that the random baseline payoff Π_b is statistically independent from that of the specific game, which for an mutant individual i denote $\tilde{\Pi}_\tau$. Then

$$\begin{aligned}\mathbb{E}_1\left[w(\Pi_i, \boldsymbol{\Pi}_{-i}, \Pi^*)\right] \quad &= \quad 1 \; + \; \delta \cdot \mathbb{E}_1\left[w_1(\boldsymbol{\Pi}_b)\right] \cdot \mathbb{E}_1\left[\tilde{\Pi}_i\right] \\ &+ \; \delta \cdot \sum_{j=2}^{n} \mathbb{E}_1\left[w_j(\boldsymbol{\Pi}_b)\right] \cdot \mathbb{E}_1\left[\tilde{\Pi}_j\right] \; + \; \delta \cdot \mathbb{E}_1\left[w_{n+1}\left(\boldsymbol{\Pi}_b\right)\right] \cdot \mathbb{E}_1\left[\Pi^*\right] \; + \; \mathcal{O}(\delta^2), \end{aligned} \tag{51}$$

which leads to the same results as obtained by eq. (68) here below, but with partial derivatives $w_j(\boldsymbol{\Pi}_b)$ replaced by their expectation.

6.3 Proof of Proposition 1

We show first that $X_u \subseteq \hat{X}(\mathcal{P})$ is a sufficient condition for u to be uninvadable. Suppose that $X_u \subseteq \hat{X}(\mathcal{P})$. Then for each $\tilde{x} \in X_u$, (8) is satisfied for any strategy $y \in X$ played by mutants. In other words, there exists no $v \in \mathcal{F}$ for which some $(x, y) \in B_{\mathrm{NE}}(u, v)$ does not satisfy the inequality in (10). Hence, the condition (10) for u to be uninvadable in $\Theta = \mathcal{F}$ is satisfied.

We now show that $X_u \subseteq \hat{X}(\mathcal{P})$ is a necessary condition for u is uninvadable. Suppose to the contrary that u is uninvadable and that there exists some $\tilde{x} \in X_u$ such that $\tilde{x} \notin \hat{X}(\mathcal{P})$. Then there exists some $\tilde{y} \in X$ for which the inequality in (10) is not satisfied for $(\tilde{x}, \tilde{y})$. Consider the mutant utility function $v(x_i, \boldsymbol{x}_{-i}) \equiv \|x_i - \tilde{y}\|^2$; it induces mutants to play the strategy $\tilde{y}$ whichever strategy the residents play. Hence, there exists $(x, y) \in B_{\mathrm{NE}}(u, v)$ for which (10) is not satisfied. Since $v \in \mathcal{F}$, this means that u is invadable in $\Theta = \mathcal{F}$.

6.4 Proof of Proposition 2

Consider some uninvadable strategy $\hat{x} \in \hat{X}(\mathcal{P})$. Then

$$\hat{x} \in \arg\max_{y \in X} \sum_{k=0}^{n-1} p_k(y, \hat{x}) \cdot \tilde{w}\left(y, \mathbf{y}^{(k)}, \hat{\mathbf{x}}^{(n-1-k)}, \hat{x}\right). \tag{52}$$

Suppose now that $u_{\hat{x}}$ is the resident utility function. To see that $\hat{x}$ is then a resident strategy, note that given that an individual i's opponents in the group play $\hat{x}$, $u_{\hat{x}}$ writes:

$$u_{\hat{x},\boldsymbol{p}(x_i,\hat{x})}\left(x_i, \hat{\mathbf{x}}^{(n-1)}\right) = \sum_{k=0}^{n-1} p_k(x_i, \hat{x}) \cdot \tilde{w}\left(x_i, \mathbf{x}_i^{(k)}, \hat{\mathbf{x}}^{(n-1-k)}, \hat{x}\right), \tag{53}$$

so that $\hat{x}$ is a resident strategy iff

$$\hat{x} \in \arg\max_{x_i \in X} \sum_{k=0}^{n-1} p_k(x_i, \hat{x}) \cdot \tilde{w}\left(x_i, \mathbf{x}_i^{(k)}, \hat{\mathbf{x}}^{(n-1-k)}, \hat{x}\right), \tag{54}$$

which is true (to see this, compare this expression to (52)).

However, the fact that $\hat{x}$ is the unique strategy satisfying (52) does not preclude existence of other resident strategies under $u_{\hat{x}}$. Indeed, consider some strategy $\tilde{x}$. This is a resident strategy if

$$\tilde{x} \in \arg\max_{x_i \in X} \sum_{k=0}^{n-1} p_k(x_i, \hat{x}) \cdot \tilde{w}\left(x_i, \mathbf{x}_i^{(k)}, \tilde{\mathbf{x}}^{(n-1-k)}, \hat{x}\right). \tag{55}$$

Lastly, if $\hat{x}$ is the unique resident strategy under $u_{\hat{x}}$, then the set of resident strategies under $u_{\hat{x}}$ is a subset of $\hat{X}(\mathcal{P})$. This together with Proposition 1 implies that $u_{\hat{x}}$ is uninvadable in $\mathcal{F}$.

6.5 Proof of Proposition 3

For x to be uninvadable it must be that, given x, $y = x$ is a local maximum of

$$W(y, x) = \sum_{k=0}^{n-1} p_k(y, x)\, \tilde{w}\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big), \tag{56}$$

where $\mathbf{y}^{(k)}$ is the k-dimensional vector whose components all equal y, and $\mathbf{x}^{(n-1-k)}$ is the $(n-1-k)$-dimensional vector whose components all equal x, or $\left.\frac{\partial W(y,x)}{\partial y}\right|_{y=x} = 0$. To evaluate this first-order condition, we follow the same calculations as in Lehmann, Alger, and Weibull (2015) Appendix B. In particular, writing $\tilde{w}_j$ for the partial derivative of $\tilde{w}$ with

respect to its j-th argument,

$$\frac{\partial W(y,x)}{\partial y} = \sum_{k=0}^{n-1}\left[\frac{\partial p_k(y,x)}{\partial y}\tilde{w}\big(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\big)\right] + \sum_{k=0}^{n-1}\left[p_k(y,x)\sum_{j=1}^{k+1}\tilde{w}_j\big(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\big)\right]. \tag{57}$$

Noting that for $y=x$, $\tilde{w}\left(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\right)=\tilde{w}\left(x,\mathbf{x}^{(n-1)},x\right)=1$, which is independent of k so that it can be factored out in the first term, and that

$$\sum_{k=0}^{n-1}\left(\frac{\partial p_k(y,x)}{\partial y}\bigg|_{y=x}\right) = \frac{\partial}{\partial y}\left(\sum_{k=0}^{n-1}p_k(y,x)\right)\bigg|_{y=x} = \frac{\partial}{\partial y}(1)\bigg|_{y=x} = 0, \tag{58}$$

the expression simplifies to

$$\frac{\partial W(y,x)}{\partial y}\bigg|_{y=x} = \sum_{k=0}^{n-1}\left[p_k(y,x)\sum_{j=1}^{k+1}\tilde{w}_j\big(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\big)\right]\bigg|_{y=x}. \tag{59}$$

Permutation invariance further implies that for any $j\geq 2$, $\tilde{w}_j\big(x,\mathbf{x}^{(n-1)},x\big)=\tilde{w}_n\big(x,\mathbf{x}^{(n-1)},x\big)$ (it's as if the individual whose marginal type change is under consideration were systematically labeled to appear as the last component in the vector $\mathbf{x}^{(n-1)}$). Noticing also that $\sum_{k=0}^{n-1}\left[p_k(y,x)\tilde{w}_1\big(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\big)\right]\big|_{y=x}=\tilde{w}_1\big(x,\mathbf{x}^{(n-1)},x\big)$, we can write:

$$\begin{aligned}\frac{\partial W(y,x)}{\partial y}\bigg|_{y=x} &= \tilde{w}_1\big(x,\mathbf{x}^{(n-1)},x\big)+\sum_{k=1}^{n-1}\left[p_k(y,x)\sum_{j=2}^{k+1}\tilde{w}_j\left(y,\mathbf{y}^{(k)},\mathbf{x}^{(n-1-k)},x\right)\right]\bigg|_{y=x} \\ &= \tilde{w}_1\left(x,\mathbf{x}^{(n-1)},x\right)+\sum_{k=1}^{n-1}\left[p_k(x,x)k\tilde{w}_n\big(x,\mathbf{x}^{(n-1)},x\big)\right] \\ &= \tilde{w}_1\left(x,\mathbf{x}^{(n-1)},x\right)+(n-1)\,\tilde{w}_n\big(x,\mathbf{x}^{(n-1)},x\big)\sum_{k=1}^{n-1}\left[\frac{kp_k(x,x)}{(n-1)}\right] \\ &= \tilde{w}_1\big(x,\mathbf{x}^{(n-1)},x\big)+r(x,x)\tilde{w}_n\big(x,\mathbf{x}^{(n-1)},x\big),\end{aligned} \tag{60}$$

which owing to permutation invariance can also be written

$$\frac{\partial W(y,x)}{\partial y}\bigg|_{y=x} = \tilde{w}_1\big(x,\mathbf{x}^{(n-1)},x\big)+r(x,x)\sum_{j=2}^{n}\tilde{w}_j\big(x,\mathbf{x}^{(n-1)},x\big). \tag{61}$$

6.6 Proof of Proposition 4

The proof begins by deriving a lemma under strategy evolution, which is a generalization of Appendix B of Lehmann, Alger, and Weibull (2015), and will be a stepping stone towards

the result on preference evolution stated in the proposition. For this purpose, we define the *lineage payoff-advantage* of a mutant strategy $y \in X$ in a population of residents using strategy $x \in X$ as

$$\Pi(y,x) = \sum_{k=0}^{n-1} p_k^0 \cdot \tilde{\pi}^{(k)}(y,x), \tag{62}$$

where $\tilde{\pi}^{(k)}(y,x)$ is the mutant's *payoff advantage* when there are k other mutants in her or his island, defined by

$$\tilde{\pi}^{(k)}(y,x) = \pi(y|k) - \lambda_0 \cdot \left[\frac{k}{n-1}\pi(y|k) + \frac{n-1-k}{n-1}\pi(x|k)\right]. \tag{63}$$

The first term in (63) is the payoff of a descendant of the initial mutant who finds herself in an island with k other such descendants. The term in square brackets is the average material payoff earned by the other members in the island.

Lemma 1 *A strategy $\hat{x} \in X$ is uninvadable under weak selection if and only if*

$$\Pi(y,\hat{x}) \leq \Pi(\hat{x},\hat{x}) \qquad \forall y \in X. \tag{64}$$

Moreover, $1-n \leq \lambda_0 \leq 1$.

Proof of Lemma 1: Let $w : \mathbb{R}^{n+1} \to \mathbb{R}$ be any continuously differentiable fitness function, let $b \in \mathbb{R}$, and let $\mathbf{b}$ denote the vector in $\mathbb{R}^{n+1}$ that has all components equal to b. Then, by virtue of (86),

$$w_1(\mathbf{b}) + \sum_{j=2}^{n} w_j(\mathbf{b}) + w_{n+1}(\mathbf{b}) = 0, \tag{65}$$

where an index $k = 1, ..., n+1$ stands for the partial derivative of w with respect to its k-th argument.

Recalling the definition of $\bar{\pi}$ (see (16)), and omitting for notational simplicity the term $(1-\delta)\pi_0$, for any given payoff vector $(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) \in \mathbb{R}^{n+1}$ a first-order Taylor expansion of w with respect to δ evaluated at δ_0 writes

$$\begin{aligned} w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = {} & w(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*) + (\delta-\delta_0)\cdot w_1(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*)\cdot\pi_i \\ & + (\delta-\delta_0)\cdot\sum_{j=2}^{n}\left[w_j(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*)\cdot\pi_j\right] \\ & + (\delta-\delta_0)\cdot w_{n+1}(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*)\cdot\pi^* + \mathcal{O}(\delta^2). \end{aligned} \tag{66}$$

Evaluated at $\delta_0 = 0$, this expression writes

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = w(\pi_0) + \delta\cdot w_1(\pi_0)\cdot\pi_i + \delta\cdot\sum_{j=2}^{n} w_j(\pi_0)\cdot\pi_j + \delta\cdot w_{n+1}(\pi_0)\cdot\pi^* + \mathcal{O}(\delta^2), \tag{67}$$

where $w(\pi_0) = 1$, and π_0 is the vector in $\mathbb{R}^{n+1}$ whose components all equal π_0. By permutation invariance of $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ with respect to the components of $\boldsymbol{\pi}_{-i}$, we may for each $j = 2, ..., n$ write $w_n(\pi_0)$ instead of $w_j(\pi_0)$. Letting $\beta = w_1(\pi_0)$ and $\gamma = -(n-1)w_n(\pi_0)$, using (65), and rearranging terms, (67) can thus be written

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = 1 + \delta \cdot \left[\beta \cdot (\pi_i - \pi^*) - \frac{\gamma}{n-1}\sum_{j\neq i}(\pi_j - \pi^*)\right] + \mathcal{O}(\delta^2). \tag{68}$$

Letting

$$\lambda_0 = \frac{\gamma}{\beta} = -\frac{(n-1)w_n(\pi_0)}{w_1(\pi_0)}, \tag{69}$$

and factoring out $\beta > 0$ from (68), and simply omitting to write the factor δ in the fitness function, we conclude that for small $\delta > 0$,

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = 1 + \delta \cdot \beta\left[\pi_i - \lambda_0\sum_{j\neq i}\frac{\pi_j}{n-1} - (1-\lambda_0)\pi^*\right] + \mathcal{O}(\delta^2). \tag{70}$$

This shows that λ_0 quantifies fitness interdependence among patch members (Lehmann, Alger, and Weibull, 2015; see also Frank, 1998, and Gardner and West, 2004, for a description, but without a formal derivation, of λ_0).

The next step of the proof consists in obtaining an expression for local linage fitness under weak selection. Under weak selection the evolutionary process is what in biology is called *neutral* (Crow and Kimura, 1970, Ewens, 2004, Gillespie, 2004, and, for an explicit example, Rousset, 2004). Formally, this means that we can write

$$p_k(y, \hat{x}) = p_k^0 + \mathcal{O}(\delta) \qquad \forall k, \tag{71}$$

where $\mathcal{O}(\delta)$ accounts for the deviation (relative to the neutral process) of the strategy-profile distribution induced by selection (i.e., $\delta > 0$) that is at most of order δ, where p_k^0 is strategy-independent. Second, recalling the definition of $\tilde{w}$ (see (7)) and letting $\hat{x}$ denote the resident strategy, (70) can be written

$$\tilde{w}\left(y, \left(\mathbf{y}^{(k)}, \hat{\mathbf{x}}\right), \hat{x}\right) = 1 + \delta\beta \cdot \left[\tilde{\pi}^{(k)}(y, \hat{x}) - (1-\lambda_0)\pi(\hat{\mathbf{x}})\right] + \mathcal{O}(\delta^2), \tag{72}$$

where $\left(\mathbf{y}^{(k)}, \hat{\mathbf{x}}\right)$ is the $(n-1)$-dimensional vector with k components equal to y and the remaining components equal to $\hat{x}$, and (see equation (63))

$$\tilde{\pi}^{(k)}(y, \hat{x}) = \pi(y|k) - \lambda_0\left[\frac{k}{n-1}\pi(y|k) + \frac{n-1-k}{n-1}\pi(\hat{x}|k)\right]. \tag{73}$$

Using (71) and (72), local lineage fitness (see (2)) writes

$$\begin{aligned} W(y, \hat{x}) &= \sum_{k=0}^{n-1} p_k(y, \hat{x}) \cdot \tilde{w}\left(y, \left(\mathbf{y}^{(k)}, \hat{\mathbf{x}}\right)\right) \\ &= 1 + \delta\beta \cdot \sum_{k=0}^{n-1} p_k^0 \cdot \left[\tilde{\pi}^{(k)}(y, \hat{x}) - (1-\lambda_0)\pi(\hat{\mathbf{x}})\right] + \mathcal{O}(\delta^2). \end{aligned} \tag{74}$$

Recalling the definition of lineage payoff-advantage $\Pi(y,\hat{x})$ (see (62)),this can be written as

$$W(y,\hat{x}) = 1 + \delta\beta \cdot [\Pi(y,\hat{x}) - (1-\lambda_0)\pi(\hat{\mathbf{x}})] + \mathcal{O}(\delta^2). \tag{75}$$

Neglecting higher order terms in δ in this equation, the condition for uninvadability $[W(y,\hat{x}) \leq W(\hat{x},\hat{x})$ for all $y \in X]$ under weak selection is equivalent to the condition $\Pi(y,\hat{x}) \leq \Pi(\hat{x},\hat{x})$ for all $y \in X$.

Finally, we determine the implications of Assumption **[M]** for the bounds on $\lambda_0 = -(n-1)\cdot w_n(\pi_0)/w_1(\pi_0)$, focusing on the non-trivial case $w_n(\pi_0) \neq 0$. Part *(ii)* of the assumption implies $-(n-1) \leq \lambda_0$. Moreover, recalling (65) we obtain $\lambda_0 \leq 1$, with strict inequality when either $w_{n+1}(\pi_0) < 0$ or $w_{n+1}(\pi_0) = 0$ and $n > 2$. **Q.E.D.**

We are now in a position to complete the proof of the proposition. To establish the first claim of the proposition, we note that Lemma 1 implies that $X_{v^0} = \hat{X}$. The second claim follows by noting that any utility function $u \in \mathcal{F}$ for which some $x \in X_u$ is not an element of X_{v^0}, is invadable.

6.7 Approximation of the neutral distribution

Standard populations genetics results (see e.g., Lessard, 2007, and references therein) show that the neutral distribution of types in an island model with constant group size, and with population share of mutants $\varepsilon > 0$, is well approximated by way of the hypergeometric distribution

$$\phi_j(\varepsilon) = \binom{j+\omega\varepsilon-1}{j}\binom{n-j+\omega(1-\varepsilon)-1}{n-j}\Big/\binom{n+\omega-1}{n}, \tag{76}$$

where $\phi_j(\varepsilon)$ is the probability that there are $j = 0,1,..,n$ mutants in any given group, and $\omega = r_0/(1-r_0)$ (see Lessard, 2007, equation (7)). Since $\boldsymbol{p}^0 = \left(p_0^0,...,p_{n-1}^0\right)$ is the limit distribution when $\varepsilon \to 0$ of the number of *other* mutants in a given mutant's group, we have

$$p_k^0 = \lim_{\varepsilon\to 0}\ (k+1)\,\phi_{k+1}(\varepsilon)\Big/\left(\sum_{j=1}^{n} j\phi_j(\varepsilon)\right), \tag{77}$$

for $k = 0,1,...,n-1$. Upon rearrangements, this produces

$$p_k^0 = \binom{n}{k+1}\cdot\frac{(k+1)\,\omega}{n}\cdot\frac{\Gamma(k+1)\Gamma(\omega+n-k-1)}{\Gamma(\omega+n)}, \tag{78}$$

where Γ is the gamma function. This distribution depends only on group size n and pairwise relatedness r_0.

Numerical comparison between this approximation for the above evolutionary scenarios (that can all be subsumed under the relatedness in (40)) and the exact distribution shows that the average total variation between the approximate and exact distributions is quite small. Sampling randomly 10 000 values of s and m when $n = 5$ gives an average total variation of 0.005, a variation that should diminish with n. It can also be shown that in the special case of a Moran process ($s(\pi_i) = 1/n$) the approximation is in fact exact. (Indeed it can be verified that the expression in (78) then reduces to equation D.6 in Lehmann, Alger, and Weibull, 2015.)

6.8 Proof of Proposition 5

Recalling that

$$\tilde{w}(x_i, \boldsymbol{x}_{-i}, x) = w\left(\pi\left(x_i, \boldsymbol{x}_{-i}\right), \left(\pi\left(x_j, \boldsymbol{x}_{-j}\right)\right)_{j\neq i}, \pi^*\left(x\right)\right), \tag{79}$$

we obtain

$$\begin{aligned}\tilde{w}_1\big(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right), x\big) = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-1) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right),\end{aligned} \tag{80}$$

where $\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}$ denotes the $(n-1)$-dimensional vector whose components all equal $\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)$, and

$$\begin{aligned}\tilde{w}_n\big(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right), x\big) = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-2) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right).\end{aligned} \tag{81}$$

Substituting the last two equations into the last line of (60) produces

$$\begin{aligned}0 = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-1) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)\, w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)(n-2) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)\, w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right).\end{aligned} \tag{82}$$

Noting that with the notation used in this proof, $\lambda\left(x\right)$ writes

$$\lambda\left(x\right) = -\frac{(n-1)\, w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right)}{w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right)}, \tag{83}$$

(82) can be written

$$\pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) + (n-1) \cdot \frac{r\left(x, x\right) - \lambda\left(x\right)\left[\frac{1}{n-1} + r\left(x, x\right)\frac{n-2}{n-1}\right]}{1 - \lambda\left(x\right) r\left(x, x\right)} \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) = 0, \tag{84}$$

or

$$\left[1 - \kappa\left(x\right)\right] \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) + \kappa\left(x\right) \cdot \sum_{k=1}^{n} \pi_k\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) = 0. \tag{85}$$

6.9 Proof of Proposition 6

To show that $\kappa(x) \in [-1,1]$, we begin by studying $\lambda(x)$. Note that the terms that define $\lambda(x)$ are partial derivatives evaluated in a homogenous population. Furthermore, since population size is constant in a homogenous population, each individual's fitness would remain at 1 following a marginal change in the material payoff of all the individuals in the population. Formally:

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\bigg|_{\pi_i=\pi_j=\pi^*} + \sum_{j=2}^{n} \frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\bigg|_{\pi_i=\pi_j=\pi^*} + \frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi^*}\bigg|_{\pi_i=\pi_j=\pi^*} = 0. \tag{86}$$

By permutation invariance, and using a more compact notation, this writes $w_1(\cdot)+(n-1)w_n(\cdot)+w_{n+1}(\cdot)=0$. Using this and the assumption $w_1(\cdot)>0$,

$$\begin{aligned} \lambda(x) < 1 &\Leftrightarrow -(n-1)w_n(\cdot) < w_1(\cdot) \\ &\Leftrightarrow w_1(\cdot) + w_{\pi^*}(\cdot) < w_1(\cdot), \end{aligned} \tag{87}$$

which is true by Assumption **[M]** *(iii)*.

Since $r(x,x) \in [0,1]$ for all x this implies that $\lambda(x)\, r(x,x) < 1$, and hence

$$\begin{aligned} \kappa(x) \leq 1 &\Leftrightarrow r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right] \leq 1 - \lambda(x)\, r(x,x) \\ &\Leftrightarrow \lambda(x)\left[r(x,x) - \frac{1}{n-1} - r(x,x)\frac{n-2}{n-1}\right] \leq 1 - r(x,x) \\ &\Leftrightarrow \lambda(x)\left[\frac{r(x,x)-1}{n-1}\right] \leq 1 - r(x,x) \\ &\Leftrightarrow \lambda(x) \geq -(n-1) \\ &\Leftrightarrow -\frac{(n-1)w_n(\cdot)}{w_1(\cdot)} \geq -(n-1) \\ &\Leftrightarrow w_n(\cdot) \leq w_1(\cdot), \end{aligned} \tag{88}$$

which is true by virtue of Assumption **[M]** *(ii)*.

We now show that $\kappa(x) \geq -1$. For any $\lambda(x) < 1$, $\kappa(x)$ is increasing in $r(x,x)$. Indeed, the partial derivative of the expression for $\kappa(x)$ with respect to $r(x,x)$ has the same sign as (in this expression $r \equiv r(x,x)$ and $\lambda \equiv \lambda(x)$)

$$\begin{aligned} &[(n-1)(1-\lambda)+\lambda](n-1)(1-\lambda r) + \lambda(n-1)[r(n-1)(1-\lambda) - \lambda(1-r)] \\ &= (n-1)(1-\lambda)(n-1+\lambda). \end{aligned} \tag{89}$$

For the inequality $\kappa(x) \geq -1$ to hold, it is thus sufficient that $\kappa(x) \geq -1$ for $r(x,x)=0$, a condition which reduces to

$$-\lambda(x) \geq -(n-1), \tag{90}$$

which is true for any $n \geq 2$ since $\lambda(x) < 1$.

Finally,

$$\begin{aligned}\kappa(x) \leq r(x,x) &\Leftrightarrow \frac{r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right]}{1 - \lambda(x) r(x,x)} \leq r(x,x) \\ &\Leftrightarrow r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right] \leq r(x,x)\left[1 - \lambda(x) r(x,x)\right] \\ &\Leftrightarrow \lambda(x)\left[r(x,x)\right]^2 \leq \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right] \\ &\Leftrightarrow \lambda(x) r(x,x)\left[1 - \left[1 - r(x,x)\right](n-1)\right] \leq \lambda(x).\end{aligned} \tag{91}$$

This inequality is true if and only if $\lambda(x) \geq 0$ by virtue of the fact that for all $r(x,x) \in [0,1)$ we have $r(x,x)\left[1 - \left[1 - r(x,x)\right](n-1)\right] \leq 1$. Likewise, it is clear that $\kappa(x) > r(x,x)$ if and only if $\lambda(x) < 0$.

Finally, the last result stated in the proposition is implied by (20) together with Assumption **[M]** *(i)*.

6.10 Calculating the coefficients of fitness interdependence and pairwise relatedness

6.10.1 Scenario A: Genes

To calculate $\lambda(x)$ we begin by calculating the partial derivatives needed for this purpose. Here, from the individual fitness function (26):

$$\begin{aligned}\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j} &= -\frac{\partial s(\pi_j)}{\partial \pi_j} \cdot \frac{(1-m) f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &\quad - \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2}\end{aligned} \tag{92}$$

and

$$\begin{aligned}\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i} &= \frac{\partial s(\pi_i)}{\partial \pi_i} - \frac{\partial s(\pi_i)}{\partial \pi_i} \cdot \frac{(1-m) f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &\quad + \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)\left[\partial f(\pi_i)/\partial \pi_i\right]}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &\quad - \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2} \\ &\quad + \left[1 - s(\pi^*)\right] \cdot \frac{m\left[\partial f(\pi_i)/\partial \pi_i\right]}{f(\pi^*)}\end{aligned} \tag{93}$$

Writing $s'(\pi^*)$ for $\left.\frac{\partial s(\pi_j)}{\partial \pi_j}\right|_{\pi_j=\pi^*}$ and $f'(\pi^*)$ for $\left.\frac{\partial f(\pi_j)}{\partial \pi_j}\right|_{\pi_j=\pi^*}$, we obtain (upon simplification)

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\right|_{\pi_i=\pi_j=\pi^*} = -s'(\pi^*) \cdot \frac{1-m}{n} - [1-s(\pi^*)] \cdot \frac{(1-m)^2 f'(\pi^*)}{n f(\pi^*)} \tag{94}$$

and

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\right|_{\pi_i=\pi_j=\pi^*} = \frac{1}{n}(n+m-1)s'(\pi^*) + [1-s(\pi^*)] \frac{f'(\pi^*)}{n f(\pi^*)} \left[n-(1-m)^2\right]. \tag{95}$$

Upon simplification, we thus obtain

$$\lambda(x) = \frac{(n-1)(1-m)\left\{(1-m)[1-s(\pi^*)]\frac{f'(\pi^*)}{f(\pi^*)} + s'(\pi^*)\right\}}{\left[n-(1-m)^2\right][1-s(\pi^*)]\frac{f'(\pi^*)}{f(\pi^*)} + (n+m-1)s'(\pi^*)}. \tag{96}$$

The expression in (28) obtains by setting $s'(\pi^*) = 0$.

To calculate $r(x,x)$, one uses a recursion equation (this is a standard technique for calculating probabilities of identity by descent; see Nagylaki, 1992, and Rousset, 2004, for a background). In the scenario at hand, this equation writes

$$\begin{aligned} r(x,x) = {} & [s(\pi^*)]^2 r(x,x) + 2s(\pi^*)[1-s(\pi^*)](1-m)\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right] \\ & + [1-s(\pi^*)]^2 (1-m)^2 \left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right]. \end{aligned} \tag{97}$$

The left-hand side is the average probability that, in a monomorphic population in which all individuals play x, the neighbor of a randomly drawn member of a certain local lineage is also a member of this local lineage. The terms on the right-hand side details the events in which this happens. The first term on the right hand side corresponds to the event that both the individual at hand and the randomly drawn neighbor survived from the previous period. The second term on the right hand side corresponds to the two events in which either the individual at hand or the randomly drawn neighbor survived from the previous period while the other didn't, and the one who didn't survive from the previous period did not migrate in from another island. In this case, there is a probability $1/n$ that one is the offspring of the other, in which case they are both members of the same local lineage with certainty; with the complementary probability, they are not parent and child, in which case the probability that they are both members of the same local lineage equals $r(x,x)$. The third term on the right hand side corresponds to the event in which neither the individual at hand nor the randomly drawn neighbor survived from the previous period, and neither of them migrated in from another island. In this case, there is a probability $1/n$ that they have the same parent, in which case they are both members of the same local lineage with certainty; with the complementary probability, they have different parents, in which case the probability that they are both members of the same local lineage equals $r(x,x)$. Solving (97) for $r(x,x)$ yields

$$r(x,x) = \frac{(1-m)^2 + (1-m^2)s(\pi^*)}{n-(n-1)(1-m)^2 + [1+(n-1)m^2]s(\pi^*)}. \tag{98}$$

The expression in (27) obtains by replacing $s(\pi^*)$ by s_0.

6.10.2 Scenario B: Guns

In the biological scenario with wars, we obtain from the individual fitness function (30):

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j} = 2\rho \left[\partial g(\boldsymbol{\pi}, \pi^*)/\partial \pi_j\right] \cdot \left[m \cdot \frac{f(\pi_i)}{f(\pi^*)} + (1-m)\, n \cdot \frac{f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \right] - \left[(1-\rho) + 2\rho g(\boldsymbol{\pi}, \pi^*)\right] \cdot (1-m)\, n \cdot \frac{(1-m) f(\pi_i) \left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2} \tag{99}$$

and

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i} = 2\rho \left[\partial g(\boldsymbol{\pi}, \pi^*)/\partial \pi_i\right] \cdot \left[m \cdot \frac{f(\pi_i)}{f(\pi^*)} + (1-m)\, n \cdot \frac{f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \right] + \left[(1-\rho) + 2\rho g(\boldsymbol{\pi}, \pi^*)\right] \cdot (1-m)\, n \cdot \frac{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right] \partial f(\pi_i)/\partial \pi_i - (1-m) f(\pi_i) \left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2} + m \frac{\partial f(\pi_i)/\partial \pi_i}{f(\pi^*)}. \tag{100, 101}$$

By permutation invariance of g, write $g_n(\boldsymbol{\pi}^*, \pi^*)$ for $\left.\frac{\partial g(\boldsymbol{\pi}, \pi^*)}{\partial \pi_j}\right|_{\pi_j = \pi^*}$ for all $j = 1, ..., n$. Since, moreover, $g(\boldsymbol{\pi}, \pi^*) = 1/2$, upon simplification we obtain:

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\right|_{\pi_i = \pi_j = \pi^*} = 2\rho g_n(\boldsymbol{\pi}^*, \pi^*) - \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)} \tag{102}$$

and

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\right|_{\pi_i = \pi_j = \pi^*} = 2\rho g_n(\boldsymbol{\pi}^*, \pi^*) - \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)} + \frac{f'(\pi^*)}{f(\pi^*)}, \tag{103}$$

so that

$$\lambda(x) = -\frac{(n-1)\left[2\rho g_n(\boldsymbol{\pi}^*, \pi^*) - \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)}\right]}{2\rho g_n(\boldsymbol{\pi}^*, \pi^*) + \left[n - (1-m)^2\right] \frac{f'(\pi^*)}{nf(\pi^*)}}. \tag{104}$$

The recursion equation to calculate $r(x, x)$ writes

$$r(x, x) = (1-m)^2 \left[\frac{1}{n} + \frac{n-1}{n} r(x, x)\right]. \tag{105}$$

In this scenario, the only event in which a randomly drawn individual can belong to the same local lineage as a randomly drawn neighbor, is when both stayed in their natal island. In this case, there is a probability $1/n$ that they have the same parent, in which case they belong to the same local lineage with certainty; with the complementary probability, they have different parents, in which case the probability that they belong to the same local lineage is $r(x,x)$. Solving for $r(x,x)$ yields

$$r(x,x) = \frac{(1-m)^2}{n-(n-1)(1-m)^2}. \tag{106}$$

6.10.3 Scenario B: Wars (weak selection)

Recall that under weak selection we write the individual fitness of individual i as $w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)$, where $(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*) = ((1-\delta)\pi_0 + \delta\pi_i, (1-\delta)\boldsymbol{\pi}_0 + \delta\boldsymbol{\pi}_{-i}, (1-\delta)\pi_0 + \delta\pi^*)$ (here $\boldsymbol{\pi}_0$ is the $(n-1)$-dimensional vector whose components all equal π_0) and $\delta \geq 0$ represents the intensity of selection (see (16)). From (69) in the proof of Proposition 5, we have

$$\lambda_0 = -\frac{\sum_{j\neq i} \partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)/\partial \bar{\pi}_j|_{\delta=0}}{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)/\partial \bar{\pi}_i|_{\delta=0}}. \tag{107}$$

Since, for $\delta = 0$, $\bar{\pi}_i = \bar{\pi}_j = \bar{\pi}^*$, we obtain from (102) and (103) that

$$\left.\frac{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)}{\partial \bar{\pi}_j}\right|_{\delta=0} = 2\rho g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) - \frac{(1-m)^2 f'(\bar{\pi}^*)}{n f(\bar{\pi}^*)} \tag{108}$$

and

$$\left.\frac{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)}{\partial \bar{\pi}_i}\right|_{\delta=0} = 2\rho g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) - \frac{(1-m)^2 f'(\bar{\pi}^*)}{n f(\bar{\pi}^*)} + \frac{f'(\bar{\pi}^*)}{f(\bar{\pi}^*)}. \tag{109}$$

With the expressions for f and g given in (33) and (34), and the assumption that $V(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}) = (1-\delta)n\pi_0 + \delta\left(\pi_i + \sum_{j\neq i}\pi_j\right)$ (note that we assume that the intensity of selection is the same for fecundity and for the probability of winning wars; one can also allow for different selection intensities), we have:

$$\frac{f'(\bar{\pi}^*)}{f(\bar{\pi}^*)} = 1 \tag{110}$$

and

$$g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) = \frac{1}{4}. \tag{111}$$

Hence, we get

$$\lambda_0 = -\frac{(n-1)\left[\rho/2 - \frac{(1-m)^2}{n}\right]}{\rho/2 - \frac{(1-m)^2}{n} + 1}, \tag{112}$$

which upon simplification gives the expression in (35). It can then be verified that, together with the fact that r_0 is given by (106), this gives the expression for κ_0 in (36).

6.10.4 Scenario C: Culture

In the cultural scenario, we have from (38):

$$\begin{aligned}\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j} &= -\frac{\partial s\left(\pi_j\right)}{\partial \pi_j} \cdot \frac{(1-m) f\left(\pi_i\right)}{\sum_{j=1}^n f\left(\pi_j\right)} \\ &\quad - \left(n - \sum_{j=1}^n s\left(\pi_j\right)\right) \cdot \frac{(1-m) f\left(\pi_i\right)\left[\partial f\left(\pi_j\right) / \partial \pi_j\right]}{\left[\sum_{j=1}^n f\left(\pi_j\right)\right]^2}\end{aligned} \tag{113}$$

and

$$\begin{aligned}\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i} &= \frac{\partial s\left(\pi_i\right)}{\partial \pi_i} - \frac{\partial s\left(\pi_i\right)}{\partial \pi_i} \cdot \frac{(1-m) f\left(\pi_i\right)}{\sum_{j=1}^n f\left(\pi_j\right)} \\ &\quad + (1-m)\left(n - \sum_{j=1}^n s\left(\pi_j\right)\right) \cdot \frac{\left[\partial f\left(\pi_i\right) / \partial \pi_i\right]\left[\sum_{j \neq i} f\left(\pi_j\right)\right]}{\left[\sum_{j=1}^n f\left(\pi_j\right)\right]^2} \\ &\quad + \left[1 - s\left(\pi^*\right)\right] \cdot \frac{m\left[\partial f\left(\pi_i\right) / \partial \pi_i\right]}{f\left(\pi^*\right)}.\end{aligned} \tag{114}$$

Upon simplification, we obtain:

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j}\right|_{\pi_i = \pi_j = \pi^*} = -\frac{(1-m)}{n}\left[s'\left(\pi^*\right) + \left[1 - s\left(\pi^*\right)\right] \cdot \frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}\right] \tag{115}$$

and

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i}\right|_{\pi_i = \pi_j = \pi^*} = \left(\frac{n-1+m}{n}\right)\left[s'\left(\pi^*\right) + \left[1 - s\left(\pi^*\right)\right] \cdot \frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}\right]. \tag{116}$$

Hence:

$$\lambda(x) = \frac{(n-1)(1-m)}{n-1+m}. \tag{117}$$

For $r(x,x)$ the recursion equation writes

$$\begin{aligned}r(x,x) &= \left[s\left(\pi^*\right)\right]^2 r(x,x) + 2(1-m)\, s\left(\pi^*\right)\left[1 - s\left(\pi^*\right)\right]\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right] \\ &\quad + (1-m)^2\left[1 - s\left(\pi^*\right)\right]^2 \cdot \left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right].\end{aligned} \tag{118}$$

The first term on the right-hand side corresponds to the event that both the individual at hand and the randomly drawn neighbor have been loyal to their parents, where the neighbor's parent belongs to the individual's lineage with probability $r(x,x)$. The second term on the right hand side corresponds to the event that either the individual at hand was loyal to its parent but the randomly drawn neighbor was not loyal to its parent, or the other way around. In this case, there is a probability $1/n$ that the non-loyal child acquired its trait from the

loyal child's parent, in which case they both belong to the same local lineage with certainty, while with the complementary probability this did not happen, in which case the probability that the randomly neighbor belongs to the same local lineage is $r(x,x)$. The third term on the right hand side corresponds to the event that neither the individual at hand nor the randomly drawn neighbor were loyal to their parents but both of them acquired their trait from someone in the island. In this case, there is a probability $1/n$ that they acquired their type from the same adult, in which case they belong to the same local lineage with certainty; with the complementary probability they have different cultural parents, in which case the probability that the randomly drawn neighbor belongs to the same local lineage as the individual at hand is $r(x,x)$. We note that the equation simplifies to

$$r(x,x) = [s(\pi^*)]^2 r(x,x) + 2s(\pi^*)[1-s(\pi^*)](1-m)\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right] \tag{119}$$
$$+ [1-s(\pi^*)]^2 (1-m)^2 \left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right].$$

The expression in the text obtains upon observing that this equation is identical to the one in (97).

References

[1] Akçay, E., and J. van Cleve (2013): "Behavioral responses in structured populations pave the way to group optimality," *American Naturalist* 179, 257-269.

[2] Alger, I. and J. Weibull (2010): "Kinship, Incentives and Evolution," *American Economic Review*, 100, 1725-1758.

[3] Alger, I. and J. Weibull (2012): "A Generalization of Hamilton's Rule—Love Others How Much?" *Journal of Theoretical Biology*, 299, 42-54.

[4] Alger, I., and J.W. Weibull (2013): "Homo moralis—Preference evolution under incomplete information and assortative matching," *Econometrica* 81, 2269-2302.

[5] Alger, I., and J.W. Weibull (2016): "Evolution and Kantian morality," *Games and Economic Behavior* 98, 56-67.

[6] Alger, I., and J.W. Weibull (2019): "Evolution and Kantian morality," forthcoming *Annual Review of Economics.*

[7] Aoki, K. (1982): "A condition for group selection to prevail over counteracting individual selection," *Evolution* 36, 832–842.

[8] Bao, M., and G. Wild (2012): "Reproductive skew can provide a net advantage in both conditional and unconditional social interactions," *Theoretical Population Biology* 82, 200-208.

[9] Bergstrom, T. (1995): "On the evolution of altruistic ethical rules for siblings," *American Economic Review* 85, 58-81.

[10] Bergstrom, T. (1996): "Economics in a family way," *Journal of Economic Literature* 34, 1903-1934.

[11] Bergstrom, T. (2003): "The algebra of assortative encounters and the evolution of cooperation," *International Game Theory Review* 5, 211-228.

[12] Binmore, K. (1998): *Game Theory and the Social Contract: Just Playing*. MIT Press, Massachusetts.

[13] Bisin, A., and T. Verdier (2001): "The economics of cultural transmission and the dynamics of preferences", *Journal of Economic Theory* 97: 298-319.

[14] Bowles, S. (2006): "Group competition, reproductive leveling, and the evolution of human altruism," *Science* 314, 1569-1572.

[15] Bowles S. (2009): "Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors?" *Science* 324, 1293-8.

[16] Bowles, S., and H. Gintis (1998): "The moral economy of communities: structured populations and the evolution of pro-social norms," *Evolution and Human Behavior* 19, 3-25.

[17] Bshary, R., and R. Bergmüller (2008): "Distinguishing four fundamental approaches to the evolution of helping," *Journal of Evolutionary Biology* 21, 405-20

[18] Cavalli-Sforza, L.L., and W.F. Bodmer (1971): *The Genetics of Human Populations.* W.H. Freeman, San Francisco.

[19] Clobert, J., E. Danchin, A. Dhondt, J.D. and Nichols (2001): *Dispersal*. Oxford University Press, Oxford.

[20] Crow, J. F. and M. Kimura (1970): *An Introduction to Population Genetics Theory.* Harper and Row, New York.

[21] Dekel, E., J.C. Ely, and O. Yilankaya (2007): "Evolution of preferences," *Review of Economic Studies* 74, 685-704.

[22] Dos Santos, M., and J. Peña (2017): "Antisocial rewarding in structured populations," *Scientific Reports* 7, 6212.

[23] Eshel, I. (1972): "On the neighbor effect and the evolution of altruistic traits," *Theoretical Population Biology* 11, 258–277.

[24] Ewens, W. J. (2004): *Mathematical Population Genetics*. Springer Verlag, New York.

[25] Frank, S.A. (1998): *Foundations of Social Evolution*. Princeton University Press, Princeton, NJ.

[26] Gardner, A., and S.A. West (2004), "Spite and the scale of competition," *Journal of Evolutionary Biology* 17, 1195-1203.

[27] Gardner, A., and S.A. West (2006): "Demography, altruism, and the benefits of budding," *Journal of Evolutionary Biology* 19, 1707-1716.

[28] Gillespie, J. H. (2004): *Population Genetics: a Concise Guide*. Johns Hopkins, Baltimore & London.

[29] Grafen, A. (1985): "A geometric view of relatedness," in Dawkins, R. and M. Ridley, eds., *Oxford Surveys in Evolutionary Biology*. Oxford University Press, Oxford.

[30] Grueter, C.C., B. Chapais, and D. Zinner (2012): "Evolution of multilevel social systems in nonhuman primates and humans," *International Journal of Primatology* 33, 1002–1037.

[31] Hamilton, W.D. (1964): "The genetical evolution of social behaviour," *Journal of Theoretical Biology* 7, 1-52.

[32] Hamilton, W.D. (1967): "Extraordinary sex ratios: a sex-ratio theory for sex linkage and inbreeding has new implications in cytogenetics and entomology *Science* 156, 477-88.

[33] Hamilton, W.D. (1971): "Selection of selfish and altruistic behavior in some extreme models," in J.F. Eisenberg and W. S.Dillon, eds. *Man and beast: comparative social behavior*. Smithsonian Institution Press, Washington, D.C.

[34] Hartl, D.L., and A.G. Clark (2007): *Principles of Population Genetics* (4th edition). Sinauer and Associates , Sunderland, MA.

[35] Heifetz, A., C. Shannon, and Y. Spiegel (2007a): "What to maximize if you must," *Journal of Economic Theory* 133, 31-57.

[36] Heifetz, A., C. Shannon, and Y. Spiegel (2007b): "The dynamic evolution of preferences," *Economic Theory* 32, 251-286.

[37] Hirshleifer, J. (1977): "Economics from a biological viewpoint," *Journal of Law and Economics* 20, 1-52.

[38] Johnstone, R.A., and M.A. Cant (2008) "Sex differences in dispersal and the evolution of helping and harming," *American Naturalist* 172, 318-30.

[39] Konrad, K.A. (2014) "Strategic aspects of fighting in alliances," in K. Wärneryd, ed. *The Economics of Conflict: Theory and Empirical Evidence*. MIT Press, Cambridge, MA.

[40] Layton, R., S. O'Hara, and A. Bilsborough (2012): "Antiquity and social functions of multilevel social organization among human hunter-gatherers," *International Journal of Primatology* 33, 1215–1245.

[41] Lehmann, L., I. Alger, and J.W. Weibull (2015): "Does evolution lead to maximizing behavior?" *Evolution* 69-7, 1858–1873.

[42] Lehmann, L., and Balloux (2007): "Natural selection on fecundity variance in subdivided populations: kin selection meets bet hedging," *Genetics* 176, 361-377.

[43] Lehmann, L., K., Foster, and F. Feldman (2008): "Cultural transmission can inhibit the evolution of altruistic helping," *American Naturalist* 172, 12-24

[44] Lehmann, L., C. Mullon, E. Akçay, and J. Van Cleve (2016): " Invasion fitness, inclusive fitness, and reproductive numbers in heterogeneous populations," *Evolution* 70, 1689–1702.

[45] Lehmann, L., and F. Rousset (2010): "How life history and demography promote or inhibit the evolution of helping behaviours," *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 2599-2617.

[46] Lessard, S. (2007): "An exact sampling formula for the Wright-Fisher model and a solution to a conjecture about the finite-island model," *Genetics* 177, 1249-1254.

[47] Lion, S., and S. Gandon (2010): "Life history, habitat saturation and the evolution of fecundity and survival altruism," *Evolution* 64, 1594-606.

[48] Malone, N., A. Fuentes, and F.J. White (2012): "Variation in the social systems of extant hominoids: comparative insight into the social behavior of early hominins," *International Journal of Primatology* 33, 1251–1277.

[49] Maynard Smith, J. (1964): "Group selection and kin selection," *Nature*, 201, 1145- 1147.

[50] Micheletti, Alberto J.C., D. Ruxton and A. Gardner (2017): "Intrafamily and intragenomic conflicts in human warfare", *Proceedings of the Royal Society of London Series B* (Biological Sciences) 284, issue 1849.

[51] Michod, R.E., and W.D. Hamilton (1980): "Coefficients of relatedness in sociobiology," *Nature* 288, 694–697.

[52] Michod, R.E. (1982): "The theory of kin selection," *Annual Review of Ecology and Systematics* 13, 23-55.

[53] Nagylaki, T. (1992): *Introduction to Population Genetics*. Springer, Berlin.

[54] Nagylaki, T. (1993): "The evolution of multilocus systems under weak selection," *Genetics*, 134, 627-647.

[55] Nax, H., and A. Rigos (2016): "Assortativity evolving from social dilemmas," *Journal of Theoretical Biology*, 395, 194-203.

[56] Newton, J. (2017): "The preferences of Homo Moralis are unstable under evolving assortativity," *International Journal of Game Theory* 46, 583-589.

[57] Newton, J. (2018): "Evolutionary game theory: a renaissance," *Games,* 9, 31.

[58] Ohtsuki, H. (2012): "Does synergy rescue the evolution of cooperation? An analysis for homogeneous populations with non-overlapping generations," *Journal of Theoretical Biology,* 307, 20-28.

[59] 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.

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

[61] Rogers, A.R. (1994): "Evolution of time preference by natural selection," *American Economic Review*, 84, 460-481.

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

[63] Schaffer, M.E. (1988): "Evolutionarily stable strategies for finite populations and variable contest size," *Journal of Theoretical Biology* 132, 467-478.

[64] Skaperdas, S. (1996): "Contest success functions," *Economic Theory* 7, 283-290.

[65] Taylor, P.D. (1992a) "Altruism in viscous populations - an inclusive fitness model," *Evolutionary Ecology* 6, 352-356.

[66] Taylor, P.D. (1992b) "Inclusive fitness in a homogeneous environment," *Proceedings of the Royal Society B*, 249, 299-302.

[67] Taylor, P.D., and A.J. Irwin (2000) "Overlapping generations can promote altruistic behavior," *Evolution* 54, 1135-41.

[68] Taylor, P.D., and S. Frank (1996) "How to make a kin selection model," *Journal of Theoretical Biology* 180, 27-37.

[69] Tullock, G. (1980): "Efficient rent seeking," in Buchanan, J., Tollison, R. and Tullock, G., Eds., *Toward a Theory of Rent Seeking Society.* Texas A and M University Press, College Station, 97-112.

[70] Van Cleve, J. (2015): "Social evolution and genetic interactions in the short and long term," *Theoretical Population Biology* 2013, 2-26.

[71] Van Schaik, C. P. (2016) *The Primate Origin of Human Behavior*. Wiley-Blackwell, Hoboken, NJ.

[72] West, S.A., A. Griffin, and A. Gardner (2007): "Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection," *Journal of Evolutionary Biology* 20, 415-32.

[73] Wild, G., and A. Traulsen (2007): "The different limits of weak selection and the evolutionary dynamics of finite populations," *Journal of Theoretical Biology* 247, 382-390.

[74] Wilson, D.S., G.B. Pollock, and L.A. Dugatkin (1992): "Can altruism evolve in purely viscous populations?," *Evolutionary Ecology* 6, 331-341.

[75] Wright, S. (1931): "Evolution in Mendelian populations," *Genetics* 16, 97–159.

[76] Wright, S. (1943): "Isolation by distance," *Genetics* 28, 114–138.

[77] Wu, J. (2017): "Political institutions and the evolution of character traits," *Games and Economic Behavior*, 106, 260-276.

[78] Wu, J. (2019): "Labelling, homophily, and preference evolution," *International Journal of Game Theory*, forthcoming.