J. Econ. Theory 229 (2025) 106076

Contents lists available at ScienceDirect

Journal of Economic Theory

journal homepage: www.elsevier.com/locate/jet

Spontaneous discrimination with endogenous information disclosure

Arno Apffelstaedt [1]

University of Cologne, Germany

ARTICLE INFO

JEL classification:
D83
C73
J71

Keywords:
Spontaneous discrimination
Repeated games
Social norms
Endogenous information disclosure

ABSTRACT

I introduce endogenous information disclosure into a model of "spontaneous discrimination" à la Peski and Szentes (2013). Individuals in a finite population repeatedly decide whether to engage in profitable interactions with a randomly assigned or chosen partner. Each individual has a fixed physical and a dynamic social color. Social color conveys information about the colors of past partners—but only if that information is disclosed by the decision maker or a random observer. I characterize conditions under which endogenous disclosure supports inefficient equilibria where individuals discriminate by conditioning interactions on the (payoff-irrelevant) colors of potential partners. The analysis shows how competition for being selected as partner interacts with discriminatory norms to create strict incentives for information disclosure, thereby sustaining discrimination that would otherwise break down.

1. Introduction

"Spontaneous discrimination", as developed by Peski and Szentes (2013), refers to the possibility that out-group discrimination can emerge as a self-enforcing social norm in tolerant societies—that is, societies with no inherent taste for discrimination and no payoff-relevant differences between groups. The intuition is simple: consider a population of perfectly tolerant individuals who are identical in all payoff-relevant respects but differ in a (payoff-irrelevant) physical color, which can be red or green. Suppose red individuals refuse to interact with other reds who have previously interacted with greens. If red individuals value future interactions with their in-group sufficiently, it is then rational for them not to accept green partners, that is, to discriminate. This strategy is sequentially rational when red individuals refuse to interact not only with in-group members who have interacted with greens, but also with in-group members who failed to reject those who have interacted with greens. Following nomenclature introduced by Kandori (1992), equilibrium coordination that relies on such community enforcement mechanisms may be termed a "social norm". Spontaneous discrimination is a *discriminatory* social norm.

Spontaneous discrimination provides a compelling complement to canonical preference- and belief-based theories of inefficient discrimination, but requires that individuals have ready access to reliable information on others' partner histories, including the payoff-irrelevant color of past partners and the color of those partners' prior partners. This is because sustaining the norm requires that community members identify both violators of the norm and those who failed to punish them. In many settings, such observability may not arise automatically but depend on whether individuals themselves disclose this information or are observed and reported on by others. As Kandori (1992, p.64) notes, in such settings, "the crux of the matter is information transmission among the community

E-mail address: apffelstaedt@wiso.uni-koeln.de.

[1] Faculty of Management, Economics and Social Sciences, University of Cologne, Albertus-Magnus-Platz, 50931 Cologne, Germany.

https://doi.org/10.1016/j.jet.2025.106076
Received 4 July 2023; Received in revised form 14 August 2025; Accepted 20 August 2025

Available online 29 August 2025

members." This paper studies whether and when individuals and observers have incentives to disclose the color of partners voluntarily and derives conditions for spontaneous discrimination equilibria to arise under endogenous information disclosure.

I study a finite-population, repeated random-matching game in the spirit of Peski and Szentes (2013). In the model, each individual has a fixed physical color $c \in \{red, green\}$ and a dynamic social color $s \in \{red, green\}$ that evolves over time. Both colors are payoff-irrelevant. Each period, one individual is randomly selected as principal and observes the (c, s)-type of a uniformly drawn potential partner before deciding whether to engage in a profitable pairwise interaction. Social color acts as a temporary public label that can convey information about the types of an individual's past partners. This allows interaction decisions to condition—in a simple, Markovian manner—on the partnership history of potential partners.

I first consider a benchmark in which, whenever an interaction occurs, the principal's social color updates automatically to match the physical or social color of their partner, each with positive probability, as in Peski and Szentes (2013). In this setting, discriminatory equilibria can arise in which individuals of physical color c interact only with partners of type (c, c)—excluding those of the opposite physical color as well as in-group members whose social color signals a past deviation from this norm. These equilibria build directly on the exogenous updating assumption for the principal's social color.

I then relax this assumption by making information disclosure endogenous: the principal's social color updates to match their partner's physical or social color only when that specific information is voluntarily reported by the principal or an observer. I focus on strict incentive compatibility, where disclosure occurs only if it strictly benefits the sender. Under random matching, reporting can only harm the principal and observers—by making them, or a potential partner, unacceptable in future interactions—without improving anyone's chance of being selected as a partner. Hence, when disclosure requires a strict incentive, no one reports, social color ceases to carry relevant information, and discriminatory equilibria unravel.

To restore information disclosure, I extend the model to allow principals to choose their partner from a broader set of individuals, rather than being assigned one at random. This introduces competition: being selected as partner implies others are not. Because selection is valuable, individuals benefit not only from improving their own chances but also from reducing others'—thereby creating strict incentives to report. As a result, social color becomes informative again, and discriminatory norms can re-emerge in equilibrium. I characterize conditions for both out-group discrimination, sustained by observer reports, and in-group discrimination, sustained by self-reports, to arise. Finally, I show that, in the competitive setting, out-group discrimination can strictly increase the expected payoff of the in-group relative to a colorblind baseline—despite reducing aggregate welfare—highlighting how group-level incentives may play a role in sustaining inefficient discriminatory norms under competition.

This paper makes several contributions. First, it develops a tractable framework to study the endogenous disclosure of payoff-irrelevant partner attributes in the context of spontaneous discrimination equilibria. Second, it establishes a negative (non-existence) result: when partners are randomly matched—consistent with the interaction structure in Peski and Szentes (2013)—and disclosure requires a strict incentive, discriminatory equilibria cannot be sustained. Third, it provides a positive (existence) result: when individuals can choose their partners, competition restores incentives to disclose and revives discriminatory equilibria. Fourth, it provides new insights into the range and nature of potential equilibrium outcomes by (i) establishing the possibility of spontaneous in-group (rather than out-group) discrimination, and (ii) showing how out-group discrimination can generate benefits for the in-group on average, even when it reduces total welfare. Overall, the paper highlights the role of revealing information or "snitching" in the complex interplay of social norms enforcement, group competition and discrimination.

Related literature Discrimination based on race, ethnicity, or other observable characteristics (here collectively modeled as "color") has been a central topic in economics at least since Becker (1957), with models spanning labor, product, housing, and marriage markets, among other settings (e.g., Lang and Lehmann, 2012; Holzer and Ihlanfeldt, 1998; Sethi and Somanathan, 2004; Eeckhout, 2006). The dominant explanations are taste-based discrimination (Becker, 1957), where differential treatment reflects preferences, and statistical discrimination (Arrow, 1973; Phelps, 1972), where it reflects beliefs about payoff-relevant group differences. The latter includes models in which such differences emerge endogenously through gameplay (e.g., Arrow, 1973; Coate and Loury, 1993; Mailath et al., 2000; Ramachandran and Rauh, 2018; Onuchic and Ray, 2023). For a recent overview, see Onuchic (2024).

This paper instead studies discrimination driven by reputational concerns, in particular building on Peski and Szentes (2013), who develop a model in which discrimination can arise "spontaneously" as a social norm without any inherent taste for, or statistical differences between, colors. This shifts the focus from preferences or beliefs to institutional sources of discrimination—an approach still rare in economics but increasingly recognized as important (e.g., Small and Pager, 2020; Onuchic, 2024). Related work includes Dal Bó (2007) on discriminatory coordination equilibria; Eguia (2017), who shows that norms can induce members of disadvantaged groups to assimilate into more advantaged ones (which echoes the in-group discrimination observed in my model); and Choy (2017), who models trust-building by limiting interactions to the in-group. No prior paper formally analyzes incentives for voluntary information disclosure in the presence of discriminatory norms.

The focus on information disclosure relates this paper to the broader literature on information transmission in repeated games. A seminal contribution is Kandori (1992), who studies how social cooperation can be sustained through labeling mechanisms, and characterizes the minimal information required for such norms to be enforceable in homogeneous populations. The concept of "social color" that I borrow from Peski and Szentes (2013) can be interpreted a specific version of Kandori's stigmatization device. While Kandori treats the information structure as exogenous, subsequent work, for instance Ali and Miller (2016), endogenizes the decision to transmit information. However, these models do not involve group distinctions and therefore cannot address incentives to disclose information in service of discriminatory norms, which is the focus of this paper.

While standard economic theories predict that competition reduces discrimination (e.g., Darity and Williams, 1985), this paper adds to arguments, mainly from sociology, suggesting that competition may instead reinforce it. Bobo and Hutchings (1996) and

Small and Pager (2020) argue that inter-group competition can generate exclusionary incentives when groups compete over scarce resources. A similar argument is developed in a formal model by Bramoullé and Goyal (2016). The logic developed in their paper parallels my analysis of how competition can generate group-level incentives for discrimination.

Finally, the model can be seen as one way to rationalize individual incentives to seek group status through out-group discrimination and stigmatization, as in McAdams (1995). In a similar vein, it relates to "behavioral" models of group identity and social norms such as Akerlof and Kranton (2000) and Bénabou and Tirole (2012).

The paper proceeds as follows. Section 2 begins the analysis with a benchmark model featuring exogenous information disclosure, restating the "spontaneous discrimination" result of Peski and Szentes (2013) within my framework. Section 3 and Section 4 extend the model by introducing endogenous information disclosure and a setting where principals choose from a broader set of potential partners, creating competition to be selected. Section 5 discusses central model assumptions, before Section 6 concludes. The central arguments of proofs are provided in the main text, with additional details and derivations in the appendix where necessary.

2. Benchmark: spontaneous discrimination with exogenous information disclosure

I start by restating the spontaneous discrimination result of Peski and Szentes (2013) within a starkly simplified framework. The simplified framework makes the role of information disclosure and the incentives behind it more transparent and analytically tractable, without qualitatively changing the core insights. I discuss some of the key departures from Peski and Szentes (2013) and their implications in Section 5.

2.1. The benchmark model

Consider a finite population where each individual has a publicly observable physical color, $c_i \in \{red, green\}$. The number of individuals of a given physical color c is n_c, where $n_c \geq 3$, and the total number of individuals is $n = n_{red} + n_{green}$. For any $c \in \{red, green\}$, the opposite color is denoted $-c \in \{green, red\}$. Time is discrete and infinite, $t = 0, 1, 2, 3, \ldots$, and the common discount factor is $\delta < 1$.

Interactions Payoff is produced in pairwise interactions involving one principal and one agent.[2] Opportunities for interaction arrive randomly. Each period, nature selects two individuals uniformly at random—a principal, $p(t)$, and a match, $\mu(t)$. The probability of any individual being selected as the principal is $\frac{1}{n}$, and the conditional probability of any other individual being selected as the match is $\frac{1}{n-1}$.

The principal, $p(t)$, decides whether to select the match as her agent, $a(t) = \mu(t)$, or destroy the opportunity for interaction, $a(t) = \emptyset$. The match is passive in this decision. If $a(t) = \mu(t)$, the interaction generates an immediate payoff of $H > 0$, with the principal receiving a share $\pi > 0$ and the agent receiving $1 - \pi > 0$. If $a(t) = \emptyset$, no payoff is generated that period. After payoffs are realized, the game proceeds to the next period. Individuals are forward-looking and aim to maximize the discounted stream of payoffs from both current and future interactions.

Social color In addition to their physical color c_i, each individual has a publicly observable *social* color, $s_{i,t} \in \{red, green\}$. The pair $(c_i, s_{i,t})$ defines the type $\theta_{i,t}$ of individual i. The type space is $\Theta := \{red, green\} \times \{red, green\}$.

Social color acts as a dynamic marker, temporarily linking characteristics of the agent selected in period t to the individual acting as the principal in period t. By default, an individual's social color matches their physical color: at $t = 0$, $s_{i,t} = c_i$ for all i. Moreover, for all individuals except the principal $p(t)$, social color reverts to their physical color in period $t+1$, $s_{i,t+1} = c_i$.[3]

For the principal, the evolution of social color depends on her decision. If the principal destroys the opportunity ($a(t) = \emptyset$), her social color reverts to her physical color ($s_{p(t),t+1} = c_{p(t)}$). If the principal selects the match as her agent ($a(t) = \mu(t)$), her new social color is determined probabilistically by either the agent's physical color or the agent's social color[4]:

$$s_{p(t),t+1} = \begin{cases} \text{physical color of the agent } c_{a(t)} & \text{with probability } .5, \\ \text{social color of the agent } s_{a(t),t} & \text{with probability } .5. \end{cases} \tag{1}$$

Timeline Fig. 1 summarizes the timeline of the stage game.

[2] In Peski and Szentes (2013), these roles were referred to as employer and worker, respectively.

[3] I discuss the implications of relaxing this assumption in Section 5.

[4] The probability of 0.5 is chosen for technical convenience. The crucial assumption for the discriminatory equilibria in Proposition 1 to emerge is that each option has strictly positive probability.

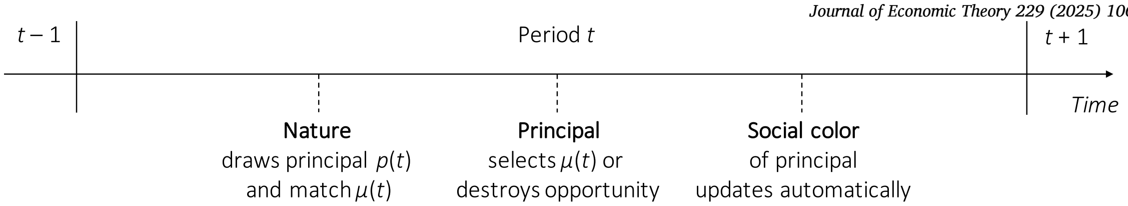

Fig. 1. Stage Game with Exogenous Information Disclosure (Benchmark).

2.2. Equilibrium concept

Following Peski and Szentes (2013), I focus on equilibria where the principal's decision to select or reject the match depends solely on the physical color of the principal $c_{p(t)} \in \{red, green\}$, and the type of the match, $(c_{\mu(t)}, s_{\mu(t),t}) \in \{red, green\} \times \{red, green\}$.[5] Given the discrete time structure and time-invariant payoffs, I use stationary Markov perfect equilibrium as the equilibrium concept, where the Markov state in period t is defined by the principal's physical color and the type of the match.[6] I focus on pure-strategy equilibria.

Social norms I characterize the (stationary Markov perfect) equilibrium by characterizing for each physical color $c \in \{red, green\}$ the set of types $(c, s) \in \Theta = \{red, green\} \times \{red, green\}$ that individuals of this color select as agent. I denote this set A_c and call it the *social norm* of color c. A pair of social norms (A_{red}, A_{green}) maps into equilibrium behavior as follows: If $(c_{\mu(t)}, s_{\mu(t),t}) \in A_{c_{p(t)}}$, then $a(t) = \mu(t)$. If $(c_{\mu(t)}, s_{\mu(t),t}) \notin A_{c_{p(t)}}$, then $a(t) = \emptyset$.

I impose the following tie-break assumption: When indifferent between selecting and rejecting a match, individuals accept. This is an unrestrictive assumption that does not affect equilibria beyond the knife-edge case of a certain parameter condition.[7]

Example 1. The following are examples of social norms.[8]

- $A_c = \Theta$: colorblind social norm; select anyone
- $A_c = \{(c, c)\}$: out-group discrimination; only select "pure" own-colored matches
- $A_c = \Theta \setminus \{(c, c)\}$: in-group discrimination; don't select "pure" own-colored matches

The model's information structure is defined by the publicly observable types of all individuals, $(c_i, s_{i,t})$, which include each individual's physical and social colors. The Markov state, consisting of the principal's physical color $c_{p(t)}$ and the type of the match, $(c_{\mu(t)}, s_{\mu(t),t})$, captures the subset of this information that is relevant for equilibrium play. Since equilibrium strategies depend only on public information about current types, and stage game payoffs are independent of private information, historical and private information do not affect the characterization of best responses. Given the discounted, time-separable payoff structure, stationary Markov perfect equilibria can be derived using the one-shot deviation principle.[9]

2.3. Equilibrium

I derive the equilibrium social norms by analyzing the incentives of individuals to select or reject different types of matches under different norms. In the benchmark setting—consistent with the spontaneous discrimination result in Peski and Szentes (2013, Proposition 1)—equilibrium behavior can take two forms: colorblind behavior or spontaneous out-group discrimination.

Consider an individual of physical color c acting in the role of principal $p(t)$. Her decision is whether to select or reject a match of type (c_a, s_a). Selecting the match yields an immediate payoff of πH, while rejecting the match and destroying the opportunity results in an immediate payoff of 0. Her continuation payoffs depend on the evolution of her social color and the social norms A_c and A_{-c}, which determine whether she will be accepted as an agent in future interactions with her own group and the other group, based on

[5] Allowing the decision of the principal to also depend on her current social color $s_{p(t),t}$ does not change the results. Consistent with Peski and Szentes (2013), the immediate and continuation payoffs of the principal are independent of her current social color. Hence, her actions do not depend on $s_{p(t),t}$. For simplicity, this dependency is omitted from the equilibrium definition.

[6] By defining the Markov state using payoff-*irrelevant* characteristics, I deviate from the standard requirement of Markov perfect equilibrium that the state can only encode payoff-*relevant* information. Being aware of this slight misuse of the Markov equilibrium concept, I use it on the basis of it being simpler and easier to apprehend than the definition of a stand-alone equilibrium concept.

[7] In particular, the assumption determines the existence of the discriminatory social norm $A_c = \{(c, c)\}$ in the specific case where parameters are chosen such that $\pi = .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1 - \pi)$. Under the imposed assumption, the norm exists if and only if $\pi < .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1 - \pi)$ (see (DC Benchmark) in Proposition 1). Assuming instead that individuals reject the agent and destroy the opportunity when indifferent leads to the same Proposition, but with the (DC Benchmark) condition expressed as a weak inequality, $\pi \leq .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1 - \pi)$.

[8] While I do not restrict behavior to the three norms highlighted in this example, only these three norms will emerge as equilibrium social norms in this paper.

[9] I omit a proof of this standard result. Because all payoff-relevant information when deriving the best-response is belief-free, the proof is essentially equivalent to a proof of the one-shot deviation principle when considering subgame perfectness in a repeated game with perfect monitoring, see, for example, Proposition 2.2.1 in Mailath and Samuelson (2006).

her physical and future social color. Let $V_p(c,s)$ denote the discounted continuation payoffs of a principal with physical color c who transitions to social color s in period $t+1$.

If the principal selects the match, her next-period social color becomes c_a (the agent's physical color) or s_a (the agent's social color), each with probability one-half. Hence, the principal's discounted present value of selecting the match is

$$\pi H + \begin{cases} V_p(c,c) & \text{if } (c_a,s_a) = (c,c) \\ V_p(c,-c) & \text{if } (c_a,s_a) = (-c,-c) \\ .5V_p(c,c) + .5V_p(c,-c) & \text{if } (c_a,s_a) \in \{(c,-c),(-c,c)\}. \end{cases} \tag{2}$$

If the principal rejects the match and destroys the opportunity, her next-period social color is equal to her own physical color c. Hence, the principal's discounted present value of rejecting the match is

$$0 + V_p(c,c). \tag{3}$$

The principal will select the match if (2) is (weakly) larger than (3).

Since the continuation payoffs from selecting or rejecting a match of type (c,c) are identical, the principal always selects a match of type (c,c). This excludes social norms featuring in-group discrimination (i.e., norms where $(c,c) \notin A_c$). In contrast, norms where the principal rejects any match other than (c,c) (out-group discrimination) can be sustained. These norms create lower future payoffs when the principal's social color transitions to $s=-c$ rather than $s=c$, implying $V_p(c,-c) < V_p(c,c)$. Out-group discrimination is thus an equilibrium if the continuation payoffs from transitioning to $s=c$ are sufficiently higher than those from transitioning to $s=-c$.

Proposition 1 *(Benchmark: spontaneous discrimination with exogenous information disclosure). The color-blind social norm, $A_c = \Theta$, is an equilibrium social norm for $c \in \{red, green\}$. In addition, if*

$$\pi < .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1-\pi), \tag{DC Benchmark}$$

the discriminatory social norm $A_c = \{(c,c)\}$ is also an equilibrium social norm for $c \in \{red, green\}$.

To build intuition for the result, note that, unlike in standard models of racial discrimination, discrimination in this model is not driven by a preference for color or by statistical differences between groups. This is reflected in the immediate stage-game payoffs, which are maximized by always selecting the match and never destroying the opportunity for interaction. The myopically optimal and efficiency-maximizing strategy, therefore, is to select any match as an agent. The profile of color-blind social norms $(A_{red}, A_{green}) = (\Theta,\Theta)$, under which agents of any type $(c_a,s_a) \in \{red, green\} \times \{red, green\}$ are accepted by both colors, implements the myopically optimal strategy, requires no enforcement, and is therefore always sustainable in equilibrium.

The key to understanding how out-group discrimination can emerge lies in how social color works. By Equation (1), selecting the match as her agent causes the social color of the principal to take on the physical or the social color of the match, each with strictly positive probability. As a result, norms are possible under which individuals are punished for carrying certain social colors, effectively discouraging interactions with specific types.

To illustrate, assume that the social norm of color c is $A_c = \{(c,c)\}$, which demands that only matches of type (c,c) be selected as agents, while all others are rejected. If someone breaks the norm by selecting a match of type $(-c,-c)$, $(-c,c)$, or $(c,-c)$, that person's social color will change to $-c$ with positive probability. This reduces continuation payoffs because, according to the norm, future principals of color c will refuse to interact with individuals of type $(c,-c)$. By destroying the opportunity for interaction, the decision-maker ensures that her next-period social color is c, avoiding the risk of being ostracized. If the expected future loss from being ostracized outweighs the immediate benefit of interaction, rejecting types $(-c,-c)$, $(-c,c)$, and $(c,-c)$ is indeed optimal.

The key to sustaining the norm in equilibrium is ensuring that individuals punish norm-deviators by refusing to interact with individuals of type $(c,-c)$. This condition is satisfied if $\pi < .5 \cdot \delta \cdot \frac{n_c-1}{n(n-1)} \cdot (1-\pi)$ (DC Benchmark). Through this condition, discrimination becomes self-enforcing: it is maintained by the credible and sequentially rational refusal to engage with those who deviate from the norm. Crucially, this refusal is itself driven by the same incentives that sustain the norm: punishers who interact with norm-deviators risk acquiring a "bad" social color, making them unacceptable as agents in future interactions. By this mechanism, spontaneous discrimination can emerge and persist even in perfectly tolerant societies where no individual harbors an inherent preference or bias for color.

3. Endogenous information disclosure

Maintaining spontaneous discrimination requires that individuals are punished for deviating from discriminatory norms—that is, for selecting agents whose physical or social color differs from their own physical color. This punishment relies on the evolution of the principal's social color, which reveals information about the type of agent the principal has accepted, see Equation (1).

Crucially, if social color were not informative about the type of agents being selected, spontaneous discrimination would lose its bite as a self-enforcing mechanism. Selecting agents whose characteristics violate the norm would carry no consequences for the

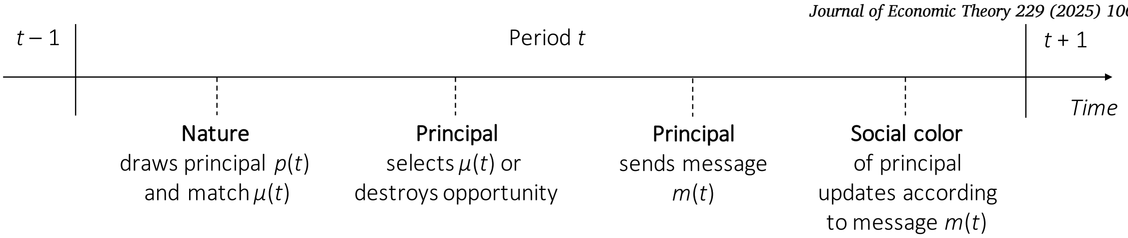

Fig. 2. Stage Game with Self-Reports.

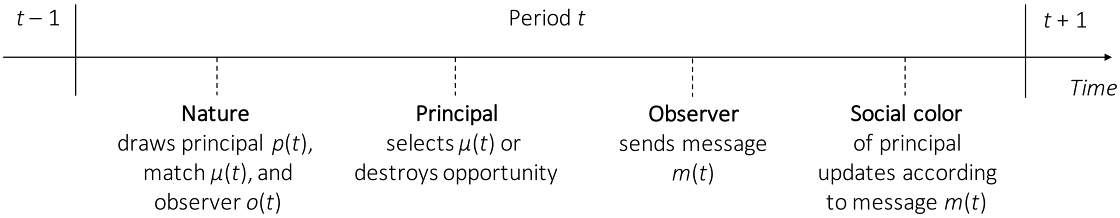

Fig. 3. Stage Game with Observer Reports.

principal, as her social color would fail to reveal norm deviations. As a result, individuals would lack incentives to reject agents based on their physical or social color, leading to a breakdown of discriminatory norms in Markov equilibria.[10]

Given that information about individual interactions is crucial to maintaining discriminatory norms, a natural question arises: Would such information be endogenously shared or disclosed, and can discriminatory norms persist if the evolution of social color depends on deliberate disclosure decisions?[11]

3.1. A model of endogenous information disclosure

I endogenize information disclosure by assuming that whenever a principal interacts with an agent ($a(t) \neq \emptyset$), her social color updates according to a message $m(t)$ that is sent by an individual who observes the interaction. I consider two distinct versions of the model, which differ in whether it is the principal herself or another individual who sends the message.

Version a): Self-Reports Whenever principal $p(t)$ interacts with an agent, $a(t) \neq \emptyset$, the principal can decide to send a message $m(t) \in \{c_{a(t)}, s_{a(t),t}\}$ or decide to stay silent. If she sends a message, her next-period social color reveals that message, $s_{p(t),t+1} = m(t)$. If the principal stays silent, or if she destroys the opportunity, $a(t) = \emptyset$, then $m(t) = \emptyset$. In that case, the next-period social color of the principal conforms to the default, $s_{p(t),t+1} = c_{p(t)}$. Fig. 2 illustrates the stage game.

Version b): Observer Reports Whenever principal $p(t)$ interacts with an agent, $a(t) \neq \emptyset$, nature privately and uniform randomly draws an observer $o(t) \neq p(t)$. The observer can decide to send a message $m(t) \in \{c_{a(t)}, s_{a(t),t}\}$ or decide to stay silent. If he sends a message, the next-period social color of the principal reveals that message, $s_{p(t),t+1} = m(t)$. If the observer stays silent, or if the principal destroys the opportunity, $a(t) = \emptyset$, then $m(t) = \emptyset$. In that case, the next-period social color of the principal conforms to the default, $s_{p(t),t+1} = c_{p(t)}$. Fig. 3 illustrates the stage game.

Note that messages simply make the social color of the principal *red* or *green* without explicitly revealing whether that color conforms to the physical or social color of the agent. In that sense, the model retains the assumption of the benchmark that principals can only be punished (or rewarded) for holding a certain social color, without differentiating between more detailed interaction histories.[12]

3.2. Equilibrium concept

I extend the stationary Markov perfect equilibrium concept from the benchmark to include message strategies. The principal's decision to select or reject the match as her agent, $a(t) \in \{\mu(t), \emptyset\}$, remains determined by social norms A_c, $c \in \{red, green\}$. Under Self-Reports, the Markov state determining message $m(t) \in \{c_{a(t)}, s_{a(t),t}, \emptyset\}$ is given by the physical color of the principal, $c_{p(t)}$, and the

[10] To see this, suppose the principal's next-period social color, conditional on accepting an agent, were fixed at $s_{p(t),t+1} = c_{p(t)}$, regardless of the agent's physical or social color. The continuation payoffs in Equation (2) would hence be $V_p(c, c)$, regardless of the type of agent, and the discounted present value from accepting the match $(\pi H + V_p(c, c))$ would always exceed that of rejecting $(0 + V_p(c, c))$.

[11] In reality, information about individual interactions may often be available through means other than voluntary disclosure, such as physical observation, news, or word-of-mouth. Moreover, senders might strategically lie about interactions using cheap talk. Below, I abstract from these alternative mechanisms to focus on the strategic incentives underlying voluntary information sharing and its role in sustaining discriminatory norms.

[12] Exploring richer messages and their consequences is an interesting extension that lies outside the scope of the present paper.

type of the agent she interacts with, $\theta_{a(t),t} = (c_{a(t)}, s_{a(t),t})$. Under Observer Reports, the Markov state additionally includes the physical color of the observer, $c_{o(t)}$.[13]

In equilibrium, principals and observers choose messages $m(t)$ to maximize their continuation payoffs, taking into account their physical colors, $c_{p(t)}$ and $c_{o(t)}$, and the social norms A_{red} and A_{green}, which determine whether the principal will be accepted or rejected based on her next-period social color. That is, message choices are determined by whether revealing characteristics of the agent affect how the principal is treated and whether the sender has personal benefits from the principal being accepted or rejected in future interactions.[14] Note that a message choice exists in period t if and only if $a(t) \neq \emptyset$, that is, if and only if the principal indeed interacts with an agent. If $a(t) = \emptyset$, then $m(t) = \emptyset$ by default.

I specify equilibria under the following tie-break assumption:

Assumption 1 *(Strict incentives for information disclosure.). When indifferent between choosing $m(t) = \emptyset$ (staying silent) and revealing a characteristic of the agent, $m(t) \in \{c_{a(t)}, s_{a(t),t}\}$, the sender chooses $m(t) = \emptyset$.*

This assumption has several implications, some of which are formalized in Lemma 1 below. Most importantly, it ensures that discriminatory equilibria are not sustained by observers who are indifferent between staying silent and revealing agent characteristics.[15] By excluding such equilibria, the assumption focuses the analysis on settings where information disclosure reflects strict incentives rather than arbitrary tie-breaking. In the context of discriminatory norms, the assumption aligns with a natural tie-breaking rule for "tolerant societies," where a default of non-disclosure aligns with an aversion to enforcing discrimination without explicit incentives.

3.3. Sender incentives and message patterns

I begin with the following observation, which is a direct consequence of Assumption 1:

Lemma 1 *(Sender incentives and message patterns). Consider messages following the interaction of a principal of physical color $c \in \{red, green\}$ with an agent of type $(c_a, s_a) \in \{c, -c\} \times \{c, -c\}$. In equilibrium,*

1. *If $m(t) \neq \emptyset$, then $m(t) = -c$.*
2. *If $(c_a, s_a) = (c, c)$, then $m(t) = \emptyset$.*
3. *If $(c_a, s_a) \in \{(-c, -c), (-c, c), (c, -c)\}$, then*
   - *(i) $m(t) = -c$ with a probability q_c that is constant over time and identical for all agents of type $(c_a, s_a) \in \{(-c, -c), (-c, c), (c, -c)\}$,*
   - *(ii) $m(t) = \emptyset$ with the remaining probability.*

Lemma 1 establishes a simple rule: Senders will either remain silent ($m(t) = \emptyset$), or they will send a message that assigns the opposite color to the principal ($m(t) = -c$). Moreover, given the principal's physical color c, the probability with which the principal is assigned the opposite color is constant across periods and for all agents whose type contains the opposite color, $(c_a, s_a) \in \{(-c, -c), (-c, c), (c, -c)\}$.

To understand the lemma, consider how the message affects the principal's next-period social color and, in turn, the sender's continuation payoffs. Staying silent ($m(t) = \emptyset$) automatically results in $s_{p(t),t+1} = c$, aligning the principal's social color with her physical color. Hence, assigning the principal the opposite color, by sending message $m(t) = -c$, is the only case for which the sender might strictly prefer to reveal information. This explains Part 1 of the lemma.

Part 2 follows immediately: If the agent's type is $(c_a, s_a) = (c, c)$, the only possible message is $m(t) = c$, which does not assign the opposite color to the principal. Thus, there can never be a strict incentive to send such a message, hence $m(t) = \emptyset$.

Part 3 concerns agents whose type includes the opposite color $-c$ in either coordinate, that is, $(c_a, s_a) \in \{(-c, -c), (-c, c), (c, -c)\}$. Fixing the physical color of the sender, the incentive to send $m(t) = -c$ or remain silent depends only on how the message affects the principal's next-period social color. If the sender strictly prefers the principal to attain $s_{p(t),t+1} = -c$ for one such type, the same preference applies to all such types, implying $m(t) = -c$ is sent for each of them. If there is no strict preference for $s_{p(t),t+1} = -c$, Assumption 1 implies that the sender will also behave consistently across these types by remaining silent ($m(t) = \emptyset$). Stationarity further implies that, for a given physical color of the sender, the message choice is constant over time, yielding a constant probability q_c of sending $m(t) = -c$ for a principal of physical color c. In the case of Self-Reports, this probability is either 1 (if the principal, given social norms, strictly prefers to attain $s_{p(t),t+1} = -c$) or 0 (otherwise). In the case of Observer Reports, all observers of a given physical color behave identically in equilibrium. If observers of both physical colors strictly prefer the principal to attain $s_{p(t),t+1} = -c$, the probability that the message is $m(t) = -c$ is 1; if neither color does, it is 0. If only observers of color c strictly prefer $s_{p(t),t+1} = -c$, the probability is $\frac{n_c - 1}{n}$, which is the probability of drawing an observer of color c when the principal also has that color; if only

[13] Given Assumption 1, including the social color of the principal or observer in the Markov state does not affect equilibrium messages, as the equilibrium incentives for information disclosure are fully determined by the physical colors and the agent's type.

[14] Beyond this, messages do not have payoff relevance for the sender. In particular, senders are not directly punishable or rewardable for the messages they send.

[15] See Footnote 17.

observers of the opposite color strictly prefer $s_{p(t),t+1} = -c$, the probability is $\frac{n_{-c}}{n}$, which is the probability of drawing an observer of the opposite color.

3.4. Candidate social norms under endogenous information disclosure

For the principal's decision whether to select or reject an agent, the message pattern established by Lemma 1 partitions potential agents into two sets: those whose type includes the opposite color, $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$ and those whose type does not, $(c_a, s_a) = (c,c)$. Since accepting any agent in the set $\{(-c,-c),(-c,c),(c,-c)\}$ induces the same probability of message $m(t) = -c$, and hence the same continuation payoffs for the principal, social norms can only differ by whether all agents in this set are accepted or rejected, and by whether agents of type (c,c) are accepted or rejected. This restriction reduces the set of potential norms to three possibilities[16]:

Proposition 2 *(Candidate social norms under endogenous information disclosure). Color $c \in \{red, green\}$ follows one of three social norms:*

1. *All types are accepted, $A_c = \Theta$ (colorblind norm),*
2. *Only type (c,c) is accepted, $A_c = \{(c,c)\}$ (out-group discrimination),*
3. *Only types $\{(-c,-c),(-c,c),(c,-c)\}$ are accepted, $A_c = \Theta \setminus \{(c,c)\}$ (in-group discrimination).*

Notice, importantly, that none of the candidate norms differentiate between agents of the opposite physical color $-c$ based on their social color. Specifically, under $A_c = \{(c,c)\}$, all agents with physical color $-c$ are categorically rejected, irrespective of their social color. Similarly, under $A_c = \Theta$ and $A_c = \Theta \setminus \{(c,c)\}$, all agents with physical color $-c$ are categorically accepted, irrespective of their social color. By an analogous argument, social norms A_{-c} do not differentiate between agents of physical color c.

This lack of differentiation implies that the continuation payoffs of a principal of color c, determined by her next-period social color $s_{p(t),t+1} \in \{c,-c\}$, neither vary within nor across the potential social norms, A_{-c}, adopted by the opposite color. As a result, the principal's decision to select or reject an agent, and hence, whether a given norm A_c can be sustained for color $c \in \{red, green\}$ in equilibrium, will not depend on the social norm A_{-c} of the opposite color:

Corollary 1 *(Strategic independence of social norms). In equilibrium, the best response of a principal of color c to the pair of social norms (A_c, A_{-c}) is independent of the norm adopted by the opposite color, A_{-c}.*

Corollary 1 allows for analyzing the conditions for the existence of each social norm for color c without specifying or conditioning on the social norm of the opposite color $-c$, a technical convenience that I will be exploiting in the rest of the paper.

3.5. Breakdown of spontaneous discrimination

In the benchmark model's non-competitive interaction setting, endogenous information disclosure leads to the breakdown of discriminatory norms when strict incentives for information disclosure are required:

Proposition 3 *(Breakdown of spontaneous discrimination). Under strict incentives for information disclosure (Assumption 1), the unique equilibrium social norm for $c \in \{red, green\}$ is the color-blind social norm, $A_c = \Theta$.*

Notice first that the non-competitive setting rules out in-group discrimination, $A_c = \Theta \setminus \{(c,c)\}$. This was shown already in Section 2 for the case of exogenous information disclosure, and the same logic applies here: Not accepting an agent of type (c,c) means having to destroy the opportunity, which results in the same next-period social color, $s_{p(t),t+1} = c$ (and hence the same continuation payoffs) as accepting the agent, but sacrifices the immediate interaction payoff, πH. Since this behavior is clearly payoff-dominated, it cannot be part of an equilibrium.

We are left with out-group discrimination, $A_c = \{(c,c)\}$, which emerged as a potential equilibrium norm in the benchmark setting with exogenous information disclosure (Proposition 1). To see why this norm fails under endogenous information disclosure, recall how the norm works: The reason why the principal complies with the norm and rejects agents of type $\{(-c,-c),(-c,c),(c,-c)\}$ is the risk of receiving social color $s_{p(t),t+1} = -c$, which would change her type to $(c,-c)$ and make herself unacceptable as an agent in future interactions with her own group. Under endogenous information disclosure, however, the principal does not automatically receive the social color $s_{p(t),t+1} = -c$ after interacting with an agent of type $\{(-c,-c),(-c,c),(c,-c)\}$. Instead, this occurs only if the principal herself or an observer sends the message $m(t) = -c$.

Consider first the case of Self-Reports. Under $A_c = \{(c,c)\}$, sending $m(t) = -c$ directly reduces the principal's continuation payoff by assigning herself a social color that will lead to punishment. Effectively, the principal would be reporting herself for a deviation and

[16] I refrain from including the empty social norm, $A_c = \emptyset$, where both types (c,c) and types $\{(-c,-c),(-c,c),(c,-c)\}$ are rejected and all opportunities for interaction are destroyed, as a viable equilibrium possibility. Destroying an opportunity results in the same next-period social color, $s_{p(t),t+1} = c$ (and hence the same continuation payoffs) as accepting an agent of type (c,c), but sacrifices the immediate interaction payoff, πH. Therefore, the empty social norm cannot be sustained in equilibrium.

subjecting herself to penalties. Since the principal has no incentive to punish herself, she will always choose $m(t) = \emptyset$ after interacting with agents of type $\{(-c,-c),(-c,c),(c,-c)\}$. Consequently, the norm $A_c = \{(c,c)\}$ cannot be sustained.

Now consider the case of Observer Reports. Observers of the opposite physical color to the principal, $-c$, are indifferent to disclosing information for two reasons. First, by Proposition 2, any norm followed by the observer's own group, A_{-c}, treats agents of physical color c identically, regardless of their social color. As a result, the observer's continuation payoffs are unaffected by the principal's social color $s_{p(t),t+1}$ in interactions where an individual of color $-c$ is the principal. Second, under the norm $A_c = \{(c,c)\}$, principals of color c reject all individuals of physical color $-c$, irrespective of their social color. Consequently, the observer's continuation payoffs remain unaffected by $s_{p(t),t+1}$ in interactions involving a principal of color c. Therefore, $s_{p(t),t+1}$ does not influence the observer's incentives in either scenario. By Assumption 1, indifferent observers remain silent, meaning that observers of physical color $-c$ do not disclose information.[17]

What about observers of the same color as the principal? Under the norm $A_c = \{(c,c)\}$, sending $m(t) = -c$ assigns the principal a social color that makes her unacceptable as an agent in the next period. This action harms the observer if the observer becomes the principal in period $t+1$ and is matched with the current principal as an agent. In such a case, the observer would have to reject the agent, forfeiting the immediate payoff πH. Consequently, observers of color c have a strict incentive *not* to report the principal, and instead, will remain silent. Just as in the case of Self-Reports, information disclosure breaks down, and the social norm $A_c = \{(c,c)\}$ fails to constitute an equilibrium under Observer Reports.

4. Competition for interactions

Does the breakdown of spontaneous discrimination imply that the concept fails to explain discrimination in environments requiring endogenous information disclosure? This section argues otherwise, proposing an extension that focuses on the nature of interactions to revive discriminatory outcomes. Specifically, it demonstrates that competition for interactions can sustain discriminatory social norms, even when strict incentives for endogenous information disclosure are required.

Sections 4.1 and 4.2 introduce the extension and the extended equilibrium concept, respectively. Sections 4.3 and 4.4 study out-group and in-group discrimination, showing that both norms can be sustained in a competitive setting. Finally, Section 4.5 shows how competition generates group-level incentives for discrimination.

4.1. A model with competitive interactions

I introduce competition for interactions by expanding the principal's choice set: Instead of selecting the randomly assigned match $\mu(t)$ or destroying the opportunity, the principal can also choose any other individual $j \neq p(t)$ as her agent. This change introduces competition among agents, as each individual $j \neq p(t)$ now has an incentive to be selected by the principal. Formally, the principal's choice set consists of any individual other than $p(t)$, or the empty set $\emptyset$, where $a(t) = \emptyset$ implies that the principal destroys the opportunity by choosing no one.

As before, interacting with an agent in period t produces a positive payoff shared between the principal (receiving a share π) and the agent (receiving $1-\pi$), while destroying the opportunity results in no payoff. However, while selecting $\mu(t)$ as the agent generates payoff H, selecting another agent $j \neq \mu(t)$ only generates $L > 0$, where $L < H$. Thus, selecting another agent is preferable to destroying the opportunity but incurs a cost compared to selecting the match.

Assumption $L < H$ has two different interpretations: $\mu(t)$ may be an expert uniquely suited for the opportunity that arose in period t, with other agents being less productive on the job. Alternatively, all agents might have the same productivity, but searching for an agent other than the initial match $\mu(t)$ is costly. The ratio $\frac{L}{H} \in (0,1)$ serves as a measure of the substitutability of $\mu(t)$ and the intensity of competition for interactions.

Information disclosure is endogenous, following the rules specified in Section 3: If the principal destroys the opportunity, $a(t) = \emptyset$, her next-period social color reverts to her physical color, $s_{p(t),t+1} = c_{p(t)}$. If the principal interacts with an agent, $a(t) \neq \emptyset$, the social color of the principal updates according to a message $m(t) \in \{c_{a(t)}, s_{a(t),t}, \emptyset\}$, which is either sent by the principal herself (Self-Reports) or by a randomly chosen observer (Observer Reports).[18] Note that the results on the message pattern (Lemma 1) and on the candidate social norms in equilibrium (Proposition 3, with Corollary 1) established in Section 3 remain valid.

Fig. 4 illustrates the new sequence of events in period t.

4.2. Equilibrium concept

I continue concentrating on stationary Markov perfect equilibria that allow for a characterization of agent choice via social norms A_c. Given a social norm A_c and a draw of principal $p(t)$ of color c, let $J^{A_c}(t)$ be the period t set of potential agents who are acceptable under the norm:

[17] Adopting a tie-break rule opposite to Assumption 1, where indifferent senders disclose instead of staying silent, out-group discrimination can be sustained as an equilibrium norm under Observer Reports, given suitable parameter conditions. In that case, observers of the same color as the principal remain silent, while those of the opposite color might disclose $m(t) = -c$. If all opposite-color observers disclose, the norm $A_c = \{(c,c)\}$ is an equilibrium if and only if $\pi < \frac{n_{-c}}{n} \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1-\pi)$, where $\frac{n_{-c}}{n}$ is the probability of the observer being drawn from the opposite color and, thus, the probability of the principal being assigned $s_{p(t),t+1} = -c$.

[18] For all individuals except the principal, social color reverts back to their physical color in period $t+1$, that is: $s_{j,t+1} = c_j$ for all $j \neq p(t)$.

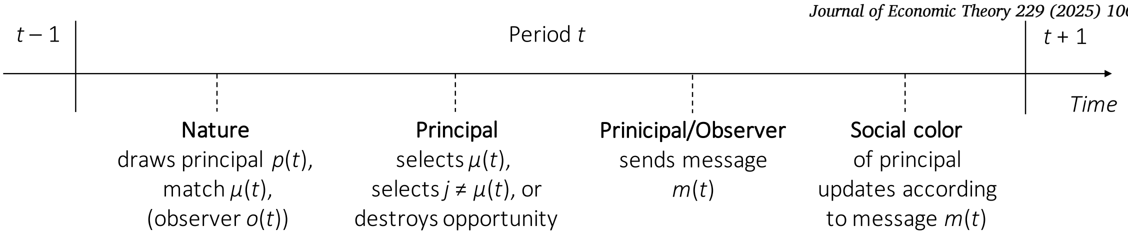

Fig. 4. Stage Game with Competition (and Endogenous Information Disclosure).

$$J^{A_c}(t) := \{\, j \neq p(t) \,|\, (c_j, s_{j,t}) \in A_c \,\}.$$

Social norm A_c then maps into the principal's choice as follows: If the match is acceptable, $\mu(t) \in J^{A_c}(t)$, the principal chooses the match ($a(t) = \mu(t)$). If the match is not acceptable, $\mu(t) \notin J^{A_c}(t)$, the principal picks another acceptable agent, $j \in J^{A_c}(t)$, where each $j \in J^{A_c}(t)$ is chosen with equal probability. If no acceptable agent exists, $J^{A_c}(t) = \emptyset$, the principal destroys the opportunity ($a(t) = \emptyset$).[19]

4.3. Out-group discrimination

The key to understanding how out-group discrimination can be sustained under competition lies in the altered reporting incentives of observers. When interactions are competitive, compliance with the social norm $A_c = (c, c)$ allows the principal to reject a match of type $(c_a, s_a) \in \{(-c,-c), (-c, c), (c,-c)\}$ without destroying productive opportunities. Individuals of type (c, c) benefit from such rejections because they gain a positive chance of being selected as the replacement and receiving a payoff $(1-\pi)L$. Importantly, the smaller the set of acceptable agents $j \in J^{A_c}(t)$ with $(c_j, s_{j,t}) = (c, c)$, the higher the chances for any such j to be selected as a replacement. This creates a snitching incentive: By reporting a principal of color c who violates the norm, observers of physical color c reduce the set of acceptable agents, thereby decreasing competition and increasing their individual chances of being selected as an agent in future interactions.

To formally derive the condition under which an observer of physical color c reports, consider a principal of physical color c who violates the norm by interacting with an agent of type $(c_a, s_a) \in \{(-c,-c), (-c, c), (c,-c)\}$. The observer can either remain silent, $m(t) = \emptyset$, which results in the principal's next-period social color becoming $s_{p(t),t+1} = c$, or report the violation by sending $m(t) = -c$, resulting in $s_{p(t),t+1} = -c$. Sending $m(t) = -c$ changes the principal's type to $(c,-c)$, rendering her unacceptable as an agent in period $t+1$. This has both negative and, under competition for interactions, positive consequences for the observer.

The negative consequences arise if the observer becomes the principal in the following period, $p(t+1) = o(t)$, and the current principal becomes the match, $\mu(t+1) = p(t)$, which happens with probability $\frac{1}{n} \cdot \frac{1}{n-1}$. In this case, the observer must reject the match, incurring a loss of $\pi(H-L)$. In the non-competitive benchmark, the probability of such a scenario caused the observer to strictly prefer remaining silent, contributing to the breakdown of spontaneous discrimination (see Section 3.5)

The positive consequences for the observer of sending $m(t) = -c$ are unique to the competitive setting. If another individual of physical color c becomes the principal in period $t+1$ (probability $\frac{n_c-2}{n}$), sending $m(t) = -c$ increases the observer's chances of being selected as a replacement. The observer stands to benefit in the case of two potential events: First, if the current principal becomes the match, which occurs with probability $\frac{1}{n-1}$, sending $m(t) = -c$ causes the next-period principal to seek a replacement, giving the observer a chance of $\frac{1}{n_c-2}$ to be selected. Second, if the match is someone from the opposite physical color $-c$, which occurs with probability $\frac{n_{-c}}{n-1}$, the observer's chances of being selected as the replacement rise from $\frac{1}{n_c-1}$ to $\frac{1}{n_c-2}$. In either event, if selected, the observer earns a payoff of $(1-\pi)L$.

The observer has a strict incentive to send $m(t) = -c$ and reveal the norm deviation if the benefits of doing so strictly outweigh the costs:

$$\frac{n_c-2}{n}\left[\frac{1}{n-1} \cdot \frac{1}{n_c-2} + \frac{n_{-c}}{n-1}\left(\frac{1}{n_c-2} - \frac{1}{n_c-1}\right)\right](1-\pi)L > \frac{1}{n} \cdot \frac{1}{n-1} \cdot \pi(H-L), \tag{4}$$

which simplifies to condition (IC Out-Group) in the following lemma:

Lemma 2 *(Reporting incentive condition for out-group discrimination). Observers of physical color c will send $m(t) = -c$ when a principal of physical color c violates the norm $A_c = \{(c, c)\}$ if and only if the following condition holds:*

$$\frac{L}{H} > \frac{\pi}{1 + \frac{n_{-c}}{n_c-1}(1-\pi)}. \tag{IC Out-Group}$$

[19] The optimality of choosing $\mu(t)$ over $j \neq \mu(t)$ (when both are acceptable) and $j \neq \mu(t)$ over $a(t) = \emptyset$ (when $\mu(t)$ is not acceptable) follows directly from the ranking of immediate payoffs. Assigning equal probability to each $j \neq \mu(t)$, $j \in J^{A_c}(t)$, aligns with the idea that the principal can identify whether an agent belongs to $J^{A_c}(t)$ but may not be able distinguish individual agents within the set. Moreover, it ensures that the choice rule is stationary optimal, as it avoids fixing probabilities on specific agents who might be unacceptable given certain norms or histories.

Lemma 2 establishes that observers of physical color c will report norm violations if the ratio $\frac{L}{H}$ is sufficiently high, reflecting a trade-off between the potential gains from being selected as a replacement when others comply with the norm and the expected loss from having to reject a now-unacceptable principal in future interactions. Since the right-hand side of condition (IC Out-Group) is strictly less than 1, a sufficiently high $\frac{L}{H}$ ensures that reporting is incentive compatible.

Hence, while observers of the opposite physical color remain silent—being indifferent to the principal's social color, as explained in Section 3.5—the norm $A_c = \{(c,c)\}$ may be sustained under Observer Reports through the messages sent by observers of physical color c. In contrast, under Self-Reports, no information is disclosed in equilibrium: a principal who interacts with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$ has no incentive to report, since this would result in social color $-c$ and expose her to rejection in future interactions.

Existence of social norm $A_c = \{(c,c)\}$ To derive the conditions under which $A_c = \{(c,c)\}$ can be sustained as an equilibrium norm, consider a principal of physical color c matched with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$. Following the norm requires the principal to reject the match and instead select an agent of type (c,c). This incurs an immediate cost of $\pi(H-L)$, but avoids the risk of acquiring social color $s_{p(t),t+1} = -c$, which would lead to rejection in future interactions.

Let q_c denote the probability that message $m(t) = -c$ follows the interaction of a principal of physical color c with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$. The principal's incentive constraint under the norm is

$$\pi(H-L) < q_c \cdot \left[V_p(c,c) - V_p(c,-c)\right], \tag{5}$$

which ensures that rejecting a match of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$ is optimal.

The loss in continuation payoffs from acquiring social color $-c$ is given by

$$V_p(c,c) - V_p(c,-c) = \delta \cdot \frac{n_c - 1}{n}\left[\frac{1}{n-1}(1-\pi)H + \frac{n_{-c}}{n-1} \cdot \frac{1}{n_c - 1}(1-\pi)L\right], \tag{6}$$

where the first term within the brackets reflects the loss from not being accepted as a match by individuals of color c, while the second term captures the loss from not being selected as a replacement when individuals of color c reject matches of the opposite color.[20]

Under Self-Reports, $q_c = 0$, implying that the norm must fail. Under Observer Reports, assuming that condition (IC Out-Group) holds, q_c is given by the probability that the principal is observed by an individual of the same physical color, hence $q_c = \frac{n_c - 1}{n-1} > 0$. As a result, with Observer Reports, the norm exists under suitable parameter conditions.

Proposition 4 *(Out-group discrimination under competition). In addition to the color-blind social norm, $A_c = \Theta$, the out-group discrimination norm $A_c = \{(c,c)\}$ is an equilibrium social norm for $c \in \{red, green\}$ under Observer Reports if and only if*

$$\frac{L}{H} > \frac{\pi}{1 + \frac{n_{-c}}{n_c - 1}(1-\pi)} \tag{IC Out-Group}$$

and

$$\pi\left(1 - \frac{L}{H}\right) < q_c \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_{-c}}{n_c - 1} \cdot \frac{L}{H}\right)(1-\pi) \tag{DC Out-Group}$$

where $q_c = \frac{n_c - 1}{n-1}$ is the probability that a principal of physical color c is observed by an individual of physical color c.

Under Self-Reports, $A_c = \{(c,c)\}$ is not an equilibrium social norm.

How does competition affect the thresholds under which out-group discrimination constitutes an equilibrium norm under Observer Reports? I study this using the ratio $\frac{L}{H}$, which enters both equilibrium conditions and captures the degree to which competition shapes reporting incentives and influences the principal's compliance with the discriminatory norm.

Note first that Proposition 4 nests the non-competitive benchmark: By taking $L \to 0$ and setting $q_c = .5$, condition (DC Out-Group) reduces to condition (DC Benchmark) in Proposition 1. Moreover, as $L \to 0$, the left-hand side of condition (IC Out-Group) approaches zero, replicating the result from Section 3.5 that observers choose not to disclose.

Under Observer Reports, competition generates strict incentives for information disclosure and increases the parameter range for which out-group discrimination is an equilibrium norm. As $\frac{L}{H}$ increases, the left-hand side of condition (IC Out-Group) rises, reflecting that with a sufficient degree of competition, the benefits of reporting norm violations exceed the costs. At the same time, higher values of $\frac{L}{H}$ reduce the cost for the principal to comply with the norm, as the relative cost of replacing the match with a norm-compatible agent decreases. This directly strengthens the principal's incentive to reject norm-violating matches, making condition (DC Out-Group) easier to satisfy. As $\frac{L}{H} \to 1$ $(L \to H)$, matches become perfectly substitutable at no cost, and the conditions

[20] To calculate $\left[V_p(c,c) - V_p(c,-c)\right]$, note that differences in continuation payoff are confined to period $t+1$, as the principal's social color permanently reverts to $s = c$ thereafter. Differences arise if the next-period principal is another individual of physical color c (probability $\frac{n_c - 1}{n}$). In that case, social color c benefits the current principal in two ways: (i) if she is drawn as the match (probability $\frac{1}{n-1}$), she is accepted and receives the discounted payoff $\delta(1-\pi)H$; (ii) if the match is someone of physical color $-c$ (probability $\frac{n_{-c}}{n-1}$), she may be selected as a replacement (probability $\frac{1}{n_c - 1}$), yielding discounted payoff $\delta(1-\pi)L$.

for out-group discrimination always hold. In this limit, competition guarantees the existence of out-group discrimination as an equilibrium norm regardless of the choice of the remaining parameter values.

4.4. In-group discrimination

In addition to restoring out-group discrimination as an equilibrium norm, competition for interactions creates conditions under which in-group discrimination, $A_c = \Theta \setminus \{(c,c)\}$, previously unviable, can emerge in equilibrium.

Unlike out-group discrimination, where information disclosure is a significant challenge, in-group discrimination fails in the non-competitive setting not because of a lack of information disclosure, but because the outside option to accepting an agent of type (c,c)—destroying the opportunity—fails to offer any advantage to the principal. Whether the principal accepts or rejects, her next-period social color equals $s_{p(t),t+1} = c$, leaving her continuation payoffs unchanged. As a result, rejection of an agent of type (c,c) serves only to forgo the immediate payoff πH, making compliance with the norm $A_c = \Theta \setminus \{(c,c)\}$ unsustainable (see Section 3.5).

What competition does is fundamentally change the principal's outside option when rejecting an agent of type (c,c). In the competitive setting, following the norm when being matched to an agent of type (c,c) no longer means destroying the opportunity. Instead, competition allows the principal to replace the rejected (c,c)-agent with one of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$, enabling her to attain the social color $s_{p(t),t+1} = -c$ and making the rejection an incentive-compatible action.

Provided that the principal selects an agent of type $(-c,-c)$, $(-c,c)$, or $(c,-c)$, information disclosure can be sustained through both Self-Reports and Observer Reports. Specifically, the norm $A_c = \Theta \setminus \{(c,c)\}$ gives principals and observers of physical color c a strict incentive to send message $m(t) = -c$. For the principal, acquiring the social color $-c$ ensures acceptance under the norm and avoids rejection in future interactions; hence, she will always choose $m(t) = -c$ following such interactions. Observers of physical color c likewise prefer that the principal acquire social color $-c$, as this spares them the enforcement costs of rejecting her in later interactions. Thus, under Observer Reports, observers of physical color c send $m(t) = -c$ whenever the principal interacts with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$. By contrast, observers of physical color $-c$ strictly prefer to remain silent ($m(t) = \emptyset$) because sending $m(t) = -c$ would reduce their chances of being chosen as replacements in future periods.

Existence of social norm $A_c = \Theta \setminus \{(c,c)\}$ To derive the conditions under which $A_c = \Theta \setminus \{(c,c)\}$ is an equilibrium norm, consider a principal of physical color c matched with an agent of type (c,c). Following the norm requires that the principal rejects the match and instead selects an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$. This action incurs an immediate loss of $\pi(H-L)$ but leads to a continuation payoff gain $\left[V_p(c,-c) - V_p(c,c)\right]$ if it induces the principal's social color to become $s_{p(t),t+1} = -c$.

Let q_c denote the probability that message $m(t) = -c$ follows the interaction of a principal of physical color c with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$. The tradeoff of the principal is then captured by the condition

$$\pi(H-L) < q_c \cdot \left[V_p(c,-c) - V_p(c,c)\right], \tag{7}$$

which ensures that rejecting a match of type (c,c) is incentive-compatible under the norm.

The difference in continuation payoffs, $\left[V_p(c,-c) - V_p(c,c)\right]$, is given by

$$V_p(c,-c) - V_p(c,c) = \delta \cdot \frac{n_c - 1}{n}\left[\frac{1}{n-1}(1-\pi)H + \frac{n_c - 2}{n-1} \cdot \frac{1}{n_{-c}+1}(1-\pi)L\right], \tag{8}$$

where the first term within the brackets reflects the expected payoff from being accepted as a match by individuals of color c, while the second term accounts for the potential payoff from being selected as a replacement for agents of type (c,c).[21]

We are now ready to state the conditions for which $A_c = \Theta \setminus \{(c,c)\}$ can be sustained in equilibrium. Under this norm, the principal has a strict incentive to induce $s_{p(t),t+1} = -c$, implying that with Self-Reports, the probability that message $m(t) = -c$ follows an interaction of a principal of physical color c with an agent of type $(c_a, s_a) \in \{(-c,-c),(-c,c),(c,-c)\}$ is $q_c = 1$. With Observer Reports, by contrast, q_c is given by the probability that the principal is observed by an individual of the same physical color, that is $q_c = \frac{n_c - 1}{n-1}$. This is because observers of physical color c have a strict incentive to send message $m(t) = -c$, while observers of the opposite color strictly prefer to remain silent. Simplifying and rearranging (8) yields the following proposition:

Proposition 5 *(In-group discrimination under competition). In addition to the color-blind social norm, $A_c = \Theta$, the in-group discrimination norm $A_c = \Theta \setminus \{(c,c)\}$ is an equilibrium social norm for $c \in \{red, green\}$ if and only if*

$$\pi\left(1 - \frac{L}{H}\right) < q_c \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_c - 2}{n_{-c}+1} \cdot \frac{L}{H}\right)(1-\pi), \tag{DC In-Group}$$

where

[21] To understand the specific payoffs in Equation (8), note that $\frac{n_c-1}{n}$ is the probability that the next-period principal is of physical color c. Within the first term, $\frac{1}{n-1}$ is the probability of the current principal being chosen as the match, leading to a payoff of $(1-\pi)H$ upon acceptance. The second term captures matches between a principal of color c and another individual of color c (excluding the current principal), which occur with conditional probability $\frac{n_c-2}{n-1}$. In period $t+1$, all individuals of physical color c, except the current principal, have social color c, i.e., are of type (c,c). By the social norm $A_c = \Theta \setminus \{(c,c)\}$, such matches are rejected. Replacements are drawn from all individuals of color $-c$ and the current principal, giving the current principal a selection probability of $\frac{1}{n_{-c}+1}$. If selected, the payoff is $(1-\pi)L$.

- $q_c = 1$ *under Self-Reports, and*
- $q_c = \frac{n_c - 1}{n-1}$ *(probability that a principal of physical color c is observed by an individual of physical color c) under Observer Reports.*

Similar to out-group discrimination, a higher degree of competition, as measured by the ratio $\frac{L}{H}$, increases the parameter range for which in-group discrimination is an equilibrium norm. In the case of in-group discrimination, where strict incentives to disclose information are always given, this effect is solely driven by the reduced compliance cost for the principal as replacing a match of type (c, c) becomes cheaper with an increased ratio $\frac{L}{H}$. In the limit, where $\frac{L}{H} \to 1$ ($L \to H$), the norm $A = \Theta \setminus \{(c, c)\}$ always exists.

An interesting aspect of the norm $A_c = \Theta \setminus \{(c, c)\}$ is that the other group (individuals of physical color $-c$) does not play any role in its enforcement or existence. While individuals of physical color $-c$ benefit from the norm, as they are selected as replacements when matches of type (c, c) are rejected, their equilibrium behavior does not contribute to sustaining the norm. Specifically, agents of physical color $-c$ neither reward principals for selecting them over their own type nor disclose information that would help uphold the norm. Conversely, agents of color c, i.e., the color who follows and enforces the norm, are disadvantaged by their own group following the norm.[22]

4.5. *Group incentives for discrimination*

From a total welfare perspective, discriminatory norms are unequivocally harmful. They lead to the destruction of productive interactions and thus destroy valuable payoffs. The unique welfare-maximizing social norm is the color-blind norm $A_c = \Theta$. Yet, from the perspective of a particular color c, competition for interactions can create conditions under which members of that color, when evaluating their average payoffs under a given norm, perceive discrimination as beneficial. This can be detrimental, as the social norms of the two physical colors $c \in \{red, green\}$ are strategically independent in equilibrium and hence, each group may find it optimal to independently coordinate on discriminating the other.

While in-group discrimination ($A_c = \Theta \setminus \{(c, c)\}$) clearly harms group c, as it involves rejecting its own members and diverts payoffs to the out-group, out-group discrimination ($A_c = \{(c, c)\}$) can appear advantageous under competition. The rejection of out-group agents implies that in-group members are selected as replacements, and if the payoffs L from these interactions are sufficiently high, the benefits for group c can outweigh the losses from rejection:

Proposition 6 *(Group incentives for discrimination). If*

$$\frac{L}{H} > \pi, \qquad \text{(DC group incentives)}$$

out-group discrimination, $A_c = \{(c, c)\}$, results in strictly higher average payoffs for individuals of color $c \in \{red, green\}$ than the color-blind social norm $A_c = \Theta$.

Proposition 6 highlights the potential for each group $c \in \{red, green\}$ to prefer and internally coordinate on a norm that discriminates against the other group. If individuals of a given color could ex-ante agree and commit to a norm (e.g., through a hypothetical election in period $t = 0$), they would choose the colorblind norm $A_c = \Theta$ when $\frac{L}{H} \leq \pi$, but the out-group discrimination norm $A_c = \{(c, c)\}$ if $\frac{L}{H} > \pi$. This preference arises despite the fact that discrimination destroys welfare. From an ex-ante perspective, the situation resembles a prisoners' dilemma: If $\frac{L}{H} > \pi$, each group benefits from unilaterally deviating from a colorblind norm. However, if both groups deviate, the resulting equilibrium with mutual discrimination yields lower average payoffs for all individuals in society compared to a colorblind equilibrium. The finding aligns with sociological literature that identifies inter-group competition as a potential driver of inefficient discrimination and in-group favoritism (see, for instance, Bobo and Hutchings (1996); and Bramoullé and Goyal (2016) for a microeconomic model that makes a similar claim).

Of course, in the absence of other commitment devices, the feasibility of out-group discrimination, $A_c = \{(c, c)\}$, depends on whether the conditions in Proposition 4 are met. The information disclosure constraint (IC Out-Group), $\frac{L}{H} > \frac{\pi}{1 + \frac{n - c}{n_c - 1}(1 - \pi)}$, is always met when $\frac{L}{H} > \pi$, ensuring that observers of physical color c report norm violations if the group prefers the norm. As a result, the norm's enforceability depends on the discrimination constraint (DC Out-Group), which determines whether the risk of ostracism from the in-group deters individuals sufficiently from interacting with the out-group.

Importantly, there are parameter regions where the group strictly prefers the colorblind norm ($\frac{L}{H} < \pi$), yet both the information disclosure and discrimination constraints in Proposition 4 are satisfied. In such cases, the out-group discrimination norm $A_c = \{(c, c)\}$ can emerge as a collectively undesirable group norm—one that harms even the in-group yet persists through individually rational collective enforcement. This mirrors the "spontaneous discrimination" equilibria identified by Peski and Szentes (2013), where out-group discrimination arises as an equilibrium outcome despite being collectively suboptimal.

[22] The fact that the other group does not play a role in enforcing the norm may seem counter-intuitive, given that we often associate in-group discrimination with strong incentives imposed by the out-group. However, in this model, the norm followed by each group is strategically independent of the behavior of the other group (Corollary 1). In-group discrimination can still emerge because every individual acts both as a principal and as an agent: By discriminating against her own group and obtaining social color $-c$, an individual acting as the principal in period t can effectively 'transform' herself into an agent of the opposite group. This allows her to benefit individually from the social norm $A_c = \Theta \setminus \{(c, c)\}$, even though her group as a whole is disadvantaged. Following the norm can therefore be individually rational due to the dual role of principal and agent, even in the absence of external incentives from the out-group.

5. Extensions and variations

I discuss some of the central assumptions of the model and the consequences of relaxing them.

5.1. Finite population

The model assumes a finite population, which is crucial for generating the result that information disclosure under norm $A = \{(c, c)\}$ can strictly benefit an observer of physical color c when interactions are competitive. By sending the message $m(t) = -c$, the observer assigns the principal the social color $s_{p(t),t+1} = -c$, rendering her unacceptable as an agent under the norm in future interactions. The key point is that only when a replacement for the rejected agent is drawn from a finite set of agents (including the observer) does this action strictly increase the observer's probability of being selected as an agent, generating an incentive for information disclosure.

In contrast, the model by Peski and Szentes (2013) assumes an infinite population, where opportunities for interaction (involving a principal and a match) arrive independently across individuals over time. To generate a similar incentive for information disclosure in this infinite-population framework, one would need to impose structure on how replacements and observers are selected, such as restricting them to finite subsets (or "neighborhoods") of the population rather than the population at large. Specifically, while initial matches may be random across the entire population, replacements could be drawn from finite neighborhoods around the initial match, and observers could be selected from a finite neighborhood surrounding the principal.[23] In this setup, observers of physical color c would again have an incentive to disclose information, despite the infinite population: In the continuation game, a principal of color c, when matched to the current principal with social color $s = -c$, would seek a replacement from the finite neighborhood around the match. Since this neighborhood includes the current observer, disclosing information strictly increases the observer's chance of being selected as the replacement. While such a "finite neighborhood" model helps illustrate how competition can generate incentives for information disclosure even in infinite populations, the current setup offers a simpler and more tractable framework for analyzing these dynamics.

5.2. One-period decay of social color

The model assumes that social color reverts to the physical color after one period of not being the principal. To study the implications of relaxing this assumption, suppose instead that for any individual other than the principal, social color reverts to the physical color ($s_{i,t+1} = c_i$) with probability γ, and remains unchanged ($s_{i,t+1} = s_{i,t}$) with probability $1 - \gamma$. The original assumption is nested as a special case when $\gamma = 1$.

Consequences for the benchmark The conclusions of the benchmark model (Section 2) remain unchanged, with a slight adjustment to the condition for the out-group discrimination norm $A_c = \{(c, c)\}$ to exist in equilibrium:

$$\pi < \frac{1}{1 - \delta \cdot \frac{n-1}{n}(1-\gamma)} \cdot .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1 - \pi), \tag{DC Benchmark*}$$

where the additional factor $\frac{1}{1-\delta\cdot\frac{n-1}{n}(1-\gamma)}$ appears on the right-hand side.

To derive this condition, note that the same losses that attain in period $t+1$, $\frac{n_c-1}{n(n-1)} \cdot (1-\pi)H$, now attain in any future period $\tau > t$ where $s_{i,\tau} = -c$. This leads to the following equation for the difference in continuation payoffs:

$$V_p(c, c) - V_p(c, -c) = \delta \left[\frac{n_c - 1}{n(n-1)} \cdot (1-\pi)H + \frac{n-1}{n} \cdot (1-\gamma) \cdot \left(V_p(c, c) - V_p(c, -c) \right) \right] \tag{9}$$

where $\frac{n-1}{n} \cdot (1-\gamma)$ is the probability of not being drawn as the principal in period $t+1$, multiplied by the probability that social color remains the same, conditional on not being the principal. Rearranging and solving for $V_p(c, c) - V_p(c, -c)$ gives:

$$V_p(c, c) - V_p(c, -c) = \frac{1}{1 - \delta \cdot \frac{n-1}{n}(1-\gamma)} \cdot \delta \frac{n_c - 1}{n(n-1)} \cdot (1-\pi)H. \tag{10}$$

Plugging this into the equilibrium condition $\pi H < .5[V_p(c, c) - V_p(c, -c)]$ yields the condition (DC Benchmark*).

For $\gamma = 1$, the factor $\frac{1}{1-\delta\cdot\frac{n-1}{n}(1-\gamma)}$ equals 1, recovering the original constraint (DC Benchmark) from Proposition 1. For $\gamma < 1$, the factor increases, relaxing the constraint and expanding the range of parameters for which spontaneous discrimination exists.[24]

[23] To illustrate how this idea relates to real-world scenarios, assume that each opportunity for interaction requires an agent from a specific profession, such as a lawyer, mechanic, or plumber. While the number of professions (and hence individuals) is infinite, the number of agents within each profession might be finite. Competition generates incentives for information disclosure if individuals within the same profession (i.e., neighborhood) are more likely to observe each other's actions and replace one another when a principal requires their services.

[24] In the limit as $\gamma \to 0$, conversions to $s_{p(t),t+1} = -c$ become effectively permanent, and the factor approaches $\frac{1}{1-\delta\cdot\frac{n-1}{n}} \approx \frac{1}{1-\delta}$, analogous to the familiar multiplier on punishments with grim-trigger strategies.

The intuition is that a lower probability of an individual's social color reverting to their physical color extends the effective punishment for attaining social color $-c$ beyond a single period, thereby strengthening incentives to follow the norm. Importantly, γ only affects the evolution of the social colors of individuals who are not the principal in a given period, while the update rule for the principal's social color remains unchanged. Hence, a principal's incentives when deciding whether to select or reject a partner are affected only through altered continuation values, not through any change in how their own social color evolves. As a result, as before, principals of a given physical color have identical incentives to follow the norm, regardless of their current social color.

Breakdown of spontaneous discrimination with endogenous information disclosure The conclusion that discriminatory norms fail in the non-competitive setting under strict incentives for endogenous information disclosure (Proposition 3) remains valid for $\gamma < 1$. In the non-competitive setting, the consequences of sending message $m(t) = -c$ are limited to making the principal an unacceptable match (type $(c, -c)$) for observers of physical color c, thereby reducing the observer's continuation payoff relative to remaining silent ($m(t) = \emptyset$). Specifically, whenever the principal is of type $(c, -c)$ in future periods, an observer matched with this principal will forgo the payoff πH, as they must reject agents of type $(c, -c)$ under the norm $A_c = \{(c, c)\}$. With $\gamma < 1$, the expected duration for which the principal retains the social color $-c$ increases, amplifying these negative consequences and further strengthening the incentives of observers of physical color c to remain silent compared to the case of $\gamma = 1$. Observers of the opposite physical color $-c$, as before, are indifferent between sending $m(t) = -c$ and remaining silent ($m(t) = \emptyset$); by Assumption 1, they choose to remain silent. Consequently, the social norm $A_c = \{(c, c)\}$ cannot be sustained under Observer Reports. Under Self-Reports, the principal herself would incur a loss by sending $m(t) = -c$, making the norm unsustainable in this case as well.

Effects of competition The effects of competition on information disclosure and discrimination become more complex to characterize when social color reverts to physical color with probability $\gamma < 1$ rather than deterministically after one period of not being the principal ($\gamma = 1$). With $\gamma < 1$, type distributions can evolve to include multiple individuals of type $(c, -c)$, complicating the incentives of both principals and observers under norm $A_c = \{(c, c)\}$ in competitive interactions. Specifically, in the competitive setting, these incentives depend on the (future) probability of finding an acceptable replacement after rejecting an agent of type $(c_a, s_a) \neq (c, c)$, as well as the (future) probability of being selected as a replacement when others reject such agents (see Section 4.3 and constraints (IC Out-Group) and (DC Out-Group) in particular). Both probabilities depend on the number of individuals of types (c, c) and $(c, -c)$ in the population, i.e., on the full profile and evolution of social colors among individuals of physical color c.

For the stationary norm $A_c = \{(c, c)\}$ to constitute an equilibrium, it must be optimal for a principal of color c to follow the norm and for an observer of color c to disclose information, given any feasible type profile at period t, the set of which is larger when $\gamma < 1$. For instance, revealing information and rejecting an agent of type $(c_a, s_a) \neq (c, c)$ must be optimal even when no agent of type (c, c) is available as a replacement and when all observers are of type $(c, -c)$. While a formal characterization of the conditions under which these requirements are met lies outside the scope of this paper, the expanded set of feasible type profiles and the associated constraints point to information disclosure and discrimination being harder to sustain under competition when $\gamma < 1$ compared to $\gamma = 1$. For one, the principal must follow norm $A_c = \{(c, c)\}$ even when no replacement of type (c, c) is available, eliminating a central component of the effect of competition on the discrimination constraint (DC Out-Group), which is to reduce the immediate cost of compliance. Additionally, when a majority of observers are of type $(c, -c)$, information disclosure may be hindered. For $\gamma < 1$, these observers have a reduced incentive to send $m(t) = -c$, as they are less likely to revert to an acceptable type and thus be considered a suitable replacement in future interactions where the principal is rejected.

That being said, the case of $\gamma = 1$ studied in the paper is not a knife-edge case. For $\gamma < 1$, but sufficiently close to 1, the continuation payoffs for principals and observers, as specified in Section 4.3, remain approximately unchanged. This implies that observers of physical color c will continue to disclose information under a constraint similar to (IC Out-Group), while principals will follow the norm (even after the most stringent histories) under a constraint (DC Out-Group*) that approaches

$$\pi H < p \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot \left(1 + \frac{n_{-c}}{n_c - 1} \cdot \frac{L}{H}\right)(1 - \pi) \tag{DC Out-Group*}$$

as $\gamma \to 1$. The difference between this condition and the original (DC Out-Group) lies in the left-hand side being πH instead of $\pi(H - L)$, reflecting the fact that in a period where no agent of type (c, c) is available, the principal must forgo the opportunity for interaction to comply with the norm. Hence, while the range of parameters under which the stationary norm $A_c = \{(c, c)\}$ exists becomes narrower, competition still restores spontaneous discrimination under endogenous information disclosure.

5.3. Empty social color

Following the framework of Peski and Szentes (2013), the model assumes that the default social color of individuals is their physical color, i.e., $s_{i,t} = c_i$. At first glance, this assumption might appear crucial to the results. However, under endogenous information disclosure, this assumption is surprisingly innocuous. It is effectively equivalent to assuming that the default social color is *empty*, i.e., contains no information about group affiliation.

To illustrate, consider the model with endogenous information disclosure and introduce an additional neutral (or empty) social color, $s_{i,t} = \emptyset$, alongside $s_{i,t} = c_i$ and $s_{i,t} = -c_i$. Suppose $s_{i,t} = \emptyset$ replaces $s_{i,t} = c_i$ as the default by implementing the following rules. In the initial period $t = 0$, all individuals are assigned $s_{i,0} = \emptyset$. For all individuals who are not the principal in a given period t, the next-period social color is empty ($s_{i,t+1} = \emptyset$). For the principal in period t, the next-period social color is empty if the opportunity is

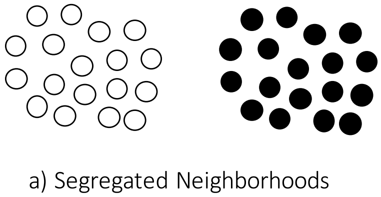

Fig. 5. A population with two neighborhoods.

destroyed; otherwise, it is determined by the message sent about the interaction, $m(t) \in \{c_{a(t)}, -c_{a(t)}, \emptyset\}$, where $m(t) = \emptyset$ results in an empty social color.

Under these modifications, we can state the following result:

Proposition 7 *(Empty social color leaves results unaffected). Consider the model with endogenous information disclosure and assume the default social color is empty. Equilibrium social norms (Propositions 3–5) remain unchanged with the addition that, for any $c_a \in \{c, -c\}$, $(c_a, \emptyset) \in A_c$ if and only if $(c_a, c_a) \in A_c$.*

The intuition for this result lies in the role of the message space in shaping continuation payoffs and social norms in a model with endogenous information disclosure. When an individual is selected as an agent, their physical color $c_{a(t)}$ allows for a message $m(t) = c_{a(t)}$ to be sent regardless of their social color. Consequently, an agent having a social color equal to their physical color, $s_{a(t),t} = c_a(t)$, does not expand the message space available to principals or observers compared to the same agent with an empty social color, $s_{a(t),t} = \emptyset$. Through Assumption 1, both types will induce the same message, and hence be treated equivalently in equilibrium. Specifically, if one type is acceptable under A_c or A_{-c}, the other must also be acceptable under the same norm. Conversely, if one type is rejected, the other is similarly rejected. This equivalence allows the two types, $(c_i, \emptyset)$ and (c_i, c_i), to be treated interchangeably in the analysis.

5.4. Assortative matching and observation

I have assumed that the draw of match $\mu(t)$—as well as in the case of endogenous information disclosure, the draw of observer $o(t)$—is uniform random from the residual population. Given a principal of physical color $c_{p(t)} = c \in \{red, green\}$, the individual drawn as match $\mu(t)$ or observer $o(t)$ is of the same color with probability $\frac{n_c-1}{n-1}$ and of the opposite color with probability $\frac{n_{-c}}{n-1}$. The probability to meet a person of a given physical color is thus equal to the share of that color in the residual population. This assumption is in line with the matching mechanism in Peski and Szentes (2013).

In some instances, however, it may be more realistic to assume that the likelihood to meet a member of a given color systematically deviates from these ratios. For instance, it may be that—for exogenous reasons such as neighborhood structure or group-level correlations of preferences and abilities—individuals $\mu(t)$ and $o(t)$ are disproportionally likely to be of the same physical color as the principal. In such a case the matching and observation processes would be assortative. How does assortativity affect the incentives for discrimination? As a concrete example, assume that a principal is more likely to be matched with (respectively, observed by) an individual in her spatial proximity. Is the propensity for discrimination then higher in a society with segregated neighborhoods or in a society with mixed neighborhoods (see Fig. 5)?

For a formal analysis of assortative matching probabilities, let the probability that $c_{\mu(t)} = c_{p(t)} = c$ be given by a constant $\rho^c \in (0,1)$. Matching is *assortative* if $\rho^c > \frac{n_c-1}{n-1}$ and *disassortative* if $\rho^c < \frac{n_c-1}{n-1}$. Under Observer Reports, the same rule applies to the observer draw, meaning the probability that $c_{o(t)} = c_{p(t)} = c$ is also ρ^c. Conditional on matching with color $c' \in \{c, -c\}$, the probability of being matched with any individual member of that color is assumed to be uniformly random. We can then observe:

Proposition 8 *((Dis-)assortative matching). Assume that matching and observation are (dis-)assortative, with the probabilities for $c_{\mu(t)} = c_{p(t)} = c$ and $c_{o(t)} = c_{p(t)} = c$ given by $\rho^c \in (0,1)$. Under this assumption, the discrimination constraints in Propositions 1, 4, and 5 are modified as follows:*

$$\pi < .5 \cdot \delta \cdot \frac{\rho^c}{n} \cdot (1-\pi) \tag{DC Benchmark}$$

$$\pi\left(1-\frac{L}{H}\right) < q_c \cdot \delta \cdot \frac{\rho^c}{n} \cdot \left(1 + \frac{1-\rho^c}{n_c - 1} \cdot \frac{L}{H}\right)(1-\pi) \tag{DC Out-Group}$$

$$\pi\left(1-\frac{L}{H}\right) < q_c \cdot \delta \cdot \frac{\rho^c}{n} \cdot \left(1 + \frac{\rho^c\left(1-\frac{1}{n_c}\right)}{n_{-c}+1} \cdot \frac{L}{H}\right)(1-\pi) \tag{DC In-Group}$$

where $q_c = \rho^c$ under Observer Reports. The information disclosure constraint in Proposition 4 becomes:

$$\frac{L}{H} > \frac{\pi}{1 + \frac{1-\rho^c}{\rho^c}\frac{n_c}{n_c-1}(1-\pi)}. \tag{IC Out-Group}$$

All other parts of Propositions 1–5 remain unchanged.

We can conclude that assortativity does not qualitatively alter the main results of the model. However, it does make discriminatory norms easier to enforce: note that ρ^c strictly increases the right-hand sides of the discrimination constraints, implying that discriminatory norms become easier to sustain under assortative matching. This result is intuitive, particularly in light of Corollary 1, which establishes that norms in equilibrium are coordinated on and enforced independently within each group $c \in \{red, green\}$. Assortativity implies a higher probability of interaction and observability within the group, which increases the reputational cost of deviation. As individuals are more likely to interact with members of their own color, assortativity also raises the probability of punishment when an individual violates the norm. In a model of neighborhood assortativity (Fig. 5), spatial segregation would similarly result in a higher propensity for norm compliance and, consequently, a higher propensity for discrimination.

6. Conclusion

Discrimination can arise in tolerant societies via the spontaneous coordination of groups on inefficient social norms that deliver reputational rewards to individuals who restrict their interactions to partners of a certain color. For such norms to be sustainable, information about the color of partners needs to be revealed to other members of society. This paper shows how competition for being selected as partner can generate strict incentives for information disclosure. In the presence of such competition, discriminatory social norms yield benefits for one group at the expense of the other. Individuals disclose information about the color of partners in order to gain access to the (preferably small) group that benefits as well as to exclude others from it. Competition can also generate group-incentives for discrimination.

When partner selection is competitive, both out-group and in-group discrimination can emerge as a social norm. While out-group discrimination requires third-party (Observer) reports to be sustainable (Proposition 4), in-group discrimination can also be sustained with Self-Reports (Proposition 5). On one hand, this result speaks for in-group discrimination to be more likely to emerge than out-group discrimination. On the other hand, group-incentives for out-group discrimination (Proposition 6) point toward the opposite conclusion.

Funding

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2126/1 - 390838866.

Declaration of competing interest

There is no conflict of interest.

Acknowledgments

I am grateful to the editor, associate editor, and two anonymous reviewers for their constructive and thoughtful feedback while revising the manuscript. I further thank Leonie Baumann, Gilat Levy, Lydia Mechtenberg, Lennart Struth, Balazs Szentes, and audiences at LSE, the Berlin Micro Theory Seminar, and at the Florence-Konstanz workshop on Social Norms in Multi-Ethnic Societies for valuable comments.

Appendix A. Proofs

Preliminaries To check whether a pair of social norms (A_{red}, A_{green}) defines an equilibrium, consider whether a one-shot deviation by the individual acting as the principal in period t is profitable. Immediate payoffs (in period t) depend solely on the period-t decision of principal $p(t)$. The principal's expected payoffs regarding any future period $\tau > t$, on the other hand, depend solely on the distribution of types in period τ. This is due to the time-invariant random nature of the matching game and equilibrium decisions being dependent only on the type of potential principals and agents. The question about how actions in period t can influence future payoffs is thus equivalent to the question about how actions in period t can influence future type distributions.

Consider the evolution of types from period t to period $t+1$. By assumption, $s_{j,t+1} = c_j$ for all individuals $j \neq p(t)$, implying that their type in period $t+1$ is fixed. The only variable type in period $t+1$ is that of the current principal $p(t)$, whose next period social color $s_{p(t),t+1} \in \{c_{p(t)}, -c_{p(t)}\}$ may depend on period t actions. Period t actions only indirectly impact types beyond $t+1$: Because future employment decisions follow type-dependent equilibrium strategies, the evolution of types from period τ to period $\tau+1$, $\tau > t+1$, depends only on the type-distribution in period τ. The impact on continuation payoffs of period t actions is thus restricted to their impact on the type distribution in period $t+1$, which is variable only in the next-period type of the current principal, $\theta_{p(t),t+1} = (c_{p(t)}, s_{p(t),t+1})$: Let $V_i(t)$ denote the discounted continuation payoff of individual i in period t. In any period t, $V_i(t)$ depends solely on the physical color and the next-period social color of the current principal $p(t)$. We write $V_i(c,s)$ for the continuation payoff of individual i as a function of type $\theta_{p(t),t+1} = (c,s) \in \Theta$.

Assume that there are two actions a and a' that yield immediate payoff $u_i(a)$ and $u_i(a')$, respectively. Assume further that action a is associated with the principal having social color $s_{p(t),t+1} = s$ and action a' is associated with the principal having social color $s_{p(t),t+1} = s'$. Exploiting the one-shot deviation principle, saying that i prefers action a over a' is then equivalent to the statement $u_i(a) + V_i(c,s) \geq u_i(a') + V_i(c,s')$. We are now ready to prove Proposition 1.

Proof of Proposition 1. Consider a period t with a principal $p(t)$ of physical color c. Her continuation payoff is $V_p(c,c)$ if $s_{p(t),t+1} = c$, and $V_p(c,-c)$ if $s_{p(t),t+1} = -c$.

Step 1: $(c,c) \in A_c$. We show that a principal of physical color c never rejects a match of type (c,c). Note that both selecting the match and destroying the opportunity yield $s_{p(t),t+1} = c$ with probability 1 (see Equations (2) and (3)). However, selecting the match provides an immediate payoff of $\pi H > 0$, whereas destroying the opportunity yields zero immediate payoff. Hence, $(c,c) \in A_c$.

Step 2: Possible equilibrium norms are restricted to $A_c \in \{\Theta, \{(c,c)\}, \Theta \setminus \{(-c,-c)\}\}$. Assume there exists a type $(c_a, s_a) \notin A_c$. By step 1, then $-c \in (c_a, s_a)$. The principal rejects a match only if destroying the opportunity yields higher continuation payoffs than selecting the match. Since destroying the opportunity yields $V_p(c,c)$, while accepting a match where $-c \in (c_a, s_a)$ yields $V_p(c,-c)$ with positive probability, it follows that $V_p(c,c) > V_p(c,-c)$.

Since selecting an agent of type $(-c,-c)$ yields $V_p(c,-c)$ with a strictly higher probability than selecting an agent of type $(c,-c)$ or $(-c,c)$, it follows that if any type $(c_a, s_a) \notin A_c$, then $(-c,-c) \notin A_c$ must also hold. Moreover, since selecting $(c,-c)$ or $(-c,c)$ imply identical continuation payoffs, either both are selected or both are rejected. Thus, $(c,-c) \in A_c \Leftrightarrow (-c,c) \in A_c$.

We thus have: If there exists a type $(c_a, s_a) \notin A_c$, then either $A_c = \Theta \setminus \{(-c,-c)\}$ or $A_c = \{(c,c)\}$. Otherwise, $A_c = \Theta$.

Step 3: Existence of $A_c = \Theta \setminus \{(-c,-c)\}$. We show that $A_c = \Theta \setminus \{(-c,-c)\}$ cannot be sustained in equilibrium. Assume, toward a contradiction, that $A_c = \Theta \setminus \{(-c,-c)\}$. For the norm to exist, the principal $p(t)$ must reject a match of type $(-c,-c)$. Fix any norm A_{-c} for color $-c$.

- Case 1: $A_{-c} \in \{\Theta, \{(-c,-c)\}\}$. In this case, agents of type (c,c) as well as agents of type $(c,-c)$ are accepted by both colors, i.e., $\{(c,c),(c,-c)\} \subset A_c \cap A_{-c}$. This implies that transitioning to social color $s = -c$ has no impact on the continuation payoffs of the principal relative to transitioning to social color $s = c$. Consequently, $V_p(c,c) = V_p(c,-c)$. But then, selecting an agent of type $(-c,-c)$ yields the same continuation payoff as rejecting the agent, implying that the principal will select $(-c,-c)$. This contradicts the assumption that $(-c,-c) \notin A_c$.
- Case 2: $A_{-c} = \Theta \setminus \{(c,c)\}$. In this case, agents of type (c,c) are rejected by principals of color $-c$, while agents of type $(c,-c)$ are accepted, i.e., $(c,c) \notin A_{-c}$ and $(c,-c) \in A_{-c}$. For principals of color c, both (c,c) and $(c,-c)$ are accepted, i.e., $\{(c,c),(c,-c)\} \subset A_c$. Under these norms, transitioning to social color $s = -c$ increases the continuation payoffs of the principal relative to transitioning to $s = c$, as agents of type (c,c) are rejected in future interactions with principals of color $-c$, but agents of type $(c,-c)$ are not. Hence, $V_p(c,c) < V_p(c,-c)$. But then, selecting an agent of type $(-c,-c)$ yields a higher continuation payoff than rejecting the agent, implying that the principal will select $(-c,-c)$. This contradicts the assumption that $(-c,-c) \notin A_c$.

Thus, $A_c = \Theta \setminus \{(-c,-c)\}$ cannot exist in equilibrium.

Step 4: Existence of $A_c = \Theta$. We show that the colorblind norm $A_c = \Theta$ always exists. Fix $A_{-c} \in \{\Theta, \{(-c,-c)\}\}$ and assume $A_c = \Theta$. By definition of $A_c = \Theta$, it follows that $\{(c,c),(c,-c)\} \subset A_c$. Since $(c,c) \in A_{-c} \Leftrightarrow (c,-c) \in A_{-c}$, the continuation payoffs from transitioning to social colors $s = c$ and $s = -c$ are identical, i.e., $V_p(c,c) = V_p(c,-c)$. Because the continuation payoff is independent of the evolution of her social color, a principal of color c selects any type of agent, confirming that $A_c = \Theta$ is an equilibrium norm.

Step 5: Existence of $A_c = \{(c,c)\}$. We show that the discriminatory norm $A_c = \{(c,c)\}$ exists if and only if (DC Benchmark) is satisfied. Fix $A_{-c} \in \{\Theta, \{(-c,-c)\}\}$ and assume $A_c = \{(c,c)\}$. We then have $(c,c) \in A_c$, but $(c,-c) \notin A_c$, and $(c,c) \in A_{-c} \Leftrightarrow (c,-c) \in A_{-c}$. This implies that $V_p(c,c) > V_p(c,-c)$.

Given $V_p(c,c) > V_p(c,-c)$, selecting an agent of type $(-c,-c)$ leads to a strictly greater loss in continuation payoffs compared to selecting types $(c,-c)$ or $(-c,c)$. Thus, the principal's decision to comply with the norm $A_c = \{(c,c)\}$ hinges on rejecting matches of types $(c,-c)$ and $(-c,c)$.

Selecting a match of type $(c,-c)$ or $(-c,c)$ provides an immediate payoff of πH and continuation payoffs of $0.5 \cdot V_p(c,c) + 0.5 \cdot V_p(c,-c)$, while rejecting the match yields no immediate payoff and continuation payoffs of $V_p(c,c)$. The principal will reject such matches if and only if

$$\pi H + 0.5 \cdot V_p(c,c) + 0.5 \cdot V_p(c,-c) < V_p(c,c) \Leftrightarrow \pi H < .5 \left[V_p(c,c) - V_p(c,-c)\right].$$

To calculate the difference in continuation payoffs $[V_p(c,c) - V_p(c,-c)]$, note that losses are confined to one period, as the principal's social color reverts permanently to $s = c$ after period $t+1$. If the current principal $p(t)$ is selected again as the principal in period $t+1$, equilibrium play ensures $s_{p(t),t+2} = c$. Otherwise, her social color reverts to her physical color by default, yielding $s_{p(t),t+2} = c$.[25]

[25] See the Discussion for how the constraint (DC Benchmark) changes when the social color of the principal does not revert to her physical color with certainty after period $t+1$.

Social norm $A_c = \{(c,c)\}$ implies that individual $p(t)$ will be rejected as a match in period $t+1$ by any principal of physical color c if $s_{p(t),t+1} = -c$. The probability that a different player is selected to be the principal in the next period and has physical color c (i.e., $c_{p(t+1)} = c$ and $p(t+1) \neq p(t)$) is $\frac{n_c - 1}{n}$. The conditional probability that principal $p(t)$ is selected to be the match in the next period is $\frac{1}{n-1}$. The foregone payoff from rejection is $(1-\pi)H$. The expected loss in period $t+1$ thus calculates to $\frac{n_c-1}{n(n-1)} \cdot (1-\pi)H$.

Thus,

$$[V_p(c,c) - V_p(c,-c)] = \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1-\pi)H.$$

Substituting this result into the inequality, the principal will reject matches of type $(c,-c)$ and $(-c,c)$, thereby sustaining norm $A_c = \{(c,c)\}$, if and only if:

$$\pi H < .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1-\pi)H,$$

or equivalently:

$$\pi < .5 \cdot \delta \cdot \frac{n_c - 1}{n(n-1)} \cdot (1-\pi). \quad \square \tag{DC Benchmark}$$

Proof of Lemma 1. Provided as part of the main text (below Lemma 1). □

Proof of Proposition 2. Provided as part of the main text (above Proposition 2). □

Proof of Proposition 3. By Proposition 2, possible equilibrium norms are restricted to $A_c \in \{\Theta, \{(c,c)\}, \Theta \setminus \{(c,c)\}\}$ for $c \in \{red, green\}$. Consider a principal of color c. Fix $A_{-c} \in \{\Theta, \{(-c,-c)\}, \Theta \setminus \{(-c,-c)\}\}$. Since $(c,c) \in A_{-c} \Leftrightarrow (c,-c) \in A_{-c}$, the principal's best response to (A_c, A_{-c}) is independent of the norm adopted by the opposite color (Corollary 1).

The color-blind norm $A_c = \Theta$ always exists. Under norm $A_c = \Theta$, the evolution of the principal's social color does not affect her continuation payoffs since $\{(c,c),(c,-c)\} \subset A_c$, making always selecting the match irrespective of the agent's type the unique best response. Since the norm does not require information disclosure, it exists independently of messages $m(t)$. For the remainder of the argument (proving that $A_c = \Theta \setminus \{(c,c)\}$ and $A_c = \{(c,c)\}$ do not exist), see the main text (below Proposition 3). □

Proof of Lemma 2. Condition (IC Out-Group), which forms the Lemma, is derived in the main text, directly preceding Lemma 2. □

Proof of Proposition 4. First note that Lemma 1 still holds (the same proof applies). Hence, by Proposition 2, possible equilibrium norms are restricted to $A_c \in \{\Theta, \{(c,c)\}, \Theta \setminus \{(c,c)\}\}$ for $c \in \{red, green\}$. Consider a principal of color c. Fix $A_{-c} \in \{\Theta, \{(-c,-c)\}, \Theta \setminus \{(-c,-c)\}\}$. Since $(c,c) \in A_{-c} \Leftrightarrow (c,-c) \in A_{-c}$, the principal's best response to (A_c, A_{-c}) is independent of the norm adopted by the opposite color (Corollary 1).

The color-blind norm $A_c = \Theta$ always exists (see the proof of Proposition 3). The remainder of the proof consists of deriving the conditions for norm $A_c = \{(c,c)\}$ to exist. These are derived in the main text, directly preceding Proposition 4. □

Proof of Proposition 5. The initial steps of the proof, up to establishing that the color-blind norm $A_c = \Theta$ always exists, are identical to those in the proof of Proposition 4, provided above. The remainder of the proof consists of deriving the conditions for norm $A_c = \Theta \setminus \{(c,c)\}$ to exist. These are derived in the main text, directly preceding Proposition 5. □

Proof of Proposition 6. Consider the expected payoff of an individual i of color c on the equilibrium path. The expected payoff of the individual in any given period t is determined from

$$E\left[u_i(t) \,|\, c_i = c\right] = \frac{n_c}{n} \cdot E\left[u_i(t) \middle| c_{p(t)} = c\right] + \frac{n_{-c}}{n} \cdot E\left[u_i(t) \middle| c_{p(t)} = -c\right],$$

where the first term is the expected payoff associated with interactions where the principal in period t is an individual of color c and the second term is that associated with interactions where the principal in period t is an individual of the opposite color $-c$.

In equilibrium, payoff component $E[u_i(t) \,|\, c_{p(t)} = -c]$ is governed by the social norm of the opposite group, which by Proposition 2 takes $A_{-c} \in \{\Theta, \{(-c,-c)\}, \Theta \setminus \{(-c,-c)\}\}$. Note that $(c,c) \in A_{-c} \Leftrightarrow (c,-c) \in A_{-c}$, and hence, these payoffs do not depend on the variable component (social color) of the type of individual i and are thus independent of the social norm that group c follows. Social norm A_c affects only payoff component $E[u_i(t) \,|\, c_{p(t)} = c]$. Denote the average (across time) of this component $\bar{u}^c$.

If group c follows a colorblind social norm, $A_c = \Theta$, then

$$\bar{u}^c(\text{color-blind}) = \frac{1}{n_c} \cdot \pi H + \frac{n_c - 1}{n_c} \cdot \frac{1}{n-1}(1-\pi)H.$$

The first term accounts for the case that individual i is the principal: If $A_c = \Theta$, she always selects her match as agent, earning payoff πH. The second term accounts for the case that another individual of group c is the principal. In that case, individual i earns payoff $(1-\pi)H$ if she happens to be the match of period t, which happens with conditional probability $\frac{1}{n-1}$.

Assume that, instead, group c follows a social norm of out-group discrimination $A_c = \{(c,c)\}$. It that case, we have

$$\bar{u}^c(\text{out-group discrimination}) = \frac{1}{n_c} \cdot \left(\pi H - \frac{n_{-c}}{n-1} \cdot \pi(H-L)\right) + \frac{n_c - 1}{n_c} \cdot \frac{1}{n-1}\left((1-\pi)H + \frac{n_{-c}}{n_c - 1}(1-\pi)L\right),$$

where the term $-\frac{n_{-c}}{n-1} \cdot \pi(H-L)$ accounts for the loss in income on the side of the principal when having to reject an out-group match and the term $+\frac{n_{-c}}{n_c-1}(1-\pi)L$ accounts for the added income on the side of in-group agents who will be selected when out-group agents are rejected. Given the social norm of the opposite color, A_{-c}, out-group discrimination *on average* yields a strictly higher payoff for individual i if and only if $\bar{u}^c(\text{out-group discrimination}) > \bar{u}^c(\text{color-blind})$, that is, if and only if $\frac{L}{H} > \pi$. □

Proof of Proposition 7. Consider any period t, and a principal of physical color $c \in \{red, green\}$. Assume the principal selects an agent, $a(t) \neq \emptyset$, with $c_a \in \{c, -c\}$. Note that types (c_a, c_a) and $(c_a, \emptyset)$ induce the same message space, $m(t) \in \{c_a, \emptyset\}$. By Assumption 1, these types generate identical messages in equilibrium, implying identical continuation payoffs for the principal. Therefore, $(c_a, c_a) \in A_c$ if and only if $(c_a, \emptyset) \in A_c$. □

Proof of Proposition 8. Omitted. The result follows directly from incorporating ρ^c into the proofs of Propositions 1–5. □

Data availability

No data was used for the research described in the article.

References

Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. Q. J. Econ. 115 (3), 715–753.
Ali, S.N., Miller, D.A., 2016. Ostracism and forgiveness. Am. Econ. Rev. 106 (8), 2329–2348.
Arrow, K.J., 1973. The theory of discrimination. In: Ashenfelter, O.C., Rees, A.E. (Eds.), Discrimination in Labor Markets. Princeton University Press, Princeton, pp. 3–33.
Becker, G.S., 1957. The Economics of Discrimination. The University of Chicago Press, Chicago.
Bénabou, R., Tirole, J., 2012. Laws and Norms. IZA Discussion Paper No. 6290.
Bobo, L., Hutchings, V., 1996. Perceptions of racial group competition: extending Blumer's theory of group position to a multiracial social context. Am. Sociol. Rev. 61 (6), 951–972.
Bramoullé, Y., Goyal, S., 2016. Favoritism. J. Dev. Econ. 122, 16–27.
Choy, J.P., 2017. Social divison with endogenous hierarchy. Econ. J. 128 (615), 2711–2742.
Coate, S., Loury, G.C., 1993. Will affirmative-action policies eliminate negative stereotypes? Am. Econ. Rev. 83 (5), 1220–1240.
Dal Bó, P., 2007. Social norms, cooperation and inequality. Econ. Theory 30 (1), 89–105.
Darity, W.A.J., Williams, R.M., 1985. Peddlers forever?: Culture, competition, and discrimination. Am. Econ. Rev. 75 (2), 256–261.
Eeckhout, J., 2006. Minorities and endogeneous segregation. Rev. Econ. Stat. 73, 31–53.
Eguia, J.X., 2017. Discrimination and assimilation. J. Public Econ. 156, 48–58.
Holzer, H.J., Ihlanfeldt, K.R., 1998. Customer discrimination and employment outcomes for minority workers. Q. J. Econ. 113 (3), 835–867.
Kandori, M., 1992. Social norms and community enforcement. Rev. Econ. Stat. 59, 63–80.
Lang, K., Lehmann, J.-Y.K., 2012. Racial discrimination in the labor market: theory and empirics. J. Econ. Lit. 50 (4), 959–1006.
Mailath, G.J., Samuelson, L., 2006. Repeated Games and Reputations. Oxford University Press, New York.
Mailath, G.J., Samuelson, L., Shaked, A., 2000. Endogenous inequality in integrated labor markets with two-sided search. Am. Econ. Rev. 90 (1), 46–72.
McAdams, R.H., 1995. Cooperation and conflict: the economics of group status production and race discrimination. Harvard Law Rev. 108 (5), 1003–1084.
Onuchic, P., 2024. Recent contributions to theories of discrimination. Unpublished Working Paper (this version: September 2024), available at: https://static1.squarespace.com/static/5a844a1b268b9698c654e9c3/t/66e7fb9725b9b9507764d426/1726479256021/ThDisc2024.pdf. (Accessed 6 August 2025).
Onuchic, P., Ray, D., 2023. Signaling and discrimination in collaborative projects. Am. Econ. Rev. 113 (1), 210–252.
Peski, M., Szentes, B., 2013. Spontaneous discrimination. Am. Econ. Rev. 103 (6), 2412–2436.
Phelps, E.S., 1972. The statistical theory of racism and sexism. Am. Econ. Rev. 62 (4), 659–661.
Ramachandran, R., Rauh, C., 2018. Discrimination without taste: how discrimination can spillover and persist. SERIEs 9, 249–274.
Sethi, R., Somanathan, R., 2004. Inequality and segregation. J. Polit. Econ. 112 (6), 1296–1321.
Small, M., Pager, D., 2020. Sociological perspectives on racial discrimination. J. Econ. Perspect. 34, 49–67.