Review of Economic Studies (2025) **00**, 1–33 doi:10.1093/restud/rdaf048

Advance access publication 18 June 2025

Overconfidence and Prejudice

Paul Heidhues
DICE, Heinrich-Heine University Düsseldorf, Germany
Botond Kőszegi
University of Bonn, Germany
and
Philipp Strack
Yale University, USA

First version received March 2024*; Editorial decision May* 2025*; Accepted June* 2025 *(Eds.)*

We develop a model of multi-dimensional misspecified learning in which an overconfident agent learns about groups in society from observations of his and others' successes. We show that the average person sees his group relative to other groups too positively, and this in-group bias 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.

Key words: Beliefs, Prejudice, Inter-group beliefs, Overconfidence, Misspecified learning

JEL codes: D01, D83, D91

1. INTRODUCTION

Individuals' beliefs about each other are crucial determinants of social and economic behaviour. 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.*, 2019; Hestermann and Le Yaouanq, 2021; Frick *et al.*, 2022; Chauvin, 2023; Bohren *et al.*, 2025). 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]

[1] We cite relevant evidence, including for empirical claims in the introduction, when presenting our formal results below.

The editor in charge of this paper was Andrea Galeotti.

We build on this research 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. Second, we allow social beliefs to be richly multi-dimensional, uncovering connections that can help account for observed empirical patterns. Third, we identify types of information that are effective in debiasing agents, and types that are not.

In Section 2, we present our model. Society is composed of individuals in disjoint groups. An agent makes many independent 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 "calibre"—*i.e.* ability, work ethic, or other measure of deservingness—of a person. But he allows for the possibility that various types of "discrimination"—*i.e.* attitudes, policies, or economic forces with group-dependent impacts—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. While the agent knows the proportions, he does not know the degrees of discrimination, so he does not know how much redistribution is going on.

Crucially, to these ingredients we add a single non-classical but empirically well-founded assumption. Namely, the agent holds stubborn, unrealistically positive—*i.e.* overconfident—views about *himself*, formalized as a point belief about his calibre that exceeds the true value. Otherwise, the agent is agnostic and rational, starting from a full-support prior about the degrees of discrimination and others' calibres, and updating his beliefs using Bayes' Rule.

Section 3 identifies 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 an overconfident agent misattributes (what appear to him) low outcomes to unfavourable external factors. In our setting, this leads him to overestimate discrimination against and underestimate discrimination in favour of his group. Consistent with opinion surveys, this implies that outsiders consider discrimination against a group as less severe than group members do. Going further, individuals' misestimates about discrimination lead them to develop excessively positive opinions about other group members, and consequently to exhibit relative in-group bias.

Beyond explaining the above basic patterns, our theory makes a rich set of comparative-statics predictions. One set of insights centres around the effects of competition. Suppose that a new type of discrimination pits an outside group against the agent, for instance because the group moves to his neighbourhood and he finds himself on opposite sides of a social or economic issue with them. Because of his misestimate of the new type of discrimination, the agent's 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, the agent's biases regarding all groups not affected by the new type of discrimination decrease. 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. All of these effects occur even if the agent competes more with members of his own group than with outsiders.

Another set of insights concerns the effects of information. While better information about a group's recognition does not lower biases, 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. This 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 calibre of the individual, giving him information about discrimination and hence lowering his bias regarding all of the individual's group. 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.

In Section 4, we consider variants of our basic model. We demonstrate that our framework's central mechanism can be operational even when the agent neither entertains the possibility of systematic discrimination, nor starts off thinking of society in terms of distinct groups. Suppose that individual j's recognition is the sum of j's calibre, a mean-zero common shock scaled by ψ_j, and a mean-zero idiosyncratic shock. The agent does not know the effects of the common shock, ψ_j, which could be different across individuals and could be positive or negative. He uses observations of everyone's recognitions to update about individuals' calibres as well as the ψ_j. We show that the agent develops a positive bias about individuals whose ψ_j has the same sign as his, and a negative bias about individuals whose ψ_j has the opposite sign. In addition, he correctly learns the signs but overestimates the absolute values of the ψ_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 consider a model in which the agent's beliefs about his calibre are not fixed, but he interprets observations about himself in a positively biased way. We show that he develops overconfidence, which has the same effect on his other beliefs as in our basic model. Finally, we investigate the extent to which our results on long-run beliefs hold in the short run.

All of the formal analysis in our article relies on general tools we have developed for studying learning under high-dimensional misspecified models. We explain these tools in Section 5. Due to the lack of such tools, prior analysis of misspecified learning has typically focused on misinferences about a single-dimensional state of the world.

We discuss related literature in Section 6. While a few theories have implications for beliefs about groups, no previous paper derives a general relative in-group bias, makes predictions regarding spillovers between multiple interdependent incorrect beliefs about others, or develops a theory of group beliefs based on overconfidence. But our theory is of course not intended to explain all social biases. Some prejudices are stoked by politicians (Glaeser, 2005); many stereotypes are about less value-laden characteristics than our notion of calibre (Bordalo *et al.*, 2016); and individuals often also have prejudices about groups they are not in tangible competition with. We conclude in Section 7 with a discussion of what our model of beliefs might imply for discriminatory behaviour.

2. INFERENCES ABOUT INDIVIDUALS AND GROUPS

2.1. *Setup*

There are I individuals in G disjoint groups subject to K types of "discrimination". Individual $j \in \{1, \dots, I\}$ has fixed "calibre" $a_j \in \mathbb{R}$ and group membership $g_j \in \{1, \dots, G\}$, and $\theta_k \in \mathbb{R}$ denotes the fixed extent of discrimination of type k. We consider society from the perspective of one member, agent $i \in \{1, \dots, I\}$; we will also compare the views of different agents, and analyse average views. Agent i repeatedly observes each individual's "recognition" $q_j \in \mathbb{R}$ as well as signals $\eta_k \in \mathbb{R}$ of θ_k. In both the true model and agent i's subjective model, these observations are generated according to

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

where $\phi_{gk} \in \mathbb{R}$ is the fixed incidence of type-k discrimination on group g, and the ϵ_j^q and ϵ_k^η are mean-zero normally distributed errors that are identically and independently drawn over time.[2] Denoting by m_g the population frequency of group g, we impose that $\sum_g m_g \phi_{gk} = 0$ for all k, *i.e.* the effect of discrimination is redistributive.

In the true model, the vector of calibres a equals A, the vector of the levels of discrimination θ equals Θ, and the errors $\epsilon_j^q, \epsilon_k^\eta$ are all independent and have variances $\upsilon_j^q, \upsilon_k^\eta$. In agent i's subjective model, g_j and ϕ_{gk} are known and the same as in the true model, but others' calibres a_{-i}, the levels of discrimination θ, and the covariance matrix Σ of the errors $(\epsilon^q, \epsilon^\eta)$ are unknowns. The agent's prior belief regarding (a_{-i}, θ) has support $\mathbb{R}^{I-1} \times \mathbb{R}^K$, and his prior belief about Σ conditional on any a_{-i}, θ_k is supported on all positive definite symmetric matrices whose eigenvalues are greater than $\underline{\lambda}$, where $\underline{\lambda}$ is chosen to be sufficiently small.[3] Crucially, the agent is overconfident about himself: his subjective model assigns probability 1 to $a_i = \tilde{a}_i > A_i$. He applies Bayes' Rule to update his beliefs. We look for the limit of his beliefs in the long run.

2.2. *Interpretation and discussion*

The calibre a_j could stand for a person's ability or general character, and recognition q_j for their income, wealth, or broader social status. Both a_j and q_j can be defined in absolute as well as relative terms. The degrees of discrimination θ_k might capture the severity of discriminatory behaviour, strength of policies, or intensity of economic forces that affect groups differently, while the signals η_k about θ_k could come from observations the agent makes in his own life, or from academic or journalistic research he hears about. For the purposes of the present paper, the θ_k are exogenous.[4]

We think of the incidences ϕ_{gk} of discrimination on groups as being determined by an underlying competition structure. For instance, affirmative action is perceived to harm Asians and whites due to competition for college spaces, and a pro-immigration policy is perceived to harm low-income natives due to competition for jobs. This perspective does not preclude—and hence our results are consistent with—the possibility that a person competes more with in-group than with out-group members.[5] Furthermore, our assumption that the agent knows the ϕ_{gk} reflects the

[2] The assumption that recognition is linear in its components is purely for tractability.

[3] For a discussion of this technical assumption, an explicit formula for $\underline{\lambda}$, and other specifications of the support of the prior, see Section 5. In particular, our results are the same if the agent knows the covariance matrix.

[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 noisy 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. For groups the agent knows little about, these observations could be very noisy. Also note that while in reality different groups often have access to different information, our basic model abstracts from this consideration. In a correctly specified model, differences in information do not by themselves generate systematic disagreement.

[5] To formalize, 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'. Denoting by $G_k \subset \{1, \ldots, G\}$ the set of groups that benefit from discrimination of type k, define $\phi_{gk} = \sum_{g' \in G \setminus G_k} f(g, g')$ if $g \in G_k$ and $\phi_{gk} = -\sum_{g' \in G_k} f(g, g')$ if $g \in G \setminus G_k$. Intuitively, the impact of discrimination of type k on an individual is determined by how many people he tends to compete with on the other side of the issue. The extent $f(g, g)$ to which

idea that he can learn competition patterns from sources such as the media or public discussions, or infer them from knowledge such as how college admissions work.

Given the above competition perspective—according to which one person's gain from discrimination comes at the expense of someone else—it is natural to assume that the effect of discrimination is redistributive ($\sum_g m_g \phi_{gk} = 0$). This assumption allows us to make statements about beliefs regarding calibres averaged across multiple groups, but plays no other role.[6]

While we focus on limiting beliefs, by definition these approximate the agent's beliefs after sufficiently long finite time. Furthermore, while we assume that the agent is certain about his calibre, the results extend to some settings in which he is slightly uncertain. Specifically, suppose that the agent starts with a Normal prior, is uncertain about the fundamental, and knows the covariance matrix of signals. At any fixed finite time, his beliefs with a sufficiently certain prior about his calibre are close to his beliefs with a degenerate prior—which, after a long time, are close to the limiting beliefs we derive. The latter is true even if the agent is correctly specified, so he eventually learns everything correctly.

There is evidence for our main premise, overconfidence, from many aspects of life (*e.g.* Malmendier and Tate, 2005; Landier and Thesmar, 2009; Spinnewijn, 2015; Augenblick and Rabin, 2019; Huffman *et al.*, 2022). Furthermore, since individuals in these and other studies have had plenty of opportunity to learn about themselves, overconfidence is stubborn: it is either not eliminated by learning, or it is eliminated very slowly. Our analysis of long-run beliefs is appropriate for a person who has had sufficient scope to learn about society but has (like most individuals observed in the empirical work) remained overconfident so far. To complement these main insights, in Section 4.3 we discuss short-run beliefs. We note that if the agent starts with a sufficiently uncertain (high-variance) prior, our results on his long-run biases approximate his average short-run biases. We also show that if $K = 1$, the directions of the agent's short-run and long-run biases are identical. Otherwise, however, short-run and long-run biases can be directionally different.

3. PATTERNS IN BELIEFS

We now analyse our model. We say that agent i's beliefs about discrimination and individuals' calibres concentrate on $(\tilde{\theta}^i, \tilde{a}^i) \in \mathbb{R}^K \times \mathbb{R}^I$ if the probability he assigns to any open set around $(\tilde{\theta}^i, \tilde{a}^i)$ converges to one. Based on a general result in Section 5, we obtain:

Theorem 1 (Long-Run Biases). *Agent i's beliefs concentrate on a single $(\tilde{\theta}^i, \tilde{a}^i)$ almost surely. His long-run bias about discrimination of type k is*

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

and his long-run bias about the calibre of individual $j \neq i$ is

$$\tilde{a}^i_j - A_j = \frac{\sum_k \phi_{g_i k} \phi_{g_j k} \upsilon^\eta_k}{\upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot (\tilde{a}_i - A_i). \tag{3}$$

individuals compete fiercely with other members of their own group does not affect ϕ_{gk}, as within-group competition does not influence the impact of between-group discrimination.

[6] More precisely, we use the assumption in Proposition 1 (all parts, except for the claim that each group overestimates itself relative to the truth), Proposition 2, Part III, and Proposition 3, Part IV. The other results hold unchanged without the assumption.

First, the direction of the agent's bias about discrimination of type k has the opposite sign from the effect of this discrimination on his group. Second, the direction of the agent's bias about an individual depends on a weighted sum of how similarly discrimination affects the agent's and the individual's groups. Types of discrimination that affect the agent and individual in the same direction contribute positively to this sum, and types of discrimination that affect the two people in opposite directions contribute negatively. We organize and discuss economic implications in the following subsections.

3.1. *In-group bias*

We start with two basic, empirically documented patterns in beliefs. Equation (2) implies that the agent overestimates discrimination that harms him ($\phi_{g_i k} < 0$), and underestimates discrimination that benefits him ($\phi_{g_i k} > 0$). Intuitively, the distorted beliefs explain to the agent why his recognition is not as high as he overconfidently expects. Though not derived formally in previous research, such views resemble misattributions in work on learning with overconfidence (Heidhues *et al.*, 2018; Hestermann and Le Yaouanq, 2021) and selective attention (Schwartzstein, 2014). Further, a person's underestimation of beneficial discrimination can be seen as a formalization of social dominance theory's notion of a "legitimizing myth"—an illusion that rationalizes a "dominant" group's advantages over "dominated" groups (*e.g.* Pratto *et al.*, 2006).

The above implies that members estimate the level of discrimination against a group as higher than non-members who are unaffected by or benefit from the discrimination. Such contrasting views are a common finding in opinion surveys.[7] Relatedly, our theory predicts that a person is biased about a type of discrimination only if it affects him. For example, a white male professor may understand discrimination in policing and firm hiring but fail to appreciate discrimination in academia. We are unaware of evidence on this prediction.

Beliefs regarding discrimination have implications for beliefs about groups. We state our results as averages over groups. To do so, we assume that v_j^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 calibre of group g, and $\tilde{a}_{g'}^g$ the average opinion of group g about (others in) group g'.

Proposition 1 (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' calibres (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, 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, since he overestimates discrimination hurting and underestimates discrimination benefiting fellow group members—who are subject to the same discrimination effects as him—he attributes too much of their recognitions to their calibres. Because the effect of

[7] See, for instance, Newport (2014) on race, Pew Research Center (2017) on gender, Pew Research Center (2018) on income, and "Weniger Respekt und wachsende Fremdenfeindlichkeit", Frankfurter Allgemeine Zeitung, September 12, 2019, on immigrants.

discrimination is redistributive, however, a group's misestimates of discrimination do not bias its average estimate of calibre in the population.

The combination of in-group overestimation and overall correct estimation generates two manifestations of in-group bias. If the average calibres 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).[8]

To connect our results to stylized facts, suppose that there are two groups. Then, if the average calibres of the groups are equal ($A_1 - A_2 = 0$), each group believes itself to be better than the other group. This kind of bias is the most basic stylized fact in the literature on stereotypes, discrimination, prejudice, and racism.[9] Furthermore, some evidence indicates that the bias reflects a mistake (Bohren *et al.*, 2019; Lambin and Palikot, 2019).

A group may, however, fail to think of itself as better if the groups' true average calibres differ, or, stepping slightly outside our model, there are other biases that affect views equally across groups. In models by Frick *et al.* (2022) and Chauvin (2023), for instance, both a dominated and a dominant group may underestimate the privileges of the dominant group. If (fixing other parameters) the difference in average calibres or underestimation of privileges is sufficiently large, then the less fortunate group will think of itself as worse than the more fortunate 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 about the same ($\tilde{a}_1^1 - \tilde{a}_2^1 > \tilde{a}_1^2 - \tilde{a}_2^2$). Indeed, when researchers do not find unanimous support for absolute in-group bias, they typically observe relative in-group bias.[10] Sometimes, however, groups do not even display relative in-group bias (*e.g.* Card *et al.*, 2020), and our theory cannot account for this evidence.

Note that in Parts II and III of Proposition 1, in-group bias holds in an average sense. In Part II, in particular, each group overestimates itself relative to the average other group. The question arises whether in-group bias holds pairwise in general when there are more than two groups. The following example shows that it does not:

Example 1. $G = 3,\ K = 1,\ v_1^\eta = 1,\ m_1 = m_2 = m_3 = 1/3,\ \phi_{11} = 3,\ \phi_{21} = -2,\ \phi_{31} = -1,$ $v_g^q = 1$ for all g, $\tilde{a}_j - A_j = 1$ for all j, and all true calibres are normalized to zero. Then, by equation (3), we obtain

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

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 (recall that our framework allows any type of discrimination to be non-existent or go the other way), benefits blacks and hurts whites and especially Asians. Then, Group 3 overestimates Group 2 more than it does itself, and more than Group 2 overestimates itself. Hence, restricting

[8] Related to our in-group bias, Hestermann and Le Yaouanq (2021) show that a person thinks too highly of an outsider who receives the same outcome in the same circumstances as he does. They do not, however, explore general implications for group-based prejudices.

[9] Classics are Allport (1954) and Tajfel (1982). Mullen *et al.* (1992) provide a meta-analysis.

[10] For instance, Shayo and Zussman (2011), Gagliarducci and Paserman (2012), Zussman (2013), De Paola and Scoppa (2015), and Mengel *et al.* (2018).

attention to this pair of groups, both absolute and relative in-group bias are violated. Intuitively, since members of Group 3 are hurt by discrimination, they overestimate it. Furthermore, since they know that Group 2 is hurt even more by discrimination, they overestimate members of Group 2 more than other members of Group 3. Nevertheless, consistent with Part II of Proposition 1, 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.

3.2. *The effects of competition*

We now consider how the development of opposing interests with another group affects a group's views. Suppose that groups g and g' are initially not affected by the same types of discrimination ($\phi_{gk}\phi_{g'k} = 0$ for all k). Then a new type of discrimination emerges, positioning groups g and g' against each other: $m_g\phi_{gK+1} + m_{g'}\phi_{g'K+1} = 0$, with $\phi_{gK+1} \neq 0$. As a potential example, northern whites experiencing an inflow of blacks could think that they are on opposite sides of local issues, such as housing, schools, and jobs.

Proposition 2. *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, g'.*

A member of group g overestimates discrimination in favour of or underestimates discrimination against group g', negatively biasing his opinion of group g' (Part I). This effect helps explain evidence that greater local ethnic diversity increases racial animus (*e.g.* Branton and Jones, 2005), and that immigration triggers hostile reactions by natives (Tabellini, 2019). More generally, the result says that a person has more negative views about groups he considers competitors. This pattern is one of the cornerstones of group conflict theory (*e.g.* Jackson, 2011). For instance, Stephan *et al.* (1999) document that the negative stereotyping of immigrants in the U.S. 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.

By Part II, new competition raises a person's (already too high) view of his own group. His bias regarding the new type of discrimination provides a new excuse for his low recognition, and means that he attributes more of group members' recognitions to their calibres.

At the same time, Part III says that bias substitution occurs: while group g's opinion of group g' decreases, its opinion of other out-groups improves. As the agent attributes his low recognition in part to the new type of discrimination, his biases regarding the other types of discrimination decrease. This means that he attributes more of the other groups' recognitions to their calibres.

In an example of bias substitution, Fouka *et al.* (2022) document that the inflow of blacks to northern U.S. cities 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 mitigates negative views domestic groups may hold about each other.

3.3. *The effects of information*

This subsection analyses the effects of information on the agent's beliefs. Note that if a correctly specified agent has sufficient information to form confident (deterministic) beliefs—as the agent does in our model—then those beliefs must be correct and hence impervious to additional information. The same is not the case for a misspecified agent, leading to the natural question: can more information mitigate such an agent's biases about others?

Theorem 1 implies that two types of information cannot. First, since we are focusing on long-run beliefs, access to more realizations of the same signals does not necessarily lower biases. Second, since υ_j^q does not appear in equation (3), an improvement in the agent's information about others' recognitions does not affect his long-run biases. Intuitively, knowing more about the successes of other groups does not help because it does not affect the central tension driving the agent's biases: the gap between his overconfident self-view and his actual outcomes. These predictions are consistent with some null effects of information on discrimination documented in the literature (*e.g.* Bertrand and Mullainathan, 2004; Boring, 2017).[11]

Instead, consider providing information about discrimination:

Proposition 3. *Suppose discrimination of type k affects agent i ($\phi_{g_i k} \neq 0$). An increase in the precision $1/\upsilon_k^\eta$ 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.*

(II) *(No-excuse effect). Lowers his view $\tilde{a}_{g_i}^i$ about others in his group.*

(III) *(Bias substitution). Raises his bias $|\tilde{\theta}_{k'}^i - \Theta_{k'}|$ 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 $\sum_{g \neq g_i} m_g \tilde{a}_g^i$ of other groups.*

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

More information about discrimination of type k has both benefits and drawbacks. It directly reduces the plausibility of a biased view about type-k discrimination, lowering the agent's bias on this dimension (Part I). Similarly, the information reduces the plausibility of a biased view about overall discrimination affecting the agent, lowering his misperceptions about other in-group members (Part II). Seeking alternative explanations for his recognition, however, bias substitution again occurs: the agent's biases about other types of discrimination affecting him increase (Part III).

The effects on the agent's views about other groups are mixed as well. Part IV says that his average view of outside groups rises, so that he improves his opinion of at least one group. By Part V, however, his bias regarding groups that are not affected by discrimination of type k rise. In particular, if he harbours any unrelated prejudices, these increase.

The above results yield a novel perspective on Allport's (1954) influential and well-documented contact hypothesis—that contact between groups reduces prejudices (for evidence, see Pettigrew and Tropp, 2006; Lowe, 2021; Corno *et al.*, 2022). Consistent with the common view that a primary channel is informational, we think of contact as providing information about the calibre of an out-group member. In a model of correctly specified learning, information about one person is likely to have a limited spillover effect on views about a large and diverse group, especially for an agent who has plenty of information to begin with. In our model, in contrast, the spillover effect can be more drastic. Suppose that agent i learns individual j's calibre (so that $\tilde{a}_j^i = a_j$), and j is subject to only one type of discrimination, k ($\phi_{g_j k} \neq 0$, but $\phi_{g_j k'} = 0$

[11] Some studies that do find a positive effect of information, such as Kaas and Manger (2012) looking at reference letters and Tjaden *et al.* (2018) looking at online reviews, involve direct information about the person's character or quality. We analyse such information below.

for all $k' \neq k$).[12] Then, *j*'s recognition q_j becomes another signal of—and hence improves agent *i*'s information about—discrimination θ_k. As a result, agent *i*'s bias about *j*'s entire group decreases.[13]

Unlike better information about a single type of discrimination, a balanced improvement in information about all types of discrimination is unambiguously beneficial:

Proposition 4. *A proportional increase in the precisions* $1/\upsilon_k^\eta$ *of the agent's signals about discrimination lowers all his (non-zero) biases regarding discrimination and others' calibres.*

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 4 says that the disadvantaged group will then have less biased beliefs.

The preceding results provide ways to distinguish our model from a "reverse-causality" alternative in which overconfidence derives from false beliefs about discrimination or others' calibres, rather than vice versa. An agent may, for instance, inherit stubborn, negatively biased beliefs about competitor groups from his parents. Observing the recognitions of these groups, he concludes that there is discrimination in favour of them and against his groups. Observing his own recognition, then, he overestimates his calibre.

While sharing the basic prediction that overconfidence and false social beliefs are related, the reverse-causality model differs in at least three ways from ours. First, it fails to predict changes in beliefs about groups in response to information about discrimination. This is because beliefs about groups are either stubborn (and hence do not change) or derive directly from stubborn beliefs about discrimination. Second, similarly, the reverse-causality model does not predict bias-substitution-type changes in beliefs about unrelated groups in response to competition with a new group. Third, by equation (3), our model predicts that a person's bias about himself is greater than his bias about his fellow group members.[14] As a result, the average person overestimates himself relative to his in-group. In the reverse-causality model, the agent's biases about himself and his average in-group member are identical—both equal his total bias about the types of discrimination affecting the group.

3.4. *Similarity bias*

In this subsection, we consider the special case of our model in which groups are defined by vectors of characteristics, such as black/white and female/male. We identify sufficient (albeit not necessary) conditions for a variant of in-group bias, similarity bias: that a person has a more positively biased opinion about more similar others.

Suppose that individual *j* has characteristics $c_j = (c_{j1}, \ldots, c_{jK}) \in \{0,1\}^K$, where $c_{jk} = 1$ means that she has characteristic *k* (*e.g.* is black). A group consists of individuals who share all characteristics, and is thus defined by a characteristic vector *c*. Furthermore, discrimination of type *k* affects individuals who have characteristic *k* and those who do not in opposite directions. We say that agent *i* is more similar to individual *j* than to individual j' if whenever j' shares a

[12] In Appendix B, we show that the logic applies also if *j* is subject to more types of discrimination.

[13] Some papers find that contact reduces prejudice only in specific environments, *e.g.* when the interaction is cooperative (*e.g.* Lowe, 2021). Our theory is consistent with such findings if these environments generate more accurate information about the out-group member, but it does not explain why this would be the case.

[14] This is immediate from observing that for a fellow group member ($g_j = g_i$), the coefficient scaling the agent's overconfidence in the equation is $\frac{\sum_k \phi^2_{g_i k} \upsilon^\eta_k}{\upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} < 1$.

characteristic with i, so does j (*i.e.* $c_{j'k} = c_{ik} \Rightarrow c_{jk} = c_{ik}$); and the relationship is strict if the characteristic vectors of j and j' are not identical.

Proposition 5 (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 individual j than to individual j', then his long-run bias regarding the calibre of j is (strictly) greater than his long-run bias regarding the calibre of j', i.e. $\tilde{a}^i_j - A_j \geq \tilde{a}^i_{j'} - A_{j'}$.*

A sufficient condition for similarity bias is that the impact of type-k discrimination depends only on whether a person has characteristic k. Then, similarity determines how much agent i believes that discrimination hurting him also hurts rather than helps individual j, so it determines how much of j's recognition i attributes to calibre.[15]

When there is a single dimension, similarity bias reduces to a two-group version of in-group bias discussed above. While the multi-attribute version of similarity bias has to our knowledge not been directly tested, some evidence does seem consistent with it. Jackson *et al.* (2022) document that students are more likely to form friendship and study links with others who match more of their attributes, and Banal-Estañol *et al.* (2023) find that grant applicants are more likely to be successful if panellists share more of their characteristics. These findings could be driven by similarity-biased beliefs, but also by taste (although, as we discuss in the conclusion, those "tastes" may actually be driven by incorrect beliefs).

4. MODEL VARIANTS

4.1. *Prejudice without discrimination or group knowledge*

We show that prejudiced beliefs can arise even if the agent does not entertain the possibility of systematic discrimination, and has no pre-existing notion of groups. Suppose that $I \geq 3$, and agent i observes a sequence of realizations of each individual j's recognition,

$$q_j = a_j + \psi_j \epsilon_g + \epsilon_j, \tag{4}$$

where a_j is j's calibre, ϵ_g and ϵ_j are independent mean-zero Normal shocks with variances υ_g and υ_j, respectively, and $\psi_j \in \mathbb{R}$ with realization $\Psi_j \neq 0$ is the incidence of the group-level shock ϵ_g on j. As in our previous model, agent i is stubbornly overconfident about himself, but agnostic about the calibres of others. Furthermore, he knows υ_g, but not the ψ_j and υ_j, with his prior supported on $\mathbb{R}^I \times [\underline{\upsilon}, \infty)^I$, where $0 < \underline{\upsilon} \leq \min_j \upsilon_j$.[16] He understands the rest of the situation correctly, and updates his beliefs using Bayes' Rule. Since models with $\psi_1, \ldots, \psi_I$ and $-\psi_1, \ldots, -\psi_I$ are equivalent, we normalize Ψ_i, $\tilde{\psi}_i \geq 0$. Then, individuals with $\Psi_j > 0$ are "in the same boat" with—*i.e.* are affected by the group-level shock ϵ_g similarly to—the agent, and in this sense belong to his in-group; and those with $\Psi_j < 0$ belong to his out-group. But the agent does not initially know who is in which group.

[15] One type of discrimination the agent may consider is "exclusive discrimination" directed only against him. This corresponds to a characteristic k that only he has, with $\phi_{ck} < 0$ for his characteristic vector c. Assuming that exclusive discrimination is actually zero ($\Theta_k = 0$), the agent develops the "paranoid" view that there is some of it ($\tilde{\theta}^i_k > 0$), believing that "the world is out to get him". So long as the agent entertains the possibility of other types of discrimination too, his social biases are qualitatively unchanged.

[16] An increase in υ_g and a rescaling of all ψ_j are observationally equivalent, so assuming that the agent correctly understands υ_g is effectively a normalization.

Proposition 6. *Agent i's beliefs concentrate on a single $(\tilde{a}^i, \tilde{\psi}^i)$. The agent's long-run belief about individual j's calibre is*

$$\tilde{a}^i_j = A_j + \frac{\Psi_i \Psi_j \upsilon_g}{\upsilon^q_i + \Psi_i^2 \upsilon_g} \cdot (\tilde{a}_i - A_i), \tag{5}$$

and his long-run belief about ψ_j is $\tilde{\psi}^i_j = \kappa \cdot \Psi_j$, where $\kappa > 1$ is a constant.

Proposition 6 says that the agent learns his in-group and out-group, develops an in-group bias and comes to exaggerate the importance of groups in determining recognition (*i.e.* overestimates $|\psi_j|$). To develop intuition for these results, suppose first that the agent knows the ψ_j. Given his overconfidence, q_i is to him often surprisingly low, so he thinks that he must be unlucky. Since part of his luck derives from the common shocks, he thinks that individuals with $\psi_j > 0$ must also have been unlucky, and those with $\psi_j < 0$ must have been lucky. Hence, given their recognitions, he overestimates the former individuals and underestimates the latter ones.

But agent i does not know the ψ_j. It turns out that he correctly infers the sign of each ψ_j, so that the above logic regarding the estimation of calibres still holds. Additionally, the agent overestimates the importance of common shocks. For an intuition, suppose that $\psi_j, \psi_{j'} > 0$. 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 $\psi_{j'}$.

4.2. *Overconfidence through biased learning*

Our main model captures stubborn overconfidence by assuming that the agent has a fixed, overly positive belief about his calibre. We now consider one possible microfoundation for stubborn overconfidence, biased learning about oneself.

We modify the model introduced in Section 2 in the following ways. The agent has a full-support prior regarding his own calibre, and observes (in addition to q_j and η_k) signals $s_i = a_i + b + \epsilon^a_i$, where ϵ^a_i is a normally distributed error with mean zero and variance υ^a_i 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 biased way.

Proposition 7. *The agent's long-run bias about his own calibre is*

$$\tilde{a}_i - A_i = \frac{\upsilon^q_i + \sum_k \phi^2_{g_i k} \upsilon^\eta_k}{\upsilon^a_i + \upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot B, \tag{6}$$

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

$$\tilde{a}^i_j - A_j = \frac{\sum_k \phi_{g_i k} \phi_{g_j k} \upsilon^\eta_k}{\upsilon^a_i + \upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot B = \frac{\sum_k \phi_{g_i k} \phi_{g_j k} \upsilon^\eta_k}{\upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot (\tilde{a}_i - A_i). \tag{7}$$

His bias regarding discrimination of type k is

$$\tilde{\theta}^i_k - \Theta_k = \frac{-\phi_{g_i k} \upsilon^\eta_k}{\upsilon^a_i + \upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot B = \frac{-\phi_{g_i k} \upsilon^\eta_k}{\upsilon^q_i + \sum_{k'} \phi^2_{g_i k'} \upsilon^\eta_{k'}} \cdot (\tilde{a}_i - A_i). \tag{8}$$

Being described by the same formulas as in Theorem 1, the relationship between the agent's social beliefs and his overconfidence is exactly the same as in our main model. Accordingly, predictions regarding his social beliefs relative to each other are unchanged. Furthermore, it readily follows that the comparative statics in Propositions 2 to 5 also hold.[17,18]

In this version of the model, however, overconfidence can depend on the learning environment. We point out one relevant implication. To motivate our result, notice that in our basic model, agent i's biases are increasing in his overconfidence $\tilde{a}_i - A_i$ (equations (2) and (3)). This implies that if an outsider can make the agent more realistic about himself, then all his social prejudices decrease. Here, in contrast:

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

Confirming the classical intuition, providing better information about the agent lowers his overconfidence. But disconfirming the insight from the basic model that lowering overconfidence helps debias the agent, 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. To reliably lower the agent's biases, one must decrease the root misspecification—overconfidence in the basic model, misinterpreting signals about himself in this variant—that he has. In practice, however, it seems difficult to determine what this root bias is.

Uncertainty regarding the cause of overconfidence also has implications for the empirical testing of our predictions. Namely, because our theory does not imply an unambiguous positive relationship between overconfidence and social biases, it cannot be tested by simply looking at correlations between the two types of distorted beliefs. At the same time, controlling for all the information a person has appears impossible in practice. Nevertheless, our theory has many predictions that can be tested—and that, as we have discussed, are consistent with existing evidence.

4.3. *Short-run beliefs*

In this section, we investigate short-run beliefs in our main model (Section 2) when agent i knows the covariance matrix of the errors, and starts off with the prior that others' calibres and the degrees of discrimination are independently and normally distributed.

First, we investigate situations with a single type of discrimination ($K = 1$). We denote the prior variances of discrimination and individual j's calibre by $\bar{\upsilon}^{\theta}$, $\bar{\upsilon}_j^a$, expected mean beliefs in period t by $\tilde{\theta}^i(t)$, $\tilde{a}_j^i(t)$, the expected mean beliefs of an agent who correctly assesses his calibre ($\tilde{a}_i = A_i$) by $\theta^i(t)$, $a_j^i(t)$, and long-run beliefs derived in Theorem 1 by $\tilde{\theta}^i$, $\tilde{a}_j^i$.

Proposition 8. *Let $K = 1$. The biases in agent i's mean beliefs in period t are given by*

$$\tilde{\theta}^i(t) - \theta^i(t) = \beta_t(\tilde{\theta}^i - \Theta) \quad \textit{and} \quad \tilde{a}_j^i(t) - a_j^i(t) = \beta_t \frac{t/\upsilon_j^q}{1/\bar{\upsilon}_j^a + t/\upsilon_j^q}\left(\tilde{a}_j^i - A_j\right),$$

[17] For Proposition 2, this requires imposing that υ_i^a is common across a group (like υ_i^q is).

[18] In a model of learning with selective memory, Fudenberg *et al.* (2024, Section IV.B and Proposition 6) establish an analogue of Proposition 7. They show that the implications of dogmatic overconfidence for long-run beliefs are identical to those of a positive memory bias that generates the same level of overconfidence. This suggests that one can also think of our model as capturing the effect of a positive memory bias, so that the exact source of overconfidence is not crucial for our main qualitative findings.

where

$$\beta_t = \frac{t/\upsilon^\eta}{1/\bar{\upsilon}^\theta + t/\upsilon^\eta + \sum_j (\phi_{g_j 1})^2 \cdot 1/\bar{\upsilon}_j^a \cdot \frac{t/\upsilon_j^q}{1/\bar{\upsilon}_j^a + t/\upsilon_j^q}}.$$

The proposition implies that the agent's short-run biases have the same sign as and a lower magnitude than his long-run biases. Thus, for $K = 1$ our qualitative results survive.

Next, we note that if the variance of the agent's prior belief is sufficiently large—*i.e.* he is sufficiently uncertain to start with—then the long-run biases we have derived approximate his average biases in every period for any K.[19] Hence, in this case all of our insights hold on average in any period (in addition to holding with probability 1 in the long run).

To conclude, we show through an example that if the agent's prior is not sufficiently uncertain and $K > 1$, then short-run and long-run biases can be qualitatively different.

Example 2. $I = 3$, $K = 2$, $i = 1$, each (representative) individual is in a separate group, $\phi_{11} = \phi_{32} = -1$, $\phi_{21} = \phi_{22} = 1$, $\phi_{12} = \phi_{31} = 0$, and all priors and errors have variance 1.

Type-1 discrimination affects Groups 1 and 2, while Type-2 discrimination affects Groups 2 and 3. Theorem 1 then implies that agent 1 is in the long run unbiased about θ_2 and a_3. Yet applying the updating formula for Normal distributions, it is easy to check that after finite time he is on average biased about both. Intuitively, due to his overconfidence, agent 1 immediately starts overestimating the degree of discrimination θ_1 against him. Consequently, individual 2's recognition—which he thinks increases in type-1 discrimination—appears to him too low. In the short run, he attributes this discrepancy partly to a_2 and partly to θ_2, thinking that individual 2 suffers from type-2 discrimination. He therefore underestimates individual 3's calibre a_3 as well. In the long run, however, agent 1 attributes individual 2's (seemingly) low recognition solely to a_2, as a biased belief about θ_2 does not help him explain other observations.

Nor does bias substitution generally hold in the short run. Indeed, suppose that the agent receives extremely precise information about θ_1. Then, his bias about θ_1 becomes small, and by the above logic, so do his biases about θ_2 and a_3. The short-run biases about θ_1 and θ_2 are complements because the latter bias derives from the former.[20]

5. MULTI-DIMENSIONAL MISSPECIFIED LEARNING

This section derives a theoretical result that we used throughout the article, and that might be useful for others studying implications of misspecifications in multidimensional settings. To the best of our knowledge, ours is the first closed-form solution for the long-run outcome of a misspecified learning process with high-dimensional interdependent beliefs.[21]

[19] To see this, denote the mean and covariance matrix of the prior by $(\bar{a}, \bar{\theta})$ and Σ_0, respectively. The agent's expected posterior mean belief in period t is given by $(\tilde{a}(t)\,\tilde{\theta}(t))^T = (\Sigma_0^{-1} + t\hat{\Sigma}^{-1})^{-1}(\Sigma_0^{-1}(\bar{a}\,\bar{\theta})^T + t\hat{\Sigma}^{-1}(\tilde{a}\,\tilde{\theta})^T)$; see the proof of Proposition 8. If the prior variance goes to infinity, Σ_0^{-1} converges to the null matrix and hence $(\tilde{a}(t), \tilde{\theta}(t))$ converges to the long-run belief $(\tilde{a}, \tilde{\theta})$.

[20] For formal simplicity, our example features unbiased long-run beliefs about θ_2 and a_3. But a modification in which ϕ_{12} is slightly negative shows that short-run and long-run biases can have strictly opposite signs. Then, agent 1 is in the long run positively biased about individual 3, but by continuity of his beliefs in ϕ_{12}, in the short run he is still negatively biased. Furthermore, in this case, biases about θ_1 and θ_2 are strict substitutes in the long run but strict complements in the short run.

[21] 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.

The agent makes inferences about a fixed vector of *fundamentals* $f \in \mathbb{R}^L$, whose realization is F. In each period t, he observes a *signal*

$$r_t = Mf + \epsilon_t \in \mathbb{R}^D,$$

where $M \in \mathbb{R}^{D\times L}$ is a matrix of rank L and $\epsilon_t \in \mathbb{R}^D$ is a vector of errors that are jointly normally distributed with mean zero and positive definite covariance matrix Σ, and that are independent over time.[22] The agent updates 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 $\mathbf{P}_t$ assigns to the set A after the sequence of signals $r = (r_1, r_2, \ldots, r_t)$ is

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

where the likelihood equals

$$\ell_t(r \mid f', \Sigma') = \prod_{z=1}^{t} \frac{1}{\sqrt{(2\pi)^L \det \Sigma'}} \exp\left(-\frac{1}{2}(r - Mf')^T \Sigma'(r - Mf')\right). \tag{9}$$

The agent is misspecified: he believes with certainty that f_i equals $\tilde{f}_i$. 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 $\mathcal{M}$ the set of positive definite symmetric matrices whose eigenvalues are all greater than $\underline{\lambda}$, where $\underline{\lambda}$ is chosen to be sufficiently small.[23] In our main 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 $\tilde{\Sigma}$:

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

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

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

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,

[22] If M had lower rank, there would be different vectors of fundamentals that entail the same distribution of signals and hence the agent could not learn the fundamentals.

[23] 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 $\tilde{f}_j = F_j + \frac{[M^T\Sigma^{-1}M]^{-1}_{ij}}{[M^T\Sigma^{-1}M]^{-1}_{ii}}(\tilde{f}_i - F_i)$ for $j \neq i$ in Case (III). The agent's long-run beliefs do not depend on the precise choice of $\underline{\lambda}$.

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

(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*

$$\tilde{f}_j - F_j = \frac{(M^T\tilde{\Sigma}^{-1}M)^{-1}_{ij}}{(M^T\tilde{\Sigma}^{-1}M)^{-1}_{ii}}(\tilde{f}_i - F_i). \tag{10}$$

(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*

$$\tilde{\Sigma} - \Sigma = (M(\tilde{f} - F))(M(\tilde{f} - F))^T. \tag{11}$$

(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*

$$\tilde{f}_j - F_j = \frac{\left[M^T\Sigma^{-1}M\right]^{-1}_{ij}}{\left[M^T\Sigma^{-1}M\right]^{-1}_{ii}}(\tilde{f}_i - F_i), \tag{12}$$

and his bias about the covariance matrix is given by Expression (11).

The initial part of the proof of Theorem 2 follows steps commonly used in econometric and statistical analyses of misspecification. First, we verify that the assumptions in Berk (1966) are satisfied. Then, applying Berk's seminal result, beliefs concentrate on the set of minimizers of the Kullback–Leibler divergence. Now it is well-known that for Normal signals, the Kullback–Leibler divergence assigned to the parameters $(\hat{f}, \hat{\Sigma})$ when the true parameters equal (F, Σ) is

$$D\left(F, \Sigma \parallel \hat{f}, \hat{\Sigma}\right) = \frac{1}{2}\left(\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + (M(\hat{f} - F))^T\hat{\Sigma}^{-1}M(\hat{f} - F) - n + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right). \tag{13}$$

In the second part of the proof, we derive the unique minimizer of (13) over the support specified in Cases (I), (II), and (III). Since our type of misspecification has not been analysed in the literature, this part of the proof is novel. Case (I) can be verified by taking first-order conditions with respect to the fundamentals. But Cases (II) and (III) are non-trivial semi-definite programming problems because (13) involves the determinant of $\hat{\Sigma}$, which is not a tractable function in general. We proceed 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.

Notwithstanding the technical nature of our proof, intuition for our results can be gleaned by looking at (13) in the special case where the covariance matrix Σ is known and errors are independent, so that Σ is diagonal. Then, the objective function reduces to $(M(\hat{f} - F))^T\Sigma^{-1}M(\hat{f} - F)$: the agent minimizes the weighted sum of the squared mean errors in his observations (the differences between his observations and his expectations), with weights equal to the precisions of his signals. Our formulas in Theorem 1 derive from this problem, and we have used properties of this problem to explain the logic behind our main results. In particular, the agent's misspecification (overconfidence) introduces errors in his observations (*e.g.* in his recognition), which is reduced by biased beliefs about other fundamentals (*e.g.* discrimination against his group). Further, if one of the agent's observations becomes more precise (*e.g.* regarding one type of discrimination), the weights in the minimization problem change, leading to bias substitution.

The trickier parts of our proofs are in establishing that the above logic works also when the agent does not know the covariance matrix. Indeed, notice that plugging $\tilde{\Sigma} = \Sigma$ into Expression (10) yields Expression (12). Hence, when the agent is initially agnostic about 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. Investigating what happens when in addition the errors are correlated (as in Proposition 6) is also much less obvious.

Our flexible theoretical apparatus provides a tool that can help researchers understand the implications of biases beyond overconfidence. Building on our results, for instance, He *et al.* (2024) analyse what a misspecified agent concludes about the biases of his information sources; Kornemann (2024) studies what happens when the agent is misspecified about the matrix M, such as when he interprets observations using a simplified, sparse model; and in Appendix B, we consider the situation in which the agent has stubborn beliefs about two fundamentals. Economic applications abound. In the political arena, for instance, a person may have the stubborn belief that Democrats/Republicans are evil, with implications for his views about a multitude of social issues. Going further, our model can serve as an input into theories of propaganda, asking what misspecified beliefs a politician wants to instil given a set of goals and constraints. And in the personal arena, an individual may misperceive an aspect of others' preferences or beliefs, and thus misinterpret a range of their behaviours. Although a multitude of such misperceptions have been documented (see Bursztyn and Yang, 2022, for a review), their ultimate sources, interrelationship, and implications for multidimensional observations have not been analysed in detail.

6. RELATED LITERATURE

In this section, we relate our theory to research not discussed so far. 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 on misspecified learning typically restricts attention to a one- or two-dimensional state of the world.[24] Unlike many others, however, we do not investigate behaviour, and assume normally distributed signals.

The agent's biased interpretation of the signal about his own calibre in Section 4.2 is naturally interpreted as driven by motivated reasoning (Bénabou and Tirole, 2016). Within discrimination settings, Rackstraw (2022), Eyting (2024), and Stoetzer and Zimmermann (2024) experimentally investigate whether motivated reasoning drives subjects' updating. In contrast, we theoretically derive what stereotypes an agent eventually holds while allowing for multiple dimensions of discrimination.

There is a large sociology and social-psychology literature on prejudice, but to our knowledge no theory is based on overconfidence, connects prejudice to opinions about discrimination, or makes precise comparative-statics predictions. Most related, social identity theory (Tajfel,

[24] Papers in this literature not mentioned previously focus on different issues than our article, including inferences by individuals who ignore some explanatory variables (Hanna *et al.*, 2014), misunderstand causal relationships (Levy *et al.*, 2022), misinterpret social observations (Bohren, 2016; Levy and Razin, 2017; Bohren and Hauser, 2019; Frick *et al.*, 2020), or draw incorrect inferences from their own past behaviour (Heidhues *et al.*, 2022).

1982) posits that individuals identify with a few relevant groups, so that thinking positively about these in-groups and negatively about out-groups leads them to think and feel positively about themselves. Our theory also connects a person's prejudices to his views about himself, but through a different—in a sense reverse—logic: he thinks positively about himself, and this leads him to develop social biases.

An influential body of research demonstrates that prejudice and discrimination can operate implicitly outside the person's awareness (*e.g.* Bertrand *et al.*, 2005). Our framework is predicated on an inferential 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 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 (Jost *et al.*, 2009). In this sense, our model is not contradictory to implicit bias.

Bordalo *et al.* (2016) model stereotypes by assuming 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.

Glaeser (2005) presents a political-economy model of hate in which beliefs about the harmfulness of others are created by politicians' messages. 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 studied beliefs, it is natural to ask what our theory implies for discriminatory behaviour. 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). Then, our model can be thought of as one of misspecified statistical discrimination—the agent uses group membership as a signal to guide behaviour (*e.g.* whom to hire), but he does so incorrectly.[25] 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 colour unless it is associated in his mind with some meaning about what such others are like.

APPENDIX

A. *Proofs*

Theorem 1 follows from Theorem 2 which we prove later in the Appendix.

[25] Others also note that it is essential to distinguish correct statistical discrimination from "error discrimination" (England and Lewin, 1989) or "inaccurate statistical discrimination" (Bohren *et al.*, 2025).

Proof of Theorem 1: Let Σ^q, Σ^η be the covariance matrices of ϵ^q and ϵ^η,

$$\Sigma^q = \mathrm{diag}(\upsilon_1^q, \ldots, \upsilon_I^q), \quad \Sigma^\eta = \mathrm{diag}(\upsilon_1^\eta, \ldots, \upsilon_K^\eta)$$

and observe that they are invertible as the variances are greater than zero. We next show that this model can be reduced into our model in Section 5. Observe that one can write the vector $(q\ \eta)^T$ in matrix notation as

$$\begin{pmatrix} q \\ \eta \end{pmatrix} = \begin{pmatrix} Id & \Phi \\ 0 & Id \end{pmatrix} \cdot \begin{pmatrix} a \\ \theta \end{pmatrix} + \begin{pmatrix} \epsilon^q \\ \epsilon^\eta \end{pmatrix}, \tag{14}$$

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, and

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

By Theorem 2, agent i's bias about the calibre of agent j is given by

$$\tilde{a}^i_j - A_j = \frac{\left[M^T \Sigma^{-1} M\right]^{-1}_{ij}}{\left[M^T \Sigma^{-1} M\right]^{-1}_{ii}} \Delta_i = \frac{\left[\Sigma^q + \Phi\Sigma^\eta\Phi^T\right]_{ij}}{\left[\Sigma^q + \Phi\Sigma^\eta\Phi^T\right]_{ii}} (\tilde{a}_i - A_i) = \frac{\sum_k \phi_{g_i k}\phi_{g_j k}\upsilon_k^\eta}{\upsilon_i^q + \sum_k \phi_{g_i k}^2 \upsilon_k^\eta} \cdot (\tilde{a}_i - A_i).$$

By a similar argument,

$$\tilde{\theta}^i_k - \Theta_k = \frac{\left[M^T \Sigma^{-1} M\right]^{-1}_{i(I+k)}}{\left[M^T \Sigma^{-1} M\right]^{-1}_{ii}} \Delta_i = \frac{\left[-\Sigma^\eta\Phi^T\right]_{ik}}{\left[\Sigma^q + \Phi\,\Sigma^\eta\,\Phi^T\right]_{ii}} (\tilde{a}_i - A_i) = \frac{-\phi_{g_i k}\upsilon_k^\eta}{\upsilon_i^q + \sum_k \phi_{g_i k}^2 \upsilon_k^\eta} \cdot (\tilde{a}_i - A_i).$$ □

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

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

so its view of group g is

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

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

Furthermore,

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

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

II. Immediate from Part I.

III. Let $\tilde{a}_g = \sum_{i \in g} \tilde{a}_i / I m_g$, and note that by Theorem 1 $\tilde{a}_g > A_g$. We have

$$\begin{aligned}\sum_{g,g'} m_g m_{g'}(\tilde{a}^g_g - \tilde{a}^g_{g'}) &= \sum_g m_g \sum_{g'} m_{g'} \tilde{a}^g_g - \sum_g m_g \sum_{g'} m_{g'} \tilde{a}^g_{g'} \\ &= \sum_g m_g \tilde{a}^g_g - \sum_g m_g \sum_{g'} m_{g'} A_{g'} = \sum_g m_g \tilde{a}^g_g - \sum_{g'} m_{g'} A_{g'} \\ &= \sum_g m_g \tilde{a}^g_g - \sum_g m_g A_g = \sum_g m_g \frac{\sum_{k=1}^K \phi^2_{gk} \upsilon^\eta_k}{\upsilon^q_g + \sum_{k=1}^K \phi^2_{gk} \upsilon^\eta_k} (\tilde{a}_g - A_g) > 0. \qquad \square\end{aligned}$$

Proof of Proposition 2: We work with $K+1$ 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 a situation with K types of discrimination, and $s = 1$ to the new situation.

I. The view of group g about g' is

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

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

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

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

III. Notice that

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

where we have used that $m_g\phi_{gK+1} + m_{g'}\phi_{g'K+1} = 0$. 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 g regarding other groups must be higher for $s = 1$ than for $s = 0$. □

Proof of Proposition 3: By Theorem 1,

$$\left|\phi_{g_i k}\right| \left|\tilde{\theta}^i_k - \Theta_k\right| = \frac{\phi^2_{g_i k}\upsilon^\eta_k}{\upsilon^q_i + \sum_{k'} \phi^2_{g_i k'}\upsilon^\eta_{k'}} \cdot |\tilde{a}_i - A_i|, \tag{15}$$

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

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

which is increasing in v_k^η. Part III is implied as for $k'' \neq k$ the term in (15) is (weakly) decreasing in v_k^η, and strictly so if $\phi_{g_i k''} \neq 0$. Part IV follows since $\sum_g m_g \tilde{a}_g^i = \sum_g m_g A_g$ and by Part (II) $\tilde{a}_{g_i}^i$ is decreasing, so that $\sum_{g \neq g_i} m_g \tilde{a}_g^i$ must be increasing. For Part V, observe that as $\phi_{gk} = 0$ for group g, Theorem 1 implies that

$$\left|\tilde{a}_g^i - A_g\right| = \left|\frac{\sum_{k' \neq k} \phi_{g_i k'} \phi_{g k'} v_{k'}^\eta}{v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta}\right| \cdot |\tilde{a}_i - A_i|,$$

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

Proof of Proposition 4: Consider a proportional change that lowers all v_k^η by some constant factor $\alpha < 1$. By Theorem 1, this implies that agent i's long-run bias about discrimination toward group k is

$$|\tilde{\theta}_k^i - \Theta_k| = \left|\frac{-\phi_{g_i k} v_k^\eta}{\frac{v_i^q}{\alpha} + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta}\right| \cdot (\tilde{a}_i - A_i) \leq \left|\frac{-\phi_{g_i k} v_k^\eta}{v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta}\right| \cdot (\tilde{a}_i - A_i),$$

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

$$\left|\tilde{a}_j^i - A_j\right| = \left|\frac{\sum_k \phi_{g_i k} \phi_{g_j k} v_k^\eta}{\frac{v_i^q}{\alpha} + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta}\right| \cdot (\tilde{a}_i - A_i) \leq \left|\frac{\sum_k \phi_{g_i k} \phi_{g_j k} v_k^\eta}{v_i^q + \sum_{k'} \phi_{g_i k'}^2 v_{k'}^\eta}\right| \cdot (\tilde{a}_i - A_i),$$

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

Proof of Proposition 5: Theorem 1 implies that the difference in agent i's long-run bias about individual j and j' is

$$(\tilde{a}_j^i - A_j) - (\tilde{a}_{j'}^i - A_{j'}) = -\sum_k (\tilde{\theta}_k^i - \Theta_k)(\phi_{c_{jk}} - \phi_{c_{j'k}}).$$

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

$$(\tilde{a}_j^i - A_j) - (\tilde{a}_{j'}^i - A_{j'}) = \sum_{k: c_{j'k} \neq c_{ik} \wedge c_{j'k} \neq c_{jk}} -(\tilde{\theta}_k^i - \Theta_k)(\phi_{c_{jk}} - \phi_{c_{j'k}}).$$

Since characteristics are binary, for any dimension k in which $c_{j'k} \neq c_{ik} \wedge c_{j'k} \neq c_{jk}$, 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 Theorem 1 (i) $\phi_{c_{ik}} > 0$ implies $-(\tilde{\theta}_k^i - \Theta_k) > 0$ and $(\phi_{c_j,k} - \phi_{c_{j'},k}) > 0$; and (ii) $\phi_{c_{ik}} < 0$ implies $-(\tilde{\theta}_k^i - \Theta_k) < 0$ and $(\phi_{c_j,k} - \phi_{c_{j'},k}) < 0$. We conclude that in any dimension k in which $c_{j'k} \neq c_{ik} \wedge c_{j'k} \neq c_{jk}$, we have $-(\tilde{\theta}_k^i - \Theta_k)(\phi_{c_j,k} - \phi_{c_{j'},k}) > 0$. Thus, $(\tilde{a}_j^i - A_j) - (\tilde{a}_{j'}^i - A_{j'}) > 0$. □

To prove Proposition 6, we solve a more general model first in which recognition $q_j = a_j + \epsilon'_j$ is an unbiased signal of calibre 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 A.1 below. All other assumptions remain unchanged. In this case, one has:

Proposition A.1 (Correlated Errors and Biases). *Agent i's long-run bias about j is*

$$\tilde{a}_j^i - A_j = \frac{\Sigma_{ij}^q}{\Sigma_{ii}^q}(\tilde{a}_i - A_i), \tag{16}$$

while his bias about the covariance matrix is given by

$$\tilde{\Sigma}^q_{jj'} - \Sigma^q_{jj'} = (\tilde{a}^i_j - A_j)(\tilde{a}^i_{j'} - A_{j'}) = \frac{\Sigma^q_{ij'}\Sigma^q_{ij}}{(\Sigma^q_{ii})^2}(\tilde{a}_i - A_i)^2. \tag{17}$$

Proof: We apply Part III of Theorem 2 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 6: Observe that the true model of Proposition 6 is a special case of the model of Proposition A.1 in which $\epsilon'_j = \psi_j \epsilon_g + \epsilon_j$, where ϵ_g and ϵ_j are independent mean-zero Normal shocks with variances υ_g and υ_j. Note that the sum of Normal random variables is Normal, and the true variance–covariance matrix of the shocks ϵ'_j has entries $\Sigma^q_{jj'} = \psi_j \psi_{j'} \upsilon_g$ for $j \neq j'$ and $\Sigma^q_{jj} = \upsilon^q_j + \psi^2_j \upsilon_g$.[26]

The agent considers the subclass of subjective covariance matrices for which $\tilde{\Sigma}^q_{jj'} = \tilde{\psi}_j \tilde{\psi}_{j'} \upsilon_g$ for $j \neq j'$ and $\tilde{\Sigma}^q_{jj} = \tilde{\upsilon}^q_j + \tilde{\psi}^2_j \upsilon_g$. 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 $\tilde{\Sigma}^q_{jj'} = \tilde{\psi}_j \tilde{\psi}_{j'} \upsilon_g$ for $j \neq j'$ and $\tilde{\Sigma}_{jj} = \tilde{\upsilon}^q_j + \tilde{\psi}^2_j \upsilon_g$, and then verify that the solution to the relaxed problem satisfies these constraints.

By Proposition A.1, we have that in the solution to the relaxed problem is given by

$$\tilde{a}^i_j = A_j + \frac{\psi_i \psi_j \upsilon_g}{\upsilon^q_i + \psi^2_i \upsilon_g} \cdot (\tilde{a}_i - A_i),$$

and

$$\tilde{\Sigma}^q_{jj'} = \Sigma^q_{jj'} + \frac{\Sigma^q_{ij'}\Sigma^q_{ij}}{(\Sigma^q_{ii})^2}(\tilde{a}_i - A_i)^2.$$

Hence,

$$\tilde{\Sigma}^q_{jj'} = \psi_j \psi_{j'} \upsilon_g \left[1 + \frac{\psi^2_i \upsilon_g}{\left(\upsilon^q_i + \psi^2_i \upsilon_g\right)^2}(\tilde{a}_i - A_i)^2 \right] \quad \text{for } j \neq j',$$

and

$$\tilde{\Sigma}^q_{jj} = \upsilon^q_j + \psi^2_j \upsilon_g + \frac{\psi^2_j \psi^2_i \upsilon^2_g}{\left(\upsilon^q_i + \psi^2_i \upsilon_g\right)^2}(\tilde{a}_i - A_i)^2 \quad \text{for } j \neq i, \tag{18}$$

and finally

$$\tilde{\Sigma}^q_{ii} = \upsilon^q_i + \psi^2_i \upsilon_g + (\tilde{a}_i - A_i)^2.$$

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

$$\tilde{\psi}_j \tilde{\psi}_{j'} \upsilon_g = \psi_j \psi_{j'} \upsilon_g \left[1 + \frac{\psi^2_i \upsilon_g}{\left(\upsilon^q_i + \psi^2_i \upsilon_g\right)^2}(\tilde{a}_i - A_i)^2 \right] \quad \text{for all } j \neq j', \tag{19}$$

[26] To see that the chosen υ_g, υ_j implies a uniform bound on the covariance matrix as required by Theorem 2, observe that the covariance matrix is given by $\upsilon_g \times (\psi \otimes \psi') + \text{diag}(\upsilon_1, \upsilon_2, \ldots, \upsilon_I)$, where $\text{diag}(\upsilon_1, \upsilon_2, \ldots, \upsilon_I)$ denotes the diagonal matrix with entries $\upsilon_1, \ldots, \upsilon_I$. The smallest eigenvector of the covariance matrix is thus greater than $\min_{x:|x|=1} x^T[\upsilon_g \times (\psi \otimes \psi') + \text{diag}(\upsilon_1, \upsilon_2, \ldots, \upsilon_I)]x \geq x^T \text{diag}(\upsilon_1, \upsilon_2, \ldots, \upsilon_I)x = \min_j \upsilon_j$.

and

$$\tilde{v}_j^q + \tilde{\psi}_j^2 v_g = v_j^q + \psi_j^2 v_g + \frac{\psi_j^2 \psi_i^2 v_g^2}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \quad \text{for } j \neq i, \tag{20}$$

and finally

$$\tilde{v}_i^q + \tilde{\psi}_i^2 v_g = v_i^q + \psi_i^2 v_g + (\tilde{a}_i - A_i)^2. \tag{21}$$

Observe that (19) to (21) are solved by

$$\tilde{\psi}_j = \psi_j \sqrt{\left[1 + \frac{\psi_i^2 v_g}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2\right]}, \tag{22}$$

and own variances

$$\tilde{v}_i^q = v_i^q + \frac{\left(v_i^q + \psi_i^2 v_g\right)^2 - \left(\psi_i^2 v_g\right)^2}{\left(v_i^q + \psi_i^2 v_g\right)^2} (\tilde{a}_i - A_i)^2 \quad \text{and} \quad \tilde{v}_j^q = v_j^q. \tag{23}$$

We now argue that for $I \geq 3$, the solution given by (22) and (23) is unique. Dividing (19) 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 (19), $\tilde{\psi}_{j'}\tilde{\psi}_{j''}$ 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 (20) and (21). □

Proof of Proposition 7: Let e_i be the ith unit row vector, and Φ the matrix with $(\Phi)_{jk} = \phi_{g_j k}$. In the notation of Theorem 2,

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

and the agent is misspecified regarding b, with $\tilde{b} - B = -B$. It is easy to check 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 $(\phi_{g_i 1}, \ldots, \phi_{g_i K})$. We thus have

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

The formulas follow by applying Theorem 2, Part III. □

Proof of Corollary 1: The result follows from taking the derivative of the respective biases in Proposition 7 with respect to v_i^q. □

Proof of Proposition 8: Recall that the agent observes the signals

$$\begin{aligned} q_j(t) &= A_j + \sum_{k=1}^{K} \phi_{g_j k}\,\Theta_k + \epsilon_j^q(t), \quad j = 1,\dots,I, \\ \eta(t) &= \Theta + \epsilon^\eta(t). \end{aligned} \tag{24}$$

We assumed that there is a single dimension of discrimination, so that *e.g.* $\Theta = \Theta_1 \in \mathbb{R}$. The agent now observes two signals that are purely about discrimination: (1) the signal directly about discrimination

$$\eta = \Theta + \epsilon^\eta$$

and (2) the signal about the agent's own calibre

$$q_i = A_i + \phi_{g_i 1}\Theta + \epsilon_i^q .$$

We transform q_i into a new signal $\hat{q}_i$, which agent *i* believes to be an unbiased signal of Θ,

$$\hat{q}_i = (q_i - \tilde{a}_i)/\phi_{g_i 1} = \Theta - \frac{\tilde{a}_i - A_i}{\phi_{g_i 1}} + \frac{1}{\phi_{g_i 1}}\epsilon_i^q .$$

The direct signal η has precision $1/\upsilon^\eta$ and the second signal $\hat{q}_i$ has precision $(\phi_{g_i 1})^2/\upsilon_i^q$. This means that the overall information of these two signals can be summarized into a single signal given by

$$\hat{\eta}_i = \frac{1/\upsilon^\eta\,\eta + (\phi_{g_i 1})^2\,1/\upsilon_i^q\,\hat{q}_i}{1/\upsilon^\eta + (\phi_{g_i 1})^2\,1/\upsilon_i^q} = \frac{1/\upsilon^\eta\,\eta + \phi_{g_i 1}\,1/\upsilon_i^q\,(q_i - \tilde{a}_i)}{1/\upsilon^\eta + (\phi_{g_i 1})^2\,1/\upsilon_i^q}.$$

The precision of this signal $\hat{\eta}_i$ is equal to $1/\upsilon^\eta + (\phi_{g_i 1})^2\,1/\upsilon_i^q$. The signal $\hat{\eta}_i$ is a sufficient statistic for Θ from the point of view of the agent, in the sense that her posterior belief about Θ will be the same after observing (η, q_i) or $\hat{\eta}_i$.[27] The objective expectation of the signal $\hat{\eta}_i$ is given by the long-run belief

$$\mathbb{E}[\hat{\eta}_i] = \Theta - \frac{\phi_{g_i 1}1/\upsilon_i^q}{1/\upsilon^\eta + (\phi_{g_i 1})^2\,1/\upsilon_i^q}(\tilde{a}_i - A_i) = \Theta - \frac{\phi_{g_i 1}\upsilon^\eta}{\upsilon_i^q + (\phi_{g_i 1})^2\upsilon^\eta}(\tilde{a}_i - A_i).$$

We can now also transform the signal about agent $j \neq i$'s ability q_j in an invertible way such that agent *i* believes it to be an unbiased signal about a_j by defining $\hat{q}_j$ as

$$\hat{q}_j = q_j - \phi_{g_j 1}\hat{\eta}_i .$$

The objective expectation of $\hat{q}_j$ is given as

$$\mathbb{E}[\hat{q}_j] = A_j + \phi_{g_j 1}\frac{\phi_{g_i 1}\upsilon^\eta}{\upsilon_i^q + (\phi_{g_i 1})^2\upsilon^\eta}(\tilde{a}_i - A_i)\,.$$

As the distribution of q_j only depends on θ and a_j, we get that $(\hat{q}_{-i}, \hat{\eta}_i)$ is a sufficient statistic for computing the agent's beliefs about (a_{-i}, θ).

By slight abuse of notation, we denote by Φ the vector $(\phi_{g_j 1})_{j \neq i}$ and by a the vector $(a_j)_{j \neq i}$ to avoid the subindices in Φ_{-i}, a_{-i}. Recall that we denote by $(\tilde{a}^i, \tilde{\theta}^i)$ the long-run belief of agent *i*. The objective expectation of $(\hat{q}, \hat{\eta}_i)$ is exactly equal to the long-run belief derived in Theorem 2 and given as

$$\mathbb{E}\left[\begin{pmatrix} \hat{q}_{-i} \\ \hat{\eta}_i \end{pmatrix}\right] - \begin{pmatrix} A \\ \Theta \end{pmatrix} = \begin{pmatrix} \tilde{a} \\ \tilde{\theta}^i \end{pmatrix} - \begin{pmatrix} A \\ \Theta \end{pmatrix} = \begin{pmatrix} -\Phi \\ 1 \end{pmatrix}\left(\tilde{\theta}^i - \Theta\right).$$

[27] This follows from the updating rules for Normal signals given a Normal prior.

According to the agent's subjective expectation

$$\tilde{\mathbb{E}}[\hat{q}_j] = A_j \quad \text{and} \quad \tilde{\mathbb{E}}[\hat{\eta}_i] = \Theta.$$

The signals $(\hat{q}, \hat{\eta}_i)$ have the subjective covariance matrix

$$\hat{\Sigma} = \begin{bmatrix} \Sigma^q + \upsilon^\eta \Phi \Phi^T & -\upsilon^\eta \Phi \\ -\upsilon^\eta \Phi^T & \upsilon^\eta \end{bmatrix}.$$

Here $\Sigma^q = \mathrm{diag}(\upsilon_1^q, \upsilon_2^q, \ldots)$ is a diagonal matrix with the variance of the outputs of the different agent's on the diagonal. Denote by $\tilde{a}^i(t), \tilde{\theta}^i(t)$ the expected posterior mean belief in period t when the agent assigns probability 1 to his own calibre being $\tilde{a}_i$, and by $\bar{a}_j, \bar{\theta}$ the prior means of individual *j*'s calibre and discrimination, respectively. By the updating formula for Normal beliefs from Normal signals,[28] we have that

$$\begin{pmatrix} \tilde{a}^i(t) \\ \tilde{\theta}^i(t) \end{pmatrix} = (\Sigma_0^{-1} + t\hat{\Sigma}^{-1})^{-1} \left(\Sigma_0^{-1} \begin{pmatrix} \bar{a} \\ \bar{\theta} \end{pmatrix} + t\hat{\Sigma}^{-1} \begin{pmatrix} \tilde{a}^i \\ \tilde{\theta}^i \end{pmatrix} \right).$$

Here

$$\Sigma_0 = diag(\bar{\upsilon}_1^a, \bar{\upsilon}_2^a, \ldots, \bar{\upsilon}^\theta) = \begin{pmatrix} \Sigma_0^a & 0 \\ 0 & \bar{\upsilon}^\theta \end{pmatrix}$$

is a diagonal matrix with the variance of the prior belief about the different agents' calibres $a_1, a_2, \ldots$ and the state Θ on the diagonal. Denote by $a(t), \theta(t)$ the expected posterior mean belief in period t when the agent is correctly specified and assigns probability 1 to his own calibre being A_i. We have that

$$\begin{pmatrix} a^i(t) \\ \theta^i(t) \end{pmatrix} = (\Sigma_0^{-1} + t\hat{\Sigma}^{-1})^{-1} \left(\Sigma_0^{-1} \begin{pmatrix} \bar{a} \\ \bar{\theta} \end{pmatrix} + t\hat{\Sigma}^{-1} \begin{pmatrix} A \\ \Theta \end{pmatrix} \right).$$

We get that the difference between the mean belief of the correctly specified agent and the agent who misestimates his own calibre is given by

$$\begin{pmatrix} \tilde{a}^i(t) \\ \tilde{\theta}^i(t) \end{pmatrix} - \begin{pmatrix} a^i(t) \\ \theta^i(t) \end{pmatrix} = (\Sigma_0^{-1} + t\hat{\Sigma}^{-1})^{-1} t\hat{\Sigma}^{-1} \left[\begin{pmatrix} \tilde{a}^i \\ \tilde{\theta}^i \end{pmatrix} - \begin{pmatrix} A \\ \Theta \end{pmatrix} \right].$$

The matrix $\hat{\Sigma}$ has an inverse equal to

$$\hat{\Sigma}^{-1} = \begin{bmatrix} \Sigma^{q\,-1} & \Sigma^{q\,-1}\Phi \\ \Phi^T \Sigma^{q\,-1} & 1/\upsilon^\eta + \Phi^T \Sigma^{q\,-1} \Phi \end{bmatrix}.$$

We observe that

$$\begin{aligned} &\begin{bmatrix} \Sigma^{q\,-1} & \Sigma^{q\,-1}\Phi \\ \Phi^T \Sigma^{q\,-1} & 1/\upsilon^\eta + \Phi^T \Sigma^{q\,-1} \Phi \end{bmatrix} \left[\begin{pmatrix} \tilde{a}^i \\ \tilde{\theta}^i \end{pmatrix} - \begin{pmatrix} A \\ \Theta \end{pmatrix} \right] \\ &= \begin{bmatrix} \Sigma^{q\,-1} & \Sigma^{q\,-1}\Phi \\ \Phi^T \Sigma^{q\,-1} & 1/\upsilon^\eta + \Phi^T \Sigma^{q\,-1} \Phi \end{bmatrix} \begin{pmatrix} -\Phi \\ 1 \end{pmatrix} (\tilde{\theta}^i - \Theta) = \begin{pmatrix} 0 \\ 1/\upsilon^\eta \end{pmatrix} (\tilde{\theta}^i - \Theta). \end{aligned}$$

Multiplying by $(\Sigma_0^{-1} + t\hat{\Sigma}^{-1})$ yields that

$$\begin{aligned} &\left[\Sigma_0^{-1} + t\hat{\Sigma}^{-1} \right] \begin{pmatrix} \tilde{a}^i(t) - a^i(t) \\ \tilde{\theta}^i(t) - \theta^i(t) \end{pmatrix} = \begin{pmatrix} 0 \\ t/\upsilon^\eta \end{pmatrix} (\tilde{\theta}^i - \Theta) \\ \Leftrightarrow &\begin{bmatrix} \Sigma_0^{a\,-1} + t\Sigma^{q\,-1} & t\Sigma^{q\,-1}\Phi \\ \Phi^T t\Sigma^{q\,-1} & 1/\bar{\upsilon}^\theta + t/\upsilon^\eta + \Phi^T t\Sigma^{q\,-1}\Phi \end{bmatrix} \begin{pmatrix} \tilde{a}^i(t) - a^i(t) \\ \tilde{\theta}^i(t) - \theta^i(t) \end{pmatrix} = \begin{pmatrix} 0 \\ t/\upsilon^\eta \end{pmatrix} (\tilde{\theta}^i - \Theta). \end{aligned}$$

[28] See *e.g.* here https://en.wikipedia.org/wiki/Conjugate_prior#When_likelihood_function_is_a_continuous_distribution.

The solution to this equation is given as

$$\tilde{a}^i(t) - a^i(t) = -\left(\Sigma_0^{a-1} + t\Sigma^{q-1}\right)^{-1} t\Sigma^{q-1}\Phi\left(\tilde{\theta}^i(t) - \theta^i(t)\right)$$

$$\tilde{\theta}^i(t) - \theta^i(t) = \frac{t/v^\eta \times (\tilde{\theta}^i - \Theta)}{1/\bar{v}^\theta + t/v^\eta + \Phi^T t\Sigma^{q-1}\Phi - \Phi^T t\Sigma^{q-1}\left(\Sigma_0^{a-1} + t\Sigma^{q-1}\right)^{-1} t\Sigma^{q-1}\Phi}$$

$$= \frac{t/v^\eta}{1/\bar{v}^\theta + t/v^\eta + \sum_j (\phi_{g_j 1})^2\, 1/\bar{v}_j^a \,\frac{t/v_j^q}{1/\bar{v}_j^a + t/v_j^q}}(\tilde{\theta}^i - \Theta).$$

Where, in the last equation we used that Σ_0^a and Σ^q are diagonal matrices. Expressing the above equations component-wise yields the formulas in the statement. □

Proof of Theorem 2: For brevity, we denote the agent's long-run bias about fundamental j by

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

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

We first verify that the assumptions of Berk (1966) are satisfied. Part I requires that the subjective density is continuous in $(f', \Sigma') \in \operatorname{supp} \mathbf{P}_0$. The subjective density is

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

which is continuous as the determinant and the inverse of a matrix are continuous functions of the coefficients of the matrix, and the determinant of a matrix whose eigenvalues are bounded from below by $\underline{\lambda}$ is bounded from below by $\underline{\lambda}^n > 0$. Part II 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 neighbourhood $U \subset \operatorname{supp} \mathbf{P}_0$ 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

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

Let $\lambda_{\max}(\Sigma'')$ be the largest and $\lambda_{\min}(\Sigma'')$ the smallest eigenvalue of Σ''. Then,

$$|\log \ell_1(r_1 \mid f'', \Sigma'')| = \frac{1}{2}\left|\log[(2\pi)^L \det \Sigma''] + (r_1 - Mf'')^T (\Sigma'')^{-1}(r_1 - Mf'')\right|$$

$$\leq \frac{1}{2}\left|L \log[(2\pi)\lambda_{\max}(\Sigma'')] + \frac{1}{\lambda_{\min}(\Sigma'')}||r_1 - Mf''||^2\right|.$$

As the eigenvalues are a continuous function of the entries of the matrix and bounded from below by $\underline{\lambda}$, we get that the above function is continuous in (f'', Σ'') and thus that the supremum is finite over every neighbourhood U.

Finally, Part IV is that for every constant $\gamma \in \mathbb{R}$ there exists a set $D \subset \operatorname{supp} \mathbf{P}_0$ with compact complement $(\operatorname{supp} \mathbf{P}_0) \setminus D$ such that

$$\mathbb{E}\left[\sup_{(f'', \Sigma'') \in D} \log \ell_1(r_1 \mid f'', \Sigma'')\right] \leq \gamma. \tag{25}$$

Fix $\delta_1, \delta_2 > 0$ and let D be the set of (f'', Σ'') such that either $||M(F - f'')|| > \delta_1$ or the covariance matrix Σ'' has its largest eigenvalue strictly greater than δ_2. For all $(f'', \Sigma'') \in D$ and $||M\epsilon_1|| \leq \delta_1/4$ the log-likelihood satisfies

$$\log \ell_1(r_1 \mid f'', \Sigma'') = -\frac{1}{2}\left(\log[(2\pi)^L \det \Sigma''] + (r_1 - Mf'')^T (\Sigma'')^{-1}(r_1 - Mf'')\right)$$

$$\leq -\frac{1}{2}\left(L \log(2\pi) + (L-1)\log \lambda_{\min}(\Sigma'') + \log(\lambda_{\max}(\Sigma'')) + \frac{1}{\lambda_{\max}(\Sigma'')}||r_1 - Mf''||^2\right).$$

Denote by $a \wedge b$ the minimum 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

$$\begin{aligned}
&\le -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log(\delta_2 \wedge ||r_1 - Mf''||^2)\right)\\
&= -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log(\delta_2 \wedge ||M(F-f'') + M\epsilon_1||^2)\right)\\
&\le -\frac{1}{2}(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'')\\
&\quad + \log(\delta_2 \wedge [||M(F-f'')||^2 + ||M\epsilon_1||^2 - 2||M(F-f'')|| \times ||M\epsilon_1||])).
\end{aligned}$$

As right-hand-side above is decreasing in $||M(F-f'')||$ for $||M\epsilon_1|| < \delta_1/2$, the maximum is attained at $||M(F-f'')|| = \delta_1$ and we obtain the following bound

$$\le -\frac{1}{2}\left(L\log(2\pi) + (L-1)\log\lambda_{\min}(\Sigma'') + \log\left(\delta_2 \vee \delta_1^2/2\right)\right).$$

As the lowest eigenvalue of all covariance matrices in $\operatorname{supp}\mathbf{P}_0$ is bounded from below by $\underline{\lambda} < 1$ we have $\log\ell_1(r_1 \mid f'', \Sigma'') \le L/2|\log(\underline{\lambda})|$ for δ_1, δ_2 large enough. That implies for δ_1, δ_2 large enough

$$\begin{aligned}
&\mathbb{E}\left[\sup_{(f'',\Sigma'')\in D}\log\ell_1(r_1 \mid f'', \Sigma'')\right] \le \mathbb{E}\left[\mathbf{1}_{||M\epsilon_1||\le\delta_1/4}\sup_{(f'',\Sigma'')\in D}\log\ell_1(r_1 \mid f'', \Sigma'')\right] + L/2|\log(\underline{\lambda})|\\
&\le -\frac{1}{2}(L\log(2\pi) + (L-1)\log\underline{\lambda} + \log\left(\delta_2 \wedge \delta_1^2/2\right))\mathbb{P}[||M\epsilon_1|| \le \delta_1/4] + L/2|\log(\underline{\lambda})|.
\end{aligned}$$

As $\lim_{\delta_1\to\infty}\mathbb{P}[||M\epsilon_1|| \le \delta_1/4] = 1$ it follows that the left-hand-side of (25) becomes arbitrarily small for δ_1 and δ_2 large enough. We are left to argue that the complement of D is compact for every δ_1, δ_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 $[\underline{\lambda}, \delta_2]$ and vectors f'' with $||M(F-f'')|| \le \delta_1$. As $||\Sigma''||$ equals the largest eigenvalue, and thus is less than δ_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 $||F - f''|| = ||M^*M(F-f'')|| \le ||M^*|| \times ||M(F-f'')|| \le \delta_1||M^*||$. 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 $\mathbf{P}_0$

$$\underset{(\hat{f},\hat{\Sigma})\in\operatorname{supp}\mathbf{P}_0}{\arg\min}\; D\left(F, \Sigma \parallel \hat{f}, \hat{\Sigma}\right), \tag{26}$$

where the Kullback–Leibler divergence is given by

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

We will argue that (26) 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 the true and subjective models are both Normal, the Kullback–Leibler divergence is given by (13).[29] Throughout, we denote by $\tilde{f}, \tilde{\Sigma}$ the agent's 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 its elements by $(B_{jk})_{j,k\in\{1,\ldots,L\}}$. 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 $B = M^T\tilde{\Sigma}^{-1}M$.

[29] See for example https://en.wikipedia.org/wiki/Multivariate_normal_distribution.

We first analyse Case (I): We solve (26) over $\Delta = \hat{f} - F$. As we can ignore all terms in (13) that do not depend on $\hat{f}$, the problem becomes

$$\begin{aligned}\underset{\hat{f}:\ \hat{f}_i=\tilde{f}_i}{\arg\min}\,(M(\hat{f}-F))^T\tilde{\Sigma}^{-1}M(\hat{f}-F) &= F + \underset{\Delta:\ \Delta_i=\tilde{f}_i-f_i}{\arg\min}\ \Delta^T\left(M^T\tilde{\Sigma}^{-1}M\right)\Delta \\ &= F + \underset{\Delta:\ \Delta_i=\tilde{f}_i-f_i}{\arg\min}\ \sum_{k=1}^{L}\sum_{j=1}^{L} B_{kj}\Delta_k\Delta_j. \end{aligned} \tag{27}$$

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 kth 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 (27) in terms of $\Delta_{-i} = \sum_{j\neq i} e_j\Delta_j$

$$\begin{aligned}\Delta^T B\Delta &= \left(e_i\Delta_i + \sum_{j\neq i} e_j\Delta_j\right)^T B\left(e_i\Delta_i + \sum_{j\neq i} e_j\Delta_j\right) = \left(e_i\Delta_i + \Delta_{-i}\right)^T B\left(e_i\Delta_i + \Delta_{-i}\right) \\ &= (e_i\Delta_i)^T B(e_i\Delta_i) + \Delta^T_{-i}B\Delta_{-i} + 2(e_i\Delta_i)^T B\Delta_{-i}. \end{aligned} \tag{28}$$

The Hessian with respect to Δ_{-i} of (28) 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

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

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

We next analyse Case (II): In this case, we minimize (13) over $\hat{\Sigma}$:

$$\underset{\hat{\Sigma}}{\arg\min}\left(\operatorname{tr}(\hat{\Sigma}^{-1}\Sigma) + (M\Delta)^T\hat{\Sigma}^{-1}(M\Delta) + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right). \tag{29}$$

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 (29) 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

$$\begin{aligned}
&\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + y^T\hat{\Sigma}^{-1}y + \log\frac{\det\hat{\Sigma}}{\det\Sigma} = \mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + \mathrm{tr}(y^T\hat{\Sigma}^{-1}y) + \log(\det\hat{\Sigma}) - \log(\det\Sigma)\\
&\quad = \mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + \mathrm{tr}(\hat{\Sigma}^{-1}[y\otimes y^T]) - \log(\det\hat{\Sigma}^{-1}) - \log(\det\Sigma)\\
&\quad = \mathrm{tr}(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) - \log(\det\hat{\Sigma}^{-1}) - \log(\det\Sigma)\\
&\quad = \mathrm{tr}(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) - \log\det(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) + \log\det(\Sigma^{-1}(\Sigma + [y\otimes y^T]))\\
&\quad = \mathrm{tr}(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) - \log\det(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) + \log\det(Id + \Sigma^{-1}[y\otimes y^T]). \qquad (30)
\end{aligned}$$

Here, we used in the first equality that a real number equals its trace and the log of the ratio equals the difference of the logs. The second equality uses that the trace of $A^T B$ equals the trace of BA^T. For third equality, we use that the trace is an additive function. In the second to last equality, 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

$$\underset{\hat{\Sigma}}{\arg\min}\quad \mathrm{tr}(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])) - \log(\det(\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])). \qquad (31)$$

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, (31) is minimized by all matrices $\hat{\Sigma}$ such that the eigenvalues of $\hat{\Sigma}^{-1}(\Sigma + [y\otimes y^T])$ minimize

$$\sum_{k=1}^{D}\lambda_k - \sum_{k=1}^{D}\log\lambda_k. \qquad (32)$$

As (32) is strictly convex, we can take the first order condition to identify the unique minimizer. This yields that (32) 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

$$\tilde{\Sigma} = \Sigma + [y\otimes y^T] = \Sigma + (M\Delta)\otimes(M\Delta)^T \qquad (33)$$

is the unique minimizer of (29) and thus the subjective long-run belief of the agent about the covariance matrix. This establishes (II).

Finally, we prove Case (III): We now solve

$$\underset{(\Delta,\hat{\Sigma}):\,\Delta_i=\tilde{f}_i-F_i}{\arg\min}\ \frac{1}{2}\left(\mathrm{tr}(\hat{\Sigma}^{-1}\Sigma) + y^T\hat{\Sigma}^{-1}y - D + \log\frac{\det\hat{\Sigma}}{\det\Sigma}\right). \qquad (34)$$

As shown in (30) this objective is equivalent to 1/2 times

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

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

$$\log\det(Id + \Sigma^{-1}[y\otimes y^T]). \qquad (35)$$

We first observe that as the determinant is the product of eigenvalues, (35) 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

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

Denoting the eigenvalues of $\Sigma^{-1}[y\otimes y^T]$ by $\lambda_1, \ldots, \lambda_D$, the objective (35) becomes

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

As eigenvalues are independent of the basis, we next choose an orthogonal basis $x_1, \dots, x_D$ 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 $\upsilon = \Sigma^{-1} y$ is an eigenvector with associated non-zero eigenvalue. Let $\upsilon = \sum_{i=1}^{D} \alpha_i x_i$ be the representation of $\upsilon = \Sigma^{-1} y$ in the basis x. We have that

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

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 $\alpha_1 = \frac{y^T \upsilon}{y^T y}$, so the non-zero eigenvalue of $\Sigma^{-1}[y \otimes y^T]$ equals

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

Consequently, the agent's long-run belief about the mean of the state satisfies

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

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

$$\begin{aligned} \Delta_k &= \frac{\left[M^T \Sigma^{-1} M\right]^{-1}_{ki}}{\left[M^T \Sigma^{-1} M\right]^{-1}_{ii}} \Delta_i \qquad \text{for } k \neq i \\ \tilde{\Sigma} &= \Sigma + (M\Delta) \otimes (M\Delta)^T \end{aligned}. \tag{36}$$

This completes the proof of (III). □

B. Two dimensions of stubborn beliefs

We consider the variant of our model in which the agent has fixed stubborn beliefs about two fundamentals, f_{i_1} and f_{i_2}. We restrict attention to the analogue of Case I in Theorem 2, supposing that the agent knows the correct covariance matrix Σ.

Using the notation $B = M^T \Sigma^{-1} M$, the agent's long-run bias about fundamental j is

$$\Delta_j = \frac{B^{-1}_{i_1 j}\left(B^{-1}_{i_2 i_2} \Delta_{i_1} - B^{-1}_{i_1 i_2} \Delta_{i_2}\right) + B^{-1}_{i_2 j}\left(B^{-1}_{i_1 i_1} \Delta_{i_2} - B^{-1}_{i_1 i_2} \Delta_{i_1}\right)}{B^{-1}_{i_1 i_1} B^{-1}_{i_2 i_2} - \left(B^{-1}_{i_1 i_2}\right)^2}. \tag{37}$$

This satisfies the first-order condition in the proof of Theorem 2, Case I.

We use (37) to prove a more general version of our result that contact with a group lowers the agent's bias regarding that group. Consider the model of Section 2 in which the fundamentals i_1, i_2, j equal i_1's, i_2's, and j's calibres, respectively, but suppose that agent i_1 knows the true calibre of individual i_2. Let individuals j and i_2 belong to the same group. Using that $\Delta_{i_2} = 0$, (37) reduces to

$$\Delta_j = \frac{B^{-1}_{i_1 j} B^{-1}_{i_2 i_2} - B^{-1}_{i_2 j} B^{-1}_{i_1 i_2}}{B^{-1}_{i_1 i_1} B^{-1}_{i_2 i_2} - \left(B^{-1}_{i_1 i_2}\right)^2} \cdot \Delta_{i_1}.$$

Since individuals i_2 and j belong to the same group, we have $B^{-1}_{i_1 i_2} = B^{-1}_{i_1 j}$, so

$$\Delta_j = \frac{B^{-1}_{i_1 j}}{B^{-1}_{i_1 i_1}} \cdot \frac{B^{-1}_{i_2 i_2} - B^{-1}_{i_2 j}}{B^{-1}_{i_2 i_2} - \left(B^{-1}_{i_1 j}\right)^2 / B^{-1}_{i_1 i_1}} \cdot \Delta_{i_1}.$$

Without contact with individual i_2, agent i_1's bias regarding individual j is $(B^{-1}_{i_1 j}/B^{-1}_{i_1 i_1}) \cdot \Delta_{i_1}$. Hence, to show that contact lowers his bias, it is sufficient to establish that

$$B^{-1}_{i_2 j} > \frac{\left(B^{-1}_{i_1 j}\right)^2}{B^{-1}_{i_1 i_1}} \quad \text{or} \quad B^{-1}_{i_2 j} B^{-1}_{i_1 i_1} > \left(B^{-1}_{i_1 j}\right)^2$$

Plugging in the expressions for the entries of B^{-1} from the proof Theorem 1, and again using that i_2 and j belong to the same group, the above inequality becomes

$$\left(\sum_k \phi^2_{g_j k} \upsilon^\eta_k\right)\left(\upsilon^q_{i_1} + \sum_k \phi^2_{g_{i_1} k} \upsilon^\eta_k\right) > \left(\sum_k \phi_{g_{i_1} k} \phi_{g_j k} \upsilon^\eta_k\right)^2,$$

which holds by the Cauchy–Schwarz inequality.

Acknowledgments. We are grateful to four anonymous referees, Aislinn Bohren, In-Koo Cho, Andrea Galeotti, Alex Imas, Jurjen Kamphorst, Robert Lieli, Ulrike Malmendier, Ameni Mehrez, Josh Schwartzstein, Tsong-Hong Tenn, Florian Zimmermann, Asaf Zussman, and seminar and conference audiences for insightful comments. Anna Vályogos provided excellent research assistance. Heidhues and Kőszegi thank the European Research Council for financial support under Grant #788918. Heidhues thanks the German Science Foundation (DFG) for financial support under project number 462020252. Also funded by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the German Excellence Strategy-EXC2126/1-390838866. This publication was supported by the Open Access Publication Fund of the University of Bonn.

REFERENCES

ALLPORT, G. W. (1954), *The Nature of Prejudice* (Cambridge, MA: Addison-Wesley).

AUGENBLICK, N. and RABIN, M. (2019), “An Experiment on Time Preference and Misprediction in Unpleasant Tasks”, *Review of Economic Studies*, **86**, 941–975.

BANAL-ESTAÑOL, A., LIU, Q., MACHO-STADLER, I., *et al.* (2023), “Similar-to-me Effects in the Grant Application Process: Applicants, Panelists, and the Likelihood of Obtaining Funds”, *R&D Management*, **53**, 819–839.

BÉNABOU, R. and TIROLE, J. (2016), “Mindful Economics: The Production, Consumption, and Value of Beliefs”, *Journal of Economic Perspectives*, **30**, 141–164.

BERK, R. H. (1966), “Limiting Behavior of Posterior Distributions when the Model Is Incorrect”, *Annals of Mathematical Statistics*, **37**, 51–58.

BERTRAND, M., CHUGH, D. and MULLAINATHAN, S. (2005), “Implicit Discrimination”, *American Economic Review*, **95**, 94–98.

BERTRAND, M. and MULLAINATHAN, S. (2004), “Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination”, *American Economic Review*, **94**, 991–1013.

BOHREN, J. A. (2016), “Informational Herding with Model Mispecification”, *Journal of Economic Theory*, **163**, 222–247.

BOHREN, J. A., HAGGAG, K., IMAS, A., *et al.* (2025), “Inaccurate Statistical Discrimination: An Identification Problem”, *Review of Economics and Statistics*, **107**, 605–620.

BOHREN, J. A. and HAUSER, D. (2019), “Misinterpreting Social Outcomes and Information Campaigns” (Working Paper).

BOHREN, J. A., IMAS, A. and ROSENBERG, M. (2019), “The Dynamics of Discrimination: Theory and Evidence”, *American Economic Review*, **109**, 3395–3436.

BORDALO, P., COFFMAN, K., GENNAIOLI, N., *et al.* (2016), “Stereotypes”, *Quarterly Journal of Economics*, **131**, 1753–1794.

BORING, A. (2017), “Gender Biases in Student Evaluations of Teaching”, *Journal of Public Economics*, **145**, 27–41.

BRANTON, R. P. and JONES, B. S. (2005), “Reexamining Racial Attitudes: The Conditional Relationship Between Diversity and Socioeconomic Environment”, *American Journal of Political Science*, **49**, 359–372.

BURSZTYN, L. and YANG, D. Y. (2022), “Misperceptions About Others”, *Annual Review of Economics*, **14**, 425–452.

CARD, D., DELLAVIGNA, S., FUNK, P., *et al.* (2020), “Are Referees and Editors in Economics Gender Neutral?”, *Quarterly Journal of Economics*, **135**, 269–327.

CHAUVIN, K. (2023), “A Misattribution Theory of Discrimination” (Working Paper).

CORNO, L., LA FERRARA, E. and BURNS, J. (2022), “Interaction, Stereotypes and Performance. Evidence from South Africa”, *American Economic Review*, **112**, 3848–3875.

DE PAOLA, M. and SCOPPA, V. (2015), “Gender Discrimination and Evaluators’ Gender: Evidence from Italian Academia”, *Economica*, **82**, 162–188.

ENGLAND, P. and LEWIN, P. (1989), “Economic and Sociological Views of Discrimination in Labor Markets: Persistence or Demise?”, *Sociological Spectrum: The Official Journal of the Mid-South Sociological Association*, **9**, 239–257.

ESSES, V. M., JACKSON, L. M. and ARMSTRONG, T. L. (1998), “Intergroup Competition and Attitudes Toward Immigrants and Immigration: An Instrumental Model of Group Conflict”, *Journal of Social Issues*, **54**, 699–724.

EYTING, M. (2024), “Why Do We Discriminate? The Role of Motivated Reasoning” (Working Paper).

FOUKA, V., MAZUMDER, S. and TABELLINI, M. (2022), From Immigrants to Americans: Race and Assimilation during the Great Migration”, *Review of Economic Studies*, **89**, 811–842.

FRICK, M., IIJIMA, R. and ISHII, Y. (2020), “Misinterpreting Others and the Fragility of Social Learning”, *Econometrica: Journal of the Econometric Society*, **88**, 2281–2328.

—— —— —— (2022), “Dispersed Behavior and Perceptions in Assortative Societies”, *American Economic Review*, **112**, 3063–3105.

FUDENBERG, D., LANZANI, G. and STRACK, P. (2024), “Selective Memory Equilibrium”, *Journal of Political Economy*, **132**, 3978–4020.

GAGLIARDUCCI, S. and PASERMAN, M. D. (2012), “Gender Interactions within Hierarchies: Evidence from the Political Arena”, *Review of Economic Studies*, **79**, 1021–1052.

GLAESER, E. L. (2005), “The Political Economy of Hatred”, *Quarterly Journal of Economics*, **120**, 45–86.

HANNA, R., MULLAINATHAN, S. and SCHWARTZSTEIN, J. (2014), “Learning Through Noticing: Theory and Evidence from a Field Experiment”, *Quarterly Journal of Economics*, **129**, 1311–1353.

HE, J., HU, L., KOVACH, M., *et al.* (2024), “Learning Source Biases: Multisource Misspecifications and Their Impact on Predictions” (Working Paper).

HEIDHUES, P., KŐSZEGI, B. and STRACK, P. (2022), “Misinterpreting Yourself” (Working Paper).

—— —— —— (2018), “Unrealistic Expectations and Misguided Learning”, *Econometrica: Journal of the Econometric Society*, **86**, 1159–1214.

HESTERMANN, N. and LE YAOUANQ, Y. (2021), “Experimentation with Self-Serving Attribution Biases”, *American Economic Journal: Microeconomics*, **13**, 198–237.

HUFFMAN, D., RAYMOND, C. and SHVETS, J. (2022), “Persistent Overconfidence and Biased Memory: Evidence from Managers”, *American Economic Review*, **112**, 3141–3175.

JACKSON, L. M. (2011), *The Psychology of Prejudice: From Attitudes to Social Action* (Washington, DC, US: American Psychological Association).

JACKSON, M. O., NEI, S. M., SNOWBERG, E., *et al.* (2022), “The Dynamics of Networks and Homophily” (Working Paper).

JOST, J. T., RUDMAN, L. A., BLAIR, I. V., *et al.* (2009), “The Existence of Implicit Bias Is Beyond Reasonable Doubt”, *Research in Organizational Behavior*, **29**, 39–69.

KAAS, L. and MANGER, C. (2012), “Ethnic Discrimination in Germany’s Labour Market: A Field Experiment”, *German Economic Review*, **13**, 1–20.

KORNEMANN, M. (2024), “The Value of Information for Misspecified Agents” (Working Paper).

LAMBIN, X. and PALIKOT, E. (2019), “The Impact of Online Reputation on Ethnic Discrimination” (Working Paper).

LANDIER, A. and THESMAR, D. (2009), “Financial Contracting with Optimistic Entrepreneurs”, *Review of Financial Studies*, **22**, 117–150.

LEVY, G. and RAZIN, R. (2017), “The Coevolution of Segregation, Polarized Beliefs, and Discrimination: The Case of Private versus State Education”, *American Economic Journal: Microeconomics*, **9**, 141–170.

LEVY, G., RAZIN, R. and YOUNG, A. (2022), “Misspecified Politics and the Recurrence of Populism”, *American Economic Review*, **112**, 928–962.

LOWE, M. (2021), “Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration”, *American Economic Review*, **111**, 1807–1844.

MALMENDIER, U. and TATE, G. (2005), “CEO Overconfidence and Corporate Investment”, *Journal of Finance*, **60**, 2661–2700.

MENGEL, F., SAUERMANN, J. and ZÖLITZ, U. (2018), “Gender Bias in Teaching Evaluations”, *Journal of the European Economic Association*, **17**, 535–566.

MULLEN, B., BROWN, R. and SMITH, C. (1992), “Ingroup Bias as a Function of Salience, Relevance, and Status: An Integration”, *European Journal of Social Psychology*, **22**, 103–122.

NEWPORT, F. (2014), “Gallup Review: Black and White Attitudes Toward Police” https://news.gallup.com/poll/175088/A%C2%ADgallup-review-black-white-attitudes-toward-police.aspx.

PETTIGREW, T. F. and TROPP, L. R. (2006), “A Meta-Analytic Test of Intergroup Contact Theory”, *Journal of Personality and Social Psychology*, **90**, 751–783.

Pew Research Center (2017), “Wide Partisan Gaps in U.S. Over How Far the Country Has Come on Gender Equality”.

—— (2018), “Partisans are Divided over The Fairness of the U.S. Economy – And Why People are Rich or Poor” Technical Report.

PRATTO, F., SIDANIUS, J. and LEVIN, S. (2006), “Social Dominance Theory and the Dynamics of Intergroup Relations: Taking Stock and Looking Forward”, *European Review of Social Psychology*, **17**, 271–320.

RACKSTRAW, E. (2022), “Bias-Motivated Updating in the Labor Market” (Working Paper).

SCHWARTZSTEIN, J. (2014), “Selective Attention and Learning”, *Journal of the European Economic Association*, **12**, 1423–1452.

SHAYO, M. and ZUSSMAN, A. (2011), “Judicial Ingroup Bias in the Shadow of Terrorism”, *Quarterly Journal of Economics*, **126**, 1447–1484.

SPIEGLER, R. (2016), “Bayesian Networks and Boundedly Rational Expectations”, *Quarterly Journal of Economics*, **131**, 1243–1290.

—— (2020), “Behavioral Implications of Causal Misperceptions”, *Annual Review of Economics*, **12**, 81–106.

SPINNEWIJN, J. (2015), “Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs”, *Journal of the European Economic Association*, **13**, 130–167.

STEPHAN, W. G., YBARRA, O. and BACHMAN, G. (1999), “Prejudice Toward Immigrants”, *Journal of Applied Social Psychology*, **29**, 2221–2237.

STOETZER, L. S. and ZIMMERMANN, F. (2024), “A Note on Motivated Cognition and Discriminatory Beliefs”, *Games and Economic Behavior*, **147**, 554–562.

TABELLINI, M. (2019), “Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration”, *Review of Economic Studies*, **87**, 454486.

TAJFEL, H. (1982), “Social Psychology of Intergroup Relations”, *Annual Review of Psychology*, **33**, 1–39.

TJADEN, J. D., SCHWEMMER, C. and KHADJAVI, M. (2018), “Ride with Me—Ethnic Discrimination, Social Markets, and the Sharing Economy”, *European Sociological Review*, **34**, 418–432.

ZUSSMAN, A. (2013), “Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars”, *Economic Journal*, **123**, F433–F468.