Heidhues, Paul; Kĩoszegi, Botond; Strack, Philipp

Working Paper

Overconfidence and prejudice

ECONtribute Discussion Paper, No. 316

Provided in Cooperation with:
Reinhard Selten Institute (RSI), University of Bonn and University of Cologne

Suggested Citation: Heidhues, Paul; Kĩoszegi, Botond; Strack, Philipp (2024) : Overconfidence and prejudice, ECONtribute Discussion Paper, No. 316, University of Bonn and University of Cologne, Reinhard Selten Institute (RSI), Bonn and Cologne

This Version is available at:
https://hdl.handle.net/10419/301022

ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 316

Overconfidence and Prejudice

Paul Heidhues Botond Kőszegi Philipp Strack

June 2024 www.econtribute.de

UNIVERSITÄT BONN Universität zu Köln

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

Overconfidence and Prejudice*

Paul Heidhues
DICE, Heinrich-Heine University Düsseldorf

Botond Kőszegi
briq

Philipp Strack
Yale University

January 13, 2023

Abstract

We develop a model in which an overconfident agent learns about groups in society from observations of his and others' successes. In our model, both the agent's information and his beliefs are multi-dimensional, allowing us to study interactions between different views. Overall, society always exhibits an in-group bias — the average person sees his group relative to other groups too positively — but this general tendency toward prejudice exhibits systematic comparative-statics patterns. First, a person is most likely to have negative opinions about other groups he competes with. Second, while information about another group's achievements does not lower a person's prejudice, information about economic or social forces affecting the group can, and personal contact with group members has a beneficial effect that is larger than in classical settings. Third, the agent's beliefs are subject to "bias substitution," whereby forces that decrease his bias regarding one group tend to increase his biases regarding unrelated other groups. Methodologically, to make our analysis of interdependent multi-dimensional beliefs possible, we develop tools for studying learning under high-dimensional misspecified models.

Keywords: beliefs, prejudice, inter-group beliefs, overconfidence, misspecified learning

JEL codes: D01, D83, D91

*We are grateful to Aislinn Bohren, In-Koo Cho, Alex Imas, Jurjen Kamphorst, Robert Lieli, Ulrike Malmendier, Josh Schwartzstein, Florian Zimmermann, and Asaf Zussman for insightful discussions, and seminar and conference audiences for comments. Anna Vályogos provided excellent research assistance. Heidhues and Kőszegi thank the European Research Council for financial support under Grant #788918. Heidhues and Strack thank briq for its hospitality.

1 Introduction

Individuals' beliefs about each other are crucial determinants of social and economic behavior and hence of the fairness and efficiency of outcomes. While the typical assumption in economics is that beliefs are correct given available information, a growing literature recognizes the possibility that individuals have incorrect beliefs about others (Bordalo et al., 2016, Heidhues et al., 2018, Bohren et al., 2019a,b, Hestermann and Le Yaouanq, 2021, Chauvin, 2020, Frick et al., 2022). In particular, theoretical work has begun to explore how false social beliefs can arise because a person makes inferences using an incorrect, "misspecified" model of the world, and empirical work documents instances of false social beliefs.[1]

We build on this research agenda to develop a theory of prejudiced inter-group beliefs, making three contributions to the economics literature. To start, we provide the first general explanation for one of the most central stylized facts about inter-group beliefs, (relative) in-group bias — that the average person sees his group relative to other groups too positively. In addition, we incorporate into our model the reality that individuals' information as well as their social beliefs are richly multi-dimensional. Indeed, people may simultaneously hold false beliefs about groups of rather different nature, such as blacks, women, rich Jews, or poor immigrants. We study how different types of information affect beliefs and how multiple beliefs interact, identifying spillovers that can help account for empirical patterns observed by researchers. A recurring theme is what we call "bias substitution," whereby forces that decrease one bias tend to increase other biases. Finally, to make our analysis of interdependent multi-dimensional beliefs possible in the first place, we develop tools for studying learning under high-dimensional misspecified models. Due to the lack of such tools, analysis of misspecified learning has typically focused on misinferences about a single-dimensional state of the world.

We begin in Section 2 with the tools mentioned above. We consider an agent who repeatedly observes, with multivariate normal noise, linear combinations of finitely many fundamentals. He has a dogmatic prior about one fundamental, but he is agnostic about the other fundamentals as well as the covariance matrix of the errors, and updates according to Bayes' Rule. Expressing the question as a semidefinite programming problem, we derive a formula for the agent's long-run bias about the fundamentals and the covariance matrix. As far as we know, this is the first closed-

[1] We cite evidence for empirical claims we make in the introduction when presenting our formal results below.

form solution to a general learning process that features a high-dimensional misspecified model and high-dimensional observations. Curiously, although the agent misinfers the covariance matrix, his long-run beliefs about the fundamentals are the same as when he correctly understands the covariance matrix. This can simplify the understanding of his beliefs about the fundamentals. Our formula is essential for all of our subsequent results, and might also be a useful general result for other researchers.

In Section 3, we turn to our model of social beliefs. Society is composed of individuals in disjoint groups. The agent makes many independent noisy observations of the "recognition" — i.e., achievement, social status, or other measure of success — of each individual, including himself. He understands that recognition depends in part on the "caliber" — i.e., ability, hard work, or other measure of deservingness — of a person. But he allows for the possibility that various types of "discrimination" — or policies or economic forces — affect recognition as well. Each type of discrimination redistributes recognition between groups according to fixed proportions, which we can think of as deriving from an underlying competition structure between individuals. While the agent knows the proportions, he does not know the degrees of discrimination, so he does not know how much of the redistribution is going on. And while he receives independent unbiased signals of the degrees of discrimination, this information may be poor.

Crucially, to these ingredients we add a single non-classical but empirically well-founded assumption. Namely, the agent observes society while maintaining stubborn, unrealistically positive views about *himself*. Formally, we model such stubborn overconfidence by assuming that the agent has a point belief about his own caliber that is above the correct one. Otherwise, however, the agent is agnostic and rational: he starts from a full-support prior, and updates his beliefs about the degrees of discrimination and individuals' calibers using Bayes' Rule.

In Section 4, we identify properties of the agent's long-run beliefs, beginning with two widely documented patterns. The first derives from a force identified by Heidhues et al. (2018) and Hestermann and Le Yaouanq (2021) in other environments: that the agent considers many of his outcomes to be worse than he expects, and misattributes the disappointing observations to external factors. In our setting, this leads him to overestimate discrimination against his group and to underestimate discrimination in favor of his group. As a result, different groups hold divergent, self-serving beliefs about discrimination. For instance, whites consider discrimination against blacks

a less serious issue than blacks do. Going further, because the agent interprets others' outcomes in light of his misestimates about discrimination, he develops excessively positive opinions about others in his group, but not of the population at large. As a result, society exhibits relative in-group bias: the opinion an average person holds about others in his group relative to the average other group is positively biased. For instance, while our model is consistent with an observation that women rate men higher than women, it predicts that then men do so even more. Previous models do not robustly predict such a pattern.

Beyond accounting for divergent views about discrimination and in-group bias, our theory makes a rich set of comparative-statics predictions on the strength and pattern of biases. One set of insights centers around the effects of competition. Suppose that a new group of immigrants moves to the agent's neighborhood, and he now finds himself on opposite sides of a social or economic issue with them. Formally, a new type of discrimination pitting the groups against each other arises. Because the agent underestimates discrimination against, or overestimates discrimination in favor of, the new group, his opinion of the group decreases. This insight helps explain why factors such as the presence of other ethnic groups in one's city, immigration to one's vicinity, and perceived competition with a group increase prejudice. More subtly, competition provides the agent another explanation for his low recognition. This "excuse effect" transfers to other members of the agent's group, so his opinion of his own group becomes even more positively biased. Furthermore, we show that the agent's bias regarding all groups not affected by the new form of discrimination decreases. Intuitively, armed with a new explanation for his low recognition, the agent's need for other explanations diminishes. This bias substitution provides a beliefs-based mechanism for how focusing on a competitor outside group — a common political tactic — can help unify a population hitherto riddled with mutual prejudice.

Another set of insights concerns the effects of information. As a benchmark, note that in a correctly specified model in which sufficient information is available to the agent, beliefs converge to the truth, so better information does not affect long-run beliefs. In our framework, better information can, depending on its nature, have neutral, beneficial, detrimental, or mixed effects on a person's long-run prejudices. On the neutral side, better information about a group's recognition does not affect biases at all, as the agent can perfectly explain groups' recognitions through his conclusions about discrimination. On the detrimental side, increasing the precision of the agent's

recognition increases all of his biases. Intuitively, the agent attributes part of his low recognition to bad luck, but as less noise makes this less plausible, the need for social explanations increases.

More promisingly, the indirect approach of providing better information about a type of discrimination that affects the agent has a range of positive effects. It lowers his bias about his own group as well as about any group also affected by the discrimination, and it improves his opinion about the average other group. But in another instance of bias substitution, the same information increases the agent's bias regarding any group not affected by the discrimination.

Relatedly, our theory provides a novel perspective on the influential and well-documented contact hypothesis (Allport, 1954), which says that contact with an individual from a different racial group can lower prejudice. Plausibly, one main effect of such contact is that the agent learns the caliber of the individual. This provides information about discrimination, and hence lowers his bias regarding all of his contact's group. Furthermore, the positive spillover to others occurs even if the agent had already had plenty of information about the group as a whole. Hence, in a sense our model predicts a stronger positive effect of contact than does a model of correctly specified learning. In such a conventional framework, information about one person often has a small effect on beliefs about a large group. But bias substitution operates here too, so that personal contact with one group exacerbates the agent's prejudice against unrelated groups.

Because it is empirically relevant and leads to additional insights, we also analyze a special case of our model in which groups are defined by vectors of socioeconomic characteristics (e.g., gender, race), and discrimination operates along the same characteristics. We identify conditions under which a stronger version of in-group bias holds: the agent has a more positive bias about individuals who share more of his characteristics. This "similarity bias" can occur even when the agent competes more strongly with more similar individuals. For example, suppose that the agent is a white man, and he competes more with whites than with blacks and with men than with women. Although the pattern of competition drives his prejudices, he will often still think most highly of white men, less highly of white women and black men, and least highly of black women.

In Section 5, we consider variants of our basic model. Most importantly, we demonstrate that our framework's central mechanism can be operational even when the agent neither entertains the possibility of discrimination, nor thinks of society in terms of distinct groups. Suppose that individual j's recognition is the sum of j's caliber, a mean-zero common shock scaled by c_j, and

a mean-zero idiosyncratic shock. The agent does not know the effects of the common shock, c_j, which could be different across individuals and could be positive or negative. He uses observations of everyone's recognitions to update about individuals' calibers as well as the c_j. We show that the agent develops a positive bias about individuals whose c_j has the same sign as his, and a negative bias about individuals whose c_j has the opposite sign. In addition, he correctly learns the signs but overestimates the absolute values of the c_j's. These results can be interpreted as saying that endogenous in- and out-groups develop based on who is in the "same boat" with the agent, and the agent exaggerates the importance of groups in determining outcomes. We also analyze a model in which the agent's beliefs about himself are not fixed, but he interprets observations about himself in a positively biased way. We show that he develops overconfidence, and all of our insights regarding his other beliefs survive. This version of the model, however, has additional comparative-statics implications regarding the level of overconfidence.

We discuss related literature in Section 6. While a few theories have implications for beliefs about groups, no previous paper derives a general in-group bias, makes predictions regarding spillovers between multiple interdependent incorrect beliefs about others, or develops a theory of group beliefs based on overconfidence. Despite these novel aspects, however, our theory of course does not explain all types of prejudices and biases. North Korean citizens hate the United States not because they have observed and interpreted the countries' outcomes, but because they are constantly bombarded with anti-American propaganda. This situation is better described by Glaeser's (2005) theory that politicians supply, and citizens often passively accept, hate-creating stories that complement desired policies. Many stereotypes, such as the view that Swedes are tall and blond, are about specific, less value-laden characteristics than our notion of caliber. Here, Bordalo et al.'s (2016) theory that individuals exaggerate distinctive true differences between groups is a compelling account. Although individuals tend to have more negative views about competing groups, they often also have prejudices about groups they are not in any tangible competition with. And as in the case of Republicans and Democrats, distorted views about each other can derive from disagreements about basic social issues.

In Section 7, we mention some questions that are unaddressed by our current framework. While we have restricted attention to studying beliefs, an obvious question is how the biases we identify affect behavior, especially discriminatory behavior. Of the two main economic approaches to

stereotypes and discrimination, our formal theory is closer to statistical models than to taste-based models, and can be thought of as generating misspecified statistical discrimination. But some of our key predictions, such as that the agent has negative biases about other groups, can lead to objectively unfounded bad treatment of others. Hence, our model can also be thought of as a statistically based microfoundation for taste-based models.

2 Theoretical Tools for Multi-Dimensional Misspecified Learning

In this section, we derive a theoretical result that we will apply in multiple ways to analyze our main models, and that might be useful for other researchers studying implications of overconfidence or other misspecifications. To the best of our knowledge, our characterization is the first closed-form solution for the long-run outcome of a general learning process with misspecification and arbitrarily high-dimensional interdependent beliefs.[2] Readers uninterested in our abstract result can skip to Section 3.

The agent makes inferences about a fixed L-dimensional vector of *fundamentals*

$$f \in \mathbb{R}^L,$$

whose realization we denote by $F \in \mathbb{R}^L$. Each period t, the agent observes a D-dimensional *signal*

$$r_t = Mf + \epsilon_t\,,$$

where $M \in \mathbb{R}^{D\times L}$ is a matrix and $\epsilon_t \in \mathbb{R}^D$ is a vector of errors that is jointly normally distributed with mean zero and positive definite covariance matrix Σ. We assume that M has rank L; otherwise, there would be different vectors of fundamentals that entail the same distribution of signals and hence the agent could not learn the fundamentals even with access to infinite data.

The agent observes a sequence of realizations of r_t, with the ϵ_t drawn independently over time. He updates his beliefs using Bayes' rule: given a prior belief $\mathbf{P}_0$ over the set of fundamentals and positive definite covariance matrices, the probability that his posterior belief $\mathbf{P}_t$ assigns to the set A after seeing the the sequence of signals $r = (r_1, r_2, \ldots, r_t)$ is given by

$$\mathbf{P}_t A = \frac{\int \mathbf{1}_{(f',\Sigma')\in A}\ell_t(r|f',\Sigma')d\mathbf{P}_0(f',\Sigma')}{\int \ell_t(r|f',\Sigma')d\mathbf{P}_0(f',\Sigma')},$$

[2] Spiegler (2016, 2020) also develops and solves in closed form models of high-dimensional interdependent misspecified inferences. These models are not based on an explicit learning process, and their economic logic and solution methods are completely different from ours. As we note in Section 6, other papers on misspecified learning typically only solve one- or two-dimensional models.

where the likelihood equals

$$\ell_t(r|f',\Sigma') = \prod_{z=1}^{t} \frac{1}{\sqrt{(2\pi)^L \det \Sigma'}} \exp\left(-\frac{1}{2}(r_z - Mf')^T \Sigma' (r_z - Mf')\right). \tag{1}$$

In making his inferences, the agent is misspecified in a particular sense: while the true value of fundamental i is F_i, he believes with certainty that it is $\tilde{f}_i$. We consider three different inference problems depending on which parts of the agent's beliefs are fixed by his prior belief, and which are derived from his observations. We denote by $\mathcal{M}$ set of positive definite symmetric matrices where all eigenvalues are greater $\underline{\lambda}$, and $\underline{\lambda}$ is chosen small enough.[3] In our preferred specification, the agent is trying to infer the fundamentals f as well as the covariance matrix Σ:

$$\operatorname{supp} \mathbf{P}_0 = \left\{(f',\Sigma') \in \mathbb{R}^L \times \mathbb{R}^{D\times D} \colon f'_i = \tilde{f}_i, \Sigma' \in \mathcal{M}\right\}. \tag{Case III}$$

Because they are potentially of interest in other applications, we also consider two simpler inference problems. We ask what the agent infers about the fundamentals when his beliefs about the covariance matrix are fixed at some positive definite $\tilde{\Sigma}$, so that

$$\operatorname{supp} \mathbf{P}_0 = \left\{(f',\Sigma') \in \mathbb{R}^L \times \mathbb{R}^{D\times D} \colon f'_i = \tilde{f}_i, \Sigma' = \tilde{\Sigma}\right\}. \tag{Case I}$$

And we ask what the agent infers about the covariance matrix when his beliefs about *all* fundamentals are fixed at $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_L)^T$, so that

$$\operatorname{supp} \mathbf{P}_0 = \left\{(f',\Sigma') \in \mathbb{R}^L \times \mathbb{R}^{D\times D} \colon f' = \tilde{f}, \Sigma' \in \mathcal{M}\right\}. \tag{Case II}$$

We say that the agent's beliefs *concentrate on a point* $(\tilde{f},\tilde{\Sigma})$ if for every open set A such that $(\tilde{f},\tilde{\Sigma}) \in A$, almost surely the agent will in the limit assign probability 1 to A: $\mathbb{P}[\lim_{t\to\infty} \mathbf{P}_t A = 1] = 1$. For stating our theorem, note that any positive definite covariance matrix $\tilde{\Sigma}$ is invertible, so the matrix $M^T\tilde{\Sigma}^{-1}M$ is well-defined; and since M has rank L, this matrix is positive definite and hence invertible.

Theorem 1 (Long-Run Beliefs)**.** *In Cases (I), (II), and (III), the agent's beliefs concentrate on a single point* $(\tilde{f},\tilde{\Sigma})$*. Furthermore:*

[3]Formally, one can choose any $\underline{\lambda}$ less than the smallest eigenvalue of $\Sigma + (M(\tilde{f} - F))(M(\tilde{f} - F))^T$, where $\tilde{f}$ is given exogenously in Case (II); and equals $\tilde{f}_j = F_j + \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{ji}}{\left[M^T\Sigma^{-1}M\right]^{-1}_{ii}}(\tilde{f}_i - F_i)$ for $j \neq i$ in Case (III). The long-run beliefs of the agent do not dependent of the precise choice of $\underline{\lambda}$ as long as it is small enough.

(I) If the agent has fixed beliefs $\tilde{\Sigma}$ about the covariance matrix but is uncertain about the fundamentals $j \neq i$, then in the limit his bias about fundamental j is

$$\tilde{f}_j - F_j = \frac{(M^T\tilde{\Sigma}^{-1}M)^{-1}_{ij}}{(M^T\tilde{\Sigma}^{-1}M)^{-1}_{ii}}(\tilde{f}_i - F_i). \tag{2}$$

(II) If the agent has fixed beliefs $\tilde{f}$ about the fundamentals but is uncertain about the covariance matrix, then in the limit his bias about the covariance matrix is

$$\tilde{\Sigma} - \Sigma = (M(\tilde{f} - F))(M(\tilde{f} - F))^T. \tag{3}$$

(III) If the agent is uncertain about both the fundamentals $j \neq i$ and the covariance matrix, then in the limit his bias about fundamental j is

$$\tilde{f}_j - F_j = \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{ji}}{[M^T\Sigma^{-1}M]^{-1}_{ii}}\,(\tilde{f}_i - F_i), \tag{4}$$

and his bias about the covariance matrix is given by Expression (3)*.*

The proof of Theorem 1 proceeds as follows. First, we verify that the assumptions necessary to apply the seminal result of Berk (1966) are satisfied for our multidimensional Normal model with an unknown covariance matrix and mean. Then, by Berk's result, beliefs concentrate on the set of minimizers of the Kullback-Leibler divergence. Intuitively, the negative of the Kullback-Leibler divergence is increasing in the objective expectation of the subjective log-likelihood the agent assigns to his observations, so it is a natural measure of how likely a vector of fundamentals is in the agent's view in the long run. Due to our assumption of normal signals, the Kullback-Leibler divergence $D\big(F, \Sigma \,\big|\big|\, \hat{f}, \hat{\Sigma}\big)$ assigned to the parameters $(\hat{f}, \hat{\Sigma})$ when the true parameters equal (F, Σ) is

$$D\big(F, \Sigma \,\big|\big|\, \hat{f}, \hat{\Sigma}\big) = \frac{1}{2}\left(\operatorname{tr}(\hat{\Sigma}^{-1}\Sigma) + (M(\hat{f} - F))^T\hat{\Sigma}^{-1}M(\hat{f} - F) - n + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right).$$

The proof then derives the unique minimizer of the above expression over the support specified in Cases (I), (II), and (III) using properties of the trace, Kronecker product, determinant, and eigenvalues of a matrix. While Case (I) can be verified by taking first-order conditions with respect to the fundamentals, in Cases (II) and (III) the objective function involves the determinant of $\hat{\Sigma}$, which is not a tractable function in general. We solve this semi-definite programming problem by looking at the eigenvalues of a well-chosen matrix in each case, greatly reducing the dimensionality of the problems as well as eliminating the determinant from the objective.

Notice that plugging $\tilde{\Sigma} = \Sigma$ into Expression (2) yields Expression (4). Curiously, therefore, when the agent is initially agnostic about both the fundamentals and the covariance matrix, then — although he misinfers the covariance matrix — his long-run beliefs about the fundamentals are the same as when he correctly understands the covariance matrix.

3 A Model of Inferences about Individuals and Groups

We now turn to our main interest, a model of how overconfidence affects social beliefs.

3.1 Setup

There are I individuals distributed between G disjoint groups. Individual $j \in \{1, \ldots, I\}$ has fixed "caliber" $a_j \in \mathbb{R}$ and group membership $g_j \in \{1, \ldots, G\}$. We consider society from the perspective of individual $i \in \{1, \ldots, I\}$, whom we call agent i; in many cases, we also compare the views of different agents who all think according to our model, or analyze average views. Agent i knows individuals' memberships g_j, and observes a sequence of realizations of

$$\begin{aligned} q_j &= a_j + \sum_{k=1}^{K} \phi_{g_j k}\, \theta_k + \epsilon_j^q, \quad j = 1, \ldots, I, \\ \eta_k &= \theta_k + \epsilon_k^\eta, \qquad\qquad\qquad\quad k = 1, \ldots, K. \end{aligned} \tag{5}$$

In the first equation, $q_j \in \mathbb{R}$ is (a signal of) individual j's "recognition," $\theta_k \in \mathbb{R}$ is the extent of discrimination of type k, and $\phi_{gk} \in \mathbb{R}$ is the effect of a unit of such discrimination on group g. Hence, recognition depends in part on caliber, but it is also affected by discrimination in society. The agent knows the ϕ_{gk}, but not a_j or θ_k. In the second equation, η_k is a signal of θ_k. Finally, ϵ_j^q and ϵ_k^η are independent normally distributed errors with mean zero and variances v_j^q and v_k^η, respectively. Denoting by m_g the population frequency of group g, we impose that $\sum_g m_g \phi_{gk} = 0$ for each k. This identity captures that the effect of discrimination is redistributive.

The crucial assumption of our model is that agent i is overconfident about himself. Formally, while his true caliber is A_i, he believes with certainty that it is $\tilde{a}_i > A_i$. Beyond having an unrealistic self-view, however, agent i is rational: he applies Bayes' Rule correctly to update his beliefs. Furthermore, he is agnostic regarding the levels of discrimination and the calibers of other individuals, with his prior having full support over all vectors of reals. Similarly, he is uncertain

about the covariance matrix of the errors, with his prior having full support over all covariance matrices whose eigenvalues are at least $\underline{\lambda} \geq 0$, where to avoid technicalities we assume $\underline{\lambda} > 0$ and sufficiently small. We look for the limit of the agent's beliefs in the long run.

3.2 Discussion

Examples. Given our context of social judgments and prejudice, we think of the variables a_j and q_j broadly. A simple interpretation is that a_j is individual j's ability and q_j is his wage or other measure of economic success. Alternatively, a_j could denote a person's deservingness of social rewards based on past work or behavior or general character, with q_j capturing the rewards he gets in the form of transfers, perks, or other recognition.[4] Furthermore, a_j and q_j could be defined in absolute or relative terms.

Discrimination θ_k could take many forms as well, including policies, economic forces, or discriminatory behavior by a group or groups. Accordingly, the signals η_k about θ_k could come from multiple sources. The agent may glean direct evidence about discrimination from his work, school, or personal life. He may, for instance, observe that some students get more opportunities to speak in class, some employees are assigned more "promotable" tasks, or some children are given better opportunities. Alternatively, the agent may hear about academic or journalistic research regarding discrimination. Finally, as we explain below, the agent may make inferences about discrimination from personal contact with others. Some or all of these types of signals might be very inaccurate; in this case, v_k^η is high.[5]

We introduce three specific examples to illustrate our model.

Example 1. There are two groups ($G = 2$), and one type of discrimination ($K = 1$). Discrimination benefits group 1, so $\phi_1 > 0 > \phi_2$.

Example 1 can be thought of as a basic, common type of discrimination often discussed in academic research or public discourse. There is discrimination favoring the "dominant" group 1 at

[4] For presentational simplicity, we refer to q_j as individual j's recognition, but our formalism also captures the case in which q_j is a signal of individual j's recognition that is observable to agent i. Furthermore, while we present the model and results by referring to individual j as a person, an equivalent model obtains if some observations q_j are average recognitions of groups or subgroups.

[5] In many situations, it seems plausible to assume that different groups have access to different information. Our results abstract from this consideration, so belief disagreements do not arise from differences in information. Furthermore, note that differences in information cannot by themselves explain systematic disagreement. In a correctly specified model, differences in information should not lead to systematic differences in beliefs, and in the long run everyone's beliefs should converge to the truth.

the expense of the "dominated" group 2.

Our other examples feature more complex situations.[6]

Example 2. There are three groups ($G = 3$), and one type of discrimination ($K = 1$). Discrimination benefits group 1 and hurts groups 2 and 3, but it hurts group 2 more ($\phi_{11} > 0 > \phi_{31} > \phi_{21}$).

This example captures one potential perception of affirmative action in college admissions. Suppose that group 1 is blacks, group 2 is Asians, and group 3 is whites. Affirmative action, if it exists (our framework allows any type of discrimination to be non-existent or go the other way), benefits blacks and hurts whites and especially Asians.

Example 3. There are three groups ($G = 3$) and two types of discrimination ($K = 2$). The first type of discrimination benefits group 1 at the expense of group 2 ($\phi_{11} > \phi_{31} = 0 > \phi_{21}$). The second type of discrimination benefits group 3 at the expense of group 2 ($\phi_{32} > \phi_{12} = 0 > \phi_{22}$).

As a potential example, suppose that group 1 is high-income natives, group 2 is low-income natives, and group 3 is (low-income) immigrants. Type-1 discrimination corresponds to the advantages high-income individuals enjoy in domestic affairs. Type-2 discrimination is a pro-immigration policy or economic force (acting, for instance, through local housing, schools, or employer hiring). If it exists, it benefits immigrants, hurts low-income natives, but does not impact high-income natives.

Competition between Groups and the Impact of Discrimination. A natural interpretation of the ϕ_{gk} derives from competition. Indeed, affirmative action harms Asians and whites due to competition for college spaces, and a pro-immigration policy harms low-income natives because they are (perceived to be) competing with low-skill immigrants. While in most of our analysis we take the ϕ_{gk} as exogenous, this perspective allows us to derive them from an underlying competition structure. Specifically, let $f(g, g')$ measure the (perceived) frequency or importance of competition for recognition that an individual with group membership g faces from individuals with group membership g'. Letting $G_k \subset \{1, \cdots, G\}$ denote the set of groups that benefit from discrimination of type k, we define

$$\phi_{gk} = \begin{cases} \sum_{g' \in G \setminus G_k} f(g, g') & \text{if } g \in G_k, \text{ and} \\ -\sum_{g' \in G_k} f(g, g') & \text{if } g \in G \setminus G_k. \end{cases} \tag{6}$$

[6] We frame our examples in the context of real-world phenomena. While the parameters are meant to be plausible, they are not based on well-established facts. Other parameters can be plugged into our framework equally easily. In our results, we emphasize predictions that can be tested without knowing the true extent of discrimination in society.

Intuitively, the impact of discrimination of type k on an individual is determined by how many people he tends to compete with him on the other side of the issue.

Notice that this microfoundation is consistent with the possibility that individuals compete fiercely with other members of their own group (e.g., that whites compete with each other for college spaces). Such competition does not affect ϕ_{gk}, as within-group competition does not influence the impact of between-group discrimination.

Overconfidence. The main premise of our framework, and the single non-classical assumption from which our results derive, is that the agent is overconfident regarding his worth in society. This premise is consistent with field evidence indicating overconfidence in some central aspects of life, including self-control (Shui and Ausubel, 2005, DellaVigna and Malmendier, 2006, Augenblick and Rabin, 2019, Chaloupka et al., 2019), productivity on the job (Park and Santos-Pinto, 2010, Hoffman and Burks, 2020, Huffman et al., 2019), likelihood of finding a job (Spinnewijn, 2015), entrepreneurial ability (Landier and Thesmar, 2009, Hyytinen et al., 2014), and managerial ability (Malmendier and Tate, 2005).[7] All individuals in these studies are adults who presumably have had plenty of opportunity to learn about themselves. Hence, overconfidence is either not eliminated by learning, or it is eliminated very slowly. Our specific assumption that the agent has a degenerate belief about his caliber is a tractable reduced-form way of capturing such stubborn overconfidence.[8] The same assumption also allows us to sidestep, for most of the paper, the question of what force generates stubborn overconfidence. In Section 5.2, we consider one microfoundation, biased learning about oneself. A disadvantage of our reduced-form approach is that it does not allow us to study effects on the agent's biases that operate at the source, overconfidence. Our result in Section 5.2 does, however, show that reducing overconfidence does not necessarily reduce other biases.

[7] Bolstering the field evidence is experimental evidence regarding overconfident beliefs about IQ (see for instance Burks et al. 2013, Charness et al. 2018, Zimmermann 2020, as well as Goette and Kozakiewicz 2018, who test our earlier paper on learning by an overconfident individual, Heidhues et al. 2018). Furthermore, although the issue is not studied with the same care, available evidence says that if anything, people have even greater illusions regarding their moral standing (e.g., Tappin and McKay, 2017).

Qualifying the evidence, Moore and Healy (2008) show that individuals' confidence depends on the task at hand and whether the measure of confidence is absolute or relative, and there are tasks for which people are on average underconfident. These task-specific distinctions are not central to our setting. Our notion of caliber is intended to capture a general capability to be a productive and worthwhile person, not an ability to perform a specific experimental task. The above evidence on overconfidence — and the corresponding lack of evidence documenting underconfidence — indicates that individuals are on average overconfident regarding these capabilities.

[8] For technical and presentational convenience, as well as to highlight that the biases are not eliminated by experience, our formal results pertain to long-run beliefs. Identical tendencies would obtain with finite data, but beliefs would then also depend on the agent's prior. Hence, to the extent that overconfidence is eventually eliminated by learning, our results apply to the long period of time before this self-discovery happens.

True Discrimination in Society. Finally, while our results are bound to have implications for discriminatory behavior, we emphasize that for the purposes of the present paper, the degrees of discrimination θ_k are exogenous. We do not make assumptions or predictions about — and our main results do not depend on — which groups face discrimination and in which direction. Instead, our main results are about agents' views relative to the truth and relative to each other.

4 Patterns in Beliefs

In this section, we analyze our model. By Theorem 1, agent i's beliefs converge to point beliefs. We denote the agent's long-run point belief about discrimination of type k by $\tilde{\theta}^i_k$, and his long-run point belief about individual j's caliber by $\tilde{a}^i_j$. We also denote the actual realizations of these variables by Θ_k and A_j, respectively.

Proposition 1 (Biases)**.** *Agent i's long-run bias about discrimination toward group k is*

$$\tilde{\theta}^i_k - \Theta_k = \frac{-\phi_{g_ik} v^\eta_k}{v^q_i + \sum_{k'} \phi^2_{g_ik'} v^\eta_{k'}} \cdot (\tilde{a}_i - A_i), \tag{7}$$

and his long-run bias about individual j's caliber is

$$\tilde{a}^i_j - A_j = \frac{\sum_k \phi_{g_ik}\phi_{g_jk} v^\eta_k}{v^q_i + \sum_{k'} \phi^2_{g_ik'} v^\eta_{k'}} \cdot (\tilde{a}_i - A_i). \tag{8}$$

We organize and discuss implications of Proposition 1 in the following subsections.

4.1 In-Group Bias

We start with basic, empirically documented patterns in the beliefs different groups develop regarding the levels of discrimination and individuals' calibers.

Equation (7) implies that the agent overestimates discrimination of type k if he suffers from it ($\phi_{g_ik} < 0$), and he underestimates discrimination of type k if he benefits from it ($\phi_{g_ik} > 0$). Intuitively, overconfidence implies that agent i's recognition is prone to falling short of his perceived caliber. A compelling explanation is that types of discrimination that harm him are strong and types of discrimination that benefit him are weak. The latter belief is especially relevant when there is a "dominant" group that has advantages over and therefore does better than other groups. In such situations, the underestimation of positive discrimination can be seen as a formalization

of social dominance theory's notion of a "legitimizing myth" — an illusion that rationalizes a social hierarchy (e.g., Pratto et al., 2006). Although these self-serving views about discrimination have not been pointed out previously as implications of overconfidence, they are analogous to the types of self-serving misattributions agents make in previous work on learning with overconfidence (Heidhues et al., 2018, Hestermann and Le Yaouanq, 2021). They are also somewhat analogous to Schwartzstein's (2014) result that an agent who ignores an important explanatory variable when trying to understand his observations may overestimate the relevance of another variable.

We point out an empirically relevant specific implication of the above. Suppose there is discrimination that harms a group g, and benefits or does not affect others. Then, a member's estimate of discrimination against group g is higher than a non-member's. Such contrasting views are a common finding in opinion surveys regarding discrimination.[9]

Beliefs regarding discrimination have implications for beliefs about groups. We state our results as averages over groups. To do so, we assume that v_i^q is the same for all individuals in group g, and denote it by v_g^q. We also let A_g be the average caliber of group g, and $\tilde{a}_{g'}^g$ the average opinion of group g about (others in) group g'. We establish an important property of these group beliefs:

Proposition 2 (In-Group Bias)**.**

I. (In-Group Overestimation). *Each group overestimates itself relative to the truth ($\tilde{a}_g^g > A_g$), but on average estimates groups correctly ($\sum_{g'} m_{g'} \tilde{a}_{g'}^g = \sum_{g'} m_{g'} A_{g'}$).*

II. (Absolute In-Group Bias). *If groups' average calibers (A_g) are equal, then each group thinks others in their group are better than the average ($\tilde{a}_g^g > \sum_{g'} m_{g'} \tilde{a}_{g'}^g$).*

III. (Relative In-Group Bias). *On average over all pairs of groups, a group's view of its fellow members relative to another group's members is positive: $\sum_{g,g'} m_g m_{g'} (\tilde{a}_g^g - \tilde{a}_{g'}^g) > 0$.*

Part I says that on average an agent overestimates other members of his group relative to the truth. Intuitively, these others are subject to the same discrimination effects as the agent. Since the agent overestimates discrimination hurting him and underestimates discrimination benefiting him, he commits the same errors regarding the effects of discrimination on fellow group members. Hence, he attributes too much of their observed recognitions to their calibers.

On the other hand, the agent understands that the effect of discrimination is redistributive.

[9] See Newport (2014) and Pew Research Center (2017a) on race, Pew Research Center (2017b) and Funk and Parker (2018) on gender, Pew Research Center (2018) on income, and "Weniger Respekt und wachsende Fremdenfeindlichkeit", Frankfurter Allgemeine Zeitung, 12.09.2019, on immigration.

Hence, he knows that if one person is harmed by discrimination, then another benefits from it. To the extent that he attributes more of the former person's recognition to caliber, therefore, he does so less for the latter person. As a result, his misestimation of the degrees of discrimination does not lead him to misestimate total caliber in the population.

The combination of in-group overestimation and overall correct estimation generates two manifestations of in-group bias. If the average calibers of groups are equal, then a person estimates his group to be above this level, and other groups to be below it on average. Hence, he thinks that his group is better than average (Part II). More generally, the average person estimates the average other member of his group to be better than average (Part III).

To illustrate our results and connect them to stylized facts, we consider the implications of Proposition 2 in the basic setting of Example 1, where the dominant group 1 benefits from discrimination at the expense of the dominated group 2.

Example 1. (cont'd) $G = 2$, $K = 1$, $v_1^q = v_2^q = v^q$, $\tilde{a}_j - A_j = 1$ for all j, and $\phi_{11} > 0 > \phi_{21}$. Then,

$$\tilde{a}_1^1 - \tilde{a}_2^1 = A_1 - A_2 + \frac{\phi_{11}(\phi_{11} - \phi_{21})v_1^\eta}{v^q + \phi_{11}^2 v_1^\eta}$$
$$\tilde{a}_2^2 - \tilde{a}_1^2 = A_2 - A_1 + \frac{\phi_{21}(\phi_{21} - \phi_{11})v_1^\eta}{v^q + \phi_{21}^2 v_1^\eta}$$

Notice that the fractions on the right-hand sides are both positive. Consequently, if the average calibers of the groups are equal ($A_1 - A_2 = 0$), then each group believes itself to be better than the other group. This two-group manifestation of absolute in-group bias is the most basic stylized fact in the literature on stereotypes, discrimination, prejudice, and racism.[10] Furthermore, although most researchers do not investigate this issue, some evidence indicates that in-group bias reflects a mistake. Bohren et al. (2019a) establish that initial opinions regarding users on an online mathematics question-and-answer platform exhibit a pro-male bias. Lambin and Palikot (2019) find that consumers discriminate against new minority drivers in a car-sharing platform in France, but they do not discriminate against minority drivers with many reviews. Furthermore, minority drivers do not select out of the platform. Together, these facts suggest learning about individual drivers starting from a systematically incorrect prior. It is natural to conjecture that these incorrect initial beliefs are based on stereotypical beliefs about the group.

[10] Classic contributions are Sumner (1906), Allport (1954), and Tajfel (1982). Mullen et al. (1992) provide a meta-analysis.

As Example 1 also makes clear, however, a group may fail to think of itself as better if the groups' true average calibers are not equal. The same might be the case if, stepping slightly outside our model, there are other biases that affect views about relative caliber equally across groups. For example, Chauvin (2020) and Frick et al. (2022) develop models in which both a dominated and a dominant group may underestimate the privileges of the dominant group.[11] In that case, the dominated group may show a bias toward the dominant group.

Our theory predicts that even then, the two groups exhibit relative in-group bias: group 1 members' opinion of group 1 relative to group 2 is more positive than group 2 members' opinion ($\tilde{a}_1^1 - \tilde{a}_2^1 > \tilde{a}_1^2 - \tilde{a}_2^2$). Indeed, when researchers do not find unanimous support for absolute in-group bias, they typically observe relative in-group bias.[12] Sometimes, however, groups do not even display relative in-group bias, and our theory cannot account for this evidence.[13]

When there are more than two groups, simple versions of in-group bias do not always obtain:

Example 2. (cont'd) $G = 3$, $K = 1$, $v_1^\eta = 1$, $m_1 = m_2 = m_3 = 1/3$, $\phi_1 = 3$, $\phi_2 = -2$, $\phi_3 = -1$, $v_g^q = 1$ for all g, $\tilde{a}_j - A_j = 1$ for all j, and all true abilities are normalized to zero. Then,

$$\tilde{a}_2^3 = 1; \quad \tilde{a}_3^3 = 1/2; \quad \tilde{a}_2^2 = 4/5; \quad \tilde{a}_3^2 = 2/5.$$

Here, group 3 overestimates group 2 more than it does itself, and more than group 2 overestimates itself. Hence, restricting attention to this pair of groups, both absolute and relative in-group bias are violated. Intuitively, group 3 is hurt by discrimination in favor of group 1, and hence overestimates discrimination. Furthermore, group 3 knows that group 2 is hurt even more by discrimination. Hence, in judging others, members of group 3 overestimate members of group 2 more than other members of group 3. Nevertheless, consistent with Part II of Proposition 2, group 3 still exhibits an absolute in-group bias relative to the average other group. Indeed, group 3 members' view regarding group 1 is $\tilde{a}_1^3 = -3/2$, so their average view of other groups is negative.

[11] More generally, many forms of discrimination, e.g., in favor of a majority or in favor of high-income individuals, take subtle forms that are difficult to pick up. To formalize such considerations in our framework, we can assume that all individuals receive biased signals about θ_1^η, but they interpret the signals as unbiased.

[12] For instance, although both Jewish- and Arab-Israeli buyers are more prone to respond to advertisements by Jewish rather than Arab car sellers, the difference is greater for Jews (Zussman, 2013). Other research documenting relative in-group bias includes Shayo and Zussman (2011), Gagliarducci and Paserman (2012), De Paola and Scoppa (2015) and Mengel et al. (2018).

[13] For example, Card et al. (2019) document that male and female referees appear to be about equally biased in favor of male authors. See also Bagues and Esteve-Volart (2010) in the context of judicial hiring decisions.

4.2 The Effects of Competition

We now consider how the development of opposing interests with another group affects a group's views. We illustrate using Example 3:

Example 3. (cont'd) $G = 3$, $K = 2$, $\tilde{a}_j - A_j = 1$ for all j, $\phi_{11} > \phi_{31} = 0 > \phi_{21}$, and $\phi_{32} > \phi_{12} = 0 > \phi_{22}$. Then, group 2's opinions about the other two groups are

$$\tilde{a}_1^2 = A_1 + \frac{\phi_{21}\phi_{11}v_1^\eta}{v_1^q + \phi_{21}^2 v_1^\eta + \phi_{22}^2 v_2^\eta} \quad \text{and} \quad \tilde{a}_3^2 = A_3 + \frac{\phi_{22}\phi_{32}v_2^\eta}{v_1^q + \phi_{21}^2 v_1^\eta + \phi_{22}^2 v_2^\eta}. \tag{9}$$

We compare group 2's views when $\phi_{22} = 0$ to those when $\phi_{22} < 0$. The former corresponds to an environment before group 2 (in the immigration example, low-income natives) competes with group 3 (immigrants), i.e., before it finds itself on opposite sides of discrimination of type 2. The latter corresponds to an environment in which competition and hence opposing interests have developed.

Most conspicuously, the above change in the environment lowers group 2's opinion of group 3. The potential for discrimination in favor of the new competitor group allows a member of group 2 to better explain his low recognition. Hence, he concludes that there is discrimination in favor of group 3, lowering his opinion of the group.

The effect of competition on a group's views helps explain evidence that greater local ethnic diversity increases racial animus, especially among individuals of low socioeconomic status (e.g., Branton and Jones, 2005), and that immigration triggers hostile reactions by local natives (Tabellini, 2019). More generally, the result says that the agent has more negative views about individuals or groups he considers competitors. This determinant of prejudice is one of the basic stylized facts that form the foundation of group conflict theory (e.g., Jackson, 2011). For instance, Stephan et al. (1999) document that the negative stereotyping of immigrants in the US is correlated with perceived competition for jobs and social transfers. Examining the direction of causality in an experiment, Esses et al. (1998) find that manipulating the sense of competition with an imaginary immigrant group leads subjects to see the group in a more negative light.

While competition with group 3 lowers group 2's view of group 3, "bias substitution" occurs: group 2's view of group 1 rises. For an intuition, consider the thinking of a member of group 2. When he faces no competition from group 3, he attributes his (in his view) low recognition partly to discrimination in favor of group 1, and partly to bad luck. But discrimination is inconsistent with his signals about it, and persistent bad luck is unlikely, so neither of these explanations is

fully compelling. When the same member of group 2 faces competition from group 3, in contrast, discrimination in favor of group 3 provides another explanation for his low recognition. This decreases the extent to which he believes in the previous, less than fully compelling explanations. And since he believes less in discrimination favoring group 1, he improves his opinion about the group.

In an example of bias substitution, Fouka et al. (2022) document that the inflow of blacks to northern U.S. cities during the great migration reduced the (previously substantial) stereotyping of Irish and Italian immigrants. Bias substitution also provides one rationale for a common political tactic, focusing citizens' attention on a competitor outside group to help unify a heterogeneous nation or constituency. In our setting, this lowers the negative views domestic groups may have about each other.

We now generalize the above insights. Suppose that groups g and g' are currently never affected by the same type of discrimination, i.e., $\phi_{gk}\phi_{g'k} = 0$ for all k. Now a new type of discrimination arises that pits groups g and g' on opposite sides: $m_g\phi_{gK+1} + m_{g'}\phi_{g'K+1} = 0$, with $\phi_{gK+1} \neq 0$. The new type of discrimination could also affect other groups.

Proposition 3. *The new type of discrimination:*

I. (Competition Effect). *Lowers the view of group g about group g'.*

II. (Excuse Effect). *Raises the view of group g about itself.*

III. (Bias Substitution). *Raises the average view of group g about groups other than g and g'.*

Part I generalizes the competition effect from our example. A member of group g overestimates discrimination in favor of or underestimates discrimination against group g'. Observing the same recognition, this lowers group g's opinion of group g'.

Part II says that competition with group g' raises a group member's view of group g, i.e., it raises members' overestimation of the group. The new competition provides a member of group g a new explanation for his low recognition. This explanation leads him to expect lower recognition for fellow group members subject to the discrimination. Given group members' recognition, therefore, he develops higher views of them.

Part III generalizes the bias substitution we have explained above. As the agent explains his low recognition in part using the new type of discrimination, his bias regarding the other types of discrimination decrease. This means that his overall bias regarding the other groups — negative

to begin with — decreases in absolute value.

4.3 The Effects of Information

In this subsection, we analyze how an improvement in the agent's information affects his beliefs. As a benchmark, we note that for a correctly specified agent, our analysis would be short. Indeed, if such an agent has sufficient information to form confident (deterministic) beliefs, then those beliefs must be correct and hence impervious to additional information. We show that for a misspecified agent, information can affect long-run beliefs, but whether and how it does depends on the type of information. While we focus on long-run beliefs, the mechanisms we identify are relevant away from the long-run limit as well.

We organize our insights from the perspective of a natural question: can more information mitigate the agent's biases about others? Proposition 1 implies that two types of information cannot. First, the very fact that we are focusing on long-run beliefs means that observing society for a longer period does not necessarily lower biases. Second, since v_j^q does not appear in Equation (8), an improvement in the agent's information about others' recognitions does not affect his long-run biases. Intuitively, demonstrating to the agent that an out-group's recognition is quite high does not improve his opinion at all, as he had already explained this by overestimating discrimination in favor of the group. These predictions are consistent with a variety of null effects of information on discrimination documented in the literature.[14]

Instead, Equation (8) suggests an indirect approach to mitigating biases: providing information about discrimination rather than the group outcomes themselves. Proposition 4 analyzes the effects on an individual's set of beliefs:

Proposition 4. *Suppose discrimination of type k affects individual i ($\phi_{g_ik} \neq 0$). An increase in the precision $1/v_k^\eta$ of information about discrimination of type k:*

I. (Direct Effect). *Lowers agent i's bias $|\tilde{\theta}_k^i - \Theta_k|$ regarding discrimination of type k.*

[14] Related to the first prediction, Boring (2017), Beck et al. (2018), and Bar and Zussman (forthcoming) find that experience does not reduce discrepancies in how different individuals treat the same groups. Related to the second prediction, many correspondence studies in which additional information about individuals does not reduce discrimination (e.g., Bertrand and Mullainathan, 2004) provide high-level information that is arguably best interpreted as being about recognition rather than caliber. In some studies that do find a positive effect of information, such as Kaas and Manger (2012) looking at reference letters in hiring and Tjaden et al. (2018) looking at online reviews of drivers, direct information about the person's character or quality is involved. We analyze the effect of such information below.

II. (No-Excuse Effect). *Lowers his average view* $\tilde{a}^i_{g_i}$ *about other members of his group.*

III. (Bias Substitution). *Raises his bias* $|\tilde{\theta}^i_{k'} - \Theta_{k'}|$ *regarding any other type of discrimination that affects him (type* $k' \neq k$ *for which* $\phi_{g_i k'} \neq 0$*).*

IV. (Indirect Benefit). *Raises his average view* $\sum_{g \neq g_i} m_g \tilde{a}^i_g$ *of other groups.*

V. (Bias Substitution). *Raises his bias* $|\tilde{a}^i_g - A_g|$ *about any group* g *not affected by discrimination of type* k *(*$\phi_{gk} = 0$*).*

More information about discrimination of type k has both beneficial and harmful effects. Quite directly, it makes a biased view about type-k discrimination less plausible, so it lowers the agent's bias about this (Part I). Furthermore, since the agent now believes less in discrimination against himself, he also believes less in discrimination against fellow group members. This lowers the part of group members' recognitions that he attributes to caliber (Part II). Looking to explain his recognition in another way, however, agent i engages in bias substitution: he comes to form more biased beliefs about discrimination of all types other than k (Part III).

The effects on the agent's views about other groups are mixed as well. Part IV says that the agent's average view of outside groups rises. This implies that the person improves his opinion of at least one group. Hence, the indirect approach of providing information about discrimination always has a beneficial effect on intergroup opinions. On the other hand, such information raises the person's bias regarding groups that are not affected by the discrimination. In particular, if he harbors any unrelated prejudices, these must necessarily increase.

For a simple illustration, take again Example 3, and suppose that group 2's information about discrimination in favor of group 3 improves (v_2^η decreases). By Equation (9), group 2's opinion about group 3 improves, but its view of group 1 declines.

The above results allow us to develop a novel perspective on one of the most influential and empirically most supported ideas in the psychology of intergroup relations. This is Allport's (1954) contact hypothesis — that contact between groups can reduce prejudices and biases.[15] Consistent with Allport's view that a primary channel is informational, we think of contact as providing information about the caliber of an out-group member. It is not obvious why this would be so

[15] Pettigrew and Tropp (2006) provide a meta-analysis of hundreds of previous studies, most of which find evidence consistent with the hypothesis. Many studies are correlational in nature, but evidence reviewed by Paluck et al. (2018) in which researchers experimentally manipulate interactions between groups shows that contact has a causal negative impact on prejudices. As a recent example in economics, Corno et al. (2019) document that being randomly paired with a roommate of a different race reduces negative stereotypes and increases inter-racial friendships (see also Lowe, 2020).

helpful. In a model of correctly specified learning, information about a single person is likely to have a limited effect on views about a large and diverse group. This is especially the case if — as in our model — the agent has sufficient other information to form confident beliefs.

In contrast, in our model the spillover effect on beliefs about others in the out-group can be more drastic. Suppose that agent i receives perfect information about individual j's caliber, fixing his belief at the truth ($\tilde{a}_j^i = a_j$). Furthermore, individual j is subject to only one type of discrimination, k ($\phi_{g_j k} \neq 0$, but $\phi_{g_j k'} = 0$ for all $k' \neq k$). Then, individual j's recognition q_j becomes another signal of discrimination θ_k. Hence, the personal contact with individual j is equivalent to an improvement in agent i's information about discrimination. Applying Proposition 4, personal contact changes agent i's view of individual j's entire group, even though he already had degenerate views about all individuals in the group. Of course, the proposition also implies that the personal contact has a downside due to bias substitution: agent i's bias regarding any group not affected by discrimination of type k increases.

Combining our insights from this subsection and the previous one, our model can be interpreted as predicting that the effect of intergroup contact is nuanced. Simple proximity — which we think of as loose contact — can increase the sense of competition and therefore increase animosity. But close, personal contact can provide information about caliber, lowering animosity. Roughly consistent with this distinction, Oliver and Wong (2003) find that greater diversity at the metropolitan level raises racial animus, but racial proximity at the neighborhood level lowers racial animus; and Laurence (2014) finds that greater diversity in one's community has a negative effect on inter-ethnic attitudes, but personal ties eliminate this effect.

We conclude this section by pointing out a type of information that is unambiguously beneficial:

Proposition 5. *A proportional decrease in all v_k^η lowers all of the agent's (non-zero) biases regarding discrimination and other individuals' calibers.*

While better information about a single type of discrimination is only partially beneficial due to bias substitution, the same is not the case for a balanced improvement in information about all types of discrimination. For example, it is plausible that members of a disadvantaged group observe discrimination with less noise. They may, for instance, see more direct evidence of discrimination, such as arbitrary searches by police, or they may be more attentive to the issue. Proposition 5 says that the disadvantaged group will then have less biased beliefs.

4.4 Similarity Bias

In our basic model, another individual is either inside or outside the agent's group. This dichotomy fails to capture the possibility that a person draws distinctions between outsiders to his group. A white male may, for instance, consider both white females and black females as out-groups, but think of the former as closer to him. In this subsection, we identify conditions for an extension of in-group bias, "similarity bias:" that a person has a more positively biased opinion about someone who is more similar to him.

To do so, we consider a special case of our model in which groups are defined by, and discrimination acts along, individuals' characteristics. Individual j has characteristics $c_j = (c_{j1}, \ldots, c_{jK}) \in \{0,1\}^K$, where $c_{jk} = 1$ means she has characteristic k (e.g., is of female gender or is black) and $c_{jk} = 0$ means that she does not. Furthermore, discrimination of type k redistributes recognition between individuals who have characteristic k and those who do not. A group, in turn, consists of individuals who share all characteristics, and is thus defined by a characteristic vector c. We say that agent i is more similar to individual j than to individual j' if whenever i and j' share a characteristic, so does j (i.e., $c_{j'k} = c_{ik} \Rightarrow c_{jk} = c_{ik}$). We also say that i is strictly more similar to j than to j' if i is more similar to j than to j', and the characteristic vectors of j and j' are not identical. As before, we denote by ϕ_{ck} the extend to which an individual with characteristics c is affected by discrimination based on characteristic k.

Proposition 6 (Similarity Bias)**.** *Suppose that ϕ_{ck} does not depend on $c_{k'}$ for any $k' \neq k$. If agent i is (strictly) more similar to j than to j', then i's long-run bias regarding the caliber of j is (strictly) greater than his long-run bias regarding the caliber of j', i.e., $\tilde{a}^i_j - A_j \geq \tilde{a}^i_{j'} - A_{j'}$ ($\tilde{a}^i_j - A_j > \tilde{a}^i_{j'} - A_{j'}$).*

Proposition 6 identifies a sufficient condition for similarity bias: that the impact of discrimination toward characteristic k depends only on whether an individual has characteristic k, and not on his other characteristics. Then, similarity determines how much agent i believes that the discrimination hurting him also hurts rather than helps individual j, so it determines how much of individual j's recognition he attributes to caliber.[16]

[16] It is worth noting that one type of discrimination the agent may consider possible is "exclusive discrimination" directed only against him. This corresponds to a characteristic k that only the agent has, and $\phi_{ck} < 0$ for the agent's characteristic vector c. Assuming, realistically, that exclusive discrimination is actually non-existent ($\Theta_k = 0$), the agent develops the view that there is some of it ($\tilde{\theta}^i_k > 0$). Hence, the agent converges on what might be called

We illustrate in a special case what the condition in Proposition 6 requires in terms of the underlying competition structure in society.

Remark 1. Suppose that $K = 2$ (so that $c \in \{(1,1),(1,0),(0,1),(0,0)\}$), $m_c = 1/4$ for all c, and $f(c,c') = f(c',c)$ for all c, c'. Then, the following are equivalent.

1. ϕ_{ck} does not depend on $c_{k'}$ for any $k' \neq k$.

2. For some $\gamma_1, \gamma_2, \gamma_3 \geq 0$,$\gamma_1 + \gamma_2 + \gamma_3 \leq 1$, we have

$$f((1,1),(0,0)) = f((1,0),(0,1)) \equiv \gamma_1$$
$$f((1,1),(0,1)) = f((1,0),(0,0)) \equiv \gamma_2$$
$$f((1,1),(1,0)) = f((0,1),(0,0)) \equiv \gamma_3.$$

An interesting, empirically plausible situation is when $\gamma_1 < \gamma_2 = \gamma_3$. Then, a person competes more with more similar than with less similar outsiders. In addition, as we have demonstrated in Section 4.2, the agent's prejudices are driven by competition with other groups: if he faces no competition with another individual, then he develops an unbiased view of her. Proposition 6 says that despite these facts, the agent develops a more positive bias of more similar outsiders. Intuitively, even if the agent competes most with the most similar out-group, he still thinks this group is hurt by discrimination along all characteristics they have in common with him. Hence, he thinks more positively of this group than of others less similar to him.

Because the similarity bias is an extension of the in-group bias, the evidence for the in-group bias discussed early is relevant when society differs mainly along a single dimension. Focusing on multiple dimensions across which applicants and panelist can differ, for applications to the UK's Engineering and Physical Sciences Research Council Banal-Estañol et al. (2021) find that an applicant's chances increase if panelists are more similar to the applicant's team.[17]

To show that without the condition the statement does not hold, we reconsider Example 3 in the language of characteristics. Suppose that there are two characteristics ($K = 2$), "high income" and "natives." This generates four possible groups: high-income natives ($c_1 = (1,1)$);

paranoid beliefs: he explains his lack of recognition in part by the belief that "the world is out to get him" and only him.

[17] Our theory is about an individual's belief regarding the caliber of others, and not committee decisions. How these caliber beliefs translate into cheap talk communication, individual influence and eventual collective decisions is, of course, non-obvious.

low-income natives ($c_2 = (0,1)$); low-income immigrants ($c_3 = (0,0)$); and high-income immigrants ($c_4 = (1,0)$). Supposing that there are no high-income immigrants (i.e. $m_{c_4} = 0$), and calling the other groups 1, 2, and 3, we get back the group structure of Example 3. Furthermore, we think of discrimination of type 1 (which favors rich natives) as being along characteristic 1, and discrimination of type 2 (which favors immigrants) as being along characteristic 2.

Example 3. (cont'd) $G = 3$, $K = 2$, $\phi_{11} > \phi_{31} = 0 > \phi_{21}$, $\tilde{a}_j - A_j = 1$ for all j, and $\phi_{32} > \phi_{12} = 0 > \phi_{22}$. Then, group 1's biases about the other two groups are

$$\tilde{a}_2^1 - A_2 = \frac{\phi_{11}\phi_{21}v_1^\eta}{v_1^q + \phi_{11}^2 v_1^\eta + \phi_{12}^2 v_2^\eta} < 0 \quad \text{and} \quad \tilde{a}_3^1 - A_3 = 0.$$

The example violates the condition of Proposition 6 because $\phi_{21} \neq \phi_{31}$ — i.e., the effect of discrimination according to income depends on whether a person is a native. And, similarity bias is violated: high-income natives are more negatively biased about low-income natives than about low-income immigrants. This occurs in our example because high-income natives compete with low-income natives but not with immigrants.

5 Model Variants

5.1 Prejudice without Discrimination or Group Knowledge

The mechanism for biased beliefs in our model above is based on two premises. First, the agent entertains the possibility of discrimination. Because of this, he can attribute his lower-than-expected recognition to discrimination. Second, the agent knows individuals' group memberships, so that he knows how each type of discrimination affects each individual. Because of this, his incorrect conclusions about discrimination translate into incorrect conclusions about individuals and groups. We now show that the mechanism of our model can be operational, and hence the agent can develop prejudice against those competing with him, even without these two assumptions.

Suppose that $I \geq 3$, and individual j's recognition is given by

$$q_j = a_j + \psi_j \epsilon_g + \epsilon_j, \tag{10}$$

where a_j is individual j's caliber, $\psi_j \in \mathbb{R}$ is a constant, and ϵ_g and ϵ_j are independent mean-zero normal shocks with variances v_g and v_j, respectively. As in our previous model, agent i observes a

sequence of q_j for each individual, being stubbornly overconfident about himself but agnostic about the calibers of all other individuals. Furthermore, the agent does not know the constants ψ_j and the variances v_j. The support of his prior is $\mathbb{R}^I \times [\underline{v}, \infty)^I$, where the smallest variance $\underline{v}$ he allows for satisfies $0 < \underline{v} \leq \min_j v_j$.[18] He understands the rest of the situation correctly, and updates his beliefs using Bayes' Rule.[19] Since models with $\psi_1, \ldots, \psi_I$ and $-\psi_1, \ldots, -\psi_I$ are equivalent, we normalize $\psi_i, \tilde{\psi}_i \geq 0$.

We interpret ϵ_g as a group-level shock that simultaneously affects many individuals. Individuals whose ψ_j has the same sign as the agent are "in the same boat" with him, and in this sense belong to his in-group; and individuals whose ψ_j has the opposite sign are in the "opposing boat," and in this sense belong to his out-group. Similarly to the case of discrimination, competition between groups can generate such a pattern of conflicting interests. Under (non-zero) discrimination, however, the conflicting interests play out unfairly in that a group's benefit or harm is systematic. Here, the competition is fair — and the agent knows that it is fair — in that neither group systematically benefits (ϵ_g has mean zero). In addition, the agent does not know the group structure.

Biases are now determined in the following way:

Proposition 7. *Agent i's long-run belief about individual j's caliber is*

$$\tilde{a}_j^i = A_j + \frac{\psi_i \psi_j v_g}{v_i^q + \psi_i^2 v_g} \cdot (\tilde{a}_i - A_i). \tag{11}$$

Furthermore, his long-run belief about ψ_j is $\tilde{\psi}_j = \kappa \cdot \psi_j$, where $\kappa > 1$ is a constant.

To develop intuition for Proposition 7, suppose first that agent i knows the group structure ψ_j — i.e., he knows how each person's output depends on the common shock. To the overconfident agent, q_i is often surprisingly low, so he thinks that he must be exceedingly unlucky. Since part of his luck derives from the common shocks, he makes inferences about others' luck as well. Namely, he thinks that individuals with $\text{sgn}(\psi_j) = \text{sgn}(\psi_i)$ must also have been unlucky, and those with $\text{sgn}(\psi_j) \neq \text{sgn}(\psi_i)$ must have been lucky. Hence, given their recognitions, he overestimates the former individuals and underestimates the latter ones.

[18] To see that this implies a uniform bound on the covariance matrix as required by Theorem 1 observe that the covariance matrix is given by $v_g \times (\psi \otimes \psi') + diag(v_1, v_2, \ldots, v_I)$. Here $diag(v_1, v_2, \ldots, v_I)$ denotes the diagonal matrix with entries $v_1, \ldots, v_I$. The smallest eigenvector of the covariabnce mattrix is thus greater $\min_{x:|x|=1} x^T[v_g \times (\psi \otimes \psi') + diag(v_1, v_2, \ldots, v_I)]x \geq x^T diag(v_1, v_2, \ldots, v_I)x = \min_j v_j$.

[19] This means that the agent knows v_g. Notice that an increase in v_g and a proportional increase in the absolute values of all ψ_j are observationally equivalent. Assuming that the agent correctly understands v_g — effectively a normalization — simplifies our presentation of his long-run beliefs.

But agent i does not know the group structure ψ_j, and is estimating it along with others' calibers. It turns out that he correctly infers the sign of each ψ_j, so that the above logic regarding the estimation of calibers still holds. But he also overestimates the importance of common shocks by a fixed factor. For an intuition, suppose that $\text{sgn}(\psi_j) = \text{sgn}(\psi_{j'}) = \text{sgn}(\psi_i)$. Then, agent i overestimates individuals j and j'. In a prototypical observation, therefore, both q_j and $q_{j'}$ seem to him unexpectedly low. Hence, agent i exaggerates the correlation between q_j and $q_{j'}$, leading him to overestimate ψ_j and $\psi_{j'}$.

Proposition 7 therefore implies that an overconfident agent is prone to prejudices even in a minimalistic situation in which he has no pre-existing notions of groups and maintains the belief that recognition is centered at caliber. Based on his observations, the agent learns his in-group and out-group, and develops an in-group bias. Furthermore, he exaggerates the importance of groups in determining recognition.

To conclude this section, we speculate on what the agent may deduce if, upon forming his long-run beliefs, he notices some obvious contradictions between his beliefs and data. What he sees is that members of his in-group receive systematically lower recognitions than their calibers, and members of his out-group receive systematically higher recognitions. We propose that even after recognizing these contradictions, it is not obvious for the agent to conclude that his ability is lower than he thought. After all, many others are not getting what they deserve either. Rather, it seems natural to conclude that discrimination is going on. In this sense, overconfident agents may be drawn towards theories allowing for discrimination.

5.2 Overconfidence through Biased Learning

In our main model, we capture stubborn overconfidence by assuming that the agent has a fixed, overly positive belief about himself. In this section, we consider one possible microfoundation for stubborn overconfidence, biased learning about oneself.

We modify the model introduced in Section 3 in the following ways. The agent has a full-support prior regarding his own caliber, and observes (in addition to q_j and η_k) signals $s_i = a_i + b + \epsilon_i^a$, where ϵ_i^a is a normally distributed error with mean zero and variance v_i^a that is independent of the other errors. In reality, $b = B > 0$, but the agent believes with certainty that it is $b = \tilde{b} = 0$: he is interpreting signals about himself in a positively biased way. This results in the following beliefs:

Proposition 8. *The agent's long-run bias about his own caliber is*

$$\tilde{a}_i - A_i = \frac{v_i^q + \sum_k \phi_{g_i k}^2 v_k^\eta}{v_i^a + v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta} \cdot B, \tag{12}$$

while his long-run bias about the caliber of individual $j \neq i$ is

$$\tilde{a}_j^i - A_j = \frac{\sum_k \phi_{g_i k} \phi_{g_j k} v_k^\eta}{v_i^a + v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta} \cdot B. \tag{13}$$

His bias regarding discrimination toward group k is

$$\tilde{\theta}_k^i - \Theta_k = \frac{-\phi_{g_i k} v_k^\eta}{v_i^a + v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta} \cdot B. \tag{14}$$

Comparing the formulas for $\tilde{a}_j^i - A_j, j \neq i$ and $\tilde{\theta}_k^i - \Theta_k$ to those in Proposition 1, the only difference is that there is an additional term v_i^a in the denominators, and overconfidence is replaced by the learning bias B. Hence, our qualitative results, as well as all of our comparative-statics predictions, are unaffected.

This version of the model, however, also has implications for how changes in the environment affect overconfidence. We point out one:

Corollary 1. *Making his own recognition a more precise signal of caliber (lowering v_i^q) lowers the agent's overconfidence and increases all his other biases.*

Confirming the classical intuition, providing better information about the agent does lower his overconfidence. Intuitively, since he finds it more difficult to reconcile his high self-view with the low recognition he obtains, he adjusts his self-view downwards. But disconfirming the classical intuition that mitigating the underlying cause of the agent's incorrect social views — overconfidence — helps debias him, all his other biases increase. Intuitively, the agent attributes his low performance partly to discrimination, and partly to bad luck. With less noise, bad luck becomes a worse explanation, raising the need for the discrimination explanation.

6 Related Literature

In this section, we relate our theory to research not discussed elsewhere in the paper. Most importantly, existing work does not derive a general in-group bias, develop a theory of group beliefs based on overconfidence, or make predictions regarding spillovers between multiple interdependent

incorrect beliefs about others. Indeed, previous research studying incorrect beliefs typically restricts attention to a one- or two-dimensional state of the world, so it cannot address such interactions. On the downside, we investigate only beliefs, not behavior, whereas much of the literature analyzes also actions that are endogenous to the incorrect beliefs.

Closest to our paper, Heidhues et al. (2018) and Hestermann and Le Yaouanq (2021) study the inferences and behavior of an overconfident whose performance depends in part on an external state. Related to our point that the agent develops biases regarding discrimination, both papers find that the agent misattributes bad outcomes to the external factor. And related in flavor to our in-group bias, Hestermann and Le Yaouanq (2021) show that the agent thinks too highly of an outsider who receives the same outcome in the same circumstances as he does. Neither paper, however, explores the implications of these insights for beliefs about discrimination or group-based prejudices.

Chauvin (2020) and Frick et al. (2022) develop models of inferences about groups in which the agent underestimates the differences in circumstances individuals face. This leads the agent to attribute differences in outcomes too much to individuals' types, thereby exaggerating differences between groups. Such models naturally explain why a fortunate person holds unrealistic negative views about a less fortunate person, but they also predict that the latter person holds similarly strong positive views about the former person. In contrast to our theory, therefore, these models do not predict a relative in-group bias, one of the basic stylized facts regarding intergroup prejudice.[20]

There is a large sociology and social-psychology literature on prejudice, but to our knowledge no theory is based on overconfidence or connects prejudice to opinions about discrimination. Furthermore, because the theories are not formalized, they do not make precise comparative-statics predictions. Most related to our framework, social identity theory (Tajfel and Turner, 1979, Tajfel, 1982) posits that individuals identify with a few relevant social groups — their in-groups. As a result, their self-esteem is bound up with their in-groups, so thinking positively about their in-groups and negatively about their out-groups leads them to think and feel positively about themselves.

[20] More generally, because the agent draws conclusions from observations while holding an incorrect view about himself, our paper belongs to the growing literature on learning with misspecified models. In the literature on misspecified learning not mentioned previously, the specific economic questions and the specific theoretical methods are different from those in our paper. Researchers have studied inferences by individuals who ignore some explanatory variables (Hanna et al., 2014, Schwartzstein, 2014), misunderstand causal relationships (Spiegler, 2016, Levy et al., 2022), misinterpret social observations (Eyster and Rabin, 2010, Bohren, 2016, Levy and Razin, 2017, Bohren and Hauser, 2019, Frick et al., 2020), make mistakes in applying Bayes' Rule (e.g., Rabin and Schrag, 1999, Rabin, 2002, Fryer et al., 2019), or draw incorrect inferences from their own past behavior (Heidhues et al., 2022).

Our theory also implies that a person's prejudices are intimately tied to his views about himself, but the connection follows a different — in a sense reverse — logic: a person thinks positively about himself, and this leads him to develop biases about his in-groups and out-groups.

An influential body of research demonstrates that prejudice and discrimination can operate implicitly outside the person's awareness (Greenwald and Banaji, 1995, Bertrand et al., 2005, and many others). Our framework — based on conclusions the agent draws from observations — is predicated on a conscious process, and hence may appear contradictory to implicit bias. But once the agent has drawn conclusions along the lines of our model, he may act on them without further conscious thought. Indeed, the idea that learned connections can unwittingly affect judgment is commonplace in psychology, and formed the basis from which the literature on implicit discrimination started in the first place (Jost et al., 2009). In this sense, our model is not contradictory to the existence of implicit bias.

Another strand of the social psychology literature conceptualizes stereotypes — i.e., generalizations about groups — as heuristic simplifications of real attributes. Bordalo et al. (2016) formalize this idea using a version of Kahneman and Tversky's (1972) representativeness heuristic. They assume that a person considers a trait more typical in a group if it is relatively more common in the group than in the relevant comparison group. This approach does not comfortably explain why stereotypes are often derogatory prejudices and why many views are self-serving, and unless different groups have different comparison groups, it also does not explain why different groups hold different views. On the other hand, our framework does not explain neutral stereotypes, such as the view that Swedes are blonde, which the framework of Bordalo et al. does.

Glaeser (2005) presents a political-economy model of hate, which he defines as beliefs about the harmfulness of others. Because voters who believe that the out-group is dangerous prefer policies that lower the out-group's resources, politicians benefit from hate-inducing messages that complement their policies. For instance, a pro-redistribution politician wants to induce hate against rich minorities. Unlike our framework, this model explains how the political environment affects people's beliefs about minorities, and which messages are communicated by which politicians. At the same time, our theory helps understand why negative attitudes often persist without politicians stoking them, or even despite politicians' attempts to debias.

7 Conclusion

While we have focused on understanding beliefs, it is natural to ask what our theory implies for discriminatory behavior. To make predictions regarding choices, we need to add an assumption about the agent's objectives. One possibility is to posit classical outcome-based preferences (e.g., earnings from one's firm). In this case, our model can be thought of as one of misspecified statistical discrimination — the agent uses group membership as a signal to guide behavior (e.g., whom to hire), but he does so incorrectly.[21] Another possibility is to assume that the agent dislikes rewarding or interacting with individuals he considers less deserving. Then, the agent treats other groups worse than his own because he has incorrectly concluded that they are less worthy. In this case, our model can be thought of as a microfoundation for taste-based discrimination. In fact, we suspect that the "pure" dislike of other groups assumed in the classical theory of taste-based discrimination is psychologically unrealistic. For instance, we do not think that a person dislikes a particular skin color unless it is associated in his mind with some meaning about what such others are like.

Going one step further, our biased-inference model can be incorporated into enrich models of discrimination with equilibrium feedback. Consider, for example, the troubling observation (Glover et al., 2017, Lavy and Sand, 2018, Carlana, 2019) that stereotypes can become self-fulfilling through the endogenous responses of interacting individuals. This is an important problem even with rational agents, but we conjecture that with biases it is especially devastating.

We also think our paper can be extended to study incorrect beliefs more broadly. For the most part, for example, our theory posits exogenously given groups that are known to individuals. What happens when groups are endogenous or not fully known is an interesting question for future research. As a simple illustration, consider a young academic who is unsure about what determines publication success but knows that he is not a member of a privileged group that accepts each other's papers at the expense of others. As he observes that his papers do not get the credit he overconfidently believes they deserve, he concludes that there must be such a privileged group. As a result, he tries to find the group and become a member of it. Since he never finds the group, he develops the conspiracy theory that it must be a secret society.

[21] Some other researchers have also noted that it would seem essential to distinguish correctly specified statistical discrimination from "error discrimination" (England and Lewin, 1989) or "inaccurate statistical discrimination" (Bohren et al., 2019b).

References

Allport, Gordon W., *The Nature of Prejudice*, Addison-Wesley Publishing Company, Inc., 1954.

Augenblick, Ned and Matthew Rabin, "An Experiment on Time Preference and Misprediction in Unpleasant Tasks," *Review of Economic Studies*, 2019, *86* (3), 941–975.

Bagues, Manuel F. and Berta Esteve-Volart, "Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment," *Review of Economic Studies*, 2010, *77* (4), 1301–1328.

Banal-Estañol, Albert, Qianshuo Liu, Inés Macho-Stadler, and David Pérez-Castrillo, "Similar-to-me Effects in the Grant Application Process: Applicants, Panelists, and the Likelihood of Obtaining Funds," September 2021. Working Paper.

Bar, Revital and Asaf Zussman, "Identity and Bias: Insights from Driving Tests," *Economic Journal*, forthcoming.

Beck, Thorsten, Patrick Behr, and Andreas Madestam, "Sex and Credit: Is There a Gender Bias in Lending?," *Journal of Banking and Finance*, 2018, *87*.

Berk, Robert H., "Limiting Behavior of Posterior Distributions when the Model Is Incorrect," *Annals of Mathematical Statistics*, 1966, *37* (1), 51–58.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 2004, *94* (4), 991–1013.

_ , Dolly Chugh, and Sendhil Mullainathan, "Implicit Discrimination," *American Economic Review*, 2005, *95* (2), 94–98.

Bohren, J. Aislinn, "Informational Herding with Model Mispecification," *Journal of Economic Theory*, 2016, *163*, 222–247.

_ , Alex Imas, and Michael Rosenberg, "The Dynamics of Discrimination: Theory and Evidence," *American Economic Review*, 2019, *109* (10), 3395–3436.

_ and Daniel Hauser, "Misinterpreting Social Outcomes and Information Campaigns," 2019. Working Paper.

_ , Kareem Haggag, Alex Imas, and Devin G. Pope, "Inaccurate Statistical Discrimination," 2019. Working Paper.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

Boring, Anne, "Gender Biases in Student Evaluations of Teaching," *Journal of Public Economics*, 2017, *145*, 27–41.

Branton, Regina P. and Bradford S. Jones, "Reexamining Racial Attitudes: The Conditional Relationship Between Diversity and Socioeconomic Environment," *American Journal of Political Science*, 2005, *49* (2), 359–372.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini, "Overconfidence and Social Signalling," *The Review of Economic Studies*, 2013, *80* (3), 949–983.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri, "Are Referees and Editors in Economics Gender Neutral?," 2019. Working Paper.

Carlana, Michela, "Implicit Stereotypes: Evidence from Teachers' Gender Bias," *The Quarterly Journal of Economics*, 2019, *134* (3), 1163–1224.

Chaloupka, Frank J., Matthew R. Levy, and Justin S. White, "Estimating Biases in Smoking Cessation: Evidence from a Field Experiment," 2019. Working Paper.

Charness, Gary, Aldo Rustichini, and Jeroen van de Ven, "Self-Confidence and Strategic Behavior," *Experimental Economics*, 2018, *21* (1), 72–98.

Chauvin, Kyle, "A Misattribution Theory of Discrimination," 2020. Working Paper.

Corno, Lucia, Eliana La Ferrara, and Justine Burns, "Interaction, Stereotypes and Performance. Evidence from South Africa," 2019. Working Paper.

De Paola, Maria and Vincenzo Scoppa, "Gender Discrimination and Evaluators' Gender: Evidence from Italian Academia," *Economica*, 2015, *82* (325), 162–188.

DellaVigna, Stefano and Ulrike Malmendier, "Paying Not to Go to the Gym," *American Economic Review*, 2006, *96* (3), 694–719.

England, Paula and Peter Lewin, "Economic and Sociological Views of Discrimination in Labor Markets: Persistence or Demise?," *Sociological Spectrum*, 1989, *9* (3), 239–257.

Esses, Victoria M., Lynne M. Jackson, and Tamara L. Armstrong, "Intergroup Competition and Attitudes Toward Immigrants and Immigration: An Instrumental Model of Group Conflict," *Journal of Social Issues*, 1998, *54* (4), 699–724.

Eyster, Erik and Matthew Rabin, "Naïve Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2010, *2* (4), 221–243.

Fouka, Vasiliki, Soumyajit Mazumder, and Marco Tabellini, "From Immigrants to Americans: Race and Assimilation during the Great Migration," *Review of Economic Studies*, 2022, *89* (2), 811–842.

Frick, Mira, Ryota Iijima, and Yuhta Ishii, "Misinterpreting Others and the Fragility of Social Learning," *Econometrica*, 2020, *88* (6), 2281–2328.

_ , _ , and _ , "Dispersed Behavior and Perceptions in Assortative Societies," *American Economic Review*, 2022, *112* (9), 3063–3105.

Fryer, Roland G., Philipp Harms, and Matthew O. Jackson, "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization," *Journal of the European Economic Association*, 2019, *17* (5), 1470–1501.

Funk, Cary and Kim Parker, "Women and Men in STEM Often at Odds Over Workplace Equity," Technical Report, Pew Research Center January 2018.

Gagliarducci, Stefano and M. Daniele Paserman, "Gender Interactions within Hierarchies: Evidence from the Political Arena," *Review of Economic Studies*, 2012, *79* (3), 1021–1052.

Glaeser, Edward L., "The Political Economy of Hatred," *Quarterly Journal of Economics*, 2005, *120* (1), 45–86.

Glover, Dylan, Amanda Pallais, and William Pariente, "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *Quarterly Journal of Economics*, 2017, *132* (3), 1219–1260.

Goette, Lorenz and Marta Kozakiewicz, "Experimental Evidence on Misguided Learning," 2018. Working Paper.

Greenwald, Anthony G. and Mahzarin R. Banaji, "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes," *Psychological Review*, 1995, *102* (1), 4–27.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning Through Noticing: Theory and Evidence from a Field Experiment," *Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 2018, *86* (4), 1159–1214.

_ , _ , and _ , "Misinterpreting Yourself," 2022. Working Paper.

Hestermann, Nina and Yves Le Yaouanq, "Experimentation with Self-Serving Attribution Biases," *American Economic Journal: Microeconomics*, 2021, *13* (3), 198–237.

Hoffman, Mitchell and Stephen V. Burks, "Worker Overconfidence: Field Evidence and Implications for Employee Turnover and Firm Profits," *Quantitative Economics*, 2020, *11* (1), 315–348.

Huffman, David, Collin Raymond, and Julia Shvets, "Persistent Overconfidence and Biased Memory: Evidence from Managers," 2019. Working Paper.

Hyytinen, Ari, Jukka Lahtonen, and Mika Pajarinen, "Forecasting Errors of New Venture Survival," *Strategic Entrepreneurship Journal*, 2014, *8* (4), 283–302.

Jackson, Lynne M., *The Psychology of Prejudice: From Attitudes to Social Action*, Washington, DC, US: American Psychological Association, 2011.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, and Curtis D. Hardin, "The Existence of Implicit Bias Is Beyond Reasonable Doubt: A Refutation of Ideological and Methodological Objections and Executive Summary of Ten Studies that No Manager Should Ignore," *Research in Organizational Behavior*, 2009, *29*, 39–69.

Kaas, Leo and Christian Manger, "Ethnic Discrimination in Germany's Labour Market: A Field Experiment," *German Economic Review*, 2012, *13* (1), 1–20.

Kahneman, Daniel and Amos Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 1972, *3* (3), 430–454.

Lambin, Xavier and Emil Palikot, "The Impact of Online Reputation on Ethnic Discrimination," 2019. Working Paper.

Landier, Augustin and David Thesmar, "Financial Contracting with Optimistic Entrepreneurs," *Review of Financial Studies*, 2009, *22* (1), 117–150.

Laurence, James, "Reconciling the Contact and Threat Hypotheses: Does Ethnic Diversity Strengthen or Weaken Community Inter-Ethnic Relations?," *Ethnic and Racial Studies*, 2014, *37* (8), 1328–1349.

Lavy, Victor and Edith Sand, "On the Origins of Gender Gaps in Human Capital: Short- and Long-Term Consequences of Teachers' Biases," *Journal of Public Economics*, 2018, *167* (C), 263–279.

Levy, Gilat and Ronny Razin, "The Coevolution of Segregation, Polarized Beliefs, and Discrimination: The Case of Private versus State Education," *American Economic Journal: Microeconomics*, 2017, *9* (4), 141–170.

_ , _ , and Alwyn Young, "Misspecified Politics and the Recurrence of Populism," *American Economic Review*, 2022, *112* (3), 928–962.

Lowe, Matt, "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration," 2020. Working Paper.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *Journal of Finance*, 2005, *60* (6), 2661–2700.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz, "Gender Bias in Teaching Evaluations," *Journal of the European Economic Association*, 2018, *17* (2), 535–566.

Moore, Don A. and Paul J. Healy, "The Trouble with Overconfidence," *Psychological Review*, 2008, *115* (2), 502–517.

Mullen, Brian, Rupert Brown, and Colleen Smith, "Ingroup Bias as a Function of Salience, Relevance, and Status: An Integration," *European Journal of Social Psychology*, 1992, *22* (2), 103–122.

Newport, Frank, "Gallup Review: Black and White Attitudes Toward Police," `https://news.gallup.com/poll/175088/{\^A}%C2%ADgallup-review-black-white-attitudes-toward-police.aspx` 2014.

Oliver, J. Eric and Janelle Wong, "Intergroup Prejudice in Multiethnic Settings," *American Journal of Political Science*, 2003, *47* (4), 567–582.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green, "The Contact Hypothesis Re-Evaluated," *Behavioural Public Policy*, 2018, pp. 1–30.

Park, Young Joon and Luis Santos-Pinto, "Overconfidence in Tournaments: Evidence from the Field," *Theory and Decision*, 2010, *69* (1), 143–166.

Pettigrew, Thomas F. and Linda R. Tropp, "A Meta-Analytic Test of Intergroup Contact Theory," *Journal of Personality and Social Psychology*, 2006, *90* (5), 751–783.

Pew Research Center, "The Partisan Divide on Political Values Grows Even Wider," Technical Report October 2017.

_ , "Wide Partisan Gaps in U.S. Over How Far the Country Has Come on Gender Equality," 2017.

_ , "Partisans are Divided over The Fairness of the U.S. Economy – And Why People are Rich or Poor," Technical Report 2018.

Pratto, Felicia, Jim Sidanius, and Shana Levin, "Social Dominance Theory and the Dynamics of Intergroup Relations: Taking Stock and Looking Forward," *European Review of Social Psychology*, 2006, *17* (1), 271–320.

Rabin, Matthew, "Inference by Believers in the Law of Small Numbers," *Quarterly Journal of Economics*, 2002, *117* (3), 775–816.

_ **and Joel Schrag**, "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 1999, *114* (1), 37–82.

Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Shayo, Moses and Asaf Zussman, "Judicial Ingroup Bias in the Shadow of Terrorism," *Quarterly Journal of Economics*, 2011, *126* (3), 1447–1484.

Shui, Haiyan and Lawrence M. Ausubel, "Time Inconsistency in the Credit Card Market," 2005. Working Paper.

Spiegler, Ran, "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

_ , "Behavioral Implications of Causal Misperceptions," *Annual Review of Economics*, 2020, *12*, 81–106.

Spinnewijn, Johannes, "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs," *Journal of the European Economic Association*, 2015, *13* (1), 130–167.

Stephan, Walter G., Oscar Ybarra, and Guy Bachman, "Prejudice Toward Immigrants," *Journal of Applied Social Psychology*, 1999, *29* (11), 2221–2237.

Sumner, William G., *Folkways*, New York: Ginn, 1906.

Tabellini, Marco, "Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration," *Review of Economic Studies*, 2019, *87* (1), 454–486.

Tajfel, Henri, "Social Psychology of Intergroup Relations," *Annual Review of Psychology*, 1982, *33* (1), 1–39.

_ **and John Turner**, "An Integrative Theory of Intergroup Conflict," in William G. Austin and Stephen Worchel, eds., *The Social Psychology of Intergroup Relations*, Brooks/Cole Pub. Co, 1979, chapter 3, pp. 33–47.

Tappin, Ben M. and Ryan T. McKay, "The Illusion of Moral Superiority," *Social Psychological and Personality Science*, 2017, *8* (6), 623–631.

Tjaden, Jasper Dag, Carsten Schwemmer, and Menusch Khadjavi, "Ride with Me — Ethnic Discrimination, Social Markets, and the Sharing Economy," *European Sociological Review*, 2018, *34* (4), 418–432.

Zimmermann, Florian, "The Dynamics of Motivated Beliefs," *American Economic Review*, 2020, *110* (2), 337–361.

Zussman, Asaf, "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars," *Economic Journal*, 2013, *123* (11), 433–468.

A Proofs

For brevity, throughout the Appendix we denote the bias of the agent's long-run beliefs about fundamental j by

$$\Delta_j = \tilde{f}_j - F_j,$$

and let $\Delta = (\Delta_1, \ldots, \Delta_L)^T$.

Proof of Theorem 1. We first verify that the assumptions of Berk (1966) are satisfied. The part (i) of the assumption stated in Berk is that the density is continuous in the parameters $(f', \Sigma') \in supp\mathbf{P}_0$. The subjective density in our model is given by

$$\frac{1}{\sqrt{(2\pi)^L \det \Sigma'}} \exp\left(-\frac{1}{2}(r - Mf)(\Sigma')^{-1}(r - Mf)\right).$$

The subjective density is continuous as the determinant and the inverse of a matrix are continuous functions of the coefficients of the matrix, and the determinant of positive definite matrix is strictly positive. Part (ii) of the assumption stated in Berk is that the above density equals zero only on a set of measure zero with respect to the true distribution, which is satisfied as the above density is always strictly positive. Part (iii) states that for some open neighborhood $U \subset \operatorname{supp} \mathbf{P}_0$ of every parameter value $(f', \Sigma') \in \operatorname{supp} \mathbf{P}_0$ the expected maximal log-likelihood is finite, i.e. for the random

first period observation r_1

$$\mathbb{E}\left[\sup_{(f'',\Sigma'')\in U}\left|\log \ell_1(r_1|f'',\Sigma'')\right|\right] < \infty\,.$$

This is satisfied as if $\lambda_{max}(\Sigma'')$ is the largest and $\lambda_{min}(\Sigma'')$ the smallest eigenvalue of Σ'' we have that

$$\begin{aligned}|\log \ell_1(r_1|f'',\Sigma'')| &= \frac{1}{2}\left|\log[(2\pi)^L \det\Sigma''] + (r_1 - Mf'')^T(\Sigma'')^{-1}(r_1 - Mf'')\right| \\ &\leq \frac{1}{2}\left|L\log[(2\pi)\lambda_{max}(\Sigma'')] + \frac{1}{\lambda_{min}(\Sigma'')}||r_1 - Mf''||^2\right|\,.\end{aligned}$$

As the eigenvalues are a continuous function of the entries of the matrix we get that the above function is continuous in (f'',Σ'') and thus that the supremum is finite over every neighborhood U.

Finally part (iv) of the assumption is satisfied if for every $c\in\mathbb{R}$ there exists a set $D\subset \operatorname{supp}\mathbf{P}_0$ with compact complement $(\operatorname{supp}\mathbf{P}_0)\setminus D$ such that

$$\mathbb{E}\left[\sup_{(f'',\Sigma'')\in D}\log \ell_1(r_1|f'',\Sigma'')\right] \leq c\,. \tag{15}$$

Fix $k_1,k_2>0$ and let D be the set of vectors f'' such that $||M(F-f'')||>k_1$ and covariance matrices Σ'' whose largest eigenvalue is strictly greater k_2. For all $(f'',\Sigma'')\in D$ and $||M\epsilon_1||\leq k_1/4$ the log-likelihood satisfies

$$\begin{aligned}\log \ell_1(r_1|f'',\Sigma'') &= -\frac{1}{2}\left(\log[(2\pi)^L\det\Sigma''] + (r_1-Mf'')^T(\Sigma'')^{-1}(r_1-Mf'')\right) \\ &\leq -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log(\lambda_{\max}(\Sigma'')) + \frac{1}{\lambda_{max}(\Sigma'')}||r_1-Mf''||^2\right)\end{aligned}$$

Denote by $a\vee b$ the maximum of a and b. As $\log(x)+y/x$ is minimized at $x=y$ with value $\log(y)+1$ we can bound the above term by

$$\begin{aligned}&\leq -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log(k_2\vee||r_1-Mf''||^2)\right) \\ &= -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log(k_2\vee||M(F-f'')+M\epsilon_1||^2)\right) \\ &\leq -\frac{1}{2}\Big(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') \\ &\qquad + \log\left(k_2\vee\left[||M(F-f'')||^2 + ||M\epsilon_1||^2 - 2||M(F-f'')||\times||M\epsilon_1||\right]\right)\Big)\end{aligned}$$

As right-hand-side above is decreasing in $||M(F-f'')||$ for $||M\epsilon_1||<k_1/2$ the minimum is attained at $||M(F-f'')||=k_1$ and we obtain the following bound

$$\leq \log -\frac{1}{2}\Big(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log\left(k_2\vee k_1^2/2\right)\Big)\,.$$

As the lowest eigenvalue of all covariance matrices in $\operatorname{supp} \mathbf{P}_0$ is bounded from below by $\underline{\lambda} < 1$ we have $\log \ell_1(r_1|f'', \Sigma'') \leq L/2|\log(\underline{\lambda})|$. That implies

$$\begin{aligned}\mathbb{E}\left[\sup_{(f'',\Sigma'')\in D} \log \ell_1(r_1|f'', \Sigma'')\right] &\leq \mathbb{E}\left[\mathbf{1}_{||M\epsilon_1||\leq k_1/4} \sup_{(f'',\Sigma'')\in D} \log \ell_1(r_1|f'', \Sigma'')\right] + L/2|\log(\underline{\lambda})| \\ &\leq -\frac{1}{2}\Big(L\log(2\pi) + (L-1)\log\underline{\lambda} + \log\left(k_2 \vee k_1^2/2\right)\Big)\mathbb{P}[||M\epsilon_1|| \leq k_1/4] + L/2|\log(\underline{\lambda})|\,.\end{aligned}$$

As $\lim_{k_1\to\infty} \mathbb{P}[||M\epsilon_1|| \leq k_1/4] = 1$ it follows that the left-hand-side of (15) becomes arbitrarily small for either k_1 or k_2 large enough. We are left to argue that the complement of D is compact for every k_1, k_2. Note, that the complement of D is the subset of $\operatorname{supp} \mathbf{P}_0$ of positive definite matrices where all eigenvalues are in $[\underline{\lambda}, k_2]$ and vectors f'' with $||M(F - f'')|| \leq k_1$. As $||\Sigma''||$ equals the largest eigenvalue, and thus is less than k_2, it follows from norm equivalence that the set of covariance matrices in the complement of D form a compact set. We can define the pseudo inverse of M as $M^* = (M^TM)^{-1}M^T$ and note that for fundamental vectors f'' in the complement of D it holds that $||F - f''|| = ||M^*M(F - f'')|| \leq ||M^*|| \times ||M(F - f'')|| \leq k_1||M^*||$, Thus, the complement of D is compact.

As shown by Berk (1966, main theorem p. 54), the support of the agent's beliefs will concentrate on the set of points that minimize the Kullback-Leibler divergence to the true model parameters (F, Σ) over the support of $\mathbf{P}_0$

$$\underset{(\hat{f},\hat{\Sigma})\in\operatorname{supp}\mathbf{P}_0}{\arg\min} \; D\big(F, \Sigma \,||\, \hat{f}, \hat{\Sigma}\big), \tag{16}$$

where the Kullback-Leibler divergence is given by

$$D\big(F, \Sigma \,||\, \hat{f}, \hat{\Sigma}\big) = \mathbb{E}\left[\log \frac{\ell_1(r_1|F, \Sigma)}{\ell_1(r_1|\hat{f}, \hat{\Sigma})}\right].$$

We will argue that the minimization problem (16) admits a unique solution when the prior $\mathbf{P}_0$ satisfies either (Case I), (Case II), or (Case III) and thus beliefs concentrate on a single point. As both the true model as well as the subjective model are Normal, we have that the Kullback-Leibler divergence simplifies to[22]

$$D\big(F, \Sigma \,||\, \hat{f}, \hat{\Sigma}\big) = \frac{1}{2}\left(\operatorname{tr}(\hat{\Sigma}^{-1}\Sigma) + (M(\hat{f} - F))^T\hat{\Sigma}^{-1}M(\hat{f} - F) - D + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right). \tag{17}$$

[22] See for example `https://en.wikipedia.org/wiki/Kullback\%E2\%80\%93Leibler_divergence#Multivariate_normal_distributions`

Throughout, we denote by $\tilde{f}, \tilde{\Sigma}$ the agents subjective long-run beliefs about the mean of the fundamentals and the covariance matrix. Define the matrix

$$B = M^T \tilde{\Sigma}^{-1} M \in \mathbb{R}^{L \times L}$$

and denote it's elements by $(B_{jk})_{j,k \in \{1,\dots,L\}}$. For future reference, note that since $\tilde{\Sigma}$ is symmetric, so is $M^T \tilde{\Sigma}^{-1} M$, and thus $B_{jk} = B_{kj}$. Furthermore, as $\tilde{\Sigma}$ is positive definite, so is $\tilde{\Sigma}^{-1}$ and $B = M^T \tilde{\Sigma}^{-1} M$.

We first analyze Case (I): By condition (Case I) the minimum in (16) is taken over means of the fundamentals $\hat{f}$ or equivalently biases $\Delta = \hat{f} - F$, taking the subjective covariance matrix $\hat{\Sigma} = \tilde{\Sigma}$ as given. By Berk's Theorem, the agent's beliefs about the fundamentals concentrate on the set that minimizes the Kullback-Leibler divergence (17). As we can ignore all terms that do not depend on $\hat{f}$, we get that the support of the subjective long-run belief about the mean of the fundamental is contained in

$$\begin{aligned} \underset{\hat{f}\colon \hat{f}_i = \tilde{f}_i}{\arg\min} (M(\hat{f} - F))^T \tilde{\Sigma}^{-1} M(\hat{f} - F) &= F + \underset{\Delta\colon \Delta_i = \tilde{f}_i - f_i}{\arg\min} \Delta^T \left(M^T \tilde{\Sigma}^{-1} M \right) \Delta \\ &= F + \underset{\Delta\colon \Delta_i = \tilde{f}_i - f_i}{\arg\min} \sum_{k=1}^{L} \sum_{j=1}^{L} B_{kj} \Delta_k \Delta_j \,. \end{aligned} \tag{18}$$

Here the sum symbolizes the addition of f to every element by element in the set of minimizers. Taking the first order conditions in the bias about fundamental Δ_h for $h \neq i$ and using that $B_{jk} = B_{kj}$ yields

$$0 = 2 \sum_{k=1}^{L} B_{kj} \Delta_k \,.$$

Dividing by 2 and plugging in $\Delta_k = \frac{B^{-1}_{ki}}{B^{-1}_{ii}} \Delta_i$ on the right-hand-side yields

$$\sum_{k=1}^{L} B_{kj} \Delta_k = \sum_{k=1}^{L} B_{kj} \frac{B^{-1}_{ki}}{B^{-1}_{ii}} \Delta_i = \frac{\Delta_i}{B^{-1}_{ii}} \sum_{k=1}^{L} B_{kj} B^{-1}_{ki} = \frac{\Delta_i}{B^{-1}_{ii}} \sum_{k=1}^{L} B_{jk} B^{-1}_{ki} = \frac{\Delta_i}{B^{-1}_{ii}} (BB^{-1})_{ji} \,,$$

which equals zero as BB^{-1} is the identity and $i \neq j$. Hence, $\Delta_k = \frac{B^{-1}_{ki}}{B^{-1}_{ii}} \Delta_i$ satisfies the first order condition.

Let e_k be the k-th unit vector, for $k \in \{1, \dots, L\}$. We next verify that the first order condition is sufficient for a global minimum. To do so, we rewrite the part of the objective (18) in terms of

$\Delta_{-i} = \sum_{j \neq i} e_j \Delta_j$

$$\Delta^T B \Delta = \left(e_i \Delta_i + \sum_{j \neq i} e_j \Delta_j \right)^T B \left(e_i \Delta_i + \sum_{j \neq i} e_j \Delta_j \right) = \left(e_i \Delta_i + \Delta_{-i} \right)^T B \left(e_i \Delta_i + \Delta_{-i} \right)$$
$$= \left(e_i \Delta_i \right)^T B \left(e_i \Delta_i \right) + \Delta_{-i}^T B \Delta_{-i} + 2 \left(e_i \Delta_i \right)^T B \Delta_{-i} \,. \tag{19}$$

The Hessian with respect to Δ_{-i} of (19) equals $2B$. As any quadratic form with a positive definite matrix Hessian has a unique global minimum that satisfies the first-order condition, it follows that indeed

$$\Delta_k = \frac{B_{ki}^{-1}}{B_{ii}^{-1}} \Delta_i = \frac{(M^T \tilde{\Sigma}^{-1} M)_{ij}^{-1}}{(M^T \tilde{\Sigma}^{-1} M)_{ii}^{-1}} \Delta_i$$

is the unique global minimizer for all $k \neq i$. This completes (I).

We next analyze Case (II): In this case the agent takes the subjective mean of the fundamentals $\tilde{f}$ and thus the bias Δ as given and estimates the covariance matrix $\tilde{\Sigma}$. Again, by Berk's Theorem the agent's beliefs about the covariance matrix concentrate on the set that minimizes the Kullback-Leibler divergence (17), which is equivalent to the set

$$\arg\min_{\hat{\Sigma}} \left(\operatorname{tr}(\hat{\Sigma}^{-1} \Sigma) + (M\Delta)^T \hat{\Sigma}^{-1} (M\Delta) + \log \frac{\det \hat{\Sigma}}{\det \Sigma} \right) . \tag{20}$$

Denote by $\cdot \otimes \cdot : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^{D \times D}$ the Kronecker product. In matrix notation, we want to show that the unique minimum of (20) is attained at

$$\hat{\Sigma} = \Sigma + (M\Delta) \otimes (M\Delta)^T$$

To simplify notation let $y = M\Delta$. We first manipulate the objective function

$$\begin{aligned}
&\operatorname{tr}(\hat{\Sigma}^{-1} \Sigma) + y^T \hat{\Sigma}^{-1} y + \log \frac{\det \hat{\Sigma}}{\det \Sigma} = \operatorname{tr}(\hat{\Sigma}^{-1} \Sigma) + \operatorname{tr}(y^T \hat{\Sigma}^{-1} y) + \log(\det \hat{\Sigma}) - \log(\det \Sigma) \\
&= \operatorname{tr}(\hat{\Sigma}^{-1} \Sigma) + \operatorname{tr}(\hat{\Sigma}^{-1} [y \otimes y^T]) - \log(\det \hat{\Sigma}^{-1}) - \log(\det \Sigma) \\
&= \operatorname{tr}\left(\hat{\Sigma}^{-1} (\Sigma + [y \otimes y^T]) \right) - \log\left(\det \hat{\Sigma}^{-1} \right) - \log(\det \Sigma) \\
&= \operatorname{tr}\left(\hat{\Sigma}^{-1} (\Sigma + [y \otimes y^T]) \right) - \log\det\left(\hat{\Sigma}^{-1} (\Sigma + [y \otimes y^T]) \right) + \log\det\left(\Sigma^{-1} (\Sigma + [y \otimes y^T]) \right) \\
&= \operatorname{tr}\left(\hat{\Sigma}^{-1} (\Sigma + [y \otimes y^T]) \right) - \log\det\left(\hat{\Sigma}^{-1} (\Sigma + [y \otimes y^T]) \right) + \log\det\left(Id + \Sigma^{-1} [y \otimes y^T] \right) .
\end{aligned} \tag{21}$$

Here we used in the first equality that a real number equals it's trace and the log of the ratio equals the difference of the logs. The second equality uses that the trace of A^TB equals the trace of BA^T. For third equality we use that the trace is an additive function. In the second to last equalities we use that the sum of logarithms equals the logarithm of the product and that the product of determinants equals the determinant of the product. Now notice that since Σ and y do not depend on $\hat{\Sigma}$, the set of minimizers equals

$$\arg\min_{\hat{\Sigma}} \ \operatorname{tr}(\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])) - \log(\det(\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])). \tag{22}$$

Let $\lambda_1, \ldots, \lambda_D$ be the eigenvalues of the matrix $\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])$. Since the trace is the sum of eigenvalues and the determinant is the product of eigenvalues, (22) is minimized by all matrices $\hat{\Sigma}$ such that the eigenvalues of $\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])$ minimize

$$\sum_{k=1}^{D} \lambda_k - \sum_{k=1}^{D} \log \lambda_k. \tag{23}$$

As (23) is strictly convex, we can take the first order condition to identify the unique minimizer. This yields that (23) uniquely minimized if and only if $\lambda_k = 1$ for all k. As all eigenvalues equal one and $\tilde{\Sigma}^{-1}(\Sigma + [y \otimes y^T])$ is symmetric—and hence diagonalizable—, $\tilde{\Sigma}^{-1}(\Sigma + [y \otimes y^T])$ is the identity matrix. This establishes that

$$\tilde{\Sigma} = \Sigma + [y \otimes y^T] = \Sigma + (M\Delta) \otimes (M\Delta)^T \tag{24}$$

is the unique minimizer of (20) and thus the subjective long-run belief of the agent about the covariance matrix. This establishes (II).

Finally, we prove Case (III): Again, by Berk's Theorem the agent's long-run bias about the fundamental and beliefs about the covariance matrix concentrate on the set that minimizes the Kullback-Leibler divergence (17)

$$\arg\min_{(\Delta,\hat{\Sigma}):\ \Delta_i = \tilde{f}_i - F_i} \frac{1}{2}\left(\operatorname{tr}(\hat{\Sigma}^{-1}\Sigma) + y^T\hat{\Sigma}^{-1}y - D + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right). \tag{25}$$

As shown in (21) this objective is equivalent to 1/2 times

$$\operatorname{tr}\left(\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])\right) - \log\det\left(\hat{\Sigma}^{-1}(\Sigma + [y \otimes y^T])\right) - D + \log\det\left(Id + \Sigma^{-1}[y \otimes y^T]\right).$$

Plugging in the minimizer for the covariance matrix $\Sigma + [y \otimes y^T]$ derived in part two simplifies the objective to

$$\log\det\left(Id + \Sigma^{-1}[y \otimes y^T]\right). \tag{26}$$

We first observe that as the determinant is the product of eigenvalues, (26) equals the sum of the logarithms of the eigenvalues of $Id+\Sigma^{-1}[y\otimes y^T]$. Furthermore, if λ is an eigenvalue of $Id+\Sigma^{-1}[y\otimes y^T]$ with associated eigenvector v then $\lambda - 1$ is an eigenvalue of $\Sigma^{-1}[y \otimes y^T]$ as

$$\lambda v = (Id + \Sigma^{-1}[y \otimes y^T])v \Rightarrow (\lambda - 1)v = \Sigma^{-1}[y \otimes y^T]v\,.$$

If we denote the eigenvalues of $\Sigma^{-1}[y \otimes y^T]$ by $\lambda_1, \ldots, \lambda_D$ then the objective (26) equals

$$\sum_{i=1}^{K} \log(\lambda_k + 1)\,.$$

As eigenvalues are independent of the basis, we next choose an orthogonal basis $x_1, \ldots, x_D$ such that $x_1 = y$ (we can always do so by picking an arbitrary basis and applying the Gram-Schmidt process). Denote, $\mathbf{1} = (1)$ the 1×1 identity matrix. As x_i is orthogonal to $y = x_1$, we have that

$$\Sigma^{-1}[y \otimes y^T]x_i = \Sigma^{-1}[y \otimes y^T][\mathbf{1} \otimes x_i] = \Sigma^{-1}[y\mathbf{1}] \otimes [y^T x_i] = \begin{cases} 0 & \text{if } i \neq 1 \\ (y^T y)(\Sigma^{-1}y) & \text{if } i = 1 \end{cases}.$$

Hence, $D - 1$ of the eigenvalues of $\Sigma^{-1}[y \otimes y^T]$ equal zero. We will next show that $v = \Sigma^{-1}y$ is an eigenvector with associated non-zero eigenvalue. Let $v = \sum_{i=1}^{D} \alpha_i x_i$ be the representation of $v = \Sigma^{-1}y$ in the basis x. We have that

$$\Sigma^{-1}[y \otimes y^T]v = \alpha_1(y^T y)(\Sigma^{-1}y) = \alpha_1(y^T y)v$$

and thus v is an eigenvector of $\Sigma^{-1}[y\otimes y^T]$ with eigenvalue $\alpha_1(y^T y)$. As α_1 is given by the projection of v on y, we have that $\alpha_1 = \frac{y^T v}{y^T y}$ and thus the non-zero eigenvalue of $\Sigma^{-1}[y \otimes y^T]$ equals

$$\alpha_1(y^T y) = y^T v = y^T \Sigma^{-1} y\,.$$

Consequently, the agents long-run belief about the mean of the state satisfies

$$\begin{aligned} \tilde{f} &= F + \underset{\Delta\colon\, \Delta_i = \tilde{f}_i - f_i}{\arg\min}\; y^T \Sigma^{-1} y \\ &= F + \underset{\Delta\colon\, \Delta_i = \tilde{f}_i - f_i}{\arg\min}\; \Delta^T \left(M^T \Sigma^{-1} M\right) \Delta\,. \end{aligned}$$

By (I) we have then have that the unique minimizer and thus the long-run belief of the agent is given by

$$\begin{aligned} \Delta_k &= \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{ki}}{\left[M^T\Sigma^{-1}M\right]^{-1}_{ii}}\Delta_i \qquad \text{for } k \neq i \\ \tilde{\Sigma} &= \Sigma + (M\Delta)\otimes(M\Delta)^T \end{aligned} . \tag{27}$$

This completes the proof of (III). □

Proof of Proposition 1. Let Σ^q, Σ^η be the variance-covariance matrices of ϵ^q and ϵ^η,

$$\begin{aligned} \Sigma^q &= \text{diag}(v_1^q, \ldots, v_I^q) \\ \Sigma^\eta &= \text{diag}(v_1^\eta, \ldots, v_K^\eta) \end{aligned}$$

and observe that they are invertible as the variances are greater than zero. We show that this model can be reduced into our old model. To see this observe that one can write the vector $(q\,\eta)^T$ in matrix notation as

$$\begin{pmatrix} q \\ \eta \end{pmatrix} = \begin{pmatrix} Id & \Phi \\ 0 & Id \end{pmatrix} \cdot \begin{pmatrix} a \\ \theta \end{pmatrix} + \begin{pmatrix} \epsilon^q \\ \epsilon^\eta \end{pmatrix} , \tag{28}$$

where the entry $(\Phi)_{jk} = \phi_{g_j k}$ of the matrix Φ is the impact of discrimination k on group g_j's output. Let

$$M = \begin{pmatrix} Id & \Phi \\ 0 & Id \end{pmatrix} .$$

As M has determinant 1, it is invertible. We have that the matrix $\left[M^T\Sigma^{-1}M\right]^{-1}$ is given by

$$\begin{aligned} \left[M^T\Sigma^{-1}M\right]^{-1} = M^{-1}\Sigma(M^{-1})^T &= \begin{pmatrix} Id & -\Phi \\ 0 & Id \end{pmatrix}\begin{pmatrix} \Sigma^q & 0 \\ 0 & \Sigma^\eta \end{pmatrix}\begin{pmatrix} Id & 0 \\ -\Phi^T & Id \end{pmatrix} \\ &= \begin{pmatrix} Id & -\Phi \\ 0 & Id \end{pmatrix}\begin{pmatrix} \Sigma^q & 0 \\ -\Sigma^\eta\Phi^T & \Sigma^\eta \end{pmatrix} = \begin{pmatrix} \Sigma^q + \Phi\,\Sigma^\eta\,\Phi^T & -\Phi\Sigma^\eta \\ -\Sigma^\eta\Phi^T & \Sigma^\eta \end{pmatrix} . \end{aligned}$$

By Theorem 1, agent i's bias about the ability of agent j is given by

$$\begin{aligned} \tilde{a}_j^i - A_j &= \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{ij}}{\left[M^T\Sigma^{-1}M\right]^{-1}_{ii}}\Delta_i = \frac{\left[\Sigma^q + \Phi\,\Sigma^\eta\,\Phi^T\right]_{ij}}{\left[\Sigma^q + \Phi\,\Sigma^\eta\,\Phi^T\right]_{ii}}(\tilde{a}_i - A_i) \\ &= \frac{\sum_k \phi_{g_i k}\phi_{g_j k} v_k^\eta}{v_i^q + \sum_k \phi_{g_i k}^2 v_k^\eta}\cdot(\tilde{a}_i - A_i) . \end{aligned} \tag{29}$$

By a similar argument we have that the estimated bias associated with characteristic k is given by

$$\begin{aligned}\tilde{\theta}_k^i - \Theta_k &= \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{i(I+k)}}{\left[M^T\Sigma^{-1}M\right]^{-1}_{ii}}\Delta_i = \frac{\left[-\Sigma^\eta\Phi^T\right]_{ik}}{\left[\Sigma^q + \Phi\,\Sigma^\eta\,\Phi^T\right]_{ii}}(\tilde{a}_i - A_i)\\ &= \frac{-\phi_{g_ik}v_k^\eta}{v_i^q + \sum_k \phi_{g_ik}^2 v_k^\eta}\cdot(\tilde{a}_i - A_i).\end{aligned}$$

This proves the result. □

Proof of Proposition 2.

I. By Proposition 1, the view of group g about group g' is

$$\tilde{a}_{g'}^g = \sum_{i\in g}\frac{\tilde{a}_{g'}^i}{Im_g} = \sum_{i\in g}\sum_{j\in g'\setminus i}\frac{\tilde{a}_j^i}{Im_g\times(Im_{g'} - \mathbb{I}_{\{g=g'\}})} = A_{g'} + \frac{\sum_{k=1}^K \phi_{gk}\phi_{g'k}v_k^\eta}{v_g^q + \sum_{k=1}^K \phi_{gk}^2 v_k^\eta}(\tilde{a}_g - A_g),$$

so its view of group g is

$$\tilde{a}_g^g = A_g + \frac{\sum_{k=1}^K \phi_{gk}^2 v_k^\eta}{v_g^q + \sum_{k=1}^K \phi_{gk}^2 v_k^\eta}(\tilde{a}_g - A_g).$$

Hence, clearly $\tilde{a}_g^g > A_g$.

Furthermore,

$$\begin{aligned}\sum_{g'} m_{g'}\tilde{a}_{g'}^g &= \sum_{g'} m_{g'}A_{g'} + \sum_{g'} m_{g'}\frac{\sum_{k=1}^K \phi_{gk}\phi_{g'k}v_k^\eta}{v_g^q + \sum_{k=1}^K \phi_{gk}^2 v_k^\eta}(\tilde{a}_g - A_g)\\ &= \sum_{g'} m_{g'}A_{g'} + \frac{\sum_{k=1}^K \phi_{gk}(\sum_{g'} m_{g'}\phi_{g'k})v_k^\eta}{v_g^q + \sum_{k=1}^K \phi_{gk}^2 v_k^\eta}(\tilde{a}_g - A_g) = \sum_{g'} m_{g'}A_{g'},\end{aligned}$$

where in the last step we have used that $\sum_{g'} m_{g'}\phi_{g'k} = 0$.

II. Immediate from part I.

III. Let $\tilde{a}_g = \sum_{i\in g} {}^{\tilde{a}_i}/_{Im_g}$, and note that by Proposition 1 $\tilde{a}_g > A_g$. We have

$$\begin{aligned}\sum_{g,g'} m_g m_{g'}(\tilde{a}_g^g - \tilde{a}_{g'}^g) &= \sum_g m_g\sum_{g'} m_{g'}\tilde{a}_g^g - \sum_g m_g\sum_{g'} m_{g'}\tilde{a}_{g'}^g = \sum_g m_g\tilde{a}_g^g - \sum_g m_g\sum_{g'} m_{g'}A_{g'}\\ &= \sum_g m_g\tilde{a}_g^g - \sum_{g'} m_{g'}A_{g'} = \sum_g m_g\tilde{a}_g^g - \sum_g m_g A_g\\ &= \sum_g m_g\frac{\sum_{k=1}^K \phi_{gk}^2 v_k^\eta}{v_g^q + \sum_{k=1}^K \phi_{gk}^2 v_k^\eta}(\tilde{a}_g - A_g) > 0.\end{aligned}$$

□

Proof of Proposition 3. We work in the context of $K+1$ types of discrimination, with type $K+1$ having effects $s\phi_{gK+1}$ and $s\phi_{g'K+1}$ on the two groups. Then, $s=0$ corresponds to the ex-ante situation with K types of discrimination, and $s=1$ to the new situation.

I. The view of group g about g' is

$$\tilde{a}^g_{g'} = A_{g'} + \frac{\sum_{k=1}^K \phi_{gk}\phi_{g'k}v^\eta_k + s^2\phi_{gK+1}\phi_{g'K+1}v^\eta_{K+1}}{v^q_g + \sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2\phi^2_{gK+1}v^\eta_{K+1}}(\tilde{a}_g - A_g) = A_{g'} + \frac{s^2\phi_{gK+1}\phi_{g'K+1}v^\eta_{K+1}}{v^q_g + s^2\phi^2_{gK+1}v^\eta_{K+1}}(\tilde{a}_g - A_g),$$

where we have used that $\phi_{gk}\phi_{g'k}=0$ for all $k \leq K$. Since $\phi_{gK+1} \neq 0$, this immediately implies that the bias of group g about g' is negative when $s=1$ and zero when $s=0$, establishing Part I.

II. The view of group g about group g is

$$\tilde{a}^g_g = A_g + \frac{\sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2\phi^2_{gK+1}v^\eta_{K+1}}{v^q_g + \sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2\phi^2_{gK+1}v^\eta_{K+1}}(\tilde{a}_g - A_g).$$

This is higher for $s=1$ than for $s=0$, proving Part II.

III. Notice that

$$\begin{aligned} m_g\tilde{a}^g_g + m_{g'}\tilde{a}^g_{g'} &= m_gA_g + m_{g'}A_{g'} + \frac{\sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2(m_g\phi^2_{gK+1} + m_{g'}\phi_{gK+1}\phi_{g'K+1})v^\eta_{K+1}}{v^q_g + \sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2\phi^2_{gK+1}v^\eta_{K+1}}(\tilde{a}_g - A_g) \\ &= m_gA_g + m_{g'}A_{g'} + \frac{\sum_{k=1}^K \phi^2_{gk}v^\eta_k}{v^q_g + \sum_{k=1}^K \phi^2_{gk}v^\eta_k + s^2\phi^2_{gK+1}v^\eta_{K+1}}(\tilde{a}_g - A_g), \end{aligned}$$

where we have used that $m_g\phi_{gK+1} + m_{g'}\phi_{g'K+1} = 0$. Clearly, the above is lower for $s=1$ than for $s=0$. Since group g has an average bias over all groups equal to zero, the average view of group g regarding other groups must be higher for $s=1$ than for $s=0$. □

Proof of Proposition 4. By Proposition 1, agent i's bias about discrimination of type k satisfies

$$|\phi_{g_ik}|\left|\tilde{\theta}^i_k - \Theta_k\right| = \frac{\phi^2_{g_ik}v^\eta_k}{v^q_i + \sum_{k'}\phi^2_{g_ik'}v^\eta_{k'}} \cdot |\tilde{a}_i - A_i|\,, \tag{30}$$

As the above term is increasing in v^η_k, Part (I) follows. Part (II) is implied as for an individual j who is a member of agent i's group

$$\tilde{a}^i_j - A_j = \frac{\sum_{k'}\phi^2_{g_ik'}v^\eta_{k'}}{v^q_i + \sum_{k'}\phi^2_{g_ik'}v^\eta_{k'}} \cdot (\tilde{a}_i - A_i)$$

which is increasing in v^η_k. Part (III) is implied as for $k'' \neq k$ the term in (30) is (weakly) decreasing in v^η_k, and strictly so if $\phi_{g_ik''} \neq 0$. Part (IV) follows since $\sum_g m_g\tilde{a}^i_g = \sum_g m_gA_g$ and by Part (ii)

$\tilde{a}^i_{g_i}$ is decreasing, so that $\sum_{g \neq g_i} m_g \tilde{a}^i_g$ must be increasing. For Part (V), observe that as $\phi_{gk} = 0$ for group g, Proposition 1 implies that

$$\left|\tilde{a}^i_g - A_g\right| = \left| \frac{\sum_{k' \neq k} \phi_{g_i k'} \phi_{g_j k'} v^\eta_{k'}}{v^q_i + \sum_{k'} \phi^2_{g_i k'} v^\eta_{k'}} \right| \cdot \left|\tilde{a}_i - A_i\right|,$$

and the first term on the right-hand side is (weakly) decreasing in v^η_k, and strictly so whenever the bias about group g is non-zero.. □

Proof of Proposition 5. Consider a proportional change that lowers all v^η_k by some constant factor $\alpha < 1$. By Proposition 1, this implies that agent i's long-run bias about discrimination toward group k is

$$|\tilde{\theta}^i_k - \Theta_k| = \left| \frac{-\phi_{g_i k} v^\eta_k}{\frac{v^q_i}{\alpha} + \sum_{k'} \phi^2_{g_i k'} v^\eta_{k'}} \right| \cdot (\tilde{a}_i - A_i) \leq \left| \frac{-\phi_{g_i k} v^\eta_k}{v^q_i + \sum_{k'} \phi^2_{g_i k'} v^\eta_{k'}} \right| \cdot (\tilde{a}_i - A_i),$$

with the inequality strict whenever $\phi_{g_i k} \neq 0$. Similarly, his long-run bias about individual j's caliber becomes

$$\left|\tilde{a}^i_j - A_j\right| = \left| \frac{\sum_k \phi_{g_i k} \phi_{g_j k} v^\eta_k}{\frac{v^q_i}{\alpha} + \sum_{k'} \phi^2_{g_i k'} v^\eta_{k'}} \right| \cdot (\tilde{a}_i - A_i) \leq \left| \frac{\sum_k \phi_{g_i k} \phi_{g_j k} v^\eta_k}{v^q_i + \sum_{k'} \phi^2_{g_i k'} v^\eta_{k'}} \right| \cdot (\tilde{a}_i - A_i),$$

with the inequality strict whenever $\sum_k \phi_{g_i k} \phi_{g_j k} v^\eta_k \neq 0$. □

Proof of Proposition 6. Proposition 1 implies that the difference in agent i's long-run bias about individual j and j' is

$$(\tilde{a}^i_j - A_j) - (\tilde{a}^i_{j'} - A_{j'}) = -\sum_k (\tilde{\theta}^i_k - \Theta_k)(\phi_{c_j,k} - \phi_{c_{j'},k}).$$

Consider an agent i who is more similar to agent j than to agent j'. Then $c_{j'k} = c_{ik}$ implies that $c_{jk} = c_{ik}$ and hence that $\phi_{c_{j'k}} = \phi_{c_{jk}} = \phi_{c_{ik}}$. Furthermore, if $c_{j'k} = c_{jk}$ then $\phi_{c_{j'k}} = \phi_{c_{jk}}$. Using these facts the above equation simplifies to

$$(\tilde{a}^i_j - A_j) - (\tilde{a}^i_{j'} - A_{j'}) = \sum_{k: c_{j'k} \neq c_{ik} \wedge c_{j'k} = c_{jk}} -(\tilde{\theta}^i_k - \Theta_k)(\phi_{c_j,k} - \phi_{c_{j'},k}).$$

Since characteristics are binary, for any dimension k in which $c_{j'k} \neq c_{ik} \wedge c_{j'k} = c_{jk}$, one has $c_{jk} = c_{ik}$ and thus $\phi_{c_{jk}} = \phi_{c_{ik}}$. Furthermore sgn $\phi_{c_{j'k}} \neq$ sgn $\phi_{c_{ik}} =$ sgn $\phi_{c_{jk}}$. Using these facts and Proposition 1 (i) $\phi_{c_{ik}} > 0$ implies $-(\tilde{\theta}^i_k - \Theta_k) > 0$ and $(\phi_{c_j,k} - \phi_{c_{j'},k}) > 0$; and (ii) $\phi_{c_{ik}} < 0$

implies $-(\tilde{\theta}^i_k - \Theta_k) < 0$ and $(\phi_{c_j,k} - \phi_{c_{j'},k}) < 0$. We conclude that in any dimension k in which $c_{j'k} \neq c_{ik} \wedge c_{j'k} = c_{jk}$, we have $-(\tilde{\theta}^i_k - \Theta_k)(\phi_{c_j,k} - \phi_{c_{j'},k}) > 0$. Thus, $(\tilde{a}^i_j - A_j) - (\tilde{a}^i_{j'} - A_{j'}) > 0$. □

To prove Proposition 7, we solve a more general model first in which recognition $q_j = a_j + \epsilon'_j$ is an unbiased signal of caliber that allows the error terms ϵ_j to have any positive definite covariance matrix Σ^q for which all eigenvalues are greater than some sufficiently small $\underline{\lambda}$ that is less than the solution stated in the Proposition 9 below. All other assumptions remain unchanged. In this case, one has:

Proposition 9 (Correlated Errors and Biases)**.** *Agent i's long-run bias about individual j is*

$$\tilde{a}^i_j = A_j + \frac{\Sigma^q_{ij}}{\Sigma^q_{ii}}(\tilde{a}_i - A_i), \tag{31}$$

while his bias about the covariance matrix is given by

$$\tilde{\Sigma}^q_{jj'} - \Sigma^q_{jj'} = (\tilde{a}^i_j - A_j)(\tilde{a}^i_{j'} - A_{j'}) = \frac{\Sigma^q_{j'i}\Sigma^q_{ji}}{(\Sigma^q_{ii})^2}(\tilde{a}_i - A_i)^2\,, \tag{32}$$

Proof of Proposition 9. We apply Part III Theorem 1 to $f = a$, $M = Id$. Then, $[M^T\Sigma^{-1}M]^{-1} = \Sigma$, and $M(\tilde{f} - F) = \tilde{a} - A$, yielding the formulas in the proposition. □

Proof of Proposition 7. Observe that the true model of Proposition 7 is a special case of the model of Proposition 9 in which $\epsilon'_j = \psi_j\epsilon_g + \epsilon_j$, where ϵ_g and ϵ_j are independent mean-zero normal shocks with variances v_g and v_j. Note that the sum of normal random variables is normal, and the true variance-covariance matrix of the shocks ϵ'_j has entries $\Sigma^q_{jj'} = \psi_j\psi_{j'}v_g$ for $j \neq j'$ and $\Sigma_{jj} = v^q_j + \psi^2_j v_g$.

The agent considers the subclass of subjective covariance matrices for which $\tilde{\Sigma}^q_{jj'} = \tilde{\psi}_j\tilde{\psi}_{j'}v_g$ for $j \neq j'$ and $\tilde{\Sigma}_{jj} = \tilde{v}^q_j + \tilde{\psi}^2_j v_g$. Note that this class of subjective models satisfies the assumptions of Berk's Theorem, and hence by Berk (1966, main theorem p. 54), the support of the agent's beliefs will concentrate on the set of points that minimize the Kullback-Leibler divergence to the true model parameters (A, Σ) over the support of the agent's subjective models. To solve this minimization problem, we minimize a relaxed problem in which we ignore the restriction that there must exist $\tilde{\psi}_j$'s such that $\tilde{\Sigma}^q_{jj'} = \tilde{\psi}_j\tilde{\psi}_{j'}v_g$ for $j \neq j'$ and $\tilde{\Sigma}_{jj} = \tilde{v}^q_j + \tilde{\psi}^2_j v_g$, and then verify that the solution to the relaxed problem satisfies these constraints.

By Proposition 9, we have that in the solution to the relaxed problem is given by

$$\tilde{a}_j^i = A_j + \frac{\psi_i \psi_j v_g}{v_i^q + \psi_i^2 v_g} \cdot (\tilde{a}_i - A_i),$$

and

$$\tilde{\Sigma}_{jj'}^q = \Sigma_{jj'}^q + \frac{\Sigma_{j'i}^q \Sigma_{ji}^q}{(\Sigma_{ii}^q)^2} (\tilde{a}_i - A_i)^2.$$

Hence,

$$\tilde{\Sigma}_{jj'}^q = \psi_j \psi_{j'} v_g \left[1 + \frac{\psi_i^2 v_g}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \right] \text{ for } j \neq j' \,,$$

and

$$\tilde{\Sigma}_{jj}^q = v_j^q + \psi_j^2 v_g + \frac{\psi_j^2 \psi_i^2 v_g^2}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \text{ for } j \neq i \,, \tag{33}$$

and finally

$$\tilde{\Sigma}_{ii}^q = v_i^q + \psi_i^2 v_g + (\tilde{a}_i - A_i)^2 \,.$$

To show that the solution to the relaxed problem is among the class of subjective models the agent considers, we are left to show that there exists $\tilde{\psi}_j$'s such that

$$\tilde{\psi}_j \tilde{\psi}_{j'} v_g = \psi_j \psi_{j'} v_g \left[1 + \frac{\psi_i^2 v_g}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \right] \text{ for all } j \neq j', \tag{34}$$

and

$$\tilde{v}_j^q + \tilde{\psi}_j^2 v_g = v_j^q + \psi_j^2 v_g + \frac{\psi_j^2 \psi_i^2 v_g^2}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \text{ for } j \neq i \,, \tag{35}$$

and finally

$$\tilde{v}_i^q + \tilde{\psi}_i^2 v_g = v_i^q + \psi_i^2 v_g + (\tilde{a}_i - A_i)^2. \tag{36}$$

Observe that (34) to (36) are solved by

$$\tilde{\psi}_j = \psi_j \sqrt{\left[1 + \frac{\psi_i^2 v_g}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \right]}, \tag{37}$$

and own variances

$$\tilde{v}_i^q = v_i^q + \frac{\left(v_i^q + \psi_i^2 v_g\right)^2 - \left(\psi_i^2 v_g\right)^2}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \text{ and } \tilde{v}_j^q = v_j^q. \tag{38}$$

We now argue that for $I \geq 3$, the solution given by (37) and (38) is unique. Dividing (34) for $j, j' \neq j$ by that for $j, j'' \neq j, j'$ implies that $\tilde{\psi}_{j'}/\tilde{\psi}_{j''} = \psi_{j'}/\psi_{j''}$, so that $\tilde{\psi}_{j'}/\tilde{\psi}_{j''}$ is unique. By (34),

$\tilde{\psi}_{j'}\tilde{\psi}_{j''}$ is also unique. Together with the normalization that $\tilde{\psi}_i \geq 0$, this implies that all $\tilde{\psi}_j$ are unique. With all $\tilde{\psi}_j$ uniquely given, own variances are unique by (35) and (36). □

Proof of Proposition 8. Let e_i be the i-th unit row vector. In the notation of Theorem 1,

$$f = \begin{pmatrix} b \\ a \\ \theta \end{pmatrix}, \quad r = \begin{pmatrix} s_i \\ q \\ \eta \end{pmatrix}, \quad M = \begin{pmatrix} 1 & e_i & 0 \\ 0 & Id & \Phi \\ 0 & 0 & Id \end{pmatrix}, \quad \Sigma = \begin{pmatrix} v_i^a & 0 & 0 \\ 0 & \Sigma^q & 0 \\ 0 & 0 & \Sigma^\eta \end{pmatrix},$$

the entry $(\Phi)_{jk} = \phi_{g_j k}$ of the matrix Φ is the impact of discrimination k on group g_j's output and where the agent is misspecified regarding b, with $\tilde{b} - B = -B$.

It is straightforward to verify that

$$M^{-1} = \begin{pmatrix} 1 & -e_i & \phi_i \\ 0 & Id & -\Phi \\ 0 & 0 & Id \end{pmatrix},$$

where ϕ_i is the row vector $(\phi_{g_i 1}, \cdots, \phi_{g_i K})$. To obtain the biases, we need to calculate the first column of the matrix $M^{-1}\Sigma(M^{-1})^T$. We have

$$\begin{aligned} M^{-1}\Sigma(M^{-1})^T &= \begin{pmatrix} 1 & -e_i & \phi_i \\ 0 & Id & -\Phi \\ 0 & 0 & Id \end{pmatrix} \begin{pmatrix} v_i^a & 0 & 0 \\ 0 & \Sigma^q & 0 \\ 0 & 0 & \Sigma^\eta \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -e_i^T & Id & 0 \\ \phi_i^T & -\Phi^T & Id \end{pmatrix} \\ &= \begin{pmatrix} 1 & -e_i & \phi_i \\ 0 & Id & -\Phi \\ 0 & 0 & Id \end{pmatrix} \begin{pmatrix} v_a^i & 0 & 0 \\ -v_i^q e_i^T & \Sigma^q & 0 \\ \Sigma^\eta \phi_i^T & -\Sigma^\eta \Phi^T & \Sigma^\eta \end{pmatrix} = \begin{pmatrix} v_a^i + v_i^q + c_i \Sigma^\eta c_i^T & \dots & \dots \\ -v_i^q e_i^T - \Phi\Sigma^\eta \phi_i^T & \dots & \dots \\ \Sigma^\eta \phi_i^T & \dots & \dots \end{pmatrix}. \end{aligned}$$

The formulas follow by applying Theorem 1, Part III. □

Proof of Corollary 1. The result follows from taking the derivative of the respective biases in Proposition 8 with respect to v_i^q. □