A MISATTRIBUTION THEORY OF DISCRIMINATION

KYLE P. CHAUVIN

APRIL 5, 2020

Abstract. Individuals often make inferences about others without considering the circumstances
that other people face. I analyze the consequences of this psychological bias in a model of discrim-
inination. Each group in a population has an observable distribution of outcomes that is produced
jointly by the private traits of group members and the beliefs which all population members have
about that group. However, biased observers ignore the role of others’ beliefs in generating out-
comes, instead drawing inferences about groups under the assumption that traits lead directly to
outcomes. There is a unique equilibrium in which beliefs about different groups generate outcomes
that are misinterpreted as the original beliefs. In equilibrium, each group’s true mean trait is ex-
aggerated relative to a population-wide average; at least one group is overestimated, and at least
one group is underestimated. The ordering of groups by outcome levels matches the ordering under
correct beliefs for low levels of bias but is arbitrary for sufficiently high levels of bias. Observers
of different groups agree in their beliefs about the ordering of groups but systematically disagree
in their absolute beliefs. Large levels of bias are shown to unambiguously diminish social welfare.
De-biasing policies which lead to corrective preferential treatment for underestimated groups have
a multiplier effect in correcting beliefs. By contrast, policies which broaden the diversity of social
networks can only correct beliefs about a strict subset of groups.

JEL Classification: C72, D84, J15, J16

Keywords: Behavioral Game Theory, Belief Formation, Inequality, Statistical Discrimination
Beliefs about the innate characteristics of gender, racial, and other social groups are commonplace and powerful. Explicitly or implicitly, from the notion that men and women have different abilities to assumptions that some racial groups are inherently more intelligent, hard-working, or criminal than others, these beliefs propel discrimination and impact public welfare. Although classical economic models predict that group-trait beliefs should be accurate conditional probabilities (Arrow, 1973; Coate and Loury, 1993), such beliefs have been shown to include the influence of psychological biases such as stereotypes (Hilton and von Hippel, 1996; Schneider, 2004; Bordalo et al., 2016), in-group favoritism (Tajfel et al., 1971), or implicit associations (Greenwald et al., 2009). This paper studies a particular pathway by which cognitive biases influence beliefs that drive discrimination: people infer unobservable traits from ostensibly informative outcomes while neglecting to account for how discrimination affects those outcomes.

Many traits are inferred. For example, people cannot directly see the work ethic, intelligence, and proclivity to violence of different social groups. Instead, observers understand that having a good work ethic generally leads to more and better employment, being smart is often reflected by higher grades and test scores, and having a violent personality leads more readily to crime, arrest, and punishment. This allows observers to indirectly infer traits from outcomes. However, such outcomes are also influenced by external circumstances, including one’s treatment by others in society. An employee’s professional success depends not only on his own work ethic but also the hiring and promotional decisions of his managers. Teacher and parent investment combines with student ability to prepare students for exams. A high arrest rate can reflect aggressive policing as much as it can reflect a high crime rate.

As psychologists have long documented, observers often ignore the role of these external circumstances. They “underestimate the impact of the situation and overestimate [the impact of] the

\(^1\)Empirical studies have documented the presence of discrimination in numerous contexts. Examples include audit studies showing discrimination against racial (Bertrand and Mullainathan, 2004; Doleac and Stein, 2013; Edelman et al., 2017) and religious (Acquisti and Fong, 2019) minorities and women (Moss-Racusin et al., 2012). Other methodologies have uncovered bias against women (Goldin and Rouse, 2000; Sarsons, 2017) and ethnic subgroups (Rubinstein and Brenner, 2014). See Bertrand (2011), Bertrand and Duflo (2017), Neumark (2018), Lang and Lehmann (2012) and Fang and Moro (2010) for useful surveys of the literature. Although the degree to which such discrimination is driven by beliefs and how much by preferences remains an open question (Guryan and Charles, 2013), some studies (List, 2004; Levitt, 2004; Pope and Sydnor, 2011; Coffman et al., 2019) find levels of discrimination that are difficult to attribute to animus, indicating beliefs are a significant factor.
individual’s traits and attitudes” (Myers, 2012) when explaining features about another person. This tendency is termed the Fundamental Attribution Error (FAE) (Ross, 1977). By ignoring situational factors, people internalize a direct correspondence linking one’s internal traits with observable outcomes. In particular, the FAE blinds them to the otherwise obvious fact that different people face different circumstances. Observers may acknowledge that discrimination exists in society, engage in it themselves, and still fail to account for it when judging others. This paper models and analyzes a self-sustaining cycle in which people explain group-outcome differences as group-trait differences, develop beliefs that lead them to discriminate on the basis of group status, and perpetuate the very outcome gaps they sought to explain.

Section 2 presents a simple special case of the model. A population of students and teachers are matched together in pairs. Both students and teachers can be either male or female. Students have a privately-observed ability type, and teachers maintain beliefs about the expected abilities of male and female students. In each matched pair, the teacher first observes the student’s gender and then chooses a level of investment in the student’s education. The student observes the teacher’s investment level and chooses his own effort level. Together, the teacher’s investment and the student’s effort prepare the student for an exam. The student’s cost of effort decreases in his ability, so for a given level of teacher investment, higher ability students receive higher exam scores. Similarly, the teacher’s cost of investment decreases in the student’s effort, so for a given student ability level, higher teacher belief about the student leads to a higher score. At the population

---

2Also called the ‘correspondence bias’ (Gilbert and Malone, 1995), the FAE has a long history of experimental evidence in social psychology. Lab subjects have been found to overestimate the extent to which a fellow subject’s argument on a debate topic, despite being assigned by the experimenter, reflects fellow subject’s true opinion (e.g. Jones and Harris (1967)). Subjects viewing a silent tape of an anxious interviewee attribute her behavior to innate anxiety to an equal extent when informed either that the interview concerned a contentious or mundane topic (e.g. Snyder and Frankel (1976)). When one among two subjects is randomly selected to pose tricky questions to the other, observers believe that the questioner is smarter than the answering subject (e.g. Block and Funder (1986)). In the infamous Milgram (1963) experiment, many participants were successfully coached into delivering what they believed was a near fatal electric shock on another lab participant. In Bierbrauer (1979), subjects who learned about the earlier experiment attributed the willingness of participants to inflict pain to innate evil and not the situational power of the experiment. See Gawronski (2004) for a survey of psychology experiments on the bias. Pettigrew (1979) in fact proposed an extension of the FAE to explicitly consider social groups, but framed his ‘ultimate attribution error’ as a form of in-group preference, by which negative out-group behaviors are attributed to innate negative characteristics. Economists have long acknowledged that empathy is difficult and that people understand the world through a localized lens of experience. Adam Smith said as much in The Theory of Moral Sentiments: “As we have no immediate experience of what other men feel, we can form no idea of the manner in which they are affected, but by conceiving what we ourselves should feel in the like situation.” Smith (1759). More recently the FAE has been studied within economics. Glaeser and Ponzetto (2017) study the consequences for political economy when voters judge their politicians without accounting for circumstances. In lab experiments, Durell (2001) and Cartwright and Wooders (2014) find subjects rating the performance of other subjects without correctly adjusting for the conditions the other subjects faced.
level, average scores of male and female students therefore depend both on their true abilities as well as the beliefs which teachers hold about gender-conditional abilities.

Teachers infer their beliefs by observing average exam scores for male and female students, which – because teachers are subject to the FAE – are incorrectly assumed to depend only on student abilities. The presumed relationship between ability and score is calibrated by own-group experiences: male teachers assume all students receive the level of teacher investment which male students receive; female teachers assume all students get the investment which female students receive. In an *attributive equilibrium*, teachers hold beliefs that, together with students’ actual abilities, produce scores that are inferred as the original beliefs. As in a Perfect Bayesian Equilibrium, teachers and students choose optimal actions on the basis of their beliefs; however, the beliefs are informed not by Bayes’ rule but rather by a biased inference heuristic. The paper’s results are stated in terms of equilibrium beliefs but can also be understood in terms of the implied actions.

For all but knife-edge parameter values, there is a unique attributive equilibrium. The degree to which equilibrium beliefs differ from the objective ability values depends critically on the ratio of the marginal effect of teacher belief on exam score to the marginal effect of student ability on exam score. This ratio, denoted \( \eta \), captures the strength of complementarity, and thus the effective strength of bias, in the model. When \( \eta = 0 \), teacher beliefs have no impact on scores, scores accurately reflect student abilities, and equilibrium beliefs are correct. For \( \eta > 0 \), the equilibrium features systematically skewed beliefs. In terms of average teachers beliefs, which aggregate those of male and female teachers, both groups of students are simultaneously misestimated. The gender with objectively higher ability is on average positively misestimated, leading to inflated exam scores, while the gender with lower ability is negatively misestimated, leading to artificially low scores. Thus, absolute beliefs are inflated while the implied relative ranking of the student groups coincides with objective abilities. Individual teachers all share the same relative ranking, but their beliefs systematically disagree in absolute terms. Teachers of the higher performing gender, basing

---

3While there are many potential ability-to-score correspondence which teachers could assume, any realistic correspondence would incorporate the actual experiences of students. The general model in Section 3 allows for aggregating the experiences of different groups of students according to arbitrary weights. Section 2 specifies own-group calibration as the simplest of such aggregations. One intuitive, if informal, rationale for this conceives of attributive equilibrium as the steady state of an overlapping generations model in which students of one generation become the teachers of the subsequent generation; in the OLG reading, own-group calibration is equivalent to assuming teachers internalize the treatment they themselves received as students and use that to judge their own students. This resembles the ‘cross-situational projection’ discussed in Van Boven and Loewenstein (2005): people reason about others facing particular circumstances by first imagining themselves in the same situation.
their inference in own-group experience, presume that all students receive a relatively high level of teacher investment. In explaining observed scores, they therefore make relatively low assessments of student ability. Similarly, teachers of the other gender presume that all students receive less investment and accordingly infer relatively high assessments.

When the complementary strength $\eta$ is sufficiently large, the attributive equilibrium is unstable: whereas a dynamic process of beliefs successively revised by misattribution would converge to equilibrium for low values of $\eta$, the same process diverges to more and more extreme levels when $\eta$ is large. Because of the instability, the divergent behavior of belief sequences more realistically characterizes the model’s long-run predictions than does equilibrium itself. In such belief sequences, the average teacher belief about students of one gender increases monotonically towards positive infinity while the average belief about the other gender tends to negative infinity.\footnote{These are directional predictions. As a stylization of reality, the model omits the myriad of factors that would prevent actual beliefs from inflating without bound. The paper’s results about beliefs tending to positive or negative infinity are understood as predictions that beliefs will increase or decrease with successive iterations of misattribution.} Moreover, different initial beliefs can support either gender being positively or negatively misestimated. This illustrates how the FAE can not only warp objective differences but also, in the extreme, completely sever the link between actual trait differences and societal beliefs about them. As with the case of stable equilibrium, both groups are simultaneously misestimated, and the disagreements between teacher groups are only magnified as the sequence progresses.

Section 3 generalizes the special case model in several ways. Instead of male and female agents, there is an arbitrary finite number of social groups interacting as junior and senior. The general model allows for the influence that one group has on another to be a flexible parameter. Similarly, it also allows for flexible empathy by having seniors calibrate their subjective models of outcome production using the experiences of juniors outside their own social group. Lemma 1 verifies that equilibrium existence and uniqueness are generic outcomes of the general model. Because the general model allows for arbitrary influence and empathy parameters, junior groups are not necessarily over/underestimated if their true traits lie above/below the population mean. However, Theorem 1 establishes that, generically, all groups are misestimated, with at least one group overestimated and at least one group underestimated. Theorem 2 both provides conditions under which the relative ranking of juniors by seniors’ beliefs is correct and establishes, for a large
subclass of networks and for high values of $\eta$ that divergent belief sequences can over- or under-estimate any subset of junior groups. Finally, Theorem 3 generalizes results about systematic differences between the beliefs of different senior groups. In particular, all seniors agree in their rankings of junior groups, and a senior that has a high belief about one junior group also has high beliefs about all junior groups.

Section 4 analyzes the welfare and policy consequences of misattribution. Although some junior groups benefit from biased beliefs, a calculation of net population welfare shows that the losses to seniors and disadvantaged juniors outweigh any gains to advantaged juniors for sufficiently high levels of complementarity. The model speaks to the efficacy of several common policies intended to blunt the effects of discrimination. It predicts that ‘affirmative action’ policies which mandate the preferential treatment of different groups or institute outcome transfers across groups have a multiplier effect on biased beliefs. Intuitively, by artificially inflating some groups’ observed outcomes, a policy maker induces belief shifts that further contribute to outcome differences, and thus an appropriately calibrated policy can completely de-bias the population.\(^5\) On the other hand, ‘diversity’ policies which broaden empathy or influence in the population cannot de-bias incorrect beliefs brought about by misattribution. Because broadened empathy or influence brings different groups closer to being compared against a common standard without changing the fact that the FAE leads to exaggerated beliefs, these policies cause different observer groups to agree more, but do not lead to beliefs which are more correct overall.

The discussion in Section 5 considers how the equilibrium concept featured in this paper compares with others in behavioral game theory, particularly the Berk-Nash equilibrium of Esponda and Pouzo (2016). Beyond that connection, this paper contributes primarily to the literatures regarding discrimination and stereotypes. Explanations for discrimination have traditionally been divided between preference-based, e.g. from animus (Becker, 1957) or identity (Akerlof and Kranton, 2000; Bénabou and Tirole, 2011), and statistical discrimination (Arrow, 1973; Coate and Loury, 1993), whereby an interaction features multiple equilibria, and group status becomes a coordination device.\(^6\)

---

\(^5\)This contrasts with the ‘patronizing equilibrium’ phenomenon of Coate and Loury (1993), where policies that lower standards for minority applicants disincentivize human capital investment among those applicants. The key difference is that rational managers who engage in statistical discrimination have no incorrect beliefs to be de-biased.

\(^6\)For example, in a good equilibrium workers acquire skills and managers hire them, while in a bad equilibrium workers do not invest in human capital and managers do not hire them. Discrimination arises when one group is
More recent work has highlighted the roles of cognitive biases in producing what Bohren et al. (2019a) term ‘incorrect statistical discrimination.’ Bordalo et al. (2016) study the stereotypes that arise when observers see the trait distributions of different groups but, on account of limited working memory, disproportionately associate each group with its most representative traits (Gennaioli and Shleifer, 2010). Fryer and Jackson (2008) analyze discrimination produced by coarse categorization. Unlike these processes, the FAE does not involve distorting or simplifying one’s dataset of observations, but rather imposing an incorrect model. An application of the selective attention model in Schwartzstein (2014) shows how skewed perceptions of group traits follow from people basing their judgements on an unrepresentative sample of personal experiences, and Heidhues et al. (2019) examines a similar pathway for individuals who have stubbornly positive views of their own abilities. While the FAE can be understood as a particular instance of inattention, the psychology literature (see Footnote 2) suggests it is a ubiquitous tendency rather than an inadvertent mistake. Accordingly, the policy measures considered above are not targeted to shift which variables people attend to but rather to mitigate the consequences of their bias.

What most prominently distinguishes beliefs informed by the FAE is the prediction the model makes regarding disagreement across groups. Namely, observers agree about which groups have relatively high traits and which have relatively low traits, but disagree in absolute terms. For many parameter confirmations, the model predicts that belief complementarity ($\eta > 0$) leads to groups which are highly ranked by others having low absolute beliefs about all groups, and belief substitutability ($\eta < 0$) leads to highly ranked groups having low absolute beliefs. The discussion in

---

7This produces several predictions, such as exaggerated perceptions of group differences and a ‘kernel of truth’ in incorrect beliefs, which are common to the present paper. It also contrasts in various ways, including in predicting underestimation of within-group heterogeneity, where misattribution leads to correct estimation of higher order distribution moments, and predicting that an observer’s beliefs are independent of her own group status, while in the present paper differential empathy across observer groups produces systematic differences in beliefs.

8More broadly, the consequences of the FAE correspond with other cases of mis-inference in which an agent’s failure to understand his own circumstances leads him to misunderstand others. For example, Bénabou and Tirole (2006) model how agents can sustain their belief in the value of effort as a means to motivate themselves. Consequently, people who believe that effort readily produces wealth accordingly infer that the poor must be lazy. Frick et al. (2019) study members of a population who interact through assortative matching but mistakenly believe that they interact with a representative sample; in settings with coordination, this bias produces more extreme beliefs about others.
Section 5 compares these predictions with evidence from the General Social Survey. It demonstrates that they are largely borne out in data on beliefs about the intelligence, work ethic, and proclivity to violence of four different racial/ethnic groups and provides a foundation for future work to test how much the FAE contributes to assessments of group trait levels. The paper concludes with a discussion about how some of the model’s simplifying assumptions could be relaxed in subsequent work on this topic.

2. Illustration with Two Groups

2.1. Model

There is a continuum of students and a continuum of teachers. Both students and teachers are equally divided between male and female agents. In addition, each student has an ability level \( \theta \in \mathbb{R} \). An agent’s gender is publicly observable while ability is privately known. The true average abilities of male and female students are denoted \( \mu_M \) and \( \mu_F \), respectively. The beliefs which teachers of gender \( i \in \{M, F\} \) have about average male and female ability are denoted \( \tilde{\mu}_i^M \) and \( \tilde{\mu}_i^F \). Finally, the average or population teacher beliefs about the students are denoted \( \tilde{\mu}_M \equiv \tilde{\mu}_M^M / 2 + \tilde{\mu}_M^F / 2 \) and \( \tilde{\mu}_F \equiv \tilde{\mu}_F^M / 2 + \tilde{\mu}_F^F / 2 \).

Students and teachers are paired together. In each pair, both parties work to jointly prepare the student for an exam. First the teacher chooses an investment level \( a_{tch} \), and then the student observes that level and chooses his own investment level \( a_{std} \). On the basis of their decisions, the student’s score \( x \) is determined by the stylized production function \( x = a_{std} + a_{tch} \). The student wishes to do well but also finds effort costly; his cost of effort is driven down by higher ability while also being affected by teacher investment:

\[
\begin{align*}
    u_{std} &= x - \frac{1}{2} (a_{std} - \theta - \varphi_{std} \cdot a_{tch})^2, \\
    u_{tch} &= x - \frac{1}{2} (a_{tch} - \varphi_{tch} \cdot a_{std})^2;
\end{align*}
\]

where \( \varphi_{std} \) measures the degree of complementarity between student and teacher investments. The teacher likewise wants the student to do well, and her cost of effort depends on the student’s investment:

\[
\begin{align*}
    u_{tch} &= x - \frac{1}{2} (a_{tch} - \varphi_{tch} \cdot a_{std})^2,
\end{align*}
\]

where \( \varphi_{tch} \) is the teacher’s corresponding complementarity parameter. Because the teacher does not observe the student’s ability, she instead bases her action on her belief about the student’s ability.
conditional on gender. Therefore, when the student has actual ability \( \theta \), the teacher has belief \( \hat{\mu} \) about the student’s expected ability level, and both teacher and student choose their investment levels optimally, the resultant score is equal to

\[
x(\hat{\mu}, \theta) = x_0 + \left( \varphi_{tch} \cdot \frac{1 + \varphi_{std}}{1 - \varphi_{std} \varphi_{tch}} \right) \hat{\mu} + \theta.
\]

That is, the student’s expected score is an affine function of both his actual ability and the teacher’s perception of his ability. The intercept is \( x_0 \), and the marginal effect of the teacher’s belief, denoted \( \eta \), depends on \( \varphi_{std} \) and \( \varphi_{tch} \) and describes the relative power of bias in the model. It is assumed \( \varphi_{tch} \) and \( \varphi_{std} \) are both positive\(^9\) but not so large that \( \varphi_{std} \varphi_{tch} > 1 \), and hence \( \eta > 0 \); this assumption is relaxed in the following section.

At the population level, repeated matching of students and teachers produces distributions of exam scores for male and female students. Average scores by gender depend on the true average abilities of students, the average beliefs of teachers, and the rates at which students of each gender are matched with teachers of each gender. Suppose that students and teachers are matched together uniformly randomly, and recall the (temporary) assumption that there are an equal proportion of male and female teachers. Then the average score of students of gender \( g \) reflects both the average ability of \( g \) students, \( \mu_g \), as well as the average belief about \( g \) students, \( \hat{\mu}_g \):

\[
x_g = x_0 + \eta \cdot \frac{1}{2} \left( \hat{\mu}_M + \hat{\mu}_F \right) + \mu_g.
\]

Teachers in the population observe the average score values \( x_M \) and \( x_F \). They could accurately infer each group \( g \)’s average ability \( \mu_g \) if they correctly accounted for the values of teacher beliefs \( \hat{\mu}_g \). However, the FAE precludes this key step. In forming beliefs, teachers ignore their own role\(^{10}\) and assume that scores depend only on student ability through some direct correspondence. There are multiple ways for teachers to calibrate their assumed ability-to-score correspondences, but for the

\(^9\)Intuitively, \( \varphi_{tch} > 0 \) means the teacher wants to help the student succeed in subjects where he/she will invest greater effort. Whenever she supposes high student ability, a teacher will anticipate greater student effort and increase her own investment in response. Likewise, \( \varphi_{std} > 0 \) means the student is encouraged to try harder when the teacher invests more in him.

\(^{10}\)Recall that the FAE is the assumption that observables about a person are explained by the person’s internal traits. In overlooking the impact of their own actions, teachers in the model are no worse than subjects in psychology experiments (see Footnote 2) that disregard situational factors despite clear reasons not to.
purposes of this illustration we consider a baseline specification, namely that teachers of each group base their models according to the experiences of students in their own groups.\textsuperscript{11} For example, a female teacher assumes the relationship

\[ x_g = x_0 + \eta \tilde{\mu}_F + \mu_g, \]

which presumes that students of both genders are treated as female students actually are. A female teacher therefore infers \( \tilde{\mu}_g^F = x_g - x_0 - \eta \tilde{\mu}_F \). Likewise, male teachers infer \( \tilde{\mu}_g^M = x_g - x_0 - \eta \tilde{\mu}_M \) from the same data.

The processes described above have illustrated (1) how a given profile of teacher beliefs, in combination with actual student abilities, leads to outcomes in the form of exam scores, and (2) how observed scores, in combination with levels of teacher investment, lead to inferred beliefs about student ability. Combining these two processes yields an explicit expression for how one profile of beliefs leads to an updated, or ‘revised,’ profile:

\[ \tilde{\mu}_g' = \mu_g + \eta \left( \tilde{\mu}_g - \frac{1}{2} \tilde{\mu}_M + \frac{1}{2} \tilde{\mu}_F \right). \]

This belief revision equation says that the revised population belief \( \tilde{\mu}_g' \) depends on the comparison between the original belief \( \tilde{\mu}_g \) and a ‘benchmark’ that averages together beliefs about the two genders, \( \tilde{\mu}_M/2 + \tilde{\mu}_F/2 \). For instance, if male students are deemed to have higher ability than female students, then this comparison is positive, \( \tilde{\mu}_M > \tilde{\mu}_M/2 + \tilde{\mu}_F/2 \), and the revised belief about male students will be an over-estimation of their ability: \( \tilde{\mu}_M' > \mu_M \). The extent of the overestimation depends on the magnitude of \( \eta \).

2.2. Results

There is a unique profile of beliefs \((\tilde{\mu}_M^M, \tilde{\mu}_F^M, \tilde{\mu}_M^F, \tilde{\mu}_F^F)\) which produces scores that lead to identical revised beliefs. This is called the attributive equilibrium. Population level beliefs are given by the following formula:

\[ \tilde{\mu}_g^* = \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \frac{1}{2} \tilde{\mu}_M + \frac{1}{2} \tilde{\mu}_F \right). \]

\textsuperscript{11}Own-group calibration is used as the baseline because, as the FAE requires observers to implicitly assume a common external environment experienced by all, an observer’s most immediate environment is the one he or she has personally experienced.
The equilibrium population belief about a group $g$ consists of $g$’s true mean ability level plus a distortion term. The distortion term is proportional to the gap between the group’s true mean ability and a benchmark value, which in this case is the population level average ability. Furthermore, the proportional factor is an increasing function of the complementarity strength $\eta$. Several properties are immediately discernable from the formula above.

**Beliefs are generally wrong.** There are only two cases in which equilibrium population beliefs are correct, and both of them serve to highlight the generic quality of incorrect beliefs. First, if the teacher does not care about the student’s level of investment ($\varphi_{\text{tech}} = 0$), then the teacher’s belief about the student’s ability does not affect the teacher’s investment, so $\eta = 0$ and the discrepancy term disappears. Second, if the two groups share exactly the same average ability level, then each group’s true mean is equal to the common benchmark value, and the discrepancy term again disappears. When neither of these corner cases hold, both groups are simultaneously misestimated.

**The two groups are misestimated in opposite directions.** Because the common benchmark value $\mu_M/2 + \mu_F/2$ necessarily lies between the true mean abilities of the two genders, the discrepancy term must be strictly positive for one group of students and negative for the other. Hence one is overestimated, and the other is underestimated.

**Population-level relative beliefs are correct.** Although the absolute values of equilibrium beliefs are incorrect, the ranking of groups according to population beliefs always matches the objective ranking by mean ability level. That is, $\tilde{\mu}_M^* \geq \tilde{\mu}_F^*$ if and only if $\mu_M \geq \mu_F$. Intuitively, the reason for this ‘kernel of truth’ amid broader inaccuracy is that students of both genders are assessed according to a common benchmark. Differences with respect to the benchmark are exaggerated, but because they are exaggerated according to a constant proportion, the relative ordering of groups is preserved. Section 3.4 demonstrates how this result can fail when, with more general social networks, different groups are assessed according to different benchmarks. For example, a high ability group assessed according to an even higher benchmark may be judged lower than a low ability group assessed according to an even lower benchmark.

**Observer groups agree in relative terms.** At the level of individual observer groups, equilibrium beliefs are determined by a similar expression. Gender $i$’s belief about $g$ is:

$$\tilde{\mu}_g^{i*} = \mu_g + \frac{\eta}{1 - \eta} \left(1 + \eta\right) \left(\mu_g - \mu_i\right).$$
As with population beliefs, gender i’s belief about g reflects both g’s true mean trait level and a discrepancy. However, the discrepancy here involves a comparison of the target group g with the observer’s (true) own-group trait mean. From the expression above we see that different observer groups share the same ranking of target groups, which matches the ranking by population beliefs:  \( \tilde{\mu}_g^M \geq \tilde{\mu}_g^M \) if and only if  \( \tilde{\mu}_g^F \geq \tilde{\mu}_g^M \).

**Higher ranked observer groups have lower absolute beliefs about everyone.** Because each observer group’s beliefs are determined with respect to own-group trait values, observer groups with higher trait values must necessarily infer lower absolute beliefs about all groups in order to explain the same data as observer groups with lower trait values. That is,  \( \tilde{\mu}_g^M \geq \tilde{\mu}_g^F \) if and only if  \( \tilde{\mu}_F \geq \tilde{\mu}_M \).

**Groups are correct in assessing their own members.** With respect to group-level beliefs, the assumption that observer groups use the experiences of their own members as the basis for evaluating all groups directly implies that, rather than engaging in self-serving overestimation, each group has correct own-group beliefs:  \( \tilde{\mu}_g^g = \mu_g \). This happens because the comparison term in gender i’s belief about its own members is zero.

**The equilibrium is stable for \( \eta < 1 \).** For relatively moderate values of the complementarity strength parameter \( \eta \), the belief revision equation models incremental changes. Thus, for any initial profile of beliefs, the sequence

\[
\tilde{\mu}_g^{t+1} = \mu_g + \eta \left( \tilde{\mu}_g^t - \left[ \frac{1}{2} \tilde{\mu}_M^t + \frac{1}{2} \tilde{\mu}_F^t \right] \right)
\]

produced by iteratively applying the belief revision equation is guaranteed to converge to the attributive equilibrium. (See the proof of Lemma 1.) Alternatively, one can understand the belief sequence as tracking the opinions of a set of overlapping generations, whereby the students of generation \( t \) become the teachers of generation \( t+1 \), forming beliefs on the basis of time \( t \) data and then acting on those beliefs in time \( t+1 \).\(^{12}\) In this interpretation, attributive equilibrium beliefs are the unique steady state levels of the dynamic process.

**Levels of \( \eta > 1 \) lead to path dependent outcomes.** When the complementarity strength \( \eta \) is sufficiently large (\( \eta > 1 \)), iterative belief revision yields belief sequences that diverge away from the

\(^{12}\)In this interpretation, the assumption that teachers calibrate their subjective models using the experiences of their own group is equivalent to assuming teachers in time \( t+1 \) recall their own personal experiences as former students time \( t \).

11
equilibrium. Instead, the gap between population beliefs about male and female students grows arbitrarily large. Beliefs about one gender tend to $+\infty$ while beliefs about the other tend to $-\infty$. Critically, the direction of the belief gap between the two genders depends both on true values and on initial conditions. Specifically, beliefs about male students become increasingly more optimistic, $\tilde{\mu}_M^t \xrightarrow{t} +\infty$, if and only if $\tilde{\mu}_M^0 - \tilde{\mu}_F^0 > -1/(\eta - 1) \cdot (\mu_M - \mu_F)$. That is, the initial population estimate of male students’ relative advantage must pass a threshold level that decreases in the objective gap. If the true mean trait level of female students is higher, then the threshold is strictly positive.

Substitutability leads to mitigated assessments. In the case of substitutability between student and teacher actions – if $\varphi_{\text{tech}} < 0$ or $\varphi_{\text{std}} < -1$ – then the complementarity parameter $\eta$ can be negative. Given the formula for equilibrium beliefs, this implies that observers may underestimate differences between groups.

3. General Model and Results

This section analyzes how the predictions of the model are altered as simplifying assumptions in the special case are relaxed. The general model considers an environment with an arbitrary number of social groups and where interaction between agents need not feature the complementarity of higher beliefs leading to higher actions. Unlike in the special case, the social network governing belief updating includes asymmetries in the influence one group’s beliefs has on another’s outcomes and considers agents calibrating their subjective models with the experiences of other groups.

3.1. Model

Population. There is a continuum population of junior and senior agents. Each agent has a private trait $\theta \in \mathbb{R}$ and a publicly observable group identity $g \in \mathcal{G}$, where $\mathcal{G}$ is a finite set. For notational clarity, juniors’ groups are referenced with $g$ and $h$, and seniors’ groups are denoted by $i$ and $j$. The true mean trait of group $g$ is denoted $\mu_g$. The belief held by members of group $i$ about the mean trait of group $g$ is $\tilde{\mu}_g^i$.

Interaction Game. Agents are matched together in pairs to play an interaction game. In each match, the senior first chooses $a_S \in \mathbb{R}$, and then the junior observes that level and chooses $a_J$; these quantities jointly produce the outcome $x = a_S + a_J$. The junior’s payoff is

$$u_J = \beta_J \left( a_J + a_S \right) - \frac{1}{2} \frac{1}{x} (a_J - \theta_J - \varphi_J a_S)^2,$$
and the senior’s payoff is
\[ u_S = \beta_S (a_J + a_S) - \frac{1}{2}(a_S - \varphi_S a_J)^2. \]

Both players are expected-utility maximizers, and each observes the group identity but not the trait of the other. The senior therefore conditions her action on her belief \( \tilde{\mu}^S_J \) about the mean trait of agents in the junior’s group. When the senior best-responds to her belief and the junior best-responds to the senior’s action, the resultant outcome is
\[ x(\tilde{\mu}^S_J, \theta_J) = \left( \frac{1 + \varphi_J}{1 - \varphi_J \varphi_S} \right)^2 \beta_S + \frac{1 + \varphi_S}{1 - \varphi_J \varphi_S} \beta_J + \left( \varphi_S \cdot \frac{1 + \varphi_J}{1 - \varphi_J \varphi_S} \right) \tilde{\mu}^S_J + \theta_J. \]

The student-teacher interaction discussed in Section 2 assumed \( \beta_J = \beta_S = 1 \), as both parties preferred the student to do well, and \( \varphi_J, \varphi_S > 0 \), which imposed complementarity between the student’s trait and the teacher’s belief. In other settings, other parameter values may be more appropriate. For example, suppose the junior is a prospective worker choosing effort, the senior is a hiring manager choosing trust in the worker, and the outcome is a hiring decision; in this case it is plausible that the senior has no fundamental interest in the outcome, \( \beta_S = 0 \), but only wants to trust workers who choose high effort: \( \varphi_J > 0 \). Alternatively, suppose the junior is a suspect choosing a level of crime, the senior is a police officer choosing a surveillance level, and the outcome is a measure of arrest and punishment severity. In this application, the junior dislikes the outcome, \( \beta_J < 0 \), and furthermore \( \varphi_J < 0 \): greater surveillance increases the cost of crime.

**Outcomes.** The probability with which a group \( g \) junior is matched with a group \( i \) senior is denoted \( w_{gi} \in [0, 1] \) and describes the influence that group \( i \) has on group \( g \). At the population level, the observed mean outcome for juniors of group \( g \) is
\[ x_g = x_0 + \eta \left( \sum_{i \in G} w_{gi} \tilde{\mu}^i_g \right) + \mu_g. \]

In the special case in Section 2, students and teachers were matched together uniformly randomly, so gender \( g \)'s average outcome was an unweighted average of the outcome conditional a male teacher and the outcome conditional on a female teacher. That is, \( w_{M,M} = w_{M,F} = w_{F,M} = w_{F,F} = 1/2 \). Instead of assuming the influence weights to be equal, the general model only requires that the weights are non-negative and that, for any junior group \( g \), the weights across all possible seniors
sum to one: \( \sum_{i \in G} w_{gi} = 1 \). This accommodates the possibility that some groups comprise an outsized share of the population or that the matching between juniors and seniors is asymmetric.

Inference. Seniors forming beliefs about juniors are subject to the FAE. They disregard the impact of seniors’ beliefs in influencing juniors’ outcomes and instead completely attribute the outcomes to traits.\(^{13}\) Instead of using only own-group experiences to calibrate their subjective models, as in Section 2, seniors aggregate the experiences of juniors of different groups according to empathy weights. The extent to which a senior of group \( i \) incorporates the experiences of juniors of group \( h \) is denoted \( z_{ih} \in [0, 1] \). That is, a group \( i \) senior implicitly assumes that all juniors face an average level of senior belief equal to \( \sum_{h \in G} z_{ih} \bar{\mu}_h \). As with the influence weights, it is assumed that any senior group’s empathy weights sum to one: \( \sum_{h \in G} z_{ih} = 1 \). The empathy weights model social interactions, such as among friends and family, in contrast to the professional encounters which generate the student’s exam scores. Allowing for arbitrary empathy weights, the group \( i \)’s belief about group \( g \)’s mean trait must satisfy

\[
x_g = x_0 + \eta \left( \sum_{h \in G} z_{ih} \bar{\mu}_h \right) + \tilde{\mu}_g^i.
\]

Benchmark Weights. It is useful for equilibrium analysis to denote by \( y_{gh} \equiv \sum_{i \in G} w_{gi} z_{ih} \) the degree to which the actual experiences of group \( h \) are used in forming beliefs about group \( g \). The parameters \( \{y_{gh}\}_{g,h \in G} \) are called the benchmark weights of the social network. They capture how the influence and empathy weights aggregate to impact population beliefs. The network is said to be fully-connected when \( y_{gh} > 0 \) for all groups \( g \) and \( h \). Combining the above expressions for outcome production and belief inference yields the general belief-revision equation:

\[
\tilde{\mu}_g = \mu_g + \eta \left( \tilde{\mu}_g - \sum_{h \in G} y_{gh} \bar{\mu}_h \right).
\] (1)

3.2. Belief Equilibrium and Belief Sequences

The key object of analysis in this paper, called an attributive equilibrium, is defined by a fixed point of the revision equation. That is, an equilibrium consists of beliefs which, through the

\(^{13}\)Observers are assumed to observe the full distribution of outcomes by junior group. However, because they care only about groups’ conditional mean traits, observers do not attempt to explain the full distributions and concentrate entirely on the conditional means. The discussion in Section 5.3 considers how the model could be extended so that observers would form beliefs about the full distributions.
interaction game, produce outcomes that, through the FAE, yield inferences which coincide with the original beliefs. Note that each profile of beliefs is equivalent to its implied profile of optimal actions, but, to be consistent, all definitions and results throughout the paper are phrased in terms of beliefs.

**Definition.** An **attributive equilibrium** is a profile of beliefs \( \tilde{\mu} = \{\tilde{\mu}_g^i\}_{i,g \in G} \) such that, for all \( i, g \in \mathcal{G} \), (1) group outcomes \( x = \{x_g\}_{g \in \mathcal{G}} \) are produced according to \( \tilde{\mu} \):

\[
x_g = x_0 + \eta \left( \sum_{i \in \mathcal{G}} w_{gi} \tilde{\mu}_g^i \right) + \mu_g
\]

and (2) beliefs \( \tilde{\mu} \) are inferred from \( x \) via misattribution:

\[
x_g = x_0 + \eta \left( \sum_{h \in \mathcal{G}} z_{gh} \tilde{\mu}_h \right) + \tilde{\mu}_g^i.
\]

In addition to studying equilibria, it is also useful to consider sequences of beliefs generated by iterated application of the revision equation.

**Definition.** An **attributive belief sequence** \( \{\tilde{\mu}_g^{(n)}\}_{g \in \mathcal{G}; n=1,2,...} \) is any sequence of beliefs such that

\[
\tilde{\mu}_g^{(n+1)} = \mu_g + \eta \left( \tilde{\mu}_g^{(n)} - \sum_{h \in \mathcal{G}} y_{gh} \tilde{\mu}_h^{(n)} \right)
\]

for all \( g \in \mathcal{G} \), all \( n \geq 1 \), and some initial belief profile \( \tilde{\mu}^{(0)} \).

While any convergent belief sequence must, by definition, converge to the unique equilibrium, convergence is not guaranteed. Whenever all possible belief sequences converge to an equilibrium, it is termed **stable**; equilibria that admit non-convergent belief sequences are **unstable**. Belief sequences in populations with unstable equilibria are particularly useful to study, because unstable equilibria are a potentially misleading characterization of how the FAE warps group-trait beliefs. While unstable belief sequences do not yield predictions about the long-run absolute beliefs of observers in a population, they can still characterize relative assessments.

**Existence, Uniqueness, and Stability.** Equilibrium existence follows from solving for a fixed point of the revision equation; examining the revision equation in matrix form shows that existence as well as uniqueness are guaranteed in almost all cases. Unless otherwise indicated, all populations studied in this paper are assumed to admit a unique equilibrium. Moreover, the equilibrium
is stable so long as \( \eta \) is sufficiently close to zero. As Lemma 1 below quantifies, for any fixed network structure, the cutoff value of \( \eta \) separating stable from unstable equilibria depends on \( y_{\text{min}} \equiv \min_{g \in G} y_{gg} \), the smallest of all groups’ own-group benchmark weights. (Recall \( y_{gh} = \sum_{i \in G} w_{gi} z_{ih} \).)

**Lemma 1.** For a given profile of true mean traits \( \{ \mu_g \}_{g \in G} \) and a given network structure \( \{ y_{gh} \}_{g,h \in G} \),

1. There is a unique attributive equilibrium for all but at most \( |G| \) values of \( \eta \), and all such values satisfy \( \eta \geq 1/2 \).
2. There exists
   \[
   \bar{\eta} \geq \frac{1/2}{1 - y_{\text{min}}^{\text{min}}} \geq \frac{1}{2}
   \]
   such that the equilibrium is stable whenever \( |\eta| < \bar{\eta} \) and unstable whenever \( |\eta| > \bar{\eta} \).

*Proof in the appendix.*

Intuitively, both \( \eta \) and the \( y_{gh} \) values govern the extent to which application of the belief revision equation expands or contracts the space of potential beliefs. As \( y_{\text{min}}^{\text{min}} \) approaches 1, i.e. as each group becomes the benchmark by which its own members are compared, the equilibrium is stable even for arbitrarily extreme values of \( \eta \). Similarly, as \( |\eta| \) becomes smaller – in particular, when \( |\eta| < 1/2 \) – there is a unique attributive equilibrium, and it is stable, regardless of the network structure.

### 3.3. Examples of Network Structures

Several noteworthy special cases of network structures lead to particularly simple expressions for equilibrium beliefs. This class of examples is characterized by \( y_{gh} = y_{*h} \) for all groups \( g, h \in G \): the distribution of benchmark weights over comparison groups used in evaluating any one group \( g \) is the same as the distribution for any other group \( g' \). When this condition holds, the network is said to have **common benchmark weights**. The advantage of examining common benchmark weight networks is that equilibrium beliefs admit a simple closed-form expression, which is not possible for general networks (see Lemma 3 in the Appendix for details).

Taking a \( y_{*h} \)-weighted average of both sides of Equation (1) and rearranging yields:

\[
\tilde{\mu}_g = \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{h \in G} y_{*h} \mu_h \right). \tag{2}
\]

Equation (2) shows that the equilibrium population belief \( \tilde{\mu}_g \) is equal to the true value \( \mu_g \) plus an offset term consisting of the \( \eta/(1 - \eta) \)-weighted difference between \( \mu_g \) and a common benchmark
value $\sum_{h \in G} y_{\ast h} \mu_h$. For $\eta > 0$, all groups with true mean traits above the objective benchmark are overestimated and all groups below are underestimated; the reverse is true if $\eta < 0$. Beliefs of individual observer groups admit a similarly straightforward expression:

$$\tilde{\mu}_g^i = \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{h \in G} z_{ih} \mu_h \right).$$

Instead of comparing group $g$ to the objective benchmark $\sum_{h \in G} y_{\ast h} \mu_h$, group $i$’s belief $\tilde{\mu}_g^i$ compares $g$ to the corresponding value $\sum_{h \in G} z_{ih} \mu_h$ that averages other groups according to group $i$’s own empathy weights instead of the common benchmark weights.

**Examples.** *Own-group empathy; common influence.* The leading case of network structure in the paper, as introduced in Section 2, combines own-group empathy with common influence. That is, each observer group calibrates its model using only own-group experiences, $z_{ih} = 1_{(h=i)}$, and influence is common across groups, so $w_{gi} = w_{\ast i}$. Because each group empathizes only with its own members, the extent to which group $h$ is used as a comparison against group $g$ is simply the influence weight $w_{gh}$, which by assumption does not depend on $g$. Thus $y_{gh} = w_{\ast h}$. It follows that group $g$’s population equilibrium belief is

$$\tilde{\mu}_g = \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{h \in G} w_{\ast h} \mu_h \right).$$

The common benchmark value against which all groups are compared depends only on the influence weights. This expression can be re-expressed slightly as

$$\tilde{\mu}_g = \mu_g + \frac{\eta}{1 - \eta} (1 - w_{\ast g}) \left( \mu_g - \sum_{h \neq g} \frac{w_{\ast h}}{1 - w_{\ast g}} \mu_h \right).$$

In environments where influence is proportional to population share, the degree to which a group is misestimated thus depends both on its numerical size (minorities are misestimated more severely than majorities) and the gap between its true mean trait and the population-weighted average of other groups’ mean traits. At the level of individual group-level beliefs, the formula produces

$$\tilde{\mu}_g^i = \mu_g + \frac{\eta}{1 - \eta} (\mu_g - \mu_i).$$

Each observer group’s individual benchmark value is simply its own true mean trait. Hence, for $\eta > 0$, groups overestimate the traits of other groups whose true values exceed theirs and
underestimate those groups whose true values are lower. These directional effects reverse in the case that $\eta < 0$.

**Common empathy; common influence.** Instead of calibrating their models according to own-group influence, each observer groups weights the experiences of juniors in the same proportions, and moreover the common empathy weights align with common influence weights, that is $z_{ih} = z_{*h} = w_{ih} = w_{*h}$, then $y_{gh} = w_{*h}$ as above. Thus population equilibrium beliefs are identical to those in the own-group empathy, common influence case. However, when empathy weights are common, then all observer groups have identical equilibrium beliefs, and those coincide with the population: $\hat{\mu}_h^g = \hat{\mu}_h$ for all $g, h \in \mathcal{G}$.

**Focal group.** When one group $d$ is dominant in the population, either because its members are used by all groups to calibrate their models, $z_{gd} = 1$ for all $g \in \mathcal{G}$, or because it is the only group that has positive influence and has own-group empathy, $w_{gd} = 1$ for all $g$ and $z_{dd} = 1$, then that group’s true mean trait becomes the benchmark against which all others are evaluated. Hence, group $d$ is correctly estimated, $\tilde{\mu}_d = \mu_d$, but all other groups $g$ for whom $\mu_g \neq \mu_d$, that is who do not share the same true mean level, are misestimated:

$$\hat{\mu}_g^i = \hat{\mu}_g = \mu_g + \frac{\eta}{1 - \eta} (\mu_g - \mu_d).$$

When $\eta > 0$ ($\eta < 0$) those groups with true mean traits above the dominant group $d$’s are over-estimated (under-estimated) and those groups with true values lower than $d$’s are under-estimated (over-estimated).

**Own-group empathy; own-group influence.** As a final example, consider a special case network which does not fit into the common benchmark paradigm: all groups have complete own-group empathy and own-group influence: $z_{gg} = w_{gg} = 1$ for all $g \in \mathcal{G}$. This models a society in which multiple groups coexist without contact. As each group is separated from others, it is used as its own benchmark, which means there no common set of benchmark weights. Instead, each group acts as its own single-group model and is correctly estimated at the population level: $\tilde{\mu}_g = \mu_g$. As in the own-group empathy, common influence case, individual observer groups have benchmark values equal to their own true mean traits. Therefore, in general it follows $\tilde{\mu}_h^g \neq \mu_h$ for $g \neq h$; however, because the groups are separated, out-group beliefs have no influence on outcomes.
3.4. General Properties of Equilibrium Beliefs

There are several ways to understand how attributive equilibrium beliefs compare in general to their objective counterparts. The first perspective measures errors in absolute terms by analyzing the differences $|\tilde{\mu}_g - \mu_g|$. The second examines relative beliefs, the ranking of groups by mean trait as implied by observers’ absolute beliefs. Finally, the third perspective analyzes how the beliefs of individual observer groups compare both with the true mean traits and with the population-level equilibrium beliefs.

**Absolute Correctness.** As Theorem 1 below establishes, attributive equilibrium beliefs are generically wrong. Instead of producing a general over- or under-estimate of all groups, it simultaneously skews beliefs about some groups upwards and others downwards. Recall from Section 2 that with two groups whose true mean traits are not precisely equal, one group is guaranteed to be overestimated and the other underestimated. This pattern holds generally. Per Equation (1), the equilibrium population belief about group $g$, $\tilde{\mu}_g$, consists of the true value $\mu_g$ distorted by a comparison between $\tilde{\mu}_g$ and its equilibrium benchmark $\sum_{h \in G} y_{gh} \tilde{\mu}_h$. With common benchmark weights, the value of $\sum_{h \in G} y_{gh} \tilde{\mu}_h$ for all groups $g$ is a simple average of true traits, $\sum_{h \in G} y_{*h} \mu_h$. As detailed by Lemma 3 in Appendix A.1, in general the value of $\sum_{h \in G} y_{gh} \tilde{\mu}_h$ depends on the average of true traits computed not only with with respect to $\{y_{gh}\}_{g,h \in G}$ but according to a sequence of ‘higher order’ benchmark weights. When the true mean trait of a group $g$ is equal to the weighted average according to all of its higher order benchmark weights, it is termed *medial*. General network structures need not admit any medial groups at all, and, as Theorem 1 establishes, a non-medial group is misestimated for almost all values of $\eta$. Theorem 1 also establishes bounds on how incorrect attributive equilibrium beliefs can be by first bounding the degree to which any one application of the revision equation can distort beliefs and then computing the limit of the compounded distortions.

**Theorem 1.** (a) Suppose the network is fully-connected ($y_{gh} > 0$ for all $g,h$). In equilibrium,

(i) All non-medial groups $g$ are misestimated, $\tilde{\mu}_g \neq \mu_g$, for all values of $\eta$ outside a set of Lebesgue measure zero.

(ii) All medial groups $g$ are correctly estimated, $\tilde{\mu}_g = \mu_g$.

(iii) At least one group is overestimated and at least one group is underestimated, $\tilde{\mu}_g^+ > \mu_g^+$ and $\tilde{\mu}_g^- < \mu_g^-$ for some $g^+, g^-$, provided $\eta \neq 0$ and $\max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}| > 0$.
(b) Whenever $|\eta| < (1/2)/(1 - y_{\min}^g)$, beliefs are bounded from their true values by

$$|\tilde{\mu}_g - \mu_g| \leq \frac{|\eta|(1 - y_{\min}^g)}{1 - 2|\eta|(1 - y_{\min}^g)} \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|$$

for all $g \in \mathcal{G}$. With common benchmark weights and $|\eta| < 1$,

$$|\tilde{\mu}_g - \mu_g| \leq \frac{\eta}{1 - \eta} \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|$$

for all $g \in \mathcal{G}$. Proof in the appendix.

The discussion in Section 2 noted that equilibrium beliefs correctly estimate their true value counterparts when $\eta = 0$ or when groups have identical true mean traits. Allowing for general network structures, it is also the case that equilibrium beliefs are correct when each group serves as its own benchmark, that is when $y_{gg} = 1$ for all $g \in \mathcal{G}$. The second half of Theorem 1 adds greater nuance to these conditions by providing a quantitative bound on equilibrium belief incorrectness. An immediate corollary is that equilibrium beliefs are guaranteed to approach their true counterparts if $|\eta|$ goes to zero, if the gap between different groups’ true mean traits $\max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|$ goes to zero, or if own-group benchmark weights $y_{gg}$ all approach 1.

**Relative Correctness.** Unlike in the case of measuring absolute belief errors, relative beliefs are either entirely correct, in that $\tilde{\mu}_g \geq \tilde{\mu}_h$ if and only if $\mu_g \geq \mu_h$, or they are incorrect. The following simple example illustrates how equilibrium relative beliefs can be incorrect. Suppose there are four groups 1, 2, 3, 4, that $\mu_1 > \mu_2 > \mu_3 > \mu_4$, and that $y_{11} = y_{21} = y_{34} = y_{44} = 1$. The highest and lowest groups, 1 and 4, have own benchmark weights equal to 1 and therefore are correctly estimated in equilibrium, $\tilde{\mu}_1 = \mu_1$ and $\tilde{\mu}_4 = \mu_4$. The intermediate groups are ranked $\mu_2 > \mu_3$, but group 2 is evaluated according to a benchmark of $\mu_1$, higher than group 3’s benchmark $\mu_4$. Given the simple network structure of the example, equilibrium can be computed directly:

$$\tilde{\mu}_2 = \mu_2 + \frac{\eta}{1 - \eta} (\mu_2 - \mu_1); \quad \tilde{\mu}_3 = \mu_3 + \frac{\eta}{1 - \eta} (\mu_3 - \mu_4).$$

As group 3 is compared favorably to its benchmark, $\mu_3 > \mu_4$, and group 2 is compared unfavorably, $\mu_2 < \mu_1$, sufficiently large values of $\eta$ yield $\tilde{\mu}_2 < \tilde{\mu}_3$ despite $\mu_2 > \mu_3$. Note how such a relative belief reversal would not be possible without the presence of more than two groups or without non-common benchmark weights, both of which are not part of the special case in Section 2.
Theorem 2 provides conditions that preclude such order reversals. The example above illustrates several critical factors for incorrect relative beliefs. First, it is critical that two groups may be compared to different benchmarks. With common benchmark weights, all groups share the same equilibrium benchmark, and order reversals are therefore impossible. Furthermore, it is necessary for $\eta$ to be sufficiently large, because otherwise the correctness of absolute beliefs would guarantee equilibrium relative beliefs to match the objective ranking. Finally, a switch in ranking requires at least two groups (like 2 and 3 in the example above) to have benchmarks that lead observers to evaluate them in the opposite order relative to their true traits. For high values of $\eta$, misattribution can lead to divergent belief paths in which the population belief about each group tends either to $+\infty$ or $-\infty$, with the direction solely dependent on the initial profile of beliefs. Intuitively, iterative application of misattribution only serves to amplify initial outcome gaps across groups, leading to greater and greater disparities in observers’ beliefs. Theorem 2 establishes such divergent belief paths can always be found when benchmark weights are common and $\eta$ is sufficiently high.

**Theorem 2.** (a) In a stable equilibrium, relative beliefs are correct under the following conditions.

(i) The network has common benchmark weights.

(ii) $|\eta|$ is sufficiently small:

$$
\frac{|\eta|(1 - y_{\text{min}}^{g})}{1 - 2|\eta|(1 - y_{\text{min}}^{g})} \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}| < \frac{1}{2} \min_{h_1, h_2 \neq h_3 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|.
$$

(iii) $\eta > 0$ and higher trait groups have lower benchmark weights:

$$
\sum_{h \in \mathcal{G}: \mu_{h} \leq K} (y_{g_1 h} - y_{g_2 h}) \geq 0
$$

for all $K$ and all $g_1, g_2 \in \mathcal{G}$ such that $\mu_{g_1} \geq \mu_{g_2}$.

(b) When the network has common benchmark weights and $\eta > 1$, then for any surjective assignment $s : \mathcal{G} \rightarrow \{+\infty, -\infty\}$ there exists an initial belief profile $\tilde{\mu}^{(0)}$ such that $\tilde{\mu}_{g}^{(n)} \rightarrow s(g)$ for all $g \in \mathcal{G}$.

Proof in the appendix.

At the level of individual observer groups, equilibrium beliefs share some but not all features of their aggregate counterparts. First, all observer groups agree in relative terms about the trait ordering of different groups. Although different groups may have different empathy weights and reach different inferences, the fact that group $i$ ranks group $g$ higher than $h$ reflects that $x_{g} \geq x_{h}$,
and this is the condition which determines whether any other observer group \( j \) would also rank \( g \) above \( h \). Next, observer groups are consistent in having high beliefs or low beliefs about all other groups. Whenever an observer group \( i \) estimates a subject group \( g \)'s mean trait higher than another observer group \( j \) does, then group \( i \) is guaranteed to consistently estimate traits higher than group \( j \). This follows from observer groups effectively using a single value \( \sum_{h \in \mathcal{G}} z_{ih} \tilde{\mu}_h \) with which to evaluate others; observers with low (high) values of \( \sum_{h \in \mathcal{G}} z_{ih} \tilde{\mu}_h \) arrive at high (low) absolute beliefs when \( \eta > 0 \) (\( \eta < 0 \)). Moreover, in the case of common benchmark weights, when observer groups place sufficiently high empathy on their own group, then higher (lower) ranked groups evaluate others under the implicit assumption that all groups face high (low) levels of senior beliefs and therefore adopt lower (higher) absolute beliefs.\(^{14}\) This link between a group’s rank and the absolute level of its beliefs also holds in networks where all groups have complete own-group empathy: \( z_{ii} = 1 \) for all \( i \in \mathcal{G} \).

These results are collected in Theorem 3 below along with a bound on the correctness of individual observer group beliefs. This bound largely reflects population level bounds, although it also shows how group \( i \)'s correctness about group \( g \) depends critically on the empathy weight \( z_{ig} \). As the empathy weight approaches 1, group \( i \) becomes arbitrarily correct about group \( g \) even though the population aggregate belief about \( g \) may remain incorrect. Thus, in networks in which each group only empathizes with its own members, all groups are correct about their own group’s mean trait; in networks with a dominant group in which all groups place complete empathy, all observers are correct about the dominant group’s mean trait.

**Theorem 3.** Let \( \tilde{\mu} \) be an attributive equilibrium. Then:

(a) Groups agree in relative terms:

\[ \tilde{\mu}^i_g \geq \tilde{\mu}^i_h \text{ if and only if } \tilde{\mu}^j_g \geq \tilde{\mu}^j_h \]

for all \( i, j, g, h \in \mathcal{G} \).

(b) Groups disagree consistently in absolute terms:

\[ \tilde{\mu}^i_g \geq \tilde{\mu}^i_g \text{ if and only if } \tilde{\mu}^j_h \geq \tilde{\mu}^j_h \]

\(^{14}\)This is predicated on \( \eta > 0 \). When \( \eta < 0 \), higher ranked groups also have higher absolute beliefs.
for all \( i, j, g, h \in G \).

(c) If benchmark weights are common, all groups have different true mean traits, and own-group empathy is sufficiently high (\( \min_{i \in G} z_{ii} \geq \bar{z} \) for some threshold \( \bar{z} < 1 \)), then observers’ differences in absolute beliefs depend on population beliefs about the observers’ own groups:

\[
\tilde{\mu}^i_g \geq \tilde{\mu}^j_g \quad \text{if and only if} \quad \eta(\tilde{\mu}_i - \tilde{\mu}_j) \leq 0.
\]

(d) Whenever \( |\eta| < (1/2)/(1 - y_{\min}^{\text{min}}) \), group beliefs are bounded by

\[
|\tilde{\mu}^i_g - \mu_g| \leq (1 - z_{ig}) \cdot \frac{|\eta|}{1 - 2|\eta|(1 - y_{\min}^{\text{min}})} \cdot \max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}|
\]

for all \( i, g \in G \). With common benchmark weights and \( |\eta| < 1 \),

\[
|\tilde{\mu}^i_g - \mu_g| \leq (1 - z_{ig}) \cdot \frac{\eta}{1 - \eta} \cdot \max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}|
\]

for all \( i, g \in G \).

Proof in the appendix.

4. Welfare and Policy Implications

This section discusses the welfare consequences of incorrect beliefs and the model’s implications for policy measures intended to de-bias such errors. Welfare is first analyzed at the level of individual juniors and seniors in the interaction game. It is then aggregated to groups and finally the whole population. The following subsection considers affirmative action style policies which mandate or affect preferential treatment of juniors. The second class of policies seek to broaden diversity by equalizing the profiles of influence and/or empathy weights among groups.

4.1. Welfare Analysis

There are several ways to approach welfare analysis in the population model. This subsection first studies the interaction game in isolation, tracking how the senior’s beliefs and junior’s trait translate into payoffs under optimal actions. From this analysis, the welfare of individual groups and then the population as a whole are derived.

Payoffs in the Interaction Game. It is straightforward to calculate the junior’s and senior’s payoffs in the interaction game. The following lemma expresses those values as a combination of
base utility levels which follow when the senior’s belief is correct and difference terms that account for how the senior’s belief differs from the junior’s actual trait.

Lemma 2. When the junior has trait $\theta_J$, the senior has belief $\tilde{\mu}_S^J$, and both players choose optimal actions, the expected payoffs to the junior, $u_J(\tilde{\mu}_S^J; \theta_J)$, and senior, $u_S(\tilde{\mu}_S^J; \theta_J)$, are

$$
\begin{align*}
    u_J(\tilde{\mu}_S^J; \theta_J) &= u_J(\theta_J; \theta_J) + W_J(\tilde{\mu}_S^J - \theta_J) \\
    u_S(\tilde{\mu}_S^J; \theta_J) &= u_S(\theta_J; \theta_J) - \frac{1}{2} W_S(\tilde{\mu}_S^J - \theta_J)^2,
\end{align*}
$$

where $W_J = \eta\beta_J$ and $W_S = \varphi_S^2$. Proof in the appendix.

These expressions show that the junior’s payoff is an affine function of the gap in beliefs, $\tilde{\mu}_S^J - \theta_J$. The coefficient on that gap, $W_J$, determines whether the junior benefits or loses from the senior overestimating his trait. In the teacher-student and manager-worker paradigms this value is positive, and in the officer-suspect interpretation, it is negative. The senior’s payoff, by contrast, is quadratically decreasing in the belief gap. The senior always loses from any over- or under-estimation, and the coefficient on the quadratic loss term, $W_S$, determines how acute the loss from misestimation is.

Population Welfare. Having computed the expected payoffs to juniors and seniors in the interaction game, it is straightforward to extend this analysis to the group level.

Fact 1. When the population has beliefs $\tilde{\mu} = \{\tilde{\mu}_g\}_{i,j \in G}$, the expected payoffs to juniors and seniors of group $g$, $v_g(\tilde{\mu})$ and $v^g(\tilde{\mu})$ respectively, are

$$
\begin{align*}
    v_g(\tilde{\mu}) &= v_g(\mu) + W_J(\tilde{\mu}_g - \mu_g) \\
    v^g(\tilde{\mu}) &= v^g(\mu) - \frac{1}{2} W_S \sum_{h \in G} w_{gh}(\tilde{\mu}_h - \mu_h)^2.
\end{align*}
$$

Proof in the appendix.

This characterization implies what is perhaps the most significant welfare consequence of misattribution, namely that for juniors the bias creates both winners and losers.

Corollary. For any population that satisfies the conditions of Theorem 1 (a), there exists at least one group $g$ which is strictly harmed by the presence of the FAE, and for sufficiently low values of $\eta$, there is at least one group that is made strictly better off.
By contrast, all seniors lose from making errors on account of the FAE. A group whose juniors are negatively misestimated therefore strictly loses. A group whose juniors are positively misestimated benefits on net from small values of $\eta$, but, as $\eta$ gets larger, eventually the losses to the seniors outweigh the benefits to the juniors, and the group as a whole loses as well.

Finally, aggregate population utility $U(\tilde{\mu})$ is computed as the weighted sum over the individual expected utilities of the different groups in the two roles. Let $p_g$ denote the proportion of juniors among group $g$ members, and let $p^g$ denote the proportion of seniors. Then,

$$U(\tilde{\mu}) = U(\mu) + W_J \sum_{g \in G} p_g (\tilde{\mu}_g - \mu_g) - \frac{1}{2} W_S \sum_{g \in G} p^g \sum_{h \in G} w_{gh} (\tilde{\mu}_h^2 - \mu_h)^2. \quad (3)$$

This expression reveals several insights. First, the utilities of juniors are entirely aligned, and maximized when either $\tilde{\mu}_g = \min_{h \in G} \mu_h$ (if $W_J > 0$) or $\tilde{\mu}_g = \max_{h \in G} \mu_h$ (if $W_J < 0$). Second, seniors benefit from having relatively correct beliefs about the groups with which they most often interact. This relationship is clarified by examining the above population utility expression as simplified in the case of common benchmark weights.

**Fact 2.** If benchmark weights are common, population aggregate welfare is equal to

$$U(\tilde{\mu}) = U(\mu) + W_J \cdot \frac{\eta}{1 - \eta} \left( \sum_{g \in G} p_g \left\{ \mu_g - \sum_{h \in G} y_{gh} \mu_h \right\} \right)$$

$$- \frac{1}{2} W_S \cdot \left( \frac{\eta}{1 - \eta} \right)^2 \sum_{g \in G} p^g \left\{ \left( \sum_{h \in G} z_{gh} \mu_h - \sum_{h \in G} w_{gh} \mu_h \right)^2 + \left( \sum_{h \in G} w_{gh} \mu_h^2 - \sum_{h \in G} w_{gh} \mu_h \right)^2 \right\}.$$

*Proof in the appendix.*

This expression combines three critical factors. First is $\Delta \mu$, which is equal to the population weighted mean trait minus the common benchmark value as determined by network parameters. If $\Delta \mu > 0$ ($\Delta \mu < 0$), then the members of the population are, on average, empathizing with low (high) trait groups. Aggregate junior welfare increases or decreases in $\eta$ according to the sign of $\Delta \mu$. Second is $(\Delta \mu^g)^2$, which reflects the difference between the average trait as weighted by group $g$’s empathy parameters and by group $g$’s influence parameters. Lastly, $V_{\mu^g}$ denotes the ‘variance’ in the trait means of the different groups from the perspective of a group $g$ senior. It reflects a utility loss that seniors suffer from only observing the junior’s group identity and not his individual
trait. Aggregate senior welfare strictly decreases in $\eta$, with the rate of loss increasing with the values of $(\Delta \mu^g)^2$ and $V \mu^g$.

Finally, the closed-form expression for aggregate welfare in the common benchmark case makes it possible to compute the socially optimal level of the FAE bias.

**Corollary.** If benchmark weights are common and all agents have transferrable utility, the Pareto optimal level of $\eta$ is

$$\eta^* = \left(1 + \frac{1}{W_J \Delta \mu} \cdot W_S \sum_{g \in \mathcal{G}} p^g \left\{ (\Delta \mu^g)^2 + V \mu^g \right\} \right)^{-1}$$

when $W_J \Delta \mu > 0$ and $\eta^* = 0$ when $W_J \Delta \mu < 0$.

Crucially, $\eta^*$ is strictly positive when juniors are on average positively misestimated. In the absence of transferrable utility, no level of bias would produce a Pareto improvement. Furthermore, with or without transferrable utility, the losses to seniors outweigh any gains to juniors for sufficiently high levels of $\eta$.

4.2. Policy Analysis

This subsection considers the potential of three different kinds of policy measures to de-bias incorrect attributive beliefs. First, a governing body may seek to address disparities in group treatment by implementing a **transfer** policy. There are two ways to implement such a policy. Seniors may be compelled to treat juniors according to preferential or disadvantaging standards which differ from otherwise optimal (according to senior belief) actions. That is, seniors of group $i$ with belief $\tilde{\mu}_i^g$ about group $g$ are required to take the action corresponding to the belief $\tilde{\mu}_i^g + D_g$.

In some cases, such as employment, the physical quantity modeled by the game product is itself fungible, and thus the transfer policy can be imposed instead through direct transfers of game product. That is, a juniors of group $g$ are assigned a transfer of $(\partial x/\partial \tilde{\mu})^{-1} D_g$ units\(^{15}\) of the game product in each match. Any profile $\{D_g\}_{g \in \mathcal{G}}$ such that $\sum_{g \in \mathcal{G}} p_g D_g = 0$ is called a transfer scheme and defines a particular transfer policy.

A governing body may alternatively adopt a **diversity** policy which either shifts the composition of observers’ empathy weights or alters the population’s influence weights. Specifically,\(^{15}\) The factor $(\partial x/\partial \tilde{\mu})^{-1}$ serves to express the transfer in terms of senior beliefs and facilitate comparison with the preferential/disadvantaging treatment implementation.
empathy weights are shifted towards target distribution \( \{ z_{tg} \}_{g \in G} \), such that observer group \( i \) places weight \((1 - \alpha)z_{ig} + \alpha z_{*g} \) on group \( g \), or influence weights are shifted towards a target profile \( \{ w_{*i} \}_{i \in G} \), such that juniors of group \( g \) are matched with seniors of group \( i \) at rate \((1 - \alpha)w_{gi} + \alpha w_{*i} \). A policy may also target both distributions simultaneously. Finally, a policy of mandated equal treatment requires that each senior chooses a single action and applies it to all juniors with which she interacts. Unlike with transfer or diversity polices, equal treatment is a single specification without any parameters of variation.

The model produces starkly divergent results concerning the de-biasing potential of the different kind of policies. Any population with incorrect beliefs can be successfully de-biased by an appropriately chosen transfer scheme, while no population can be de-biased through any diversity policy, and an equal treatment policy only de-biases for a single special case of network structure.

**How Transfers Can De-Bias Beliefs.** Suppose that a transfer policy is implemented by mandating seniors to act according to \( \tilde{\mu}_g^i + D_g \). The belief-revision equation is accordingly modified to

\[
\hat{\mu}_g^i = \mu_g + \eta \left( \tilde{\mu}_g + D_g - \sum_{h \in G} z_{ih}(\tilde{\mu}_h + D_h) \right),
\]

including \( D \) terms which appear both in the treatment of the subject group \( g \) and the comparison groups indexed by \( h \). A transfer scheme that de-biases beliefs must yield

\[
\mu_g + D_g - \sum_{h \in G} z_{ih}(\mu_h + D_h) = 0
\]

simultaneously for all groups \( g \in G \). An intuitive way of accomplishing this, and, as Theorem 4(a) establishes is the only way, chooses \( D_g \) such that \( \mu_g + D_g \) is a constant. Thus, a de-biasing transfer scheme advantages or disadvantages groups according to how the group’s true mean trait differs from a fixed level. That level is uniquely determined by the constraint that transfers net to zero across the population.

**Theorem 4(a).** For any population with incorrect attributive beliefs, there exists a unique transfer policy that de-biases beliefs. The transfers are equal to

\[
D_g^* \equiv - \left( \mu_g - \sum_{h \in G} p_h \mu_h \right)
\]

for all \( g \in G \). Proof in the appendix.
How Greater Diversity Does Not De-Bias Beliefs. Under a diversity policy, the governing body chooses target profiles of \( \{w^*_{si}\}_{i \in G} \) or \( \{z^*_{sh}\}_{h \in \hat{G}} \), yielding a target profile of benchmark weights \( \{y^*_{si}\}_{i \in G} \). The belief-revision equation is modified to

\[
\tilde{\mu}_g = \mu_g + \eta \left( \tilde{\mu}_g - \left( (1 - \alpha) \sum_{h \in \hat{G}} y_{gh} \tilde{\mu}_h + \alpha \sum_{h \in \hat{G}} y^*_{sh} \tilde{\mu}_h \right) \right).
\]

Intuitively, the policy shifts equilibrium beliefs towards comparing groups against the modified benchmark value \( \sum_{h \in \hat{G}} y^*_{sh} \mu_h \), which does not de-bias beliefs but instead alters the nature of misestimation. As the policy intensity level \( \alpha \) approaches 1, the population is shifted to full weight on the common benchmark, guaranteeing that at least the objectively highest or lowest mean trait group is misestimated. When the population itself has common benchmark weights, the modified equilibrium beliefs are

\[
\tilde{\mu}_g = \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{h \in \hat{G}} [(1 - \alpha) y_{ih} + \alpha y^*_{ih}] \mu_h \right),
\]

demonstrating that not only is a diversity policy unable to de-bias beliefs, but that at least one group must be more severely misestimated under the policy than without it. As demonstrated in the proof of Theorem 4(b), even in general networks it is impossible to simultaneously de-bias all beliefs through any choice of diversity policy.

**Theorem 4(b).** For any population with incorrect attributive beliefs, there is no diversity policy that de-biases beliefs. Proof in the appendix.

How Equal Treatment Typically Does Not De-Bias Beliefs. An equal treatment policy impacts equilibrium beliefs in a different manner than do transfer and diversity policies. A senior of group \( i \) who is mandated to set a fixed action to apply to all juniors would optimally adopt

\[
a^*_{si} = a_S(\tilde{\mu}^i_s), \text{ where } \tilde{\mu}^i_s = \sum_{g \in \hat{G}} w_{ih} \cdot \tilde{\mu}^i_g, \text{ where } w_{ig} \text{ is the probability with which the senior is matched to a group } g \text{ junior. The resulting aggregate effective belief that determines group } g \text{'s treatment is no longer the } w_{gi}^-\text{-weighted average of the senior groups' beliefs. Instead group } g \text{ faces the } w_{gi}^-\text{-weighted average of observer groups' values of } \tilde{\mu}^i_s. \text{ The modified belief-revision equation is modified to}
\]

\[
\tilde{\mu}^i_s = \mu^i_s + \frac{\eta}{1 - \eta} \left( \mu^i_s - \sum_{h \in \hat{G}} [(1 - \alpha) y_{ih} + \alpha y^*_{ih}] \mu^i_h \right),
\]

demonstrating that not only is an equal treatment policy unable to de-bias beliefs, but that at least one group must be more severely misestimated under the policy than without it. As demonstrated in the proof of Theorem 4(b), even in general networks it is impossible to simultaneously de-bias all beliefs through any choice of diversity policy.
equation can therefore be expressed in two steps as

\[
\begin{align*}
\hat{\mu}^i_g &= \mu_g + \eta \left( \hat{\mu}_g - \sum_{h \in \mathcal{G}} z_{ih} \hat{\mu}_h \right) \\
\hat{\mu}_g &= \sum_{i \in \mathcal{G}} w_{gi} \sum_{h \in \mathcal{G}} w_{ih} \hat{\mu}^i_h,
\end{align*}
\]

As a special case, when all groups empathize with others to the extent that they influence them, that is \( w_{ih} = z_{ih} \) for all \( i, h \in \mathcal{G} \), equilibrium beliefs can be expressed simply as

\[
\tilde{\mu}_g = \sum_{h \in \mathcal{G}} \left( \sum_{i \in \mathcal{G}} w_{gi} w_{ih} \right) \mu_h.
\]

Each group \( g \) is thus associated with a ‘local average’ of true mean traits determined by which juniors are also affected by the seniors who affect group \( g \). If all of the local averages are the same as, for example, is the case when influence weights are common (\( w_{gi} = w_{*i} \) for all \( g, i \in \mathcal{G} \)), then observers are not skewed by mistakenly comparing groups against incorrect benchmarks, and equilibrium beliefs are correct. As Theorem 4(c) below shows, the equal local averages condition is both necessary and sufficient for de-biasing regardless of the network structure. Note that even when groups are correctly estimated by seniors, however, they still face the local average level of treatment and not the level of treatment corresponding to their true mean values.

**Theorem 4(c).** An equal-treatment policy de-biases beliefs if and only if

\[
\sum_{i \in \mathcal{G}} w_{g1i} \sum_{h \in \mathcal{G}} w_{ih} \mu_h = \sum_{i \in \mathcal{G}} w_{g2i} \sum_{h \in \mathcal{G}} w_{ih} \mu_h
\]

for all \( g_1, g_2 \in \mathcal{G} \). Proof in the appendix.

**Literal De-Biasing.** As a final point of comparison, it is instructive to consider the consequences of a policy which managed to directly mitigate observers’ FAE bias. Were it the case that, when group \( i \) evaluated group \( g \), it placed \( \alpha \) weight on the actual experiences of group \( g \) members, then the revision equation would be

\[
\hat{\mu}^i_g = \mu_g + \eta \left( \hat{\mu}_g - \left( 1 - \alpha \right) \sum_{h \in \mathcal{G}} z_{ih} \hat{\mu}_h + \alpha \hat{\mu}_g \right).
\]
Thus, equilibrium beliefs would be the same as in the original specification but with $\eta$ replaced by $\eta' = (1 - \alpha)\eta$. For $\alpha$ approaching 1, the modified equilibrium would continuously approach correct beliefs.

5. Discussion

This paper has introduced a model of discrimination in which members of a population are associated with outcomes through repeatedly interacting with others. Outcomes reflect both individuals’ own traits and the beliefs that others have about those traits. However, observers biased by the Fundamental Attribution Error neglect the role of others’ beliefs and instead infer traits under the assumption that there is a single trait-to-outcome function which is valid for all members of the population. Observers’ inferred beliefs about different groups in turn influence group outcomes. A profile of beliefs that produces outcomes which are inferred as the original beliefs constitutes an attributive equilibrium.

For all but a negligible set of parameter profiles, an attributive equilibrium exists and is unique. Population level beliefs are incorrect for all but several special cases, including when all groups share identical mean traits. Otherwise, all groups are misunderstood by the population, with at least one group overestimated and at least one underestimated. In the case of unstable equilibria, the paper studies belief sequences produced through iterated application of misattribution. In many populations, divergent belief sequences can support arbitrary orderings of groups by outcomes given sufficiently incorrect initial beliefs. For stable equilibria, members of different observer groups always agree about the relative ranking of groups by mean trait. However, they consistently disagree in absolute terms, with some observer groups having high beliefs and others low beliefs. Data from the General Social Survey, discussed below, demonstrate that these general patterns are indeed found in assessments about the innate qualities of different racial/ethnic groups.

In terms of welfare, juniors benefit from being either over- or under-estimated by seniors, while seniors always suffer from any incorrectness. At the population level, an attributive equilibrium can Pareto dominate (with utility transfers) its unbiased counterpart, but only for certain networks and small values of $\eta$. For sufficiently high $\eta$, equilibrium welfare is lower than the unbiased counterpart. Policies intended to de-bias the population have mixed effects. When seniors are required to give juniors group-dependent preferential treatment, or transfers implement such preference, beliefs can
be completely corrected. However, policies that achieve greater diversity in terms of influence or empathy cannot de-bias beliefs about all groups simultaneously.

The remainder of this section discusses the model more broadly. First, the discussion compares the model’s predictions regarding beliefs held by members of different social groups with data from the General Social Survey. Next, it ties misattribution to other equilibrium concepts that incorporate incorrect learning in games. Finally, I consider how relaxing the model’s key assumptions could influence the results and propose several particular extensions.

5.1. **Evidence from the General Social Survey**

Administered by the National Data Program for the Social Sciences at the University of Chicago every one to two years since 1972, the General Social Survey reports the responses of one to two hour in-person interviews. Each wave of the survey also includes cohort-specific modules, and in 2000 the GSS queried $N = 1,322$ respondents\textsuperscript{16} on their views regarding White, Black, Asian and Hispanic Americans. Respondents who self identified as one of the four groups provided numeric assessments regarding the intelligence, work ethic, and proclivity to violence of ‘typical members’ of each of the four groups. Each trait was measured on a scale from 1 to 7. The data thus provide an opportunity for testing the predictions of Theorem 3 regarding the relationships between the average assessments of different groups. The figures below (1, 2, and 3) show the average ratings given by each respondent group about all four groups for each trait.

Comparing the ordering of these mean beliefs reveals support for the predictions of Theorem 3 and illustrates the value of further testing the model empirically in future work. Concerning relative beliefs, in 91.7% of specifications of an observer-group pair with a target-group pair, the two observer groups rank the two target groups in the same order with respect to intelligence (80.6% with respect to work ethic, 83.3% with respect to proclivity to violence). In 69.4% of such specifications, the observer group which gives a higher average assessment of one of the target groups with respect to intelligence also gave a higher average assessment of the other (50.0% with respect to work ethic, 61.1% with respect to proclivity to violence). Finally, in 79.2% of specifications of an observer-group pair with a given target group, the higher ranked (according to the most common ranking across different observer groups) observer group had a lower assessment of the target group

\textsuperscript{16}Per the GSS, respondents are a ‘sample of English-speaking persons 18 years of age or over, living in non-institutional arrangements within the United States.’ See http://gss.norc.org/ for details.
relative to the lower ranked observer group with respect to intelligence (66.7% with respect to work ethic). With respect to proclivity to violence, 75.0% of specifications feature the higher ranked observer group giving the higher average assessment.$^{17}$

![Figure 1. GSS Assessments of Intelligence](chart)

**Figure 1. GSS Assessments of Intelligence**

Note: Solid dots mark the mean rating given by respondents of the group listed at far left concerning the intelligence (1-7; 7 most intelligent) of the group listed second to left. Light dots show means plus and minus one standard deviation.

$^{17}$The data are thus most consistent with $\eta > 0$ with respect to intelligence and work ethic and $\eta < 0$ for proclivity to violence, echoing the predictions in Section 3.4.
Figure 2. GSS Assessments of Work Ethic

Note: Solid dots mark the mean rating given by respondents of the group listed at far left concerning the work ethic (1-7; 7 hardest working) of the group listed second to left. Light dots show means plus and minus one standard deviation.

Figure 3. GSS Assessments of Violence

Note: Solid dots mark the mean rating given by respondents of the group listed at far left concerning the proclivity to violence (1-7; 7 most violent) of the group listed second to left. Light dots show means plus and minus one standard deviation.
Additional Evidence from the Literature. In addition to the data from the General Social Survey, the model’s predictions also echo Gershenson et al. (2016), who find that non-black teachers have lower expectations vis-a-vis black teachers about black students in their classes, and Antonovics and Knight (2009), who analyze the probabilities of police officers searching a suspect and show that while all officers search non-white suspects more often than white suspects, this effect is more pronounced for white officers. It is also worth noting how the model predicts that the level of absolute disagreement between observer groups should hinge on the difference in observer groups’ empathy weights, which comports with the general pattern that agreement between men and women with respect to beliefs about gender is higher than agreement between racial groups with respect to beliefs about race.

On the other hand, empirical studies often document men and women holding relatively similar beliefs about gender. For example, Eckel and Grossman (2008) find male and female lab subjects uniformly rating the risk tolerance of men above women, and Bordalo et al. (2019) find that both men and women in a laboratory experiment predict men to outperform women in male-typed trivia categories but underperform relative to women in female-typed categories. Reuben et al. (2014), documenting that laboratory subjects acting as managers were less likely to choose a female over male employee, also found that female employers did not choose female employees at a significantly higher rate. Nonetheless, there is no clear consensus. Carrell et al. (2010), for one, finds that assignment to female professors helps improve outcomes in female students’ STEM subject performance levels. In a similar vein, Alan et al. (2018) study the effect of teacher attitudes on student performance and find that, while both male and female teachers include those who subscribe to more traditional gender stereotypes, the proportion is greater among male teachers.

5.2. Relationships With Other Solution Concepts

Relationships with Other Equilibrium Concepts. The equilibrium concept defined in this paper is closely related to Berk-Nash Equilibrium (Esponda and Pouzo, 2016). In their framework, players in a game are endowed with a set of possible subjective models; as the game generates feedback signals, players select and best-respond to the model whose implied distribution over feedback minimizes the Kullback-Leibler divergence with the observed distribution. In the present paper, players likewise form beliefs by fitting a subjective model with generated data and then best-respond accordingly. However, as they care only about mean traits and, as they are able to perfectly
explain any observed data (mean group outcomes) with a unique model specification (mean group traits), players need not engage in the KL divergence minimizing exercise that Berk-Nash requires. In a more general setup, such as the extension proposed in the following subsection, observers may find that outcome distributions do not perfectly correspond to any single distribution over traits. In this extension a natural generalization of the equilibrium concept specifies that agents’ beliefs, in conjunction with their subjective trait-to-outcome models, minimize KL divergence with the observed distribution.

A looser connection exists between this paper’s model of inference and the literature on Self Confirming Equilibrium (Battigalli, 1987; Fudenberg and Levine, 1993; Dekel et al., 2004). In such an equilibrium, players correctly understand the rules of the game but can maintain incorrect beliefs about other players’ actions and types because game feedback is too coarse to correct them. By contrast, observers with the FAE fundamentally misunderstand the production of outcomes. On account of the FAE, they assume that outcomes are produced directly from one’s trait, which ignores the role of seniors’ actions. Accordingly, observers have no beliefs about what actions seniors take against juniors. Given the model’s particular specification, however, any affine trait-to-outcome function is mathematically equivalent to assuming that all seniors apply some common action to all juniors. In a more general model this connection need not hold. The presence of a dogmatic reliance on an incorrect model of game structure also has parallels in the simplified models of opponent play in Eyster and Rabin (2005) and Jehiel (2005), as well as naive social learning (Eyster and Rabin, 2010), in which players develop beliefs and conjectures that cannot withstand strategic introspection.

5.3. **Relaxing Model Assumptions and Future Work**

In this subsection I describe several particular directions in which the model’s key assumptions might be relaxed. In particular, I consider extending the interaction game structure by employing sequential moves instead of simultaneous moves, by considering a pipeline of multiple seniors who sequentially contribute to the junior’s outcome, and by allowing for non-linear specifications of the best response and game production functions. Next, I discuss how the population level results would be changed if matching between seniors and juniors were not a primitive of the model but rather endogenously determined. Finally, there are several assumptions of the model for which generalizations would not impact the paper’s main results.
Beyond Single-Shot Interactions. In the model, the senior and junior each take a single, simultaneous action. While the production function \( x = x_0 + \eta \tilde{\mu} + \theta \) can be understood as a reduced-form simplification of a more complicated game in which parties take multiple actions and observe interim histories,\(^{18}\) it is conceivable that repeated interaction would diminish the effect of observation. For instance, Bohren et al. (2019b) develop a model showing that with initial beliefs driven by stereotypes, subsequent interactions gradually correct them, and correct them faster for groups that were initially more extremely stereotyped. Still, empirical studies such as Grossman (2013) find that additional information may be downplayed relative to initial beliefs, so the total effect may be weak.

There are two intuitive ways to consider how increased interim feedback between the players would affect the model. First, the senior-then-junior action sequence could be extended to allow for more involved interaction A more dramatic way in which the model could be made more realistic is allowing for multiple seniors to influence the outcome of any one junior. This echoes the way that successive grades of teachers influence children's development, how multiple bosses influence workers' promotions, and how multiple kinds of officials determine how suspects move through the justice system. Formally, the interaction game would be modified so that a junior is matched, sequentially and independently over time, with \( T > 1 \) seniors. In each period \( t \) of the game, the junior and his time-\( t \) senior would observe the history of period \( s \) products \( x^s, s < t \), and then choose \( a^t_J \) and \( a^t_S \), which would jointly produce \( x^t \) according to

\[
x^t(a^t_S, a^t_J) = x^t_0 + \eta^t \tilde{\mu}^t_S + \theta^t.
\]

This assumes that juniors myopically optimize their action in each period, although it would also be useful to consider how juniors might strategically over- or under-invest in early period actions to influence the beliefs of later period seniors. In the modified population, observers would form both \( \theta \mapsto x^T \) subjective correspondences and also interim \( \theta \mapsto x^t, t < T \), subjective models. Analysis of this extension could answer whether longer sequences lead to more or less incorrect discrimination, and whether seniors matched to a given junior become more or less correct over the course of the sequence.

\(^{18}\)Provided there remained a unique equilibrium conditional on ex-ante beliefs and trait, there would be some relationship between fundamentals and the distribution of game products; one can understand \( x = x_0 + \eta \tilde{\mu} + \theta \) as its linear approximation.
Beyond Linear Functional Forms. The most critical simplifying assumption in the model is the linearity of both best response and production functions. This assumption is made primarily in the interests of tractability and transparency. It guarantees that, despite non-compact action spaces, equilibrium nearly always exists and is unique. Furthermore, it simplifies the analysis of belief formation. A generalized version would take \( u_J(a_S, a_J; \theta_J) \), \( u_S(a_S, a_J) \), and \( x(a_S, a_J) \) as primitives functions. This would require a general form of inference that uses full distributions of outcomes and minimizes KL-divergence rather than just matching means. Denoting group \( g \)'s distribution of traits by \( p_g \), distribution of outcomes by \( q_g \), and trait-conditional distribution of outcomes by \( q_g(\cdot | \theta) \), observer group \( i \) would select beliefs \( \bar{p}^i_g \) to maximize

\[
\int_{x \in \mathbb{R}} q_g(x) \cdot \log \left( \int_{\theta \in \mathbb{R}} \bar{p}^i_g(\theta) \cdot \sum_{h \in G} z_{ih} q_h(x | \theta) \, d\theta \right) \, dx.
\]

This is equivalent to observers minimizing the Kullback-Leibler divergence between the observed distributions of outcomes and those predicted by different trait distributions, given the trait-to-outcome function. In this general framework, neither equilibrium existence nor uniqueness would be guaranteed, and observers could have incorrect beliefs about the higher order moments of groups' trait distributions.

Beyond Exogenous Matching. The model specifies that agents are matched together according to fixed interaction weights, ruling out the possibility for matching to depend on beliefs. In the manager-worker interpretation of the model in particular, seniors might exploit their relative beliefs by seeking out juniors from highly assessed groups. To account for this possibility, the model could allow the weights \( w_{ig} \) to depend on \( \bar{\mu}_g \), such as \( w_{ig}(\bar{\mu}) \propto \exp\{\bar{w}_{ig} + \rho \cdot \bar{\mu}_g\} \). This extension would address equilibrium existence and whether equilibria or limits of belief sequences would produce complete or partial segmentation of the population.

Other Model Assumptions. Observers in the population use an extreme form of the FAE, using outcomes to infer traits without any reference to treatment. This specification is made in the interest of highlighting how the FAE warps equilibrium beliefs, but as demonstrated in Section 4.2, it does not qualitatively alter any results. When observers place weight \( \alpha \) on the actual experiences of the groups they evaluate, the modified equilibrium beliefs correspond to those of the baseline model with factor \( \eta \) replaced by \((1 - \alpha)\eta \). In other words, the assumption can be gradually relaxed in
the style of Eyster and Rabin (2005) with the straightforward consequence of gradually weakening the main effects.

The observation structure of the model is specified in a way that focuses attention on group-level beliefs. Namely, observers see the full distributions of outcomes by group, but seniors do not see juniors’ outcomes and instead rely on the junior’s group as an informative signal. If instead seniors could perfectly see juniors’ outcomes, then seniors would directly form an individual-specific belief completely on the basis of the junior’s outcome. However this modification would not alter the equilibrium value of $\tilde{\mu}$. Owing to the linearity and uniformity specifications, the belief of a group $i$ member about the trait of a randomly selected group $g$ member would remain unchanged. Similarly, the model treats seniors of the same group as homogenous in calibrating their trait-to-outcome models, but this assumption is made solely to simplify exposition. For example, instead of a single empathy weight $z_{ih}$ that held for all seniors of group $i$, there could be a distribution of individual empathy weights whose mean is $z_{ih}$; this granular specification of individual beliefs, however, would not alter the mean belief of the group.

5.4. Concluding Thoughts

From a broad perspective, this paper illustrates how discrimination driven by agents with a cognitive bias differs substantially from classical statistical discrimination in which beliefs reflect actual circumstances and serve as coordination devices for multiple equilibria. Beliefs about groups are not only wrong, but systematically misestimated. Individual teachers, managers, or police officers maintain biased beliefs and therefore stand to benefit from de-biasing policies. These departures from classical theory evidence both the particular bias of the FAE and the population equilibrium framework within which it is studied. The population framework could accordingly serve as a template for future work examining how other biases, for example, the stereotyping of Bordalo et al. (2016) or the categorical thinking of Fryer and Jackson (2008), might analogously sustain systematically incorrect discriminatory beliefs.


Guryan, Jonathan and Kerwin Kofi Charles, “Taste-based or Statistical Discrimination: The


Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, “How Stereotypes Impair Women’s


**Smith, Adam**, *The Theory of Moral Sentiments*, 1759.


Appendix A. Proofs

A.1. Formulae for Equilibrium Beliefs and Beliefs Sequences

The equilibrium belief formula in Equation 2 has a counterpart for general networks, but the structure is more complicated. Group $g$ is compared not only against its own objective benchmark $\sum_{h \in G} y_{gh} \mu_h$, but also against the $y_{gh}$-weighted average of other groups’ objective benchmarks and all successive ‘higher order’ averages. Formally, the $k$’th order benchmark weights $y_{gh}^{(k)}$ are defined by

$$y_{gh}^{(k)} = \sum_{j_1, \ldots, j_k \in G} y_{gj_1} \cdot y_{j_1j_2} \cdot \ldots \cdot y_{j_{k-2}j_{k-1}} \cdot y_{j_{k-1}h}.$$  

As Lemma 3 below relates, an alternating sum of the true group means as weighted by these coefficients defines each group’s ultimate level of comparison. What makes common benchmark weights simpler is that, for such networks, all higher order benchmark weights reduce to the original values: $y_{gh}^{(k)} = y_{*h}$ for all $g, h \in G$ and $k \geq 0$.

Lemma 3. In any attributive equilibrium,

$$\tilde{\mu}_g = \mu_g + \frac{\eta}{1-\eta} \left( \mu_g - \sum_{h \in G} \bar{y}_{gh} \mu_h \right),$$

where

$$\bar{y}_{gh} = \sum_{k=1}^{\infty} \frac{(-\eta)^{k-1}}{(1-\eta)^k} y_{gh}^{(k)}.$$  

$\sum_{h \in G} \bar{y}_{gh} = 1$ for all $g \in G$. For general networks, $\bar{y}_{gh} > 0$ for all $g, h \in G$ is guaranteed when

$$|\eta| < \frac{1}{1 - \min_{g,h \in G} y_{gh}} \cdot \frac{y_{*h}^{\min}}{1 - 2y_{*h}^{\min}}.$$  

When benchmark weights are common, $\bar{y}_{gh} = y_{h}$ for all $g, h \in G$.

Proof of Lemma 3. As a preliminary matter, note that the Taylor expansion of $g(\eta) = (1 - \eta)^{-(k+1)}$ around $\eta = 0$ is

$$g(\eta) = \sum_{l=0}^{\infty} \frac{\eta^l}{l!} g^{(l)}(0) = \sum_{l=0}^{\infty} \frac{\eta^l}{l!} (1 - 0)^{-(k+l+1)} \frac{(k+l)!}{k!} = \sum_{n=k}^{\infty} \frac{\eta^{n-k}}{(n-k)!} \cdot \frac{n!}{k!} = \sum_{n=k}^{\infty} \eta^{n-k} \binom{n}{k},$$
where the last line substitutes \( n \equiv l + k \). From this we obtain
\[
\sum_{n=k}^{\infty} \eta^n \binom{n}{k} = \frac{\eta^k}{(1 - \eta)^{k+1}}.
\]

Now, in matrix form, equilibrium beliefs are given by
\[
\tilde{\mu} = \sum_{n=0}^{\infty} \eta^n (I - \mathcal{Y})^n \mu.
\]

Leveraging the Binomial identity and the equivalence shown above, this can be re-arranged as follows:
\[
\tilde{\mu} = \mu + \sum_{k=1}^{\infty} (-1)^{k+1} \left( \sum_{n=k}^{\infty} \eta^n \binom{n}{k} \right) (I - \mathcal{Y}^k) \mu
\]
\[
= \mu + \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\eta^k}{(1 - \eta)^{k+1}} (I - \mathcal{Y}^k) \mu
\]
\[
= \mu + \frac{\eta}{1 - \eta} \left( I - \sum_{k=1}^{\infty} \frac{(-\eta)^{k-1}}{(1 - \eta)^k} \mathcal{Y}^k \right) \mu
\]
\[
= \mu + \frac{\eta}{1 - \eta} (I - \mathcal{Y}) \mu.
\]

Further rearranging \( \mathcal{Y} \) in terms of \( \mathcal{Y} \) yields
\[
\mathcal{Y} = \mathcal{Y} (I - \eta (I - \mathcal{Y}))^{-1},
\]
so for any \( \mu \),
\[
\bar{\mathcal{Y}} \mu = \mathcal{Y} \tilde{\mu}(\mu),
\]
where \( \tilde{\mu}(\mu) \) is the attributive equilibrium corresponding to \( \mu \). Decomposing \( \tilde{\mu}(\mu) = \mu + \varepsilon(\mu) \) yields
\( (\bar{\mathcal{Y}} - \mathcal{Y}) \mu = \mathcal{Y} \varepsilon(\mu) \). Now leveraging the bound on population beliefs in Theorem 1, by setting \( \mu = e = (1, \ldots, 1) \) we guarantee \( \varepsilon(\mu) = 0 \), which then implies \( \bar{\mathcal{Y}} e = \mathcal{Y} e \), so \( \sum_{h \in \mathcal{G}} y_{gh} = \sum_{h \in \mathcal{G}} y_{gh} = 1 \) for all \( g \in \mathcal{G} \). Finally, let \( \mu = e_g = (0, \ldots, 1, \ldots, 0)' \). From the bound in Theorem 1,
\[
\bar{y}_{gh} > \min_{g, h \in \mathcal{G}} y_{gh} - \frac{|\eta|(1 - y_{gh}^{\min})}{1 - 2|\eta|(1 - y_{gh}^{\min})}.
\]
Rearranging this expression shows \( \bar{y}_{gh} > 0 \) whenever \( |\eta| \) satisfies the bound in the Lemma statement.
Next, the lemma below explicitly characterizes of belief sequences as a function of initial beliefs.

**Lemma 4.** If \( \{\tilde{\mu}(n)\}_{n \in \mathbb{N}} \) is an attributive belief sequence, then

\[
\tilde{\mu}(n) = \mu + \eta (\eta(I - Y) - I)^{-1} \left[ [\eta(I - Y)]^{n-1} \left( \delta + (\eta(I - Y) - I)\tilde{\delta}(0) \right) - \delta \right],
\]

and if the network has common benchmark weights,

\[
\tilde{\mu}(n) = \mu + \frac{\eta}{\eta - 1} \left[ \eta^{n-1} \left( \delta + (\eta - 1)\tilde{\delta}(0) \right) - \delta \right],
\]

where \( \delta = (I - Y)\mu \) and \( \tilde{\delta}(0) = (I - Y)\tilde{\mu}(0) \).

Long-run behavior of a belief sequence is influenced by the key term \( \delta + (\eta(I - Y) - I)\tilde{\delta}(0) \) or, in the common benchmark weight case, \( \delta + (\eta - 1)\tilde{\delta}(0) \), a combination of groups’ true differences with respect to their objective benchmarks and the differences in terms of initial population beliefs.

**Proof of Lemma 4.** Iteratively applying the revision equation to an initial belief profile \( \tilde{\mu}(0) \),

\[
\tilde{\mu}(n) = \mu + \eta (\eta(I - Y))\tilde{\mu}(n-1)
= \mu + \eta (\eta(I - Y))(\mu + [\eta(I - Y)]\tilde{\mu}(n-2))
= \mu + \eta (\eta(I - Y)) \left( \sum_{k=0}^{n-2} [\eta(I - Y)]^k \mu + [\eta(I - Y)]^{n-1}\tilde{\mu}(0) \right)
= \mu + \eta (\eta(I - Y)) \left( [\eta(I - Y)](\mu + [\eta(I - Y)] - I)^{-1} \left( [\eta(I - Y)]^{n-1} - I \right) \mu + [\eta(I - Y)]^{n-1}([\eta(I - Y)] - I)\tilde{\mu}(0) \right)
= \mu + \eta (\eta(I - Y)) \left( [\eta(I - Y)](\mu + [\eta(I - Y)] - I)^{-1} \left( [\eta(I - Y)]^{n-1} \left( \mu + ([\eta(I - Y)] - I)\tilde{\mu}(0) \right) - [\eta(I - Y)]\mu \right) \right)
= \mu + \eta (\eta(I - Y) - I)^{-1} \left[ [\eta(I - Y)]^{n-1} \left( \delta + (\eta(I - Y) - I)\tilde{\delta}(0) \right) \right].
\]

In the case of common benchmark weights, the fact that \( Y^2 = Y \) implies \( (I - Y)^k = I - Y \) for all \( k \geq 1 \). Returning to the third line of the derivation above,

\[
\tilde{\mu}(n) = \mu + \eta \left( 1 - \eta^{n-1} \right) \left( 1 - \eta^{-1} \right) (I - Y)\mu + \eta^{n-1}(I - Y)\tilde{\mu}(0)
= \mu + \frac{\eta}{1 - \eta} \left[ (1 - \eta^{n-1})\delta + \eta^{n-1}(1 - \eta)\tilde{\delta}(0) \right]
= \mu + \frac{\eta}{\eta - 1} \left[ \eta^{n-1} \left( \delta + (\eta - 1)\tilde{\delta}(0) \right) - \delta \right].
\]
Proof of Lemma 1. The equilibrium characterizing Equation (1) in matrix form is \( \tilde{\mu} = \mu + \eta(\mathcal{I} - \mathcal{Y})\tilde{\mu} \), so whenever \( \mathcal{I} - \eta(\mathcal{I} - \mathcal{Y}) \) is invertible, it must follow that
\[
\tilde{\mu} = (\mathcal{I} - \eta(\mathcal{I} - \mathcal{Y}))^{-1}\mu.
\]
Invertibility of \( \mathcal{I} - \eta(\mathcal{I} - \mathcal{Y}) \) fails only when the matrix admits 0 as an eigenvalue, which is equivalent to \( \mathcal{Y} \) having an eigenvalue of \( 1 - 1/\eta \). Since \( \mathcal{Y} \) can have at most \( |\mathcal{G}| \) distinct eigenvalues, there are at most \( |\mathcal{G}| \) values of \( \eta \) such that the model lacks a unique equilibrium. Furthermore, since \( \mathcal{Y} \) is row-stochastic, all of its eigenvalues have modulus less than or equal to 1. As \( 1 - 1/\eta > 1 \) for \( \eta < 0 \) and \( 1 - 1/\eta < -1 \) for \( \eta < 1/2 \), any populations without equilibrium existence must have \( \eta \geq 1/2 \).

To show the existence of a cutoff value of \( \eta \) that determines stability, fix a population network, and suppose the equilibrium is stable for some value of \( \eta \). By Lemma 4, it follows that
\[
\tilde{\mu}^{(n)} = \eta^n(\mathcal{I} - \mathcal{Y})^n\tilde{\mu}^{(0)} + \sum_{k=0}^{n-1} \eta^k(\mathcal{I} - \mathcal{Y})^k\mu \longrightarrow (\mathcal{I} - \eta(\mathcal{I} - \mathcal{Y}))^{-1}\mu
\]
for all \( \tilde{\mu}^{(0)} \). In particular this holds for \( \tilde{\mu}^{(0)} = 0 \), so
\[
\lim_{n \to \infty} \tilde{\mu}^{(n)} = \lim_{n \to \infty} \sum_{k=0}^{n-1} \eta^k(\mathcal{I} - \mathcal{Y})^k\mu \implies \lim_{n \to \infty} \eta^n(\mathcal{I} - \mathcal{Y})^n\tilde{\mu}^{(0)} = 0.
\]
Now suppose \( |\hat{\eta}| < |\eta| \). As
\[
|\hat{\eta}^n(\mathcal{I} - \mathcal{Y})^n| < |\eta^n(\mathcal{I} - \mathcal{Y})^n|,
\]
it follows that \( \hat{\eta}^n(\mathcal{I} - \mathcal{Y})^n \longrightarrow 0 \). For all \( g \in \mathcal{G} \), and as a function of \( \eta \),
\[
\sum_{k=0}^{n-1} \hat{\eta}^k[(\mathcal{I} - \mathcal{Y})^k\mu]_g
\]
is a power series centered around 0; on account of the series converging for \( \hat{\eta} = \eta \), it follows \( \hat{\eta} \) is in the radius of convergence. This shows \( \tilde{\mu}^{(n)}_{\hat{\eta}} \) is convergent. Now define
\[
\overline{\eta} \equiv \sup \left\{ \eta : \tilde{\mu}^{(n)}_{\eta} \text{ convergent } \forall \tilde{\mu}^{(0)} \right\}.
\]
If \( |\eta| > \overline{\eta} \), then by the definition of \( \overline{\eta} \) the sequence \( \tilde{\mu}^{(n)}_{\eta} \) is not convergent for some \( \tilde{\mu}^{(0)} \) and therefore the equilibrium is not stable. If \( |\eta| < \overline{\eta} \), then by the definition of \( \overline{\eta} \) there exists \( \hat{\eta} \) such that \( |\eta| < |\hat{\eta}| \leq \overline{\eta} \) and \( \tilde{\mu}^{(n)}_{\hat{\eta}} \) is convergent for all \( \tilde{\mu}^{(0)} \). By the first part of the proof, it follows from
that the equilibrium is stable for $\eta$.

To establish a general bound on $\eta$, note that by the Contraction Mapping Theorem, the attributive equilibrium is stable whenever the revision equation, here denoted $f$, is a contraction. Using the metric given by the supremum norm, it suffices to show there exists $K < 1$ such that

$$
\| f(\tilde{\mu}^{(1)}) - f(\tilde{\mu}^{(2)}) \|_{\infty} = \| \eta(I - \mathcal{Y})(\tilde{\mu}^{(1)} - \tilde{\mu}^{(2)}) \|_{\infty} < K \| \tilde{\mu}^{(1)} - \tilde{\mu}^{(2)} \|_{\infty}
$$

for all $\tilde{\mu}^{(1)}, \tilde{\mu}^{(2)}$. To that end, note that for any $v \in \mathbb{R}^N$,

$$
\| (I - \mathcal{Y})v \|_{\infty} = \max_{g \in \mathcal{G}} \left| (1 - y_{gg})v_g - \sum_{h \neq g} y_{gh}v_h \right|
$$

$$
\leq (1 - y_{\text{min}}^{\text{min}}) \cdot \max_{g \in \mathcal{G}} \left| v_g - \frac{1}{\sum_{h \neq g} y_{gh}} \sum_{h \neq g} y_{gh}v_h \right|
$$

$$
\leq (1 - y_{\text{min}}^{\text{min}}) \cdot 2\| v \|_{\infty}.
$$

Thus, $f$ is a contraction and the equilibrium is stable whenever $|\eta| \cdot 2(1 - y_{\text{min}}^{\text{min}}) < 1$. In the case of common benchmark weights, let $\{\tilde{\mu}^{(n)}\}_{n \in \mathbb{N}}$ be any attributive belief sequence. As, for all $n \geq 1$,

$$
\sum_{g \in \mathcal{G}} y_{*g}\tilde{\mu}^{(n)}_{*g} = \sum_{g \in \mathcal{G}} y_{*g}\mu_{*g},
$$

so, for $n \geq 2$,

$$
\tilde{\mu}^{(n)}_{*g} - \mu_{*g} = \left[ \mu_{*g} + \eta \left( \tilde{\mu}^{(n-1)}_{*g} - \sum_{h \in \mathcal{G}} y_{*h}\tilde{\mu}^{(n-1)}_{*h} \right) \right] - \left[ \mu_{*g} + \frac{\eta}{1 - \eta} \left( \mu_{*g} - \sum_{h \in \mathcal{G}} y_{*h}\mu_{*h} \right) \right]
$$

$$
= \eta \left( \tilde{\mu}^{(n-1)}_{*g} - \frac{1}{1 - \eta} \mu_{*g} - \sum_{h \in \mathcal{G}} y_{*h} \left( \tilde{\mu}^{(n-1)}_{*h} - \frac{1}{1 - \eta} \mu_{*h} \right) \right)
$$

$$
= \eta \left( \tilde{\mu}^{(n-1)}_{*g} - \tilde{\mu} - \sum_{h \in \mathcal{G}} y_{*h} \left( \tilde{\mu}^{(n-1)}_{*h} - \mu_{*h} \right) \right)
$$

$$
= \eta(\tilde{\mu}^{(n-1)}_{*g} - \tilde{\mu}_{*g}).
$$

From this expression it is clear that $\tilde{\mu}^{(n)} \longrightarrow \tilde{\mu}$ for all $\tilde{\mu}^{(0)}$ if and only if $|\eta| < 1$. 

\[\Box\]
Proof of Theorem 1.

(a-i,ii) By Lemma 3, the equilibrium population misestimation of $g$ as a function of $\eta$ is

$$-rac{\eta^2}{1-\eta} \sum_{k=1}^{\infty} \left( \frac{\eta}{1-\eta} \right)^k \left( \mu_g - \sum_{h \in G} y_{gh} \mu_h \right).$$

If group $g$ is medial, then all of the difference terms are zero and the function is identically zero, proving claim (ii). Otherwise, let

$$k = \min_{k \geq 1} \left\{ k : \mu_g \neq \sum_{h \in G} y_{gh} \mu_h \right\},$$

and note that the population misestimation is zero if and only if $\rho(\varphi) = 0$ where $\varphi = \eta/(1-\eta)$ and

$$\rho(\varphi) = \sum_{k=1}^{\infty} \varphi^k \left( \mu_g - \sum_{h \in G} y_{gh} \mu_h \right).$$

As an analytic function, $\rho$ is either identically zero or admits only isolated zeros. Let $k$ be the smallest value of $k$ such $\mu_g \neq \sum_{h \in G} y_{gh} \mu_h$; then the $k$'th derivative at $\varphi = 0$ is

$$\rho^{(k)}(0) = k! \left( \mu_g - \sum_{h \in G} y_{gh} \mu_h \right) \neq 0,$$

so $\rho$ is not identically zero. Hence the set of $\varphi$, and equivalently $\eta$, such that population misestimation is zero has Lebesgue measure zero, proving claim (i).

(a-iii) Suppose $\eta > 0$ and denote $\tilde{\mu}_{\max} = \max_{g \in G} \tilde{\mu}_g$ and $\tilde{\mu}_{\min} = \min_{g \in G} \tilde{\mu}_g$. It must be that $\tilde{\mu}_{\max} > \tilde{\mu}_{\min}$; otherwise

$$\begin{cases} \tilde{\mu}_g \leq \mu_g + \eta (\tilde{\mu}_{\max} - \tilde{\mu}_{\min}) = \mu_g & \text{if } \tilde{\mu}_g \leq \mu_g + \eta (\tilde{\mu}_{\max} - \tilde{\mu}_{\min}) = \mu_g, \\ \tilde{\mu}_g \geq \mu_g + \eta (\tilde{\mu}_{\min} - \tilde{\mu}_{\max}) = \mu_g, \end{cases}$$

which implies $\tilde{\mu}_g = \mu_g$ for all $g \in G$, which would then imply

$$\tilde{\mu}_{\max} - \tilde{\mu}_{\min} = \max_{g \in G} \mu_g - \min_{g \in G} \mu_g > 0,$$

a contradiction. Thus

$$\begin{cases} \tilde{\mu}_{\max} > \mu_{\max} + \eta (\tilde{\mu}_{\max} - \tilde{\mu}_{\max}) = \mu_{\max} \\ \tilde{\mu}_{\min} < \mu_{\min} + \eta (\tilde{\mu}_{\min} - \tilde{\mu}_{\min}) = \mu_{\min}. \end{cases}$$
An analogous proof holds when \( \eta < 0 \).

(b) Note that
\[
\tilde{\mu} - \mu = \sum_{k=0}^{\infty} [\eta(I - \mathcal{Y})]^k (\eta(I - \mathcal{Y}) \mu).
\]

Combining the bound in Lemma 1 with the relation
\[
\| (I - \mathcal{Y}) \mu \|_\infty \leq (1 - y_{\ast*}) \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|,
\]

it follows
\[
\| \tilde{\mu} - \mu \|_\infty \leq \sum_{k=0}^{\infty} \left[ |\eta(I - \mathcal{Y})| \cdot (1 - y_{\ast*}) \cdot |\eta| \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}| \right]^k
\]
\[
\leq \sum_{k=0}^{\infty} (2|\eta|(1 - y_{\ast*}))^k \cdot |\eta| \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|
\]
\[
= \frac{|\eta|(1 - y_{\ast*})}{1 - 2|\eta|(1 - y_{\ast*})} \cdot \max_{h_1, h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|.
\]

The bound on absolute errors in the case of common benchmark weights follows by inspection from the formula for equilibrium beliefs.

Proof of Theorem 2. (a-i) When the network has common benchmark weights,
\[
(\tilde{\mu}_g - \tilde{\mu}_h) = (1 + \eta)(\mu_g - \mu_h),
\]

and as stability requires \( \eta > -1 \), it follows that relative beliefs are correct.

(a-ii) By definition, relative beliefs must be correct when
\[
\max_{g \in \mathcal{G}} |\tilde{\mu}_g - \mu_g| < \frac{1}{2} \min_{h_1 \neq h_2 \in \mathcal{G}} |\mu_{h_1} - \mu_{h_2}|,
\]

i.e. absolute beliefs are so close to their objective counterparts that their potential regions are non-overlapping. Applying the bound in Theorem 1 yields condition (a-ii).

(a-iii) Suppose the condition holds, and let \( \tilde{\mu}^{(0)} \) be any initial belief profile such that \( \tilde{\mu}^{(0)}_{g_1} \geq \tilde{\mu}^{(0)}_{g_2} \) for all \( g_1, g_2 \in \mathcal{G} \) such that \( \mu_{g_1} \geq \mu_{g_2} \). Then by the revision equation,
\[
\tilde{\mu}^{(1)}_{g_1} - \tilde{\mu}^{(1)}_{g_2} = (\mu_{g_1} - \mu_{g_2}) + \eta(\tilde{\mu}^{(0)}_{g_1} - \tilde{\mu}^{(0)}_{g_2}) - \eta \left( \sum_{h \in \mathcal{G}} (y_{g_1 h} - y_{g_2 h}) \tilde{\mu}^{(0)}_h \right).
\]
The first two terms are clearly positive. Concerning the third, note that since the ordering of $\mu_h^{(0)}$ follows the order of $\mu_h$, condition (a-iii) guarantees that $\sum_{h \in \mathcal{G}} (y_{gh} - y_{gh}) \mu_h^{(0)} \leq 0$, so $\hat{\mu}_g(1) \geq \hat{\mu}_g(1)$. Iteratively applying the revision equation, $\hat{\mu}_g^{(n)} \rightarrow \hat{\mu}_g$ for all $g \in \mathcal{G}$ and $\hat{\mu}_g(1) \geq \hat{\mu}_g(1)$ for all $n \geq 1$. Thus $\hat{\mu}_g \geq \hat{\mu}_g$.

(b) By Lemma 4, $\hat{\mu}_g^{(n)}$ diverges in the direction of $\delta_g + (\eta - 1)\delta_g^{(0)}$, so it suffices to find values of $\hat{\mu}_g^{(0)}$ such that

$$\delta_g^{(0)} > -\frac{1}{\eta - 1} \delta_g^{(0)}$$

if and only if $s(g) = +\infty$. Set $m_g = 1_{(s(g) = +\infty)} - 1_{(s(g) = -\infty)}$, so $m_g - \sum_{h \in \mathcal{G}} y_{gh} m_h > 0$ if and only if $s(g) = +\infty$. Then $\hat{\mu}_g^{(0)} = K \cdot m$ satisfies the above inequality for all $g$ when $K > 0$ is sufficiently large.

---

**Proof of Theorem 3.** (a) For any $i, j, g, h \in \mathcal{G}$,

$$\langle \hat{\mu}_g - \hat{\mu}_h \rangle \langle \hat{\mu}_g - \hat{\mu}_h \rangle = [\langle \mu_g - \mu_h \rangle + \eta (\hat{\mu}_g - \hat{\mu}_h)]^2 \geq 0.$$  

(b) Similarly,

$$\langle \hat{\mu}_g - \hat{\mu}_h \rangle \langle \hat{\mu}_g - \hat{\mu}_h \rangle = \eta^2 \left( \sum_{h \in \mathcal{G}} (z_{ih} - z_{jh}) \hat{\mu}_h \right)^2 \geq 0.$$  

If benchmark weights are common, 

$$\hat{\mu}_g = \mu_g + \eta \left( \mu_g - \sum_{h \in \mathcal{G}} z_{ih} \hat{\mu}_h \right)$$

$$= \mu_g + \eta \left( \mu_g + \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{j \in \mathcal{G}} y_{sj} \hat{\mu}_j \right) - \sum_{h \in \mathcal{G}} z_{ih} \left( \mu_h + \frac{\eta}{1 - \eta} \left( \mu_h - \sum_{j \in \mathcal{G}} y_{sj} \hat{\mu}_j \right) \right) \right)$$

$$= \mu_g + \frac{\eta (1 + \eta)}{1 - \eta} \left( \mu_g - \sum_{h \in \mathcal{G}} z_{ih} \mu_h \right).$$

For any $\bar{z} \leq \min \{ z_{ii}, z_{jj} \}$, let $\tilde{z}_{ii} = (z_{ii} - \bar{z})/(1 - \bar{z})$, $\tilde{z}_{jj} = (z_{jj} - \bar{z})/(1 - \bar{z})$, and $\tilde{z}_{ih} = z_{ih}/(1 - \bar{z})$ for all $h \neq i$ and $\tilde{z}_{jh} = z_{jh}/(1 - \bar{z})$ for all $h \neq j$. Note that $\sum_{h \in \mathcal{G}} \tilde{z}_{ih} = \sum_{h \in \mathcal{G}} \tilde{z}_{jh} = 1$. Then

$$\hat{\mu}_g - \hat{\mu}_g = -\eta \frac{(1 + \eta)}{1 - \eta} \left( \sum_{h \in \mathcal{G}} z_{ih} \mu_h - \sum_{h \in \mathcal{G}} z_{jh} \mu_h \right)$$

$$= -\eta \frac{(1 + \eta)}{1 - \eta} \left( \bar{z} (\mu_i - \mu_j) + (1 - \bar{z}) \left( \sum_{h \in \mathcal{G}} \tilde{z}_{ih} \mu_h - \sum_{h \in \mathcal{G}} \tilde{z}_{jh} \mu_h \right) \right).$$

51
so if
\[
\bar{z} > \frac{\max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}|}{\min_{h_1 \neq h_2 \in G} |\mu_{h_1} - \mu_{h_2}| + \max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}|}
\]
it follows
\[
\tilde{\mu}_i^g - \tilde{\mu}_j^g \geq 0 \iff \eta(\mu_i - \mu_j) \leq 0 \iff \eta(\tilde{\mu}_i - \tilde{\mu}_j) \leq 0.
\]
(c) By the revision equation,
\[
\tilde{\mu}_i^g - \mu_g = \eta \left( \tilde{\mu}_g - \sum_{h \in G} z_{ih}\tilde{\mu}_h \right)
\]
\[
= \eta(1 - z_{ig}) \left( \tilde{\mu}_g - \frac{1}{\sum_{h \neq g} z_{ih}} \sum_{h \neq g} z_{ih}\tilde{\mu}_h \right)
\]
\[
= \eta(1 - z_{ig}) \left( \tilde{\mu}_g - \frac{1}{\sum_{h \neq g} z_{ih}} \sum_{h \neq g} z_{ih}\tilde{\mu}_h + (\tilde{\mu}_g - \mu_g) - \frac{1}{\sum_{h \neq g} z_{ih}} \sum_{h \neq g} z_{ih}(\tilde{\mu}_h - \mu_h) \right).
\]
Applying the bound from Theorem 1, it follows
\[
|\tilde{\mu}_i^g - \mu_g| \leq |\eta|(1 - z_{ig}) \cdot \left( 1 + 2 \frac{|\eta|(1 - y_{\text{min}})}{1 - |\eta|(1 - y_{\text{min}})} \right) \cdot \max_{h_1, h_2 \in G} |\mu_{h_1} - \mu_{h_2}|,
\]
which simplifies to the expression in the theorem statement.

Proof of Lemma 2. Recall that the senior’s optimal action is
\[
a^*_J(\tilde{\mu}_J^S) = \frac{1 + \varphi_J}{(1 - \varphi_J)\varphi_S} \beta_S + \frac{\varphi_S}{1 - \varphi_J\varphi_S} (\beta_J + \tilde{\mu}_J^S),
\]
and the junior’s optimal action is
\[
a^*_S(a_S, \theta_J) = \beta_J + \theta_J + \varphi_J a_J.
\]
Therefore the junior’s payoff is
\[
u_J(\tilde{\mu}_J^S; \theta_J) = \frac{1}{2} \beta_J^2 + \beta_J (1 + \varphi_J) a_S(\tilde{\mu}_J^S) + \beta_J \theta_J
\]
\[
= u_J(\theta_J; \theta_J) + (1 + \varphi_J) \frac{\varphi_S}{1 - \varphi_S\varphi_J} \beta_J (\tilde{\mu}_J^S - \theta_J),
\]
\[
52
\]
and the senior’s payoff is

\[
    u_S(\tilde{\mu}_J; \theta_J) = \frac{1}{2} \left( \frac{\beta_S(1 + \varphi_J)}{1 - \varphi_J} \right)^2 + \beta_S \frac{1 + \varphi_S}{1 - \varphi_J \varphi_S} (\beta_J + \theta_J) - \frac{1}{2} \varphi_S^2 (\tilde{\mu}_J - \theta_J)^2 \\
= u_S(\theta_J; \theta_J) - \frac{1}{2} \varphi_S^2 (\tilde{\mu}_J - \theta_J)^2.
\]

**Proof of Fact 1.** By Lemma 2, the expected utility of a junior of group \( g \) is

\[
    v_g(\tilde{\mu}) = \sum_{i \in G} w_{gi} (u_J(\mu_g; \mu_g) + W_J(\tilde{\mu}_g - \mu_g)) \\
= v_g(\mu) + W_J(\tilde{\mu}_g - \mu_g).
\]

Similarly, the expected utility of a senior of group \( g \) is

\[
    v^g(\tilde{\mu}) = \sum_{h \in G} w_{gh} E \left[ u_S(\tilde{\mu}^g_h; \theta_J) | J \in h \right] \\
= \sum_{h \in G} w_{gh} E \left[ u_S(\theta_J; \theta_J) - \frac{1}{2} W_S(\tilde{\mu}^g_h - \theta_J)^2 | J \in h \right] \\
= \sum_{h \in G} w_{gh} E \left[ u_S(\theta_J; \theta_J) - \frac{1}{2} W_S [ (\tilde{\mu}^g_h - \mu_h)^2 + (\mu_h - \theta_J)^2 + 2(\mu_h - \tilde{\mu}^g_h)(\theta_J - \mu_h) ] | J \in h \right] \\
= \sum_{h \in G} w_{gh} E \left[ u_S(\theta_J; \theta_J) - \frac{1}{2} W_S(\mu_h - \theta_J)^2 | J \in h \right] - \frac{1}{2} W_S(\tilde{\mu}^g_h - \mu_h)^2 \\
= v^g(\mu) - \frac{1}{2} W_S \sum_{h \in G} w_{gh} (\tilde{\mu}^g_h - \mu_h)^2.
\]
Proof of Fact 2. In the case of common benchmark weights,

\[ \tilde{\mu}_g - \mu_g = \frac{\eta}{1 - \eta} \left( \mu_g - \sum_{h \in G} y_{gh} \mu_h \right) \]

and

\[ \sum_{h \in G} w_{gh} (\tilde{\mu}_h^g - \mu_h)^2 = \left( \frac{\eta}{1 - \eta} \right)^2 \sum_{h \in G} w_{gh} \left( \mu_h - \sum_{f \in G} z_{gf} \mu_f \right)^2 \]

\[ = \left( \frac{\eta}{1 - \eta} \right)^2 \left( \sum_{h \in G} w_{gh} \mu_h^2 - 2 \sum_{h \in G} w_{gh} \mu_h \sum_{f \in G} z_{gf} \mu_f + \left( \sum_{f \in G} z_{gf} \mu_f \right)^2 \right) \]

\[ = \left( \frac{\eta}{1 - \eta} \right)^2 \left( \sum_{h \in G} w_{gh} \mu_h^2 - \left( \sum_{h \in G} w_{gh} \mu_h \right)^2 \right) \]

\[ + \left( \sum_{h \in G} w_{gh} \mu_h \right)^2 - 2 \sum_{h \in G} w_{gh} \mu_h \sum_{f \in G} z_{gf} \mu_f + \left( \sum_{f \in G} z_{gf} \mu_f \right)^2 \]

\[ = \left( \frac{\eta}{1 - \eta} \right)^2 \left( \sum_{h \in G} w_{gh} \mu_h^2 - \sum_{h \in G} z_{gh} \mu_h \right)^2 + \sum_{h \in G} w_{gh} \mu_h^2 - \left( \sum_{h \in G} w_{gh} \mu_h \right)^2 , \]

which when substituted into Equation 3 yield the expression in the fact statement.

Proof of Theorem 4. In the case of transfer policies, any transfer values \( \{D_g\}_{g \in G} \) which de-bias beliefs must satisfy

\[ \mu_g = \mu_g + \eta \left( \mu_g + D_g - \sum_{h \in G} y_{gh} (\mu_h + D_h) \right) , \]

implying \( \mu_g + D_g = \sum_{h \in G} y_{gh} (\mu_h + D_h) = \bar{\mu} \), for all \( g \in G \). As a transfer policy it must be that \( \sum_{g \in G} p_g D_g = 0 \). Thus \( \bar{\mu} = \sum_{g \in G} p_g \mu_g \), and

\[ D^*_g = - \left( \mu_g - \sum_{h \in G} p_h \mu_h \right) . \]

To show that no diversity policy can de-bias beliefs, suppose for the sake of contradiction that one does, and denote \( y_{*h} \) to be the benchmark weights implied by either the target weights \( w_{it*} \) or \( z_{i*} \). It must be that

\[ \mu_g = \alpha \sum_{h \in G} y_{*h} \mu_h + (1 - \alpha) \sum_{h \in G} \left( \mu_h \right)_{G} , \]

54
where \(c_g\) denotes the benchmark comparison (in true value terms) for group \(g\). Let \(\mu_{\text{min}}, \mu_{\text{max}}\) refer to two groups with the lowest and highest true mean traits, respectively. We know \(\mu_{\text{max}} > \mu_{\text{min}},\) as otherwise, by Theorem 1, equilibrium beliefs would be correct. Since the above equation is satisfied simultaneously for all groups, the value of \(\alpha\) must be

\[
\alpha = \frac{(c_{\text{max}} - c_{\text{min}}) - (\mu_{\text{max}} - \mu_{\text{min}})}{c_{\text{max}} - c_{\text{min}}}.
\]

For this value to lie in the range \([0, 1]\), it first must be that \(c_{\text{max}} - c_{\text{min}} > 0\), as otherwise \(\alpha \leq 1\) would demand \(\mu_{\text{max}} - \mu_{\text{min}} \leq 0\), a contradiction. Second, it therefore follows that

\[
\mu_{\text{max}} - \mu_{\text{min}} \leq c_{\text{max}} - c_{\text{min}} \leq \mu_{\text{max}} - \mu_{\text{min}},
\]

implying \(\alpha = 0\). This proves the theorem statement. In the case of the equal treatment policy, it is straightforward to verify that correct beliefs satisfy the equilibrium conditions under the modified belief-revision equation when true mean traits satisfy the condition in the theorem statement. To show sufficiency, note that the policy de-biases beliefs only if

\[
\tilde{\mu}_g - \sum_{h \in G} z_{ih} \tilde{\mu}_h = 0
\]

for all \(g \in G\), which holds only if \(\tilde{\mu}_{g_1} = \tilde{\mu}_{g_2}\) for all \(g_1, g_2 \in G\), which holds only if

\[
\sum_{i \in G} w_{g_1 i} \sum_{h \in G} w_{ih} \tilde{\mu}_h^i = \sum_{i \in G} w_{g_2 i} \sum_{h \in G} w_{ih} \tilde{\mu}_h^i
\]

for all \(g_1, g_2 \in G\). Now, if the policy de-biases beliefs, \(\tilde{\mu}_h^i = \mu_h\) for all \(i, h \in G\) and the above equality simplifies to the condition in the theorem statement.