International Journal of Game Theory (2020) 49:1–22
https://doi.org/10.1007/s00182-019-00672-6

ORIGINAL PAPER

# Labelling, homophily and preference evolution

**Jiabin Wu**[1]

Accepted: 3 April 2019 / Published online: 9 April 2019


**Abstract**
We consider a population of agents whose preference types are unobservable but imperfectly correlated with certain observable labels such as customs, languages, and origins. In addition, the matching process exhibits homophily: agents tend to interact with those who share the same labels. We show that labelling and homophily interact in a non-trivial way to influence the evolution of preferences, which cannot be accounted for in the extant literature.

**Keywords** Homophily · Labelling · Assortative matching · Preference evolution · Evolutionary game theory

**JEL Classifications** C73 · D80

## 1 Introduction

The indirect evolutionary approach pioneered by Güth and Yaari (1992) provides a useful methodology for studying preference evolution: preferences govern behavior, behavior determines fitness successes, and fitness successes regulate the evolution of preferences.[1] Early models of preference evolution assume that (1) individuals in the population are uniformly randomly matched in pairs or groups to engage in strategic

[1] See also Güth (1995), Bester and Güth (1998), Fershtman and Weiss (1998), Huck and Oechssler (1999), McNamara et al. (1999), Sethi and Somanathan (2001), Koçkesen et al. (2000), Ely and Yilankaya (2001), Ok and Vega-Redondo (2001), Van Veelen (2006), Dekel et al. (2007), Heifetz et al. (2007a, b), Akçay et al. (2009), Herold and Kuzmics (2009), Alger (2010) and Alger (2010); Alger and Weibull (2012, 2013, 2016, 2019). See Newton (2018) and Alger and Weibull (2019) for surveys of recent contributions to preference evolution. See also Robson (2001) and Robson and Samuelson (2011) for a survey of another important approach for studying preference evolution.

Jiabin Wu sincerely thanks the associate editor and two anonymous referees for the comments and suggestions that greatly improved the paper.

✉ Jiabin Wu
jwu5@uoregon.edu

[1] Department of Economics, University of Oregon, 515 PLC, 1285, Eugene, OR 97408, USA

interactions and (2) they can perfectly observe their opponents' preferences. Recent works realize that these two assumptions may be unrealistic and need to be relaxed.

Ok and Vega-Redondo (2001) and Dekel et al. (2007) consider unobservable or partially observable preferences with uniformly random matching. They find that observability is crucial for the emergence of non-materialistic preferences such as altruism and reciprocity, while only materialistic preferences are evolutionarily stable when preferences are unobservable. Alger and Weibull (2013) consider unobservable preferences with assortative matching. That is, agents with the same preference types are matched with higher probability than those with different preference types. They find that instead of materialistic preferences, a certain preference type called *homo-moralis*, which concerns both materialistic goals and moral values, is evolutionarily stable.

In this paper, we propose an alternative model in which preferences may be (partially) observable and matching is not uniformly random. To do so, we assume that (1) agents' preferences are correlated with certain observable labels and (2) agents with the same labels are matched with a higher probability than those with different labels. The latter assumption is called *homophily*, an important sociological concept for describing the observation that people tend to interact with those who are similar in certain observable labels such as dressing codes, customs, languages, organizational affiliations, geographic locations, religions, and origins (see for example, Mcpherson et al. 2001; Ruef et al. 2003; Currarini et al. 2009, 2010).

In the model, a population consists of two preference types, $\theta$ and $\tau$, and two labels, $\theta$ and $\tau$. A proportion $\alpha \in [\frac{1}{2}, 1]$ of $\theta$ agents are correctly labeled as the $\theta$ type and $1-\alpha$ of them are incorrectly labeled as the $\tau$ type. A proportion $\beta \in [\frac{1}{2}, 1]$ of $\tau$ agents are correctly labeled as the $\tau$ type and $1-\beta$ of them are incorrectly labeled as the $\theta$ type. Both $\alpha$ and $\beta$ are exogenous variables measuring how imperfectly labels are correlated with preference types. All agents are matched in pairs and play a two-person game. The matching process exhibits homophily. That is, agents with the same labels are more likely to be matched in pairs than those with different labels. Agents' preferences may differ from the material payoffs of the game (i.e. fitness). Each pair of agents plays a Bayesian Nash equilibrium according to their preference types and beliefs about their opponents' preference types. The equilibrium outcomes determine the fitness successes of the two preference types. Correspondingly, the composition of the population evolves, as the preference type with the higher fitness success is adopted by more agents and the other one is adopted by fewer agents. We adopt the generalized version of evolutionary stability of Maynard Smith and Price (1973) from Alger and Weibull (2013) as our main solution concept and focus on studying the stability of the homogeneous population. That is, we investigate whether a preference type $\theta$ as the incumbent can resist the invasion of any arbitrarily small mutant group carrying a different preference type $\tau$ from a general set of preferences.

First, we consider the case in which both preference types are incorrectly labeled with positive probabilities ($\alpha, \beta < 1$). In this case, regardless of the degree of homophily in the matching, the probability that two mutants are matched in a pair goes to zero as the size of the mutant group goes to zero. As long as the incumbents always play a strict symmetric Nash equilibrium strategy (if one exists), they will have

a higher fitness than the mutants, implying that only the *homo-oeconomicus* type of preferences can be evolutionarily stable.

Second, suppose that both types are correctly labeled ($\alpha = \beta = 1$). In this case, observability allows the incumbents to treat themselves and the mutants differently. At the same time, mutants are protected from being too "discriminated" against by incumbents because of homophily. We find that a "Kantian-discriminating" preference type is evolutionaily stable. A "Kantian-discriminating" agent acts "morally" (in the sense of Alger and Weibull 2013) when matched with their own type agents and acts "spitefully" (playing the minimax strategy) when matched with agents with different types.[2]

Next, we consider the scenario in which the incumbents are correctly labeled ($\alpha = 1$) whereas the mutants are not ($\beta < 1$). The mutants are considered to have an informational advantage over the incumbents, which implies that they can better resist latter's "discriminating" behavior compared with the previous case of $\alpha = \beta = 1$. We find that when the incumbents play a strict and efficient symmetric Nash equilibrium strategy (if one exists) when matched with other incumbents and play spitefully when matched with the mutants, they can resist the invasion of the mutants.

Finally, suppose the incumbents are not correctly labeled ($\alpha < 1$) while the mutants are ($\beta = 1$). In this case, the incumbents are considered to have informational advantage over the mutants. We find that an interesting form of "discriminating" type arises: if incumbents with such a preference type act "cooperatively" (meaning that they coordinate on some symmetric strategy profile that yields higher payoff than some symmetric Nash equilibrium) only when they are correctly labeled and matched with agents with the same label, but act selfishly (playing the symmetric Nash equilibrium that is Pareto dominated by the "cooperative" symmetric strategy profile) otherwise, the incumbents can resist the invasion of any mutants.

The results we obtain demonstrate that the interplay between labels and homophily generates predictions that cannot be accounted for by the models in the extant literature. In particular, we find that when the mutants are correctly labeled, the incumbents are able to "discriminate" against the mutants; when the incumbents are correctly labeled, mutants can resist the incumbents' "discriminating" behavior because of homophily.

This paper is closely connected to the literature on assortative matching and label recognition. Assortative matching has been an important type of population structure for explaining other-regarding behavior in the literature on evolutionary biology dating back to Hamilton (1964a, b). Recently, Bergstrom (2003, 2013) formalizes a framework for modeling assortative matching, which is adopted by Alger and Weibull (2012, 2013, 2016, 2019) to the study of preference evolution and by Bilancini et al. (2018) to a model of cultural intolerance.[3] Bergstrom (2003, 2013) essentially treats the degree of assortativity as an exogenous parameter. A related strand of the literature considers assortativity to be a result of partner choice. See Frank (1987), McNamara et al. (2008), Izquierdo et al. (2010), Rivas (2013), Hopkins (2014) and Izquierdo

[2] Herold and Kuzmics (2009) also find stable "discriminating" types. However, since there is no homophily in their model, agents with "discriminating" types are not required to act "morally" when matched with their own type agents.

[3] See also Van Veelen (2011), Newton (2017b) and Jensen and Rigos (2018).

Millán et al. (2014), among others. Assortativiy can also be endogenously determined through political processes. See Nax and Rigos (2016) and Wu (2016, 2018). In addition, Newton (2017a) extends Alger and Weibull (2013) by subjecting the degree of assortativity under evolutionary pressure.

Label recognition is considered to be an important mechanism for promoting cooperation in prisoner's dilemma through evolution. The basic intuition is that cooperators can direct their help to other cooperators more effectively by identifying each other through labels. It is first suggested by Hamilton (1964b) and is commonly called the "green beard" effect since Dawkins (1976). Garciá et al. (2014) is the first paper that studies the interaction of assortativity and label recognition on strategy evolution in prisoner's dilemma, while ours focuses on preference evolution.

The remainder of the paper is organized as follows. Section 2 provides the model. Section 3 conducts the analysis. Section 4 concludes.

## 2 The model

Consider a continuum of agents constituting a population who are randomly matched in pairs to engage in asymmetric two-person game $\Gamma$ with the common strategy set $X$. An agent playing strategy $x \in X$ against another agent playing strategy $y \in X$ receives a material payoff (fitness), $\pi(x, y)$, where $\pi : X^2 \to \mathbb{R}$. $X$ is a nonempty, compact and convex set in a topological vector space and $\pi$ is continuous.

Each agent is characterized by his or her preference type $\theta \in \Theta$, where $\Theta$ is a general set of preference types. For the subsequent analysis, it is sufficient to consider a population with two different preference types $\theta$ and $\tau$, where $\theta, \tau \in \Theta$. A proportion $1 - \epsilon$ of the agents carry $\theta$ and the remainder carry $\tau$, where $\epsilon \in (0, 1)$. We refer to $\theta$ as the incumbent type and call $\tau$ the mutant type. Define $s = (\theta, \tau, \epsilon)$ as a population state. The set of population state is denoted by $S \in \Theta^2 \times (0, 1)$. Two labels $\theta$ and $\tau$ are available. A proportion $\alpha \in [\frac{1}{2}, 1]$ of the $\theta$ agents are correctly labeled as the $\theta$ type and $1 - \alpha$ of them are incorrectly labeled as the $\tau$ type. A proportion $\beta \in [\frac{1}{2}, 1]$ of the $\tau$ agents are correctly labeled as the $\tau$ type and $1 - \beta$ of them are incorrectly labeled as the $\theta$ type.[4] Both $\alpha$ and $\beta$ are exogenous variables measuring how perfectly labels are correlated with these preference types.

Each preference type $\theta \in \Theta$ defines a set of continuous utility functions for an agent based on his/her preference type and label as well as the matched opponent's preference type and label. In the population with two different type,s $\theta$ and $\tau$, we define

$$U\left[c_d|a_b\right](\cdot) : X^2 \to \mathbb{R}, \tag{1}$$

as the utility function of an agent with the preference type $a \in \{\theta, \tau\}$, who is labeled $b \in \{\theta, \tau\}$ and is matched with another agent with the preference type $c \in \{\theta, \tau\}$ and labeled $d \in \{\theta, \tau\}$.

[4] $\alpha = 1$ denotes the case in which the label $\theta$ is the most informative signal representing the preference type $\theta$. $\alpha = \frac{1}{2}$ denotes the case in which the label $\theta$ is the least informative signal representing the preference type $\theta$. The same explanations apply to $\beta$.

We impose no relation between $U(c_d|a_b)$ and the material payoff function $\pi$. A special example is the materialistic preference type, by which we mean that the utility function $U(c_d|a_b) = \pi$.

In the population, $\alpha(1-\epsilon)+(1-\beta)\epsilon$ agents have the $\theta$ label and $(1-\alpha)(1-\epsilon)+\beta\epsilon$ agents have the $\tau$ label. Matching may not be uniformly random. Let $\sigma \in [0, 1)$ denote the level of homophily in matching. For each agent, he or she has a probability $\sigma$ of matching with those who share the same label and probability $1-\sigma$ of randomly matching with any one in the population.[5]

Let $P[c_d|a_b, \epsilon]$ denote the probability that an agent with the preference type $a$ and label $b$ is matched with another agent with the preference type $c$ and label $d$, where $a, b, c, d \in \{\theta, \tau\}$. There are in total 16 of these probabilities given as follows:

$$
\begin{aligned}
P\Big[\theta_\theta|\theta_\theta,\epsilon\Big] &= P\Big[\theta_\theta|\tau_\theta,\epsilon\Big] = \sigma\frac{\alpha\Big(1-\epsilon\Big)}{\alpha\Big(1-\epsilon\Big)+\Big(1-\beta\Big)\epsilon}+\Big(1-\sigma\Big)\alpha\Big(1-\epsilon\Big);\\
P\Big[\tau_\theta|\theta_\theta,\epsilon\Big] &= P\Big[\tau_\theta|\tau_\theta,\epsilon\Big] = \sigma\frac{\Big(1-\beta\Big)\epsilon}{\alpha\Big(1-\epsilon\Big)+\Big(1-\beta\Big)\epsilon}+\Big(1-\sigma\Big)\Big(1-\beta\Big)\epsilon;\\
P\Big[\theta_\tau|\theta_\theta,\epsilon\Big] &= P\Big[\theta_\tau|\tau_\theta,\epsilon\Big] = \Big(1-\sigma\Big)\Big(1-\alpha\Big)\Big(1-\epsilon\Big);\\
P\Big[\tau_\tau|\theta_\theta,\epsilon\Big] &= P\Big[\tau_\tau|\tau_\theta,\epsilon\Big] = \Big(1-\sigma\Big)\beta\epsilon;\\
P\Big[\theta_\theta|\theta_\tau,\epsilon\Big] &= P\Big[\theta_\theta|\tau_\tau,\epsilon\Big] = \Big(1-\sigma\Big)\alpha\Big(1-\epsilon\Big);\\
P\Big[\tau_\theta|\theta_\tau,\epsilon\Big] &= P\Big[\tau_\theta|\tau_\tau,\epsilon\Big] = \Big(1-\sigma\Big)\Big(1-\beta\Big)\epsilon;\\
P\Big[\theta_\tau|\theta_\tau,\epsilon\Big] &= P\Big[\theta_\tau|\tau_\tau,\epsilon\Big] = \sigma\frac{\Big(1-\alpha\Big)\Big(1-\epsilon\Big)}{\Big(1-\alpha\Big)\Big(1-\epsilon\Big)+\beta\epsilon}+\Big(1-\sigma\Big)\Big(1-\alpha\Big)\Big(1-\epsilon\Big);\\
P\Big[\tau_\tau|\theta_\tau,\epsilon\Big] &= P\Big[\tau_\tau|\tau_\tau,\epsilon\Big] = \sigma\frac{\beta\epsilon}{\Big(1-\alpha\Big)\Big(1-\epsilon\Big)+\beta\epsilon}+\Big(1-\sigma\Big)\beta\epsilon. \qquad (2)
\end{aligned}
$$

Given a population state $s = (\theta, \tau, \epsilon)$, all agents play a Bayesian Nash equilibrium (assuming that agents with the same preference type and the same label choose the same strategy):

**Definition 1** A strategy profile consisting of eight strategies, $(x^*(c|a_b, \epsilon)$ with $a, b, c \in \{\theta, \tau\})$, constitutes a Bayesian Nash equilibrium if

[5] Note that the definition of homophily here shares some similarities with the definition of assortativity in Bergstrom (2003, 2013) and Alger and Weibull (2012, 2013). However, they are essentially different because the former operates on labels and the latter operates on types.

$$x^*\Big(c|a_b,\epsilon\Big) \in \arg\max_{x\in X} \frac{P\Big[\theta_c|a_b\Big]}{P\Big[\theta_c|a_b\Big]+P\Big[\tau_c|a_b\Big]}U\Big[\theta_c|a_b\Big]\Big(x,x^*\Big(b|\theta_c\Big)\Big) + \frac{P\Big[\tau_c|a_b\Big]}{P\Big[\theta_c|a_b\Big]+P\Big[\tau_c|a_b\Big]}U\Big[\tau_c|a_b\Big]\Big(x,x^*\Big(b|\tau_c,\epsilon\Big)\Big), \quad (3)$$

where $x^*(c|a_b,\epsilon)$ is interpreted as the optimal strategy chosen by an $a$-type agent with the label $b$ who is matched with an agent with the label $c$.

Given a Bayesian Nash equilibrium ($x^*(c|a_b,\epsilon)$ with $a,b,c\in\{\theta,\tau\}$), the resulting average material payoffs (average fitnesses) of the two types are given as follows:

$$\Pi_\theta\Big(\epsilon\Big) = \sum_{a,b,c\in\{\theta,\tau\}} P\Big[b_c|\theta_a,\epsilon\Big]\pi\Big(x^*\Big(c|\theta_a,\epsilon\Big),x^*\Big(a|b_c,\epsilon\Big)\Big); \quad (4)$$

$$\Pi_\tau\Big(\epsilon\Big) = \sum_{a,b,c\in\{\theta,\tau\}} P\Big[b_c|\tau_a,\epsilon\Big]\pi\Big(x^*\Big(c|\tau_a,\epsilon\Big),x^*\Big(a|b_c,\epsilon\Big)\Big). \quad (5)$$

We define evolutionary stability based on the average material payoffs as in the literature on the indirect evolutionary approach.

**Definition 2** A preference type $\theta\in\Theta$ is **evolutionarily stable against** another preference type $\tau\in\Theta$ if there exists an $\overline{\epsilon}>0$ such that $\Pi_\theta(\epsilon)>\Pi_\tau(\epsilon)$ in all Bayesian Nash equilibria ($x^*(c|a_b,\epsilon)$ with $a,b,c\in\{\theta,\tau\}$) in all states $s=(\theta,\tau,\epsilon)$ with $\epsilon\in(0,\overline{\epsilon})$. A preference type $\theta\in\Theta$ is **evolutionarily stable** if it is evolutionarily stable against all types $\tau\neq\theta$ in $\Theta$.

This definition formalizes the notion that a homogeneous population with a certain preference type would resist a small scale invasion of mutants carrying another preference type. It is a generalization of the definition of evolutionary stability by Alger and Weibull (2013) and a further generalization of the definition of evolutionary stability by Maynard Smith and Price (1973). We also introduce the notion of evolutionary unstability as follows.

**Definition 3** A preference type $\theta\in\Theta$ is **evolutionarily unstable** if there exists another preference type $\tau\in\Theta$ and an $\overline{\epsilon}>0$ such that $\Pi_\theta(\epsilon)<\Pi_\tau(\epsilon)$ in all Bayesian Nash equilibria ($x^*(c|a_b,\epsilon)$ with $a,b,c\in\{\theta,\tau\}$) in all states $s=(\theta,\tau,\epsilon)$ with $\epsilon\in(0,\overline{\epsilon})$.

## 3 Analysis

Define $B^{NE}(s)\subseteq X^8$ as the set of Bayesian Nash equilibria in population state $s=(\theta,\tau,\epsilon)$. It defines an equilibrium correspondence $B^{NE}(\theta,\tau,\cdot):(0,1)\rightrightarrows X^8$. We extend the domain of $B^{NE}(\theta,\tau,\cdot)$ to [0, 1) and standard arguments (see proof of Lemma 1 in Alger and Weibull 2013) immediately give us the following result:

**Lemma 1** *$B^{NE}(s)$ is compact for any $s = (\theta, \tau, \epsilon) \in S$. If $U[c_d|a_b](\cdot)$ is concave in its first argument, for $a, b, c, d, \in \{\theta, \tau\}$, then $B^{NE}(s) \neq \emptyset$. The equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : (0, 1) \rightrightarrows X^8$ is upper hemi-continuous.*

Since we extend the domain for the equilibrium correspondence to [0, 1), we use the following notation to denote a Bayesian Nash equilibrium as $\epsilon \to 0$: $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$.

For any $\theta \in \Theta$, let $\Theta_\theta$ be the set of types $\tau$ such that as $\epsilon \to 0$ (a $\tau$ type being the mutant type), a $\tau$-type agent is behaviorally indistinguishable from a $\theta$-type agent:

$$\Theta_\theta = \Big\{\tau \in \Theta : \exists\Big(x^*\Big(c|a_b\Big) \text{ with } a, b, c \in \Big\{\theta, \tau\Big\}\Big)\Big) \in B^{NE}\Big(\theta, \tau, 0\Big), \\ x^*\Big(c|\theta_b\Big) = x^*\Big(c|\tau_b\Big) \text{ for any } b, c \in \Big\{\theta, \tau\Big\}\Big)\Big\}. \tag{6}$$

As mentioned by Alger and Weibull (2013), an example of "behaviorally indistinguishable" types of agents is those whose utility functions are positive affine transformation of the utility function of the incumbents. In the rest of the analysis, we will exclude the consideration of types belonging to $\Theta_\theta$ when we consider the evolutionary stability of type $\theta$.

### 3.1 Case 1: $\alpha, \beta \neq 1$

We start our analysis by considering the scenario in which both $\alpha$ and $\beta$ are less than 1. We have the following result. To make the discussion concise, most of the proofs in this paper are relegated to the Appendix.

**Theorem 1** *For any $\alpha, \beta \in [\frac{1}{2}, 1)$, if there exists a symmetric strict Nash equilibrium $(x, x)$ for the fitness game $\Gamma$ and $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$, then $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$.*

Theorem 1 shows that whenever labels are not perfectly informative, incumbents who always play the same strict Nash equilibrium strategy of the fitness game $\Gamma$ (if one exists), regardless of the labels of themselves and their opponents, can resist invasion from any mutants. The intuition of Theorem 1 is straightforward: as the mutant group shrinks ($\epsilon \to 0$), because of imperfect labelling, each mutant must almost always be matched with an incumbent regardless of the degree of homophily. In addition, a mutant agent can never tell for sure the type of his/her matched opponent even if he/she is matched with another mutant. This prevents mutants behaving differently when they are matched with other mutants compared with when they are matched with incumbents. In this case, as long as the incumbents always play the same strict Nash equilibrium strategy, they always have a higher average material payoff than the mutants.

We have the following immediate corollary:

**Corollary 1** *For any $\alpha, \beta \in [\frac{1}{2}, 1)$, if the fitness game $\Gamma$ is dominance-solvable, then $\theta$ representing* ***homo-oeconomicus*** *preferences, i.e., $U[b_c|\theta_a](x, y) = \pi(x, y)$, for any*

*$a, b, c \in \{\theta, \tau\}$ and $x, y \in X$, is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$. If $\theta$ is behaviorally distinguishable from the homo-oeconomicus preference type, then there exists $\alpha, \beta \in [\frac{1}{2}, 1)$, such that $\theta$ is not evolutionarily stable.*

The intuition of Corollary 1 is as follows. When $\Gamma$ is dominance-solvable, the fitness game $\Gamma$ has a unique Nash equilibrium which is strict. Therefore, if the incumbents are homo-oeconomicus, they always play the strict Nash equilibrium strategy. According to Theorem 1, the *homo-oeconomicus* preference type is evolutionarily stable. On the contrary, if the incumbents deviate from the strict Nash equilibrium strategy, *homo-oeconomicus* as the mutant type can invade the incumbent population.

Theorem 1 and Corollary 1 highlight some important differences between our model and those in the extant literature. Let us elaborate. Two mechanisms have been found in the literature to support the stability of preference types that are behaviorally distinguishable from *homo-oeconomicus*. First, Dekel et al. (2007) demonstrate the importance of the observability of preference types.[6] In their model, when preference types are observable (even with a low probability), agents always have a positive probability of recognizing the type of their matched opponent. Hence, when two mutants are matched and they recognize each other, they can play a symmetric strategy profile that Pareto dominates the strict Nash equilibrium played by the incumbents if such a strategy profile exists.[7] Such a possibility, although rare, allows the mutants to invade the incumbent population.

Second, Alger and Weibull (2013) show that even when preference types are completely unobservable, positive assortativity according to preference types in matching makes it possible for incumbent who play a strict Nash equilibrium strategy to be destabilized. The rationale is that given positive assortativity in types, mutants have non-negligible probabilities of being matched with other mutants even when they become vanishingly rare. Hence, if the mutants commit to play a strategy that leads to a symmetric strategy profile which Pareto dominates the strict Nash equilibrium played by the incumbents, the mutants can potentially have a higher average material payoff than the incumbents. More specifically, Alger and Weibull (2013) show that only a certain preference type called *homo-moralis*, which attaches weight to both a self-interest goal and a moral goal, can be evolutionarily stable.[8]

Our model with both labels being imperfect excludes the two possibilities considered by Dekel et al. (2007) and Alger and Weibull (2013), respectively. First, since both types are incorrectly labeled, no agent can for sure tell the type of his/her matched opponent. Hence, preference types are essentially not observable (although agents have more information than when labels do not exist). This also explains why Theorem 1 shares some similarities with Dekel et al.'s (2007) result for the case with no observability. Second, we consider homophily in labels instead of assortativity in preference types; hence, when both types are incorrectly labeled, the fraction of pairs

[6] Ok and Vega-Redondo (2001) make a similar observation.

[7] As discussed in Dekel et al. (2007), their logic is reminiscent of the "secret handshake" result of Robson (1990).

[8] Note that Corollary 3 in Alger and Weibull (2013) shows that homo-oeconomicus is evolutionarily stable if agents instead engage in non-strategic activities.

**Table 1** $2 \times 2$ Anti-coordination game

| | Player 2 | |
|---|---|---|
| | $A$ | $B$ |
| Player 1 | | |
| $A$ | 0, 0 | 1, 3 |
| $B$ | 3, 1 | 0, 0 |

of matched mutants goes to zero, which contrasts to what Alger and Weibull (2013) assume.

Theorem 1 and Corollary 1 consider games with a symmetric strict Nash equilibrium. However, many games do not have such an equilibrium. In what follows, we provide an example in which Theorem 1's result does not apply.

***Example 1*** Consider the game in Table 1. Let $x \in [0, 1]$ denote a player's probability of playing the pure strategy $A$ and $X$ is the set of mixed strategies. The game has two asymmetric pure strategy Nash equilibria $(1, 0)$ and $(0, 1)$, and one symmetric mixed strategy Nash equilibrium $(\tilde{x}, \tilde{x}) = (\frac{1}{4}, \frac{1}{4})$.

Consider a $\theta$ type such that $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\tau) = \tilde{x}$, $x^*(\tau|\theta_\theta) = 0$, and $x^*(\theta|\theta_\tau) = 1$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$. That is, a $\theta$ agent tries to play the mixed strategy Nash equilibrium with his/her opponents when their labels are matched and an asymmetric pure strategy Nash equilibrium when their labels are mismatched. Given the $\theta$ agents' strategies, as $\epsilon \to 0$, a $\tau$ type agent is indifferent to any strategy when his/her own label matches his/her opponent's label and the best he/she can do when their labels are mismatched is to play $x^*(\tau|\tau_\theta) = 0$ and $x^*(\theta|\tau_\tau) = 1$ which are the best responses to $x^*(\theta|\theta_\tau)$ and $x^*(\tau|\theta_\theta)$, respectively. Hence, as long as either $x^*(\theta|\tau_\theta)$ or $x^*(\tau|\tau_\tau)$ or both do not equal $\tilde{x}$, and assuming that $x^*(\tau|\tau_\theta) = 0$ and $x^*(\theta|\tau_\tau) = 1$, then such a $\tau$ type does not belong to $\Theta_\theta$ and is considered to be the strongest mutant type against the $\theta$ type.

The average material payoffs of these two types of agents as $\epsilon \to 0$ are given by[9]

$$\lim_{\epsilon\to 0} \Pi_\theta(\epsilon) = \alpha\Big[\Big(\sigma + \big(1-\sigma\big)\alpha\Big) \times \frac{3}{4} + \big(1-\sigma\big)\big(1-\alpha\big) \times 3\Big] + \big(1-\alpha\big)\Big[\big(1-\sigma\big)\alpha \times 1 + \Big(\sigma + \big(1-\sigma\big)\big(1-\alpha\big)\Big) \times \frac{3}{4}\Big]; \quad (7)$$

$$\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) = \big(1-\beta\big)\Big[\Big(\sigma + \big(1-\sigma\big)\alpha\Big) \times \frac{3}{4} + \big(1-\sigma\big)\big(1-\alpha\big) \times 3\Big] + \beta\Big[\big(1-\sigma\big)\alpha \times 1 + \Big(\sigma + \big(1-\sigma\big)\big(1-\alpha\big)\Big) \times \frac{3}{4}\Big]. \quad (8)$$

Let $\sigma = 0.5, \alpha = 0.5$, and $\beta = 0.6$. We have $\lim_{\epsilon\to 0} \Pi_\theta(\epsilon) = \frac{17}{16}$ and $\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) = \frac{81}{80}$, respectively. Since the former is larger than the latter, from

[9] Please refer to the proof of Theorem 1 in the Appendix for the detailed expressions of the average material payoffs.

**Table 2** $3 \times 3$ Rock–Paper–Scissor game

| | Player 2 | | |
|---|---|---|---|
| | $R$ | $P$ | $S$ |
| Player 1 | | | |
| $R$ | $0, 0$ | $-1, 10$ | $1, -1$ |
| $P$ | $10, -1$ | $0, 0$ | $-1, 1$ |
| $S$ | $-1, 1$ | $1, -1$ | $0, 0$ |

the standard argument of continuity, we can conclude that $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$. The rationale is as follows. Because $\alpha > \beta$, a $\theta$ type agent is more likely to be labeled as a $\theta$ type than a $\tau$ type agent. Hence, a $\theta$ type agent has a higher probability of obtaining the highest payoff of three than a $\tau$ type agent.

In Example 1, a $\theta$ agent tries to play an asymmetric pure strategy Nash equilibrium with his/her opponents when their labels do not match. However, in games without a pure strategy Nash equilibrium, would a similar $\theta$ type be stable? We next investigate another game in which only mixed strategy Nash equilibria exist:

***Example 2*** Consider the game in Table 2. Let $x = (\alpha, \beta, \gamma) \in [0, 1] \times [0, 1] \times [0, 1]$, with $\alpha + \beta + \gamma = 1$, denoting a player's mixed strategies and $X$ being the set of mixed strategies. The game has no pure strategy Nash equilibrium and $(\tilde{x}, \tilde{x}) = \left(\left(\frac{1}{12}, \frac{1}{3}, \frac{7}{12}\right), \left(\frac{1}{12}, \frac{1}{3}, \frac{7}{12}\right)\right)$ is a symmetric mixed strategy Nash equilibrium.

Consider a $\theta$ type such that $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\tau) = \tilde{x}$, $x^*(\tau|\theta_\theta) = (0, 1, 0)$, and $x^*(\theta|\theta_\tau) = (1, 0, 0)$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$. As $\epsilon \to 0$, a $\tau$ type agent is indifferent to any strategy when his/her own label matches his/her opponent's label and the best he/she can do when their labels are mismatched is to play $x^*(\tau|\tau_\theta) = (0, 1, 0)$ and $x^*(\theta|\tau_\tau) = (0, 0, 1)$ which are the best responses to $x^*(\theta|\theta_\tau)$ and $x^*(\tau|\theta_\theta)$, respectively. Since $x^*(\theta|\tau_\tau) \neq x^*(\theta|\theta_\tau)$, $\tau$ type does not belong to $\Theta_\theta$ and is considered to be the strongest mutant type against the $\theta$ type.

The average material payoffs of the two types of agents as $\epsilon \to 0$ are given by

$$\lim_{\epsilon \to 0} \Pi_\theta(\epsilon) = \alpha\left[\left(\sigma + \left(1-\sigma\right)\alpha\right) * \frac{1}{4} + \left(1-\sigma\right)\left(1-\alpha\right) * 10\right] + \left(1-\alpha\right)\left[\left(1-\sigma\right)\alpha * \left(-1\right) + \left(\sigma + \left(1-\sigma\right)\left(1-\alpha\right)\right) * \frac{1}{4}\right]; \tag{9}$$

$$\lim_{\epsilon \to 0} \Pi_\tau(\epsilon) = \left(1-\beta\right)\left[\left(\sigma + \left(1-\sigma\right)\alpha\right) * \frac{1}{4} + \left(1-\sigma\right)\left(1-\alpha\right) * 10\right] + \beta\left[\left(1-\sigma\right)\alpha * 1 + \left(\sigma + \left(1-\sigma\right)\left(1-\alpha\right)\right) * \frac{1}{4}\right]. \tag{10}$$

Let $\sigma = 0.5, \alpha = 0.5$, and $\beta = 0.9$. We have $\lim_{\epsilon \to 0} \Pi_\theta(\epsilon) = \frac{21}{16}$ and $\lim_{\epsilon \to 0} \Pi_\tau(\epsilon) = \frac{53}{80}$, respectively. Since the former is larger than the latter, from the standard argument of continuity, we can conclude that $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$. The rationale is similar to that for Example 1. Because $\alpha > \beta$,

a $\theta$ type agent is more likely to be labeled as a $\theta$ type than a $\tau$ type agent. Hence, a $\theta$ type agent has a higher probability of obtaining the highest payoff of 10 than a $\tau$ type agent even though he/she runs the risk of receiving the lowest payoff of $-1$.

### 3.2 Case 2: $\alpha = \beta = 1$

Next, we consider the scenario in which $\alpha = \beta = 1$. In this case, both labels are perfectly correlated with the two types. Therefore, the types are observable. In addition, homophily in labels is equivalent to assortativity in types. Let $(x^e, x^e) \in \arg\max_{(x,x)\in X\times X} \pi(x, x)$ denote an ***efficient*** symmetric strategy profile as in Dekel et al. (2007). Let $x^m \in \arg\min_{x\in X} \max_{y\in X} \pi(y, x)$ be a ***minimax*** strategy. Let $\pi_{efficient} = \pi(x^e, x^e)$ denote the efficient symmetric outcome of the game $\Gamma$. Let $\pi_{minimax} = \max_{y\in X} \pi(y, x^m)$ denote the minimax value of the game $\Gamma$. We define a specific preference type as follows:

**Definition 4** $\theta \in \Theta$ is a **Kantian-discriminating** type if $x^*(\theta|\theta_\theta) = x^e$, $x^*(\tau|\theta_\theta) = x^m$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$.

We have the following result:

**Theorem 2** *When $\alpha = \beta = 1$, if $\pi_{efficient} > \pi_{minimax}$ for the fitness game $\Gamma$, a Kantian-discriminating type is evolutionarily stable for any $\sigma \in [0, 1)$.*

The intuition behind Theorem 2 is as follows. In the case with perfectly correct labels and homophily, two opposite forces affect the average material payoff of the mutants. First, given observability, as discovered by Herold and Kuzmics (2009), an agent can treat different types of opponents differently. Therefore, the incumbents can play spitefully (a minimax strategy) against mutants to minimize the mutants' average material payoffs. This is exactly the reason for the stability of the "discriminating" types in Herold and Kuzmics (2009).[10] Second, mutants are protected from being "discriminated" against by the incumbents to a certain extent because of homophily. They now have a non-negligible probability of matching with their own type of agents. Hence, when they are matched in pairs, they can play the efficient symmetric strategy profile (being "Kantian" in the sense of Alger and Weibull 2013) to increase their own group's average material payoffs. Therefore, incumbents need to play the efficient symmetric strategy profile in self-matching pairs and play the minimax strategy when matched the mutants to ensure the evolutionary stability of their types.

The necessary condition for evolutionary stability is given as follows.

**Corollary 2** *When $\alpha = \beta = 1$, if $x^*(\theta|\theta_\theta) \notin argmax_{x\in X}\pi(x, x)$, then there exists a $\sigma \in [0, 1)$, such that $\theta$ is not evolutionarily stable.*

Corollary 2 shows that non-Kantian preference types cannot be evolutionarily stable because they can be invaded by Kantian preference types when the degree of

[10] In Herold and Kuzmics (2009), agents are said to have a "discriminating" type if they play any symmetric strategy profile that yields higher material payoff than the minimax value of the game when they are matched in pairs and play a minimax strategy when they are matched with mutants.

**Table 3** $2 \times 2$ Dominant solvable game

| | Player 2 | |
|---|---|---|
| | $A$ | $B$ |
| Player 1 | | |
| $A$ | 2, 2 | 0, 4 |
| $B$ | 4, 0 | 3, 3 |

homophily is sufficiently high. This demonstrates that in the case with observability, being "discriminating" alone is not sufficient to guarantee evolutionarily stability given a positive degree of homophily.

Theorem 2 applies to games satisfying $\pi_{efficient} > \pi_{minimax}$ . Next, we consider an example in which $\pi_{efficient} = \pi_{minimax}$.

***Example 3*** Consider the game in Table 3. Let $x \in [0, 1]$ denote a player's probability of playing the pure strategy $A$ and $X$ is the set of mixed strategies. Strategy $x = 1$ is the strictly dominant strategy and $(1, 1)$ is the unique Nash equilibrium. We also have $x^e = x^m = 1$ and $\pi_{efficient} = \pi_{minimax}$. Therefore, the Kantian-discriminating type is indistinguishable from the *homo-oeconomicus* type and it is evolutionarily stable against any non-*homo-oeconomicus* type.

Note that if there exists a game in which $\pi_{efficient} < \pi_{minimax}$, then no $\theta$ type is evolutionarily stable because the mutants can obtain a higher payoff than the incumbents even when the latter play spitefully against them.

### 3.3 Case 3: $\alpha = 1, \beta \neq 1$

Third, let us consider the case in which $\alpha = 1$ and $\beta \neq 1$. In this case, the incumbents are correctly labeled whereas the mutants are not. We have the following results:

**Theorem 3** *For $\alpha = 1$ and any $\beta \in [\frac{1}{2}, 1)$. Suppose there exists a symmetric and efficient strict Nash equilibrium $(x^{e*}, x^{e*})$ for the fitness game $\Gamma$, that is, $x^{e*} = arg\max_{x \in X} \pi(x, x)$ and $x^{e*} \in arg\max_{x \in X} \pi(x, x^{e*})$ and $\pi_{efficient} > \pi_{minimax}$. If $x^*(\theta|\theta_\theta) = x^{e*}$, $x^*(\tau|\theta_\theta) = x^m$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$, then $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$.*

In the case of $\alpha = 1$ and $\beta \neq 1$, mutants are considered to have an informational advantage over incumbents, which implies that they can better resist the latter's "discriminating" behavior compared with the previous case of $\alpha = 1$ and $\beta = 1$. Hence, "Kantian-discriminating" types are no longer necessarily evolutionarily stable. Instead, stricter conditions are required for evolutionary stability: the incumbents need to be both "Kantian" and "*homo-oeconomicus*" (playing the symmetric and efficient strict Nash equilibrium strategy) when matched with agents with the $\theta$ label, and being spiteful when matched agents with the $\tau$ label to resist the invasion of any mutants as indicated in Theorem 3. The necessary condition for evolutionary stability is given as follows, which is identical to that in Corollary 2:

**Table 4** Prisoner's dilemma game

| | Player 2 | |
|---|---|---|
| | *A* | *B* |
| Player 1 | | |
| *A* | 2, 2 | 4, 0 |
| *B* | 0, 4 | 3, 3 |

**Corollary 3** *When* $\alpha = 1$ *and* $\beta \in [\frac{1}{2}, 1)$*, if* $x^*(\theta|\theta_\theta) \notin argmax_{x\in X}\pi(x, x)$*, then there exists a* $\sigma \in [0, 1)$*, such that* $\theta$ *is not evolutionarily stable.*

Corollary 2 together with Corollay 3 demonstrates the importance of efficiencywhen the incumbents are correctly labeled.

Theorem 3 considers games with a symmetric and efficient strict Nash equilibrium. Next, we consider an example without such an equilibrium.

***Example 4*** Consider the game in Table 4. Let $x \in [0, 1]$ denote a player's probability of playing the pure strategy $A$ and $X$ is the set of mixed strategies. Strategy $x = 0$ is the strictly dominant strategy and $(0, 0)$ is the unique Nash equilibrium. We also have $x^e = 1$ and $x^m = 0$ and $\pi_{efficient} = 3 > \pi_{minimax} = 2$.

First, consider a $\theta$ type such that $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\theta) = 0$ (the $\theta$ type can be the *homo-oeconomicus* type) and a $\tau$ type such that $x^*(\theta|\tau_\theta) = x^*(\theta|\tau_\tau) = 0$ and $x^*(\tau|\tau_\tau) = 1$. The average material payoffs of the two types of agents as $\epsilon \to 0$ are given by[11]

$$\lim_{\epsilon\to 0} \Pi_\theta(\epsilon) = 2, \tag{11}$$

$$\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) = (1-\beta) * 2 + \beta\Big[(1-\sigma) * 2 + \sigma * 3\Big]. \tag{12}$$

We have $\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) > 2$ as long as $\sigma > 0$. Hence, $\theta$ is not evolutionarily stable against $\tau$.

Next, consider a $\theta$ type such that $x^*(\theta|\theta_\theta) = 1$ and $x^*(\tau|\theta_\theta) = 0$ (the $\theta$ type is the Kantian-discriminating type) and a $\tau$ type such that $x^*(\theta|\tau_\theta) = x^*(\theta|\tau_\tau) = 0$ and $x^*(\tau|\tau_\tau) = 1$. The average material payoffs of the two types of agents as $\epsilon \to 0$ are given by

$$\lim_{\epsilon\to 0} \Pi_\theta(\epsilon) = 3, \tag{13}$$

$$\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) = (1-\beta) \times 4 + \beta\Big[(1-\sigma) \times 2 + \sigma \times 3\Big]. \tag{14}$$

We have $\lim_{\epsilon\to 0} \Pi_\tau(\epsilon) > 3$ as long as $\sigma > \frac{2\beta-1}{\beta}$. Hence, $\theta$ is not evolutionarily stable against $\tau$. The above analyzed two scenarios demonstrate that in the game

[11] Please refer to the proof of Theorem 3 in the Appendix for the detailed expressions of the average material payoffs.

depicted in Table 4, playing either (a) the strict but not efficient Nash equilibrium or (b) the efficient strategy profile that is not a Nash equilibrium, cannot be evolutionarily stable.

### 3.4 Case 4: $\alpha \neq 1, \beta = 1$

Finally, consider the situation in which $\alpha \neq 1$ and $\beta = 1$. In this case, the mutants are correctly labeled whereas the incumbents are not.

We first have the following result:

**Theorem 4** *For any $\alpha \in [\frac{1}{2}, 1)$ and $\beta = 1$, let $(x, x)$ be a symmetric Nash equilibrium for the fitness game $\Gamma$, as long as $\pi(x^*(\theta|\theta_\theta), x^*(\theta|\theta_\theta)) > \pi(x, x)$ and $x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$, then $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$.*

In the case of $\alpha \neq 1$ and $\beta = 1$, incumbents are considered to have an informational advantage over mutants. Theorem 4 shows that in such a case, an interesting form of "discriminating" type arises: if the incumbents act "cooperatively" (playing a symmetric strategy profile, if one exists, that Pareto dominates a symmetric Nash equilibrium) only when they are correctly labeled and matched with agents with the same label, but act selfishly (playing the symmetric Nash equilibrium that is Pareto dominated by the symmetric strategy profile) when they are incorrectly labeled or matched with agents with the mutant's label, the incumbents can resist the invasion of any mutants. Compared with the case of $\alpha \neq 1$ and $\beta \neq 1$, incumbents with the correct labels no longer worry about interacting with mutants who are masked as incumbents. Therefore, they are free from playing a Nash equilibrium when matched in pairs but reach a cooperative outcome with a higher fitness.

Second, in games with a symmetric and strict Nash equilibrium $(x, x)$, always playing $x$ is sufficient for a $\theta$ type to be evolutionarily stable as well. we have the following result:

**Theorem 5** *For any $\alpha \in [\frac{1}{2}, 1)$ and $\beta = 1$, let $(x, x)$ be a symmetric and strict Nash equilibrium for the fitness game $\Gamma$, as long as $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ for all Bayesian Nash equilibria in $B^{NE}(\theta, \tau, 0)$, then $\theta$ is evolutionarily stable against any $\tau \in \Theta \backslash \Theta_\theta$.*

Theorems 4 and 5 demonstrate an important difference between games with and without a symmetric and strict Nash equilibrium. When there is no such equilibrium, the $\theta$ agents receive an equal payoff to the $\tau$ agents if they always play a mixed strategy Nash equilibrium strategy, unless they can reach a cooperative outcome with higher fitness when they are correctly labeled and matched with each other. On the contrary, when there exists a symmetric and strict Nash equilibrium, always playing it can already guarantee that the $\theta$ agents have a higher payoff than the $\tau$ agents.

## 4 Discussion and conclusion

This paper proposes a model of preference evolution in which preferences are correlated with certain labels and the matching process exhibits homophily in labels. The analysis provides novel results compared with the extant literature. In sum, we find that when the mutants are correctly labeled (cases 2 and 4), the incumbents are able to "discriminate" against the mutants. When the incumbents are correctly labeled (cases 2 and 3), the mutants can resist the incumbents' "discriminating" behavior because of homophily, which drives the evolution to select preferences that incorporate the consideration of efficiency. Our results may thus offer new perspectives on understanding various cultural phenomena such as why culturally distinctive minorities are more likely subject to discrimination and how within-community social connections can help them resist assimilation.

The indirect evolutionary approach we adopt in this paper implicitly assumes that behavior adjusts arbitrarily faster than preferences evolve. Although it is reasonable to consider that the evolution of preferences proceeds slower than agents learn to reach an equilibrium, letting the relative rates to infinity seems extreme. The model of two-speed dynamics proposed by Sandholm (2001) provides a useful conceptual framework and techniques for simultaneously studying the dynamics of preference evolution and how agents learn to behave as their preferences dictate (see also Kuran and Sandholm 2008).

Preference evolution can be shaped by either natural selection or cultural selection. In the latter case, preferences are transmitted from one generation to the next. In each generation, agents first interact with one another according to their preferences and reach an equilibrium. Then, they become parents and exert efforts to transmit their own preferences to their children. If we assume that preferences that have led to economic success for the parents are more likely to be passed down to their children, then we have a model similar to the indirect evolutionary approach: preferences evolve slowly across generations while agents within each generation play an equilibrium. See Bisin and Verdier (2011) for an extensive survey of the literature on cultural transmission.

In the current model, labels are given exogenously. It would be an important research avenue to consider labels as part of the definition of types. Here we provide a simple example with endogenous label adoption for illustration purposes. Suppose there are two neutral labels $A$ and $B$ that both types of agents can adopt. Adopting label $A$ is cheaper than label $B$. Without loss of generality, assume that the cost of adopting label $A$ is 0 and the cost of adopting label $B$ is $c > 0$. First, consider a $\theta$ type that adopts label $A$ against a $\tau$ type that also adopts label $A$. Since both types have identical labels, they are uniformly randomly matched in pairs. In this scenario, if there exists a symmetric and strict Nash equilibrium, then playing it would guarantee that the $\theta$ type is evolutionarily stable against the $\tau$ type. Second, consider a $\theta$ type that adopts label $A$ against a $\tau$ type that adopts label $B$. This scenario is similar to case 2 ($\alpha = 1$, $\beta = 1$) in the present paper because the two types can be perfectly differentiated. When $c$ is sufficiently large, the $\theta$ type is automatically evolutionarily stable against the $\tau$ type. Otherwise, Corollary 2 applies, which requires the $\theta$ agents to play efficiently. Third, consider a $\theta$ type that adopts label $B$ against a $\tau$ type that also adopts label $B$. This scenario is identical to the first one that requires the $\theta$ type to play a strict Nash

equilibrium. Finally, consider a $\theta$ type that adopts label $B$ against a $\tau$ type that adopts label $A$. In this scenario, since the $\theta$ type agents need to pay the extra cost of $c$, even a Kantian-discriminating type may not be stable.

There are many other potential ways to endogenize labels. For example, agents may have different abilities to mimic others' labels and such abilities are subject to evolutionary selection as well. We will leave these topics to interested readers.

## Appendix

***Proof of Theorem 1*** When $\alpha, \beta \in [\frac{1}{2}, 1)$, a strategy profile $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, 0)$ if

$$\begin{aligned}
&x^*\Big(\theta|a_\theta\Big) \in \arg\max_{x\in X} U\Big[\theta_\theta|a_\theta\Big]\Big(x, x^*\Big(\theta|\theta_\theta\Big)\Big);\\
&x^*\Big(\tau|a_\theta\Big) \in \arg\max_{x\in X} U\Big[\theta_\tau|a_\tau\Big]\Big(x, x^*\Big(\theta|\theta_\tau\Big)\Big);\\
&x^*\Big(\theta|a_\tau\Big) \in \arg\max_{x\in X} U\Big[\theta_\theta|a_\tau\Big]\Big(x, x^*\Big(\tau|\theta_\theta\Big)\Big);\\
&x^*\Big(\tau|a_\tau\Big) \in \arg\max_{x\in X} U\Big[\theta_\tau|a_\tau\Big]\Big(x, x^*\Big(\tau|\theta_\tau\Big)\Big), \text{ for } a \in \{\theta, \tau\}. \qquad (15)
\end{aligned}$$

The average material payoffs corresponding to the two preference types as $\epsilon \to 0$, are given as:

$$\begin{aligned}
\lim_{\epsilon\to 0}\Pi_\theta\Big(\epsilon\Big) = \alpha\Big[&\Big(\sigma + \Big(1-\sigma\Big)\alpha\Big)\pi\Big(x^*\Big(\theta|\theta_\theta\Big), x^*\Big(\theta|\theta_\theta\Big)\Big)\\
&+\Big(1-\sigma\Big)\Big(1-\alpha\Big)\pi\Big(x^*\Big(\tau|\theta_\theta\Big), x^*\Big(\theta|\theta_\tau\Big)\Big)\Big]\\
&+\Big(1-\alpha\Big)\Big[\Big(1-\sigma\Big)\alpha\pi\Big(x^*\Big(\theta|\theta_\tau\Big), x^*\Big(\tau|\theta_\theta\Big)\Big)\\
&+\Big(\sigma + \Big(1-\sigma\Big)\Big(1-\alpha\Big)\Big)\pi\Big(x^*\Big(\tau|\theta_\tau\Big), x^*\Big(\tau|\theta_\tau\Big)\Big)\Big]; \qquad (16)\\
\lim_{\epsilon\to 0}\Pi_\tau\Big(\epsilon\Big) = \Big(1-\beta\Big)\Big[&\Big(\sigma + \Big(1-\sigma\Big)\alpha\Big)\pi\Big(x^*\Big(\theta|\tau_\theta\Big), x^*\Big(\theta|\theta_\theta\Big)\Big)\\
&+\Big(1-\sigma\Big)\Big(1-\alpha\Big)\pi\Big(x^*\Big(\tau|\tau_\theta\Big), x^*\Big(\theta|\theta_\tau\Big)\Big)\Big]\\
&+\beta\Big[\Big(1-\sigma\Big)\alpha\pi\Big(x^*\Big(\theta|\tau_\tau\Big), x^*\Big(\tau|\theta_\theta\Big)\Big)\\
&+\Big(\sigma + \Big(1-\sigma\Big)\Big(1-\alpha\Big)\Big)\pi\Big(x^*\Big(\tau|\tau_\tau\Big), x^*\Big(\tau|\theta_\tau\Big)\Big)\Big]. \qquad (17)
\end{aligned}$$

If there exists a symmetric strict Nash equilibrium $(x, x)$ for the fitness game $\Gamma$ and the incumbents play $x$, i.e., $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ in any Bayesian Nash equilibrium in $B^{NE}(\theta, \tau, 0)$, then as long as at least one of the mutants' four equilibrium strategies are not $x$, i.e., the mutants are behaviorally distinguishable from the incumbents ($\tau \notin \Theta_\theta$), we have

$$\lim_{\epsilon \to 0} \Pi_\theta(\epsilon) > \lim_{\epsilon \to 0} \Pi_\tau(\epsilon). \tag{18}$$

By continuity of $\Pi_\theta$ and $\Pi_\tau$, the strict inequality in average material payoffs holds in a neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$ and 0. Given that the equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : [0, 1) \rightrightarrows X^8$ is closed-valued and upper hemi-continuous according to Lemma 1, if for all $t \in \mathbb{N}$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, \epsilon_t)$ converges to some limit point as $\epsilon_t$ converges to zero, the limit point must belong to $B^{NE}(\theta, \tau, 0)$. This implies that there exists a $T$, such that for all $t > T$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\})$ and $\epsilon_t$ are in the neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$ and 0, implying that $\Pi_\theta(\epsilon_t) > \Pi_\tau(\epsilon_t)$. □

***Proof of Corollary 1*** When $\Gamma$ is dominance-solvable, it has a unique Nash equilibrium, which is strict and symmetric. Hence, as long as the incumbent type $\theta$ is *homo-oeconomicus*, i.e., $U[b_c|\theta_a](x, y) = \pi(x, y)$, for any $a, b, c \in \{\theta, \tau\}$ and $x, y \in X$, according to the Bayesian Nash equilibrium defined in (15), the incumbents always play the symmetric strict Nash equilibrium strategy. Hence, according to Theorem 1, $\theta$ is evolutionarily stable.

Suppose instead $\theta$ is behaviorally distinguishable from *homo-oeconomicus*. Let the mutant type $\tau$ be *homo-oeconomicus*, i.e., $U[b_c|\tau_a](x, y) = \pi(x, y)$, for any $a, b, c \in \{\theta, \tau\}$ and $x, y \in X$. Then the mutants will play the strict Nash equilibrium strategy, which also strictly dominates all other strategies. In this case, according to (16) and (17), as long as $\beta = 1 - \alpha$, we have $\lim_{\epsilon \to 0} \Pi_\theta(\epsilon) < \lim_{\epsilon \to 0} \Pi_\tau(\epsilon)$. Following the same argument as in Theorem 1, one can show that there exists a $\overline{\epsilon}$, such that for $\epsilon < \overline{\epsilon}$, $\Pi_\theta(\epsilon_t) < \Pi_\tau(\epsilon_t)$, implying the instability of $\theta$ type. □

***Proof of Theorem 2*** When $\alpha = \beta = 1$, a $\theta$-type agent with label $\tau$ or a $\tau$-type agent with label $\theta$ does not exist. Therefore, a Bayesian Nash equilibrium strategy profile consists of only 4 strategies instead of 8. We also redefine the equilibrium correspondence $B^{NE}(\theta, \tau, \cdot)$ as a correspondence from (0,1) to $X^4$ instead of $X^8$. Lemma 1 can readily be applied to this new definition.

Now, a strategy profile $(x^*(b|a_a)$ with $a, b \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, 0)$ if

$$\begin{aligned}
x^*\Big(\theta|\theta_\theta\Big) &\in \arg\max_{x \in X} U\Big[\theta_\theta|\theta_\theta\Big]\Big(x, x^*\Big(\theta|\theta_\theta\Big)\Big);\\
x^*\Big(\tau|\theta_\theta\Big) &\in \arg\max_{x \in X} U\Big[\tau_\tau|\theta_\theta\Big]\Big(x, x^*\Big(\theta|\tau_\tau\Big)\Big);\\
x^*\Big(\theta|\tau_\tau\Big) &\in \arg\max_{x \in X} U\Big[\theta_\theta|\tau_\tau\Big]\Big(x, x^*\Big(\tau|\theta_\theta\Big)\Big);\\
x^*\Big(\tau|\tau_\tau\Big) &\in \arg\max_{x \in X} U\Big[\tau_\tau|\tau_\tau\Big]\Big(x, x^*\Big(\tau|\tau_\tau\Big)\Big).
\end{aligned} \tag{19}$$

The average material payoffs corresponding to the two preference types as $\epsilon \to 0$, are given as:

$$\lim_{\epsilon\to 0}\Pi_\theta\left(\epsilon\right)=\pi\left(x^*\left(\theta|\theta_\theta\right),x^*\left(\theta|\theta_\theta\right)\right), \tag{20}$$

$$\lim_{\epsilon\to 0}\Pi_\tau\left(\epsilon\right)=\left(1-\sigma\right)\pi\left(x^*\left(\theta|\tau_\tau\right),x^*\left(\tau|\theta_\theta\right)\right)+\sigma\pi\left(x^*\left(\tau|\tau_\tau\right),x^*\left(\tau|\tau_\tau\right)\right). \tag{21}$$

Given that $\theta$ is Kantian-discriminating, i.e., $x^*(\theta|\theta_\theta)=x^e$, $x^*(\tau|\theta_\theta)=x^m$ in any Bayesian Nash equilibrium in $B^{NE}(\theta,\tau,0)$, as long as $\pi_{efficient}>\pi_{minimax}$, we have

$$\lim_{\epsilon\to 0}\Pi_\theta(\epsilon)=\pi_{efficient}>(1-\sigma)\pi_{minimax}+\sigma\pi_{efficient}\geq\lim_{\epsilon\to 0}\Pi_\tau(\epsilon),\ \text{for any}\ \sigma\in[0,1). \tag{22}$$

By continuity of $\Pi_\theta$ and $\Pi_\tau$, the strict inequality in average material payoffs holds in a neighborhood of $(x^*(b|a_a)$ with $a,b\in\{\theta,\tau\})$ and 0. Given that the equilibrium correspondence $B^{NE}(\theta,\tau,\cdot):[0,1)\rightrightarrows X^4$ is closed-valued and upper hemi-continuous according to a similar argument as in Lemma 1, if for all $t\in\mathbb{N}$, $(x^*(b|a_a,\epsilon_t)$ with $a,b\in\{\theta,\tau\})\in B^{NE}(\theta,\tau,\epsilon_t)$ converges to some limit point as $\epsilon_t$ converges to zero, the limit point must belong to $B^{NE}(\theta,\tau,0)$. This implies that there exists a $T$, such that for all $t>T$, $(x^*(b|a_a,\epsilon_t)$ with $a,b\in\{\theta,\tau\})$ and $\epsilon_t$ are in the neighborhood of $(x^*(b|a_a)$ with $a,b\in\{\theta,\tau\})$ and 0, implying that $\Pi_\theta(\epsilon_t)>\Pi_\tau(\epsilon_t)$. □

***Proof of Corollary 2*** Suppose $x^*(\theta|\theta_\theta)\notin\arg\max_{x\in X}\pi(x,x)$, let $x^*(\tau|\tau_\tau)\in\arg\max_{x,\in X}\pi(x,x)$. In this case, according to (20) and (21), as long as $\sigma$ is sufficiently close to one, we have $\lim_{\epsilon\to 0}\Pi_\theta(\epsilon)<\lim_{\epsilon\to 0}\Pi_\tau(\epsilon)$. Following the same argument as in Theorem 2, one can show that there exists a $\overline{\epsilon}$, such that for $\epsilon<\overline{\epsilon}$, $\Pi_\theta(\epsilon_t)<\Pi_\tau(\epsilon_t)$, implying the instability of $\theta$ type. □

***Proof of Theorem 3*** When $\alpha=1$ and $\beta\neq 1$, a $\theta$-type agent with label $\tau$ does not exist. Therefore, a Bayesian Nash equilibrium strategy profile consists of only six strategies instead of eight. We also redefine the equilibrium correspondence $B^{NE}(\theta,\tau,\cdot)$ as a correspondence from (0,1) to $X^6$ instead of $X^8$. Lemma 1 can readily be applied to this new definition.

Now, a strategy profile $(x^*(c|a_b)$ with $a,b,c\in\{\theta,\tau\})\in B^{NE}(\theta,\tau,0)$ if

$$\begin{aligned}
&x^*\left(\theta|\theta_\theta\right)\in\arg\max_{x\in X}U\left[\theta_\theta|\theta_\theta\right]\left(x,x^*\left(\theta|\theta_\theta\right)\right);\\
&x^*\left(\tau|\theta_\theta\right)\in\arg\max_{x\in X}U\left[\tau_\tau|\theta_\theta\right]\left(x,x^*\left(\theta|\tau_\tau\right)\right);\\
&x^*\left(\theta|\tau_\theta\right)\in\arg\max_{x\in X}U\left[\theta_\theta|\tau_\theta\right]\left(x,x^*\left(\theta|\theta_\theta\right)\right);\\
&x^*\left(\tau|\tau_\theta\right)\in\arg\max_{x\in X}U\left[\tau_\tau|\tau_\theta\right]\left(x,x^*\left(\theta|\tau_\tau\right)\right);\\
&x^*\left(\theta|\tau_\tau\right)\in\arg\max_{x\in X}U\left[\theta_\theta|\tau_\tau\right]\left(x,x^*\left(\tau|\theta_\theta\right)\right);\\
&x^*\left(\tau|\tau_\tau\right)\in\arg\max_{x\in X}U\left[\tau_\tau|\tau_\tau\right]\left(x,x^*\left(\tau|\tau_\tau\right)\right).
\end{aligned} \tag{23}$$

The average material payoffs corresponding to the two preference types as $\epsilon \to 0$, are given as:

$$\lim_{\epsilon\to 0}\Pi_\theta\Big(\epsilon\Big) = \pi\Big(x^*\Big(\theta|\theta_\theta\Big), x^*\Big(\theta|\theta_\theta\Big)\Big), \tag{24}$$

$$\begin{aligned}\lim_{\epsilon\to 0}\Pi_\tau\Big(\epsilon\Big) = &\Big(1-\beta\Big)\pi\Big(x^*\Big(\theta|\tau_\theta\Big), x^*\Big(\theta|\theta_\theta\Big)\Big)\\ &+\beta\Big[\Big(1-\sigma\Big)\pi\Big(x^*\Big(\theta|\tau_\tau\Big), x^*\Big(\tau|\theta_\theta\Big)\Big)\\ &+\sigma\pi\Big(x^*\Big(\tau|\tau_\tau\Big), x^*\Big(\tau|\tau_\tau\Big)\Big)\Big].\end{aligned} \tag{25}$$

Given that $x^*(\theta|\theta_\theta) = x^{e*}, x^*(\tau|\theta_\theta) = x^m$ in any Bayesian Nash equilibrium in $B^{NE}(\theta, \tau, 0)$, as long as $\pi_{efficient} > \pi_{minimax}$, we have

$$\begin{aligned}\lim_{\epsilon\to 0}\Pi_\theta(\epsilon) = \pi_{efficient} > (1-\beta)\pi_{efficient} + \beta\Big[\Big(1-\sigma\Big)\pi_{minimax}\\ +\sigma\pi_{efficient}\Big] > \lim_{\epsilon\to 0}\Pi_\tau(\epsilon),\\ \text{for any } \sigma \in [0,1).\end{aligned} \tag{26}$$

By continuity of $\Pi_\theta$ and $\Pi_\tau$, the strict inequality in average material payoffs holds in a neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$ and 0. Given that the equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : [0, 1) \rightrightarrows X^6$ is closed-valued and upper hemi-continuous according to a similar argument as in Lemma 1, if for all $t \in \mathbb{N}$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, \epsilon_t)$ converges to some limit point as $\epsilon_t$ converges to zero, the limit point must belong to $B^{NE}(\theta, \tau, 0)$. This implies that there exists a $T$, such that for all $t > T$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\})$ and $\epsilon_t$ are in the neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$ and 0, implying that $\Pi_\theta(\epsilon_t) > \Pi_\tau(\epsilon_t)$. □

***Proof of Corollary 3*** Suppose $x^*(\theta|\theta_\theta) \notin \arg\max_{x\in X}\pi(x, x)$. Let $x^*(\theta|\tau_\theta) \in \arg\max_{x\in X}\pi(x, x^*(\theta|\theta_\theta))$ and $x^*(\tau|\tau_\tau) \in \arg\max_{x,\in X}\pi(x, x)$. In this case, according to (24) and (25), as long as $\sigma$ is sufficiently close to 1, we have $\lim_{\epsilon\to 0}\Pi_\theta(\epsilon) < \lim_{\epsilon\to 0}\Pi_\tau(\epsilon)$. Following the same argument as in Theorem 3, one can show that there exists a $\bar{\epsilon}$, such that for $\epsilon < \bar{\epsilon}$, $\Pi_\theta(\epsilon_t) < \Pi_\tau(\epsilon_t)$, implying the instability of $\theta$ type. □

***Proof of Theorem 4*** When $\alpha \neq 1$ and $\beta = 1$, a $\tau$-type agent with label $\theta$ does not exist. Therefore, a Bayesian Nash equilibrium strategy profile consists of only 6 strategies instead of 8. We also redefine the equilibrium correspondence $B^{NE}(\theta, \tau, \cdot)$ as a correspondence from (0,1) to $X^6$ instead of $X^8$. Lemma 1 can readily be applied to this new definition.

Now, a strategy profile $(x^*(b|a_c)$ with $a, b, c \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, 0)$ if

$$x^*(\theta|\theta_\theta) \in \arg\max_{x\in X} U[\theta_\theta|\theta_\theta](x, x^*(\theta|\theta_\theta));$$

$$x^*(\tau|\theta_\theta) \in \arg\max_{x\in X} U[\theta_\tau|\theta_\theta](x, x^*(\theta|\theta_\tau));$$

$$
\begin{aligned}
x^*(\theta|\theta_\tau) &\in \arg\max_{x\in X} U[\theta_\theta|\theta_\tau](x, x^*(\tau|\theta_\theta));\\
x^*(\tau|\theta_\tau) &\in \arg\max_{x\in X} U[\theta_\tau|\theta_\tau](x, x^*(\tau|\theta_\tau));\\
x^*(\theta|\tau_\tau) &\in \arg\max_{x\in X} U[\theta_\theta|\tau_\tau](x, x^*(\tau|\theta_\theta));\\
x^*(\tau|\tau_\tau) &\in \arg\max_{x\in X} U[\theta_\tau|\tau_\tau](x, x^*(\tau|\theta_\tau)).
\end{aligned} \tag{27}
$$

The average material payoffs corresponding to the two preference types as $\epsilon \to 0$, are given as:

$$
\begin{aligned}
\lim_{\epsilon\to 0}\Pi_\theta(\epsilon) = \alpha\Big[&(\sigma+(1-\sigma)\alpha)\pi(x^*(\theta|\theta_\theta), x^*(\theta|\theta_\theta))\\
&+(1-\sigma)(1-\alpha)\pi(x^*(\tau|\theta_\theta), x^*(\theta|\theta_\tau))\Big]\\
&+(1-\alpha)\Big[(1-\sigma)\alpha\pi(x^*(\theta|\theta_\tau), x^*(\tau|\theta_\theta))\\
&+(\sigma+(1-\sigma)(1-\alpha))\pi(x^*(\tau|\theta_\tau), x^*(\tau|\theta_\tau))\Big];
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
\lim_{\epsilon\to 0}\Pi_\tau(\epsilon) = &(1-\sigma)\alpha\pi(x^*(\theta|\tau_\tau), x^*(\tau|\theta_\theta))\\
&+(\sigma+(1-\sigma)(1-\alpha))\pi(x^*(\tau|\tau_\tau), x^*(\tau|\theta_\tau)).
\end{aligned} \tag{29}
$$

Let $(x, x)$ be a symmetric Nash equilibrium for the fitness game $\Gamma$. If $\pi(x^*(\theta|\theta,\theta), x^*(\theta|\theta,\theta)) > \pi(x,x)$ and $x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ in any Bayesian Nash equilibrium in $B^{NE}(\theta,\tau,0)$, then as long as at least one of the mutants' two equilibrium strategies are not $x$, i.e., the mutants are behaviorally distinguishable from the incumbents ($\tau \notin \Theta_\theta$), we have

$$
\lim_{\epsilon\to 0}\Pi_\theta(\epsilon) > \lim_{\epsilon\to 0}\Pi_\tau(\epsilon). \tag{30}
$$

By continuity of $\Pi_\theta$ and $\Pi_\tau$, the strict inequality in average material payoffs holds in a neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta,\tau\})$ and 0. Given that the equilibrium correspondence $B^{NE}(\theta,\tau,\cdot) : [0,1) \rightrightarrows X^6$ is closed-valued and upper hemi-continuous according to a similar argument as in Lemma 1, if for all $t \in \mathbb{N}$, $(x^*(c|a_b,\epsilon_t)$ with $a, b, c \in \{\theta,\tau\}) \in B^{NE}(\theta,\tau,\epsilon_t)$ converges to some limit point as $\epsilon_t$ converges to zero, the limit point must belong to $B^{NE}(\theta,\tau,0)$. This implies that there exists a $T$, such that for all $t > T$, $(x^*(c|a_b,\epsilon_t)$ with $a, b, c \in \{\theta,\tau\})$ and $\epsilon_t$ are in the neighborhood of $(x^*(b|a_a)$ with $a, b \in \{\theta,\tau\})$ and 0, implying that $\Pi_\theta(\epsilon_t) > \Pi_\tau(\epsilon_t)$. □

***Proof of Theorem 5*** Let $(x, x)$ be a symmetric and strict Nash equilibrium for the fitness game $\Gamma$. If $x^*(\theta|\theta_\theta) = x^*(\tau|\theta_\theta) = x^*(\theta|\theta_\tau) = x^*(\tau|\theta_\tau) = x$ in any Bayesian Nash equilibrium in $B^{NE}(\theta,\tau,0)$, then as long as either $x^*(\theta|\tau_\tau) \neq x$ or $x^*(\tau|\tau_\tau) \neq x$ or both, then $\tau \notin \Theta_\theta$, and according to (28) and (29), we have

$$
\lim_{\epsilon\to 0}\Pi_\theta(\epsilon) > \lim_{\epsilon\to 0}\Pi_\tau(\epsilon). \tag{31}
$$

By continuity of $\Pi_\theta$ and $\Pi_\tau$, the strict inequality in average material payoffs holds in a neighborhood of $(x^*(c|a_b)$ with $a, b, c \in \{\theta, \tau\})$ and 0. Given that the equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : [0, 1) \rightrightarrows X^6$ is closed-valued and upper hemi-continuous according to a similar argument as in Lemma 1, if for all $t \in \mathbb{N}$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\}) \in B^{NE}(\theta, \tau, \epsilon_t)$ converges to some limit point as $\epsilon_t$ converges to zero, the limit point must belong to $B^{NE}(\theta, \tau, 0)$. This implies that there exists a $T$, such that for all $t > T$, $(x^*(c|a_b, \epsilon_t)$ with $a, b, c \in \{\theta, \tau\})$ and $\epsilon_t$ are in the neighborhood of $(x^*(b|a_a)$ with $a, b \in \{\theta, \tau\})$ and 0, implying that $\Pi_\theta(\epsilon_t) > \Pi_\tau(\epsilon_t)$. □

## References


Akçay E, Van Cleve J, Feldman M, Roughgarden J (2009) A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc Natl Acad Sci 106:19061–19066

Alger I (2010) Public goods games, altruism, and evolution. J Public Econ Theory 12(4):789–813

Alger I, Weibull JW (2010) Kinship, incentives, and evolution. Am Econ Rev 100:1725–1758

Alger I, Weibull JW (2012) A generalization of Hamilton's rule—love the sibling how much? J Theor Biol 299:42–54

Alger I, Weibull JW (2013) Homo moralis—preference evolution under incomplete information and assortative matching. Econometrica 81(6):2269–2302

Alger I, Weibull JW (2016) Evolution and kantian morality. Games Econ Behav 98:56–67

Alger I, Weibull JW (2019) Evolutionary models of preference formation. Annual Review of Economics, forthcoming

Bergstrom TC (2003) The algebra of assortative encounters and the evolution of cooperation. Int Game Theory Rev 5(3):211–228

Bergstrom TC (2013) Measures of assortativity. Biol Theory 8:133–141

Bester H, Güth W (1998) Is altruism evolutionarily stable? J Econ Behav Organ 34:193–209

Bilancini E, Bocinelli L, Wu J (2018) The interplay of cultural intolerance and action-assortativity for the emergence of cooperation and homophily. Eur Econ Rev 102:1–18

Bisin A, Verdier T (2011) The economics of cultural transmission and socialization. In: Benhabib J, Bisin A, Jackson M (eds) Handbook of social economics, vol 1. Elsevier, pp 339–416

Currarini S, Jackson MO, Pin P (2009) An economic model of friendship: homophily, minorities, and segregation. Econometrica 77(4):1003–1045

Currarini S, Jackson MO, Pin P (2010) Identifying the roles of race-based choice and chance in high school friendship network formation. Proc Natl Acad Sci 107(11):4857–4861

Dawkins R (1976) The selfish gene. Oxford University Press, New York

Dekel E, Ely JC, Yilankaya O (2007) Evolution of preferences. Rev Econ Stud 74:685–704

Ely JC, Yilankaya O (2001) Nash equilibrium and the evolution of preferences. J Econ Theory 97:255–272

Fershtman C, Weiss Y (1998) Social rewards, externalities and stable preferences. J Public Econ 70:53–73

Frank RH (1987) If homo economicus could choose his own utility function, would he want one with a conscience? Am Econ Rev 77:593–604

Garciá J, van Veelen M, Traulsen A (2014) Evil green beard: tag recognition can also be used to withhold cooperation in structured populations. J Theor Biol 360:181–186

Güth W (1995) An evolutionary approach to explaining cooperative behavior by reciprocal incentives. Int J Game Theory 24:323–344

Güth W, Yaari M (1992) An evolutionary approach to explain reciprocal behavior in a simple strategic game. In: Witt U (ed) Explaining process and change—approaches to evolutionary economics. University of Michigan Press, Ann Arbor

Hamilton WD (1964a) The genetical evolution of social behaviour. I. J Theor Biol 7:1–16

Hamilton WD (1964b) The genetical evolution of social behaviour. II. J Theor Biol 7:17–52

Heifetz A, Shannon C, Spiegel Y (2007a) The dynamic evolution of preferences. Econo Theory 32:251–286

Heifetz A, Shannon C, Spiegel Y (2007b) What to maximize if you must. J Econ Theory 133:31–57

Herold F, Kuzmics C (2009) Evolutionary stability of discrimination under observability. Games Econ Behav 67:542–551
Hopkins E (2014) Competitive altruism, mentalizing and signaling. Am Econ J Microecon 6:272–292
Huck S, Oechssler J (1999) The indirect evolutionary approach to explaining fair allocations. Games Econ Behav 28:13–24
Izquierdo Millán LR, Izquierdo SS, Vega-Redondo F (2014) Leave and let leave: a sufficient condition to explain the evolutionary emergence of cooperation. J Econo Dyn Control 46:91–113
Izquierdo SS, Izquierdo Millán LR, Vega-Redondo F (2010) The option to leave: conditional dissociation in the evolution of cooperation. J Theor Biol 267:76–84
Jensen MK, Rigos A (2018) Evolutionary games and matching rules. Int J Game Theory 47:707–735
Koçkesen L, Ok EA, Sethi R (2000) The strategic advantage of negatively interdependent preferences. J Econ Theory 92:274–299
Kuran T, Sandholm W (2008) Cultural integration and its discontents. Rev Econ Stud 75:201–228
Maynard Smith J, Price GR (1973) The logic of animal conflicts. Nature 246:15–18
McNamara JM, Gasson CE, Houston AI (1999) Incorporating rules for responding into evolutionary games. Nature 401:368–371
McNamara JM, Barta Z, Fromhage L, Houston AI (2008) The coevolution of choosiness and cooperation. Nature 451:189–192
Mcpherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27:415–444
Nax HH, Rigos A (2016) Assortativity evolving from social dilemmas. J Theor Biol 395:194–203
Newton J (2017a) The preferences of Homo Moralis are unstable under evolving assortativity. Int J Game Theory 46:583–589
Newton J (2017b) Shared intentions: the evolution of collaboration. Games Econ Behav 104:517–534
Newton J (2018) Evolutionary game theory: a renaissance. Games. https://doi.org/10.3390/g9020031
Ok EA, Vega-Redondo F (2001) On the evolution of individualistic preferences: an incomplete information scenario. J Econ Theory 97:231–254
Rivas J (2013) Cooperation, imitation and partial rematching. Games Econ Behav 79:148–162
Robson AJ (1990) Efficiency in evolutionary games: Darwin, Nash and the secret handshake. J Theor Biol 144:379–396
Robson A (2001) The biological basis of economic behavior. J Econ Lit 29:11–33
Robson A, Samuelson L (2011) The evolutionary foundations of preferences. In: Benhabib J, Bisin A, Jackson M (eds) Handbook of social economics, vol 1. Elsevier, pp 221–310
Ruef M, Aldrich HE, Carter NM (2003) The structure of founding teams: homophily, strong ties, and isolation among US entrepreneurs. Am Sociol Rev 68:195–222
Sandholm W (2001) Preference evolution, two-speed dynamics, and rapid social change. Rev Econ Dyn 4:637–679
Sethi R, Somanathan E (2001) Preference evolution and reciprocity. J Econ Theory 97:273–297
Van Veelen M (2006) Why kin and group selection models may not be enough to explain human other-regarding behaviour. J Theor Biol 242:790–797
Van Veelen M (2011) The replicator dynamics with n players and population structure. J Theor Biol 276:78–85
Wu J (2016) Evolving assortativity and social conventions. Econ Bull 36:936–941
Wu J (2018) Entitlement to assort: democracy, compromise culture and economic stability. Econ Lett 163:146–148