American Economic Review 2023, 113(1): 210–252
https://doi.org/10.1257/aer.20211729

Signaling and Discrimination in Collaborative Projects[†]

By Paula Onuchic ⓡ Debraj Ray*

We study collaborative work in pairs when potential collaborators are motivated by the reputational implications of (*joint or solo*) *projects. In equilibrium*, *individual collaboration strategies both influence and are influenced by the public assignment of credit for joint work across the two partners. We investigate the fragility of collaboration to small biases in the public's credit assignment. When collaborators are symmetric*, *symmetric equilibria are often fragile*, *and in nonfragile equilibria individuals receive asymmetric collaborative credit based on payoff-irrelevant "identities." We study payoff distributions across identities within asymmetric equilibria*, *and compare aggregate welfare across symmetric and asymmetric equilibria.*
(*JEL* A11, D82, I23)

Research is increasingly conducted in teams. In economics, coauthored papers make up over 70 percent of all published research, up from 20 percent in 1960.[1] The prominence of teamwork extends to other fields, academic or otherwise. For instance, large technology companies are known to foster collaborative group environments. Obviously, such collaboration can be beneficial as it allows workers to combine their skills. However, by its very nature, a team subsumes the person, thereby hindering an individual's ability to build reputation. This gives rise to a fundamental tension in collaborative activity, one that pits the direct gains from teamwork against the difficulty of revealing personal ability to the lens of public evaluation.

We build a theory that incorporates both these aspects. At its heart are *public perceptions* of individual ability implied by collaborative work, based on conjectures about the circumstances that led to the observed collaboration. In turn, collaboration decisions are endogenously determined by those perceptions. Our theory incorporates this fundamental circularity in a model of collaboration with reputational concerns.

In our setting, prematched pairs of individuals choose whether or not to collaborate. An individual has one of two types, good or bad. When a pair meet, each

* Onuchic: University of Oxford (email: paula.onuchic@economics.ox.ac.uk); Ray: New York University and University of Warwick (email: debraj.ray@nyu.edu). Jeffrey Ely was the coeditor for this article. Ray acknowledges funding under NSF grant SES-1851758. We thank three anonymous referees, as well as Daghan Carlos Akkar, Axel Anderson, Piotr Dworczak, Yingni Guo, Li Hao, Sam Kapon, Navin Kartik, Erik Madsen, Meg Meyer, Anja Prummer, Mauricio Ribeiro, Ran Spiegler, Ludvig Sinander, and Joshua Weiss for useful comments.

† Go to https://doi.org/10.1257/aer.20211729 to visit the article page for additional materials and author disclosure statements.

[1] See Jones (2021), who also reports that in 2010, a team was three times more likely to produce a highly cited paper than a solo author, an advantage that has also grown steadily with time.

agent draws an idea from a type-dependent distribution, with good types drawing stochastically better ideas. Both persons see both ideas, and choose whether to work together or separately. In making this decision, each person seeks to maximize a combination of the *direct value* and the *reputational value* of the project.

The direct value of a project depends just on the agent's idea if the work is completed alone, and on both agents' ideas if there is collaboration. Reputational value is generated by an observer (the "public"), who observes the project outcome and updates its beliefs about the agents' types. But in the event of collaboration, the public sees the joint outcome, and not each individual contribution. To interpret what a joint outcome implies about each agent's type, the observer uses a conjecture—to be justified in equilibrium—about which pairs of ideas might have led agents to collaborate. That conjecture is then coupled with Bayesian updating to assign credit across the two partners. Such conjectures and updates affect reputational value, and therefore the agents' collaboration decisions.

Proposition 1 characterizes equilibrium collaboration decisions. These resolve the trade-off between *direct value*, which always improves with collaboration, and *reputational value*, which is garbled in joint work. The observed project outcome and the conjectured collaboration strategies pin down the reputational payoff from collaboration, while the reputational payoff from working alone rises with the value of the agent's own idea. The proposition establishes the existence of a nonempty collaboration set which reflects this trade-off: in any such equilibrium, each agent benefits from collaboration if and only if their contribution to a project is below some endogenously determined threshold—or equivalently, if their partner's contribution is above a related threshold.

Our model anchors a potentially rich theory of collaboration that links collaborative decisions to public priors on ability. For instance, we might ask if those with established reputations are more willing to collaborate than their less-tested but possibly more ambitious counterparts, or study assortative collaborations across reputations, or explore the dynamics of collaboration as reputation evolves. In this paper, we choose to pursue a different line of inquiry, focusing on the observation that equilibrium collaboration patterns may also depend on individuals' payoff-*irrelevant* characteristics, such as gender, nationality, age, or race.

To build this theme, imagine that the two individuals are symmetric—the public has the same prior about their types. However, each person has a distinct payoff-irrelevant *identity*, one that is salient in the public eye. Now suppose that the public is "biased" and allocates reputational value in favor of one identity. That is, it thinks the favored identity contributes better ideas to collaboration, thereby assigning higher credit to that identity. Then individual collaborative strategies will react to this bias. Incentivized by the credit allocation, the favored identity is then relatively more willing to collaborate. Specifically, a favored person shares better ideas than a disfavored person would be willing to do, were those ideas her own. So, at least to some degree, this reaction actually confirms the initial bias. Given the described collaboration strategies, the public should indeed "rationally" allocate reputational value the way they do.

These echo effects of biases in public perception and optimal collaboration strategies can lead to multiple equilibria, some of them asymmetric even if the underlying collaborating agents are symmetric along all payoff-relevant dimensions. We

study these discriminatory forces, and the fragility of nondiscriminatory outcomes. We define equilibrium fragility in Section III. Informally, an equilibrium is fragile when small biases in public perceptions are amplified by the strategic collaborative responses of individuals. Propositions 2 and 3 set the background for our analysis: the former states that *nonfragile* equilibria always exist, and the latter states that symmetric equilibria always exist. Do these two sets of equilibria overlap? Proposition 4 and Corollary 1 are central results that run against that presumption. They establish conditions under which the symmetric equilibrium is indeed fragile. Put another way, nonfragile equilibria involve asymmetric treatment: they feature discrimination.

The main condition that generates these results is that agents sufficiently value the reputational aspect of their output, relative to their direct production value. This means that discrimination is a real possibility when *career concerns* are important. This is particularly salient in individuals' early careers, when their reputations are not yet established; and professions such as academia, where most value in research does not monetarily accrue to the researcher. Rather, the researcher is rewarded for establishing a reputation for their underlying quality—say, by receiving promotions and prizes that are explicitly conditioned on the perceived creativity and relevance of their past work, and that might influence their (possibly pecuniary) payoffs in the future.

These results are more than mere theoretical abstractions. For instance, Sarsons et al. (2021) use data on academic economists to argue that the public responds to joint work between women and men by attributing more credit to men. In related research, Ductor, Goyal, and Prummer (2021) document homophily in coauthorship networks, as well as gender disparities in collaboration patterns in economic research. From the perspective of our model, these empirical observations are two sides of the same coin.[2]

In Section V, we consider the payoff implications of such discrimination. Specifically, we compare payoffs across favored and disfavored identities within the same asymmetric equilibrium in Sections VA–VC; and compare aggregate welfare across symmetric and asymmetric equilibria (when they exist) in Section VD. To begin with, in an asymmetric equilibrium across symmetric partners, the "favored identity" is perceived as contributing better ideas to a collaboration—thereby receiving higher collaborative credit. But Proposition 5 argues that in such equilibria, the expected *direct* payoff to a disfavored person is higher that of a favored person. This result is a consequence of our collaborative setting, in which agents directly transfer value to each other when they share ideas. The very fact that the favored identity contributes better ideas to collaborations implies a relative gain in direct payoff for the disfavored identity. A particularly stark corollary applies when reputational utility is linear in public beliefs. Then the expected reputational payoff is constant, as an implication of Bayes plausibility. Therefore *overall* expected payoffs move in tandem with direct payoffs, so that the disfavored person is better off in terms

[2] Ong et al. (2018) also document that the decision to form coauthorships responds to expected credit assignment. Specifically, they compare coauthorship behavior between authors with surname initials earlier in the alphabet, who receive more credit, and those with later initials. Lissoni, Montobbio, and Zirulia (2013) document a different pattern of discrimination in credit attribution: they study patent-publication pairs and show that women and young scholars who are credited in publications are more likely to be excluded from the generated patents.

of expected overall payoff, even though they receive a lower payoff conditional on collaboration happening (Proposition 6).

If reputational utility is not linear, then expected reputational payoffs could vary, and overall payoffs are not so clearly ranked across favored and disfavored agents. Still, our model makes sharp predictions about the *distribution* of reputational outcomes. Consider "target posteriors" that an agent might wish to attain: for instance, a reputational threshold for retention or promotion. Proposition 7 argues that in an asymmetric equilibrium, the disfavored identity is more likely to reach such a target if it is *extreme*: whether high or low. Conversely, the favored identity can more easily reach intermediate targets. In Section VC, we discuss how this result relates to a point recently made by Bohren, Hull and Imas (2022) regarding the measurement of discrimination.

In Section VD, we compare aggregate welfare across symmetric and asymmetric equilibria when both exist. We show numerically that aggregate welfare may be larger in either type of equilibrium, depending on model calibration. In Proposition 8 and its Corollary 3, we provide analytical conditions under which symmetric equilibria yield higher aggregate welfare than asymmetric equilibria. Finally, in Section VI, we discuss the (in)efficiency of collaborative equilibria, and study a simple policy based on random order that Pareto improves upon the equilibria of our model.

Related Literature.—We embed a theory of discrimination in a novel context, one of team formation with reputational concerns, and study the fragility of equal-treatment outcomes. To the best of our knowledge, ours is the first paper to propose a characterization of equilibrium fragility in the context of collaboration.

Our work intersects with the literature on discrimination, especially on equilibrium statistical discrimination in the tradition of Arrow (1973).[3] In that framework, an employer holds distinct beliefs about the quality of potential hires based on payoff-irrelevant identities. In turn, these differences in perceptions incentivize different identities to make unequal investments in human capital, confirming the employer's initial bias. (See, e.g., Coate and Loury 1993.) In the baseline version of that model, each identity plays its own equilibria with the employer; there is no *within*-equilibrium connection across identities.[4] In our setting, agents directly form productive teams, and the interaction across identities is central to the entire exercise. Here, discrimination is the unequal treatment of different identities *within a single equilibrium*. Unlike in the core model of statistical discrimination in labor markets, this interaction is key. No asymmetric equilibrium exists in our model if agents only work on their own, or with agents indistinguishable from themselves. The direct interaction is also central to the payoff results and testable empirical implications we discuss in Section V.

[3] More broadly and in addition to Arrow (1973) and the seminal contribution of Phelps (1972), the literature on statistical discrimination dates back to Myrdal (1944). Fang and Moro (2011) and Onuchic (2022) survey this literature. Recent contributions include Pęski and Szentes (2013); Bohren, Imas, and Rosenberg (2019); Bohren et al. (2021); and Bardhi, Guo, and Strulovici (2020).

[4] There are certainly extensions of that setting. For instance, Moro and Norman (2004) study statistical discrimination in general equilibrium. In their model, people of different identities are hired by the same firm, and in asymmetric equilibria, each identity specializes in vertically ranked tasks.

A second notable difference is that we refine our equilibrium set using a new fragility argument, based on small biases in the public's perception. The question of whether nondiscriminatory equilibria are robust is normally not invoked in models of statistical discrimination, though we note that Gu and Norman (2020) take a related approach. They study a search-theoretic model of the labor market, where workers sort into high-tech and low-tech sectors. They show (numerically) that the introduction of a payoff-irrelevant gender characteristic can render the symmetric equilibrium unstable, and generate gender-based sorting into the two occupations. Both the model and the forces that make for instability are entirely different from those we explore, but we mention this paper as an exception to the general approach taken in the literature.

Stability concepts are used in other settings with symmetric and asymmetric equilibria. Chaudhuri and Sethi (2008) and Bowles, Loury, and Sethi (2014) study the stability of segregation and social integration. In general-equilibrium models with imperfect capital markets, ex ante symmetric agents will make different occupational choices with implications for economic inequality (Mookherjee and Ray 2002, 2003).

A small literature considers unequal credit attribution in teams. Ray, Baland, and Dagnelie (2007); Ray ⓡ Robson (2018); and Ozerturk and Yildirim (2021) study team production with unequal credit to agents. In the latter two papers, the attribution of credit is endogenously based on estimates of individual contributions, which inefficiently affects individual effort decisions. But there are no reputational concerns, and credit attributed to each agent only determines their share in the physical outcome of the project. In our model, in contrast, reputational concerns occupy center stage.[5]

In an unpublished working paper,[6] Tumlinson (2012) also notes that unequal credit assignment across equal partners can persist even if credit is allocated "fairly." Though the discriminatory mechanism is similar, our analyses significantly differ. Tumlinson (2012) mainly studies a binary setting and illustrates that asymmetric collaborative equilibria may exist, while our central contribution lies in proposing a notion of fragility and characterizing environments in which symmetric equilibria are fragile, and discrimination thus "inevitable."

Our paper relates to Holmström (1982) and subsequent literature on incentive provision in teams. Winter (2004) connects team production and discrimination, arguing that unequal rewards may be unavoidable even among identical individuals. Chalioti (2016) studies career concerns in teams and the incentives of workers to support (or sabotage) the efforts of colleagues. Bar-Isaac (2007) considers the coevolution of worker and firm reputations, and effort incentives for junior and senior team members.

Our payoff function defined on reputation and direct project value allows for nonlinear returns to reputation. Often that functional form can be derived from a larger game. For instance, Anderson and Smith (2010) show how these payoff functions might emerge as endogenous value functions in a dynamic setting. However, in their

[5] The attribution of individual credit in groups has been explored in other contexts—see, for instance, Levy (2007) and Visser and Swank (2007) on decision-making in committees.

[6] This paper was brought to our attention after our final revision to this journal was submitted.

model, collaboration decisions play no role, and posterior updates are symmetric by assumption when partners are symmetric. In contrast, our main questions concern the collaboration choices of agents and the public's conjectures about collaborative patterns.[7]

I. Model

Two individuals have the opportunity to collaborate on a project. They bring ideas to the table, generated by a distribution that depends on individual ability, which is either 0 (bad) or 1 (good). There is a public prior that a person is good, shared also by her potential partner. What the individual herself knows about her ability will turn out to be irrelevant, so we presume nothing. Each agent sees both ideas, and chooses whether to work together or alone. If both prefer to work together, collaboration occurs. Otherwise, both work alone, with no plagiarism of ideas.[8] Each individual values both the project as well as her reputation, which is the updated *public* belief on her ability.

Each person's idea w is drawn from a distribution with strictly positive densities $g(w,0)$ and $g(w,1)$ for types 0 and 1, both with full support on $\mathbb{R}_+$. We assume that

$$(1) \qquad \text{the likelihood ratio } \frac{g(w,1)}{g(w,0)} \text{ is strictly increasing in } w,$$

and at a more technical level, that both densities have bounded derivatives on any compact set. With ideas revealed in initial discussion, agents decide whether to collaborate. Each person seeks to maximize a combination of the project's *direct value*, implied by the ideas, and its *reputational value*, implied by the public update on starting priors.[9]

A. *Direct and Reputational Payoffs*

Let $p \in (0,1)$ and $q \in (0,1)$ be the public priors on the pair. We will also use these letters as individual names, even when $p = q$. Suppose that p has idea x, and q has idea y. If p and q collaborate, then the joint project has direct value

$$z = f(x,y),$$

[7] Chade and Eeckhout (2020) study a different model of team formation in which teams compete against each other. In their model, agents' conjectures of the matching pattern affect their incentives to form matches in the first place. As in our model, the interplay between these conjectures and individual actions creates scope for multiple equilibria with distinct matching patterns.

[8] In our model, two individuals are "randomly matched" and choose whether to work together or separately. This is a good description of the collaboration environment of early career individuals, with limited networks of potential collaborators. It also fits situations in which individuals are assigned to teams and choose their contributions to team projects, as well as projects they conduct individually. Being assigned to a team still leaves some latitude to do this, by varying the mix of individually observable tasks, without necessarily leaving the confines of the "assigned" team environment.

[9] We assume away the possibility that agents choose to not work on any projects, but this is without loss of generality. Suppose instead that agents have the option to not work. Then, in an equilibrium where agents sometimes don't work on any project, this choice is associated with no direct value and a low signaling value. Using standard arguments, we can show that any such equilibrium would unravel.

where f is symmetric, strictly increasing and twice continuously differentiable with derivatives bounded above and below by positive numbers. Let $f(x,0) = x$ and $f(0,y) = y$ describe, respectively, the direct value of p and q's projects if they work alone.[10] In this latter event, public posteriors are found by applying ideas x and y to the functions

$$b_p(w) \equiv \frac{g(w,1)p}{g(w,p)} \quad \text{and} \quad b_q(w) \equiv \frac{g(w,1)q}{g(w,q)}, \tag{2}$$

respectively, where $g(w,r) \equiv rg(w,1) + (1-r)g(w,0)$ for $w \in \mathbb{R}_+$ and $r \in (0,1)$. By the likelihood ratio assumption (1), $b_p(w)$ and $b_q(w)$ are increasing.

If, otherwise, p and q combine their ideas into a joint project, the public posterior is calculated "in equilibrium." That is, if a collaboration happens, the outside observer sees the outcome $z = f(x,y)$, but not x and y separately. To infer these underlying ideas, the observer conjectures some *collaboration set*

$$C(z) \equiv \big\{(x,y) \,|\, f(x,y) = z \text{ and } p \text{ and } q \text{ choose to work together, given ideas } x \text{ and } y\big\},$$

which describes, for each joint outcome $z > 0$ and pair of priors, all combinations of x and y that yield z and lead to both agents agreeing to work together.

Such a set induces a probability distribution on combinations of x and y that could have led to the collaborative outcome z. Using this distribution, the public update averages equation (2) across every pair (x,y) in the conjectured collaboration set:

$$\beta_p(z) = E\big[b_p(x) \,|\, (x,y) \in C(z)\big] \quad \text{and} \quad \beta_q(z) = E\big[b_q(y) \,|\, (x,y) \in C(z)\big]. \tag{3}$$

In the Appendix, we describe in detail the distribution of ideas of each agent, given an observed joint outcome z and a conjectured collaboration pair $C(z)$. Specific properties of this distribution will be needed in some of the arguments.

B. *Overall Payoff*

Each agent values direct and reputational payoffs from projects. If a project has direct value d and yields Bayesian posterior b, the overall payoff is

$$\alpha d + u(b),$$

where $\alpha > 0$ is the weight on direct value, and u, assumed to be smooth with positive but bounded derivative, is defined on all individual reputations $b \in [0,1]$.

[10] These assumptions on the direct value production function f guarantee that, in terms solely of direct value, agents always wish to collaborate—they each receive z from the collaborative outcome, which is larger than x and y. This assumption is made mainly to present the reputational channel more starkly: here, reputational concerns are the only reason why agents may choose to not collaborate.

We make four remarks about this payoff structure. First, separability aside, the linearity of payoff in d is not an additional assumption provided we leave the joint production function f unrestricted. Second, the notation α is only useful because we will be interested in the case of "small" direct value ($\alpha \to 0$). Third, while linear u represents a convenient benchmark, our presumed generality is useful for applications. For instance, a strictly concave u can approximate career concerns in which reputation is initially useful but quickly loses relevance after some acceptable threshold is reached; e.g., in environments where research considerations are secondary after a point. On the other hand, a strictly convex u could approximate situations in which additional increments of reputation begin to generate superstar effects, as in a community where research prowess is highly valued.

Finally, much (though not all) of this paper can be read by viewing $g(w,r)$ and $b_r(w)$ as primitives of the model. The former could be viewed as the density of ideas for a person with reputation r, without presuming the additive form $g(w,r) \equiv rg(w,1) + (1-r)g(w,0)$. The latter is the reputational update on seeing an idea of quality w from type r, without presuming the application of Bayes' rule. The likelihood ratio condition would be replaced by the assumption that $b_r(w)$ is strictly increasing in w.

II. Equilibrium

A. *Definition*

Given p and q, an *equilibrium* is a nonempty-valued correspondence $z \mapsto C(z)$ describing the collaborative behavior of individuals conditional on public posterior beliefs. It is also used to calculate those beliefs on observing a joint outcome. That is, (x,y) belongs to $C(z)$—the equilibrium collaboration set—if and only if

1. Ideas x and y feasibly yield z, that is, $f(x,y) = z$;

2. The public uses the functions b_p and b_q in (2), and then forms β_p and β_q as in (3), using the set $C(z)$; and

3. Given this belief formation, both p and q willingly collaborate when ideas are (x,y):

$$(4) \quad \alpha(z-x) \geq u\big(b_p(x)\big) - u(\beta_p) \quad \text{and} \quad \alpha(z-y) \geq u\big(b_q(y)\big) - u(\beta_q).$$

For each z, an equilibrium collaboration set describes nonempty sets of ideas $\mathcal{X}$ for p and $\mathcal{Y}$ for q, with each combination generating z. We write this compactly using the notation $C(z) = \mathcal{X} \times_z \mathcal{Y}$.[11] Most of our analysis focuses on equilibrium

[11] Our definition insists on nonempty collaboration sets. If p and q refuse to collaborate no matter what ideas they have, such an arrangement must specify off-path beliefs in case a "surprise collaboration" is observed. However, given z, if those beliefs assign probability 1 to any *one* combination of x and y, then both agents would prefer collaboration when ideas are x and y. With this restriction on off-path beliefs, zero collaboration cannot occur. We ignore the empty case for the rest of the paper.

collaboration sets for a single joint outcome z,[12] but it alludes to the equilibrium as defined above, which is composed of an equilibrium collaboration set for each z.

B. *Characterization*

PROPOSITION 1: *An equilibrium exists. In any equilibrium, for each* z, $C(z)$ *is of the form* $[\underline{x},\bar{x}] \times_z [\underline{y},\bar{y}]$, *with* $0 < \underline{x} < \bar{x} < z$ *and* $0 < \underline{y} < \bar{y} < z$, *and*

$$\alpha(z - \bar{x}) = u\big(b_p(\bar{x})\big) - u(\beta_p), \tag{5}$$

$$\alpha(z - \bar{y}) = u\big(b_q(\bar{y})\big) - u(\beta_q), \tag{6}$$

where β_p *and* β_q *solve* (*3*).

To understand the proposition, suppose the public conjectures that p and q collaborate when they draw ideas in some set C, and suppose β_p and β_q satisfy (3), given this conjectured set. Then each individual faces a trade-off across the direct payoff gains of collaboration, and potential reputational losses. The former are always beneficial in our setting. But a loss in reputational payoff will occur if an individual draws a particularly high-quality idea and then agrees to collaborate. The inequalities in (4) describe when the trade-off is resolved in favor of collaboration for both parties. The resulting indifference thresholds $\bar{x}$ and $\bar{y}$ are described in (5) and (6).[13]

Because all ideas in $C(z)$ "add up" to z, an equivalent description of the equilibrium collaboration strategies is that p agrees to collaborate whenever q draws an idea $y \geq \underline{y}$, where $f(\bar{x},\underline{y}) = z$, and q similarly agrees whenever p has $x \geq \underline{x}$, where $f(\underline{x},\bar{y}) = z$. Figure 1 displays these equilibrium collaboration regions, which are subsets of the locus $\{(x,y) : f(x,y) = z\}$.

III. Fragile Equilibria

Proposition 1 guarantees that a nonempty equilibrium exists, but there is no presumption of uniqueness. For any z, several collections $(\underline{x},\bar{x},\underline{y},\bar{y},\beta_p,\beta_q)$ could lock together in the way described in the proposition. Indeed, a central theme of our paper concerns equilibrium multiplicity. Some of them could feature asymmetric treatment of individuals who are identical in all payoff-relevant characteristics. We evaluate the robustness of these various equilibria and argue that symmetric treatment of such identical individuals is often *fragile*. Before formally introducing this concept, we provide an intuitive discussion.

[12] The definition of z rests on public perception in case collaboration occurs. For example, suppose that the academic community regards all "well-published papers" as a single category. Then C contains all pairs of ideas that lead to a "well-published paper" and such that both p and q agree to collaborate.

[13] The assumption that $\alpha > 0$ is crucial in implying $\underline{x} < \bar{x}$ and $\underline{y} < \bar{y}$. If instead $\alpha = 0$, then signaling is the only concern, and by an unraveling argument, only (and all) *singleton* sets $C(z,p,q) = \{x\} \times_z \{y\}$ with $x \in [0,z]$ and $f(x,y) = z$ are equilibrium collaboration sets. We will be interested in approximating the case of "pure signaling," but always with $\alpha > 0$.

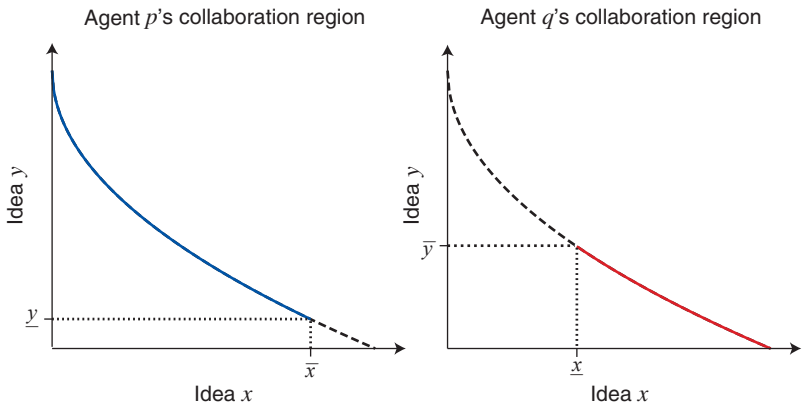

FIGURE 1. COLLABORATION REGIONS FOR AGENTS p AND q

Notes: The curves display all combinations of ideas x and y that yield a project z. The left panel shows combinations (in blue) such that p agrees to collaborate. The right panel shows combinations (in red) such that q agrees to collaborate.

Temporarily assume that $p = q$, so that both potential partners are identical in their payoff-relevant characteristics. Consider an equilibrium collaboration set at z which is also symmetric, inducing a common public update β in the event of collaboration. Now imagine that the individuals can be differentiated by some payoff-irrelevant identity, such as race, gender or nationality. Suppose that the public sees these individual identities as salient and "slightly reallocates" posterior credit in favor of p: $\beta_p > \beta > \beta_q$. This could come from some cultural bias against q's identity; perhaps a very small bias.

Anticipating this generous public update, p is now more willing to collaborate with q. Conversely, q is *less* open to collaborating with p. In short, in response to this small bias, we have $\bar{x} > \bar{y}$; person p shares ideas of higher quality than q does. Observe that to some degree, this now-asymmetric behavior confirms the public's initial bias. Furthermore, if those behavioral responses lead to new collaboration sets that "overshoot" the original bias, they may destabilize the symmetric outcome, and moreover, no bias would then be needed to shore up the asymmetric outcomes that might result. In this way, infinitesimally small identity-based biases could precipitate a discretely asymmetric outcome across functionally identical individuals.

More formally, take as given p, q and $z > 0$. Define a domain $\mathbf{B} \equiv \big[b_p(0), b_p(z)\big] \times \big[b_q(0), b_q(z)\big]$. Obviously, all pairs of equilibrium updates consistent with z must lie in $\mathbf{B}$. For $(\beta_p, \beta_q) \in \mathbf{B}$, we can describe collaboration strategies using (5) and (6):

$$(7) \quad u\big(b_p(\bar{x})\big) + \alpha(\bar{x} - z) = u(\beta_p) \quad \text{and} \quad u\big(b_q(\bar{y})\big) + \alpha(\bar{y} - z) = u(\beta_q),$$

which can be interpreted as saying that if p and q anticipate public updates (β_p, β_q) in the event of collaboration, then they will use the collaboration set generated by

$\bar{x}$ and $\bar{y}$. Call it $\tilde{C}$. But in that case, public updates conditional on collaboration will be given by (β_p', β_q') using (3), with $C(z)$ replaced by $\tilde{C}$. This iteration $(\beta_p, \beta_q) \mapsto \Theta(\beta_p, \beta_q) \equiv (\beta_p', \beta_q')$ is a mental map of the equilibrium process. Each fixed point of it corresponds to an equilibrium collaboration set.[14]

Definition.—An equilibrium collaboration set at (p, q, z), with collaborative update vector (β_p, β_q), is *fragile* if there is $\delta > 0$ and $\zeta > 0$ such that for every $\epsilon \in (-\delta, \delta)$,

$$|\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) - \beta_p| \geq |\epsilon|(1 + \zeta) \quad \text{and} \tag{8}$$

$$|\Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) - \beta_q| \geq |\epsilon|(1 + \zeta),$$

where the subscripts on Θ refer to the component functions of Θ.

The entries in these component functions perturb public perceptions by locally reallocating belief updates. The condition asks that the resulting "iterated updates" move away from the original equilibrium vector at some minimally positive rate ζ.[15] Of course, there is no one concept of fragility: we could have defined it using a weaker tendency for the updates to move away, or ask that at least one of the updates move away, or by tracing higher iterations as in the well-known concept of Lyapunov stability. The definition we provide comes close, in our view, to the intuitive discussion that prefaced it. In any event, the analysis that follows applies, possibly with minor changes, to any reasonable notion of fragility.

Fragility is a mathematical construct, which may or may not have a bearing on actual equilibrium selection. But combined with the existence of identities that might invite differential treatment for reasons of predisposed bias or historical inequality, the concept takes on real meaning. For then, even a small bias will find an anchor in the presence of our definition, ensuring that social assessments move significantly away from a fragile equilibrium. We also note that such biases can amplify real differences, even if no other salient identities exist. For instance, if p exceeds q but only by a tiny amount, a public bias that p gets discretely more credit will destabilize any near-symmetric equilibrium if (8) holds.

We note that *non*fragility is always a characteristic of some equilibrium.

PROPOSITION 2: *For every z, a nonfragile equilibrium collaboration set exists.*

[14] Indeed, this map is central to our proof of Proposition 1; see Appendix.

[15] The requirement $\zeta > 0$ in (8) ensures that small perturbations not only locally amplify but that they do so at some minimal geometric rate. Alternatively fragility could be defined by the less demanding requirement that $|\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) - \beta_p| > |\epsilon|$, etc., instead of (8). But this creates technical yet nongeneric complications of little conceptual interest in the present setting. The gap between the two definitions is analogous to that between a strictly increasing differentiable function, and a differentiable function with a strictly positive first derivative. Our definition sidesteps such issues.

IV. Equilibrium with Symmetric Partners

The fragility concept defined above will be used to assess the robustness of both symmetric and asymmetric equilibria in symmetric settings. We shall see that symmetric equilibria are often fragile; and, when that is the case, nonfragile asymmetric equilibria exist, in which ostensibly "equal" partners are treated differently. The unequal treatment of equals has received extensive attention in the literature on statistical discrimination; our theory is most similar to equilibrium statistical discrimination à la Arrow (1973). In such theories, unequal treatment is one equilibrium, but there could be an equally robust equilibrium with equal treatment. Our approach is different, in that it explicitly interrogates the fragility of the equal-treatment outcome.

For the remainder of this section, as well as in Section V, we study symmetric players, who possess identical priors ($p = q$). Sometimes, we drop the subscripts p and q, but typically retain them as names for the partners in the collaboration.

A. *Symmetric Equilibrium*

PROPOSITION 3: *Suppose that $p = q$. Then for each z, there exists a symmetric equilibrium collaboration set, with $\beta_p = \beta_q$ and $C(z) = [\underline{x},\bar{x}] \times_z [\underline{x},\bar{x}]$ for some $\underline{x} < \bar{x} \leq z$. Additionally for all α small or large enough, there is just one symmetric equilibrium.*

Proposition 3 asserts the uniqueness of symmetric equilibrium when reputational concerns are either large (a case of principal interest to us), or small. The general question of uniqueness of symmetric equilibrium remains open, though it will not interfere with our analysis of fragility.

B. *The Fragility of Symmetry*

An equilibrium is fragile when individual responses to a small bias "more than justify" that bias. This statement contains two parts: (i) the effect of the bias on collaboration strategies, and (ii) the effect of those strategies on "subsequent" perceptions. Fragility asks that either or both effects be large.

The first of these effects requires that reputational concerns be strong (or equivalently, that α be small). For if the opposite were true, then the direct value of collaborative output would dominate all other considerations, and any perturbation in public updating would induce a muted response. It also requires that the *slope* of the reputational payoff from solo work, evaluated at the symmetric equilibrium threshold $\bar{x}$, be not too large. For if it were large, the increase in collaborative range following a perturbation in that individual's favor would perforce be small.

The second of these effects concerns the observer's reaction to changes in individual collaboration strategies. Specifically, the response in the observer's perception of a particular individual will have two parts: a direct effect due to that individual's change in their own willingness to collaborate (via a change in $\bar{x}$), and an cross effect due to the change in the *other* person's willingness to collaborate (via a change in $\underline{x}$).

For each $z > 0$ and $w \in [0,z]$, define the "isoquant" $\iota_z(w)$ by $f\big(w,\iota_z(w)\big) \equiv z$. The following proposition formalizes our discussion.

PROPOSITION 4: *A symmetric equilibrium* $[\underline{x},\bar{x}] \times_z [\underline{x},\bar{x}]$ *ascribed to symmetric partners with* $p = q$ *is fragile if and only if*

$$\alpha + u'\big(b(\bar{x})\big)b'(\bar{x}) < u'\big(\beta(\underline{x},\bar{x})\big)\left[\frac{\partial\beta(\underline{x},\bar{x})}{\partial\bar{x}} - \iota_z'(\bar{x})\frac{\partial\beta(\underline{x},\bar{x})}{\partial\underline{x}}\right], \tag{9}$$

where $\beta(\underline{x},\bar{x})$ *is the symmetric Bayes' update from* (3) *conditional on collaboration.*

The terms on the left-hand side of (9) need to be small for effect (i) to be strong, as discussed above. Effect (ii) is represented on the right-hand side of (9). The first term captures the direct effect due to the person's change in their own willingness to collaborate, and the second term reflects the cross effect due to the change in the other person's willingness to collaborate, which is mediated by the effect of $\bar{x}$ on $\underline{x}$.

To build more intuition, suppose the production function is linearly additive with $f(x,y) = x + y$, so that $\iota_z'(\bar{x}) = -1$. Suppose, too, that the distribution of an individual's ideas conditional on the joint outcome z is uniform over $[0,z]$.[16] In that case, equation (9) becomes[17]

$$\alpha + u'\big(b(\bar{x})\big)b'(\bar{x}) < u'\Big[E\big(b(x)\,|\,x \in [\underline{x},\bar{x}]\big)\Big]E\big(b'(x)\,|\,x \in [\underline{x},\bar{x}]\big). \tag{10}$$

This implied fragility condition (10) compares the slope of the reputational value $u\big(b(\bar{x})\big)$ to the average slope of b in the interval $[\underline{x},\bar{x}]$, multiplied by the slope of u evaluated at the average reputation b in that same interval. Condition (9) is therefore related to the concavity of the reputation function b and of the value function u. Indeed, we show below that the fragility of symmetric equilibria can be related to the concavity of the *reputational value function* $v \equiv u \circ b$, the composition of u and b.

The following central corollary of Proposition 4 (though technically not an immediate implication) provides a necessary and sufficient condition, *depending only on the primitives of the model*, for the symmetric equilibrium to be fragile under all α positive but small. Remember that if reputational concerns are strong—i.e., if α is small—we already know from Proposition 3 that there is a unique symmetric equilibrium.

COROLLARY 1: *The following two statements are equivalent. First, there is* $\underline{\alpha} > 0$ *such that for every* $\alpha < \underline{\alpha}$, *the symmetric equilibrium is fragile. Second,*

$$-\frac{v''(e_z)e_z}{v'(e_z)} > \frac{e_z}{z - e_z} + \frac{1}{2}\iota_z''(e_z)e_z, \tag{11}$$

where e_z *is the "equal input" for* z; *i.e.,* $f(e_z,e_z) = z$.

[16] This is the case when $g(\cdot\,,p)$ is exponentially distributed.
[17] We prove this assertion in the Appendix.

On the left-hand side of (11) is a measure of the concavity of the reputational value function $v = u \circ b$; specifically, the coefficient of relative risk aversion of v, evaluated at the equal input idea e_z. On the right-hand side, we have terms related to the production technology f. Let us now briefly examine the effects of each of these objects—the reputational value function and the production function—on fragility.

To focus on the left-hand side, assume that f is linearly additive: $f(x,y) = x + y$. Then the right-hand side of (11) is just 1. By Corollary 1, symmetric equilibria are therefore fragile for small α if and only if the reputational value v is "more risk-averse than log utility;" that is, when

$$-\frac{v''(e_z)e_z}{v'(e_z)} > 1. \tag{12}$$

Let Z be the set of z-values such that (12) holds, so v is locally more risk averse than log utility. Because b is bounded above by 1, v is a bounded function, which implies that (12) cannot fail throughout, and that Z is nonempty. Indeed, as we show in the Appendix, Z is an unbounded union of open intervals. For every z in Z, (12)—and therefore (11)—holds, and so symmetric equilibria are fragile when reputational concerns are sufficiently strong. For instance, if u is linear and ideas are exponentially distributed for both good and bad types, then $Z = [\underline{z}, \infty)$, for some $\underline{z} \in \mathbb{R}_+$. The same is true when ideas are distributed according to two Weibull distributions with a common shape parameter, or according to two log-normal distributions.

Now we take a more careful look at the right-hand side of (11). The first of the two terms there reflects the extent of *synergy* in the combination of ideas. All other things being equal, if collaboration is more efficient, then the value of $e_z/(z - e_z)$ is lower, making it more likely that (8) holds. The very synergy of collaboration makes it highly desirable at the margin for an individual who is favored by bias—that marginal desirability could destabilize the symmetric equilibrium and render it fragile. In contrast, the second term captures the *complementarity* of ideas. The greater the positive curvature in the production isoquant, the larger that complementarity. This term makes for stability of a symmetric outcome, for it is less rewarding to collaborate more when the other individual is collaborating less. Conversely, if the isoquant has negative curvature, which will happen when the production function for ideas is *convex*, the symmetric equilibrium is more likely to be fragile.

The discussion above considered conditions on the primitives of the model that make fragility more or less plausible when reputational concerns are strong ($\alpha \to 0$). When these conditions hold, we know from Proposition 3 that there is a *unique* and fragile symmetric equilibrium. But Proposition 2 asserts that nonfragile equilibria exist. And so we conclude that at least one *asymmetric* and *nonfragile* equilibrium exists.

For completeness, we note a second corollary of Proposition 4: when the weight α placed on direct project value is large, then symmetric equilibria are nonfragile.

COROLLARY 2: *There is* $\bar{\alpha} > 0$ *such that if* $\alpha > \bar{\alpha}$, *no symmetric equilibrium is fragile.*

We omit the proof, as all it requires are minor technical verifications that all the derivatives in (9) are bounded above even as $\bar{x}$ and $\underline{x}$ respond endogenously to α.

The central takeaway of this section is that, if reputational concerns are uppermost, there is a real danger that symmetric players will not be treated symmetrically when their payoff-irrelevant identities are visible to the public, and when those identities are additionally associated with a salient social history of unequal treatment.

V. Payoff Implications of Asymmetric Equilibria

When symmetric equilibria are fragile, Proposition 2 assures us that other nonfragile equilibria exist. They must be asymmetric, of course. If society can distinguish between agents using payoff-irrelevant identities, functionally identical individuals will settle into such equilibria, and each identity will collaborate for distinct sets of ideas. One identity will be *favored*; that is, it will be perceived by the public as (stochastically) contributing better ideas to the collaboration, compared to the other identity.

In this section, we discuss the payoff implications of favoritism within an asymmetric equilibrium. Such a favored identity benefits—almost by definition—from the reputational aspects of collaboration. But matters are more complicated when not only reputational payoffs but also the direct payoffs of collaboration are taken into account. For those interested not so much in the distribution of payoffs but the overall implications of asymmetric treatment, Section VD compares aggregate welfare across symmetric and asymmetric equilibria, when both exist.

A. *The Direct Gains from Collaboration*

Given z, write agent p's payoffs as

$$\Pi_p(z) = R_p(z) + D_p(z), \tag{13}$$

where R_p stands for reputational payoff and D_p for direct payoff. Let $\Gamma_z(x)$ be the distribution of person p's ideas, conditional on a joint outcome z. That is, Γ_z is the marginal distribution of x along the locus $f(x,y) = z$. Let γ_z be its probability density function. (The Appendix contains an explicit derivation of this object from model primitives.) Then

$$R_p(z) = \int_0^z v_p(x)\gamma_z(x)dx, \tag{14}$$

where $v_p(x) = u(\beta_p)$ if $x \in [\underline{x},\bar{x}]$, and $v_p(x) = u\big(b_p(x)\big)$ if $x \notin [\underline{x},\bar{x}]$, and

$$D_p(z) = \left[\alpha\int_0^z x\gamma_z(x)dx\right] + \alpha\int_{\underline{x}}^{\bar{x}}(z-x)\gamma_z(x)dx. \tag{15}$$

The first term in (15) is independent of the collaboration set. The second term represents the "extra" direct value produced when collaboration occurs. Analogous expressions hold for person q, using the thresholds $\{\underline{y},\bar{y}\}$. (Remember that $p = q$, so that the original distributions of ideas are the same for both agents.) This formulation takes an *ex interim* stance: it supposes that z has already been realized and will

be observed by the public if there is collaboration. One could additionally assess payoffs from an ex ante perspective by integrating all ex interim payoffs over z.

PROPOSITION 5: *Consider a pair of symmetric agents, and an asymmetric equilibrium at z, with p favored. Then*

$$D_p(z) < D_q(z),$$

so that p, despite being favored, receives a lower direct payoff from collaboration.

Proposition 5 shows that in an asymmetric equilibrium with symmetric agents, the disfavored identity receives *higher direct payoff* than their favored counterpart, while at the same time they suffer a lower reputational payoff conditional on collaboration.[18] The two effects are connected in the following way. Being favored means that the public singles out individual p (or their identity) and gives them greater credit in a cross-identity collaboration. That very treatment is of course "justified" in equilibrium, with p contributing more and q less, each affected by the public bias. But it is precisely for this reason that the favored individual p loses out on the direct gains from collaboration: individual p shares better ideas with q than q does with p.

Next, we evaluate *overall payoffs*; that is the sum of reputational payoffs and direct payoffs weighted by α. One might imagine that the overall payoff ranking would depend on the weight α, but that isn't true for an important special case.

B. *Overall Gains with Linear Reputational Payoff*

When the reputational payoff function u is linear, then the ex interim expected reputational payoff must be equal in expectation across all equilibria. This is a consequence of the martingale property of Bayesian updates (or "Bayes plausibility"). Proposition 5 then immediately implies:

PROPOSITION 6: *Suppose $u(b) = b$. Then in any asymmetric equilibrium at z with symmetric agents with p as the favored identity, we have $u(\beta_p) > u(\beta_q)$, so that p is relatively better off conditional on collaboration, but*

$$\Pi_p(z) = R_p(z) + D_p(z) < R_q(z) + D_q(z) = \Pi_q(z),$$

so that q receives the higher unconditional expected payoff.

A favorable public disposition has two effects on payoffs to the favored identity. The direct component is unambiguously negative (Proposition 5). As for the reputational component, it is positive conditional on collaboration. But *it must be zero overall* when the utility of reputation is linear, as a direct implication of Bayes plausibility. Proposition 6 therefore records the paradoxical result that overall impact of favoritism on an agent's expected payoff is negative.

[18] In Appendix B, we show that in situations where $p \neq q$, there is also a connection between *relative* favoritism and the *relative* loss of direct payoffs.

These results contrast sharply with the literature on statistical discrimination. That literature typically finds either that discrimination does not affect the group favored by public beliefs, or that the favored group benefits from discrimination.[19] To our knowledge, the observation that the payoff ordering may be entirely reversed across the reputational and the overall perspectives is new.[20] In a similar environment, Tumlinson (2012) shows a weaker, but related, result: conditional on working individually, agents in the disfavored population outperform those in the favored population.

C. *Distributions of Posteriors*

Proposition 6 must be qualified when reputational payoffs are nonlinear. While Bayes plausibility continues to guarantee that expected posteriors are constant across equilibria, the expected *utility* from those posteriors will vary when utility is nonlinear. The higher moments of an agent's reputational outcome will now affect overall expected payoffs. To illustrate this point, Proposition 7 describes the dispersion of reputational outcomes across favored and disfavored agents within an asymmetric equilibrium. Specifically, fix some "target posterior" t. Think of it as some threshold that is relevant for career advancement. Let $P_p(t,z)$ and $P_q(t,z)$ be the probabilities that, after seeing the (joint or solo) projects, the public's posterior about agents p and q exceed t.

PROPOSITION 7: *In an asymmetric equilibrium at z with updates (β_p, β_q) ascribed to symmetric agents, where p has the favored identity,*

$$P_p(t,z) \geq P_q(t,z) \text{ if } t \in [\beta_q, \beta_p), \text{ but}$$

$$P_p(t,z) \leq P_q(t,z) \text{ if } t < \beta_q \text{ or } t \geq \beta_p.$$

Moreover, both inequalities are strict when t is sufficiently close to β_p or β_q.

When symmetric agents collaborate in an asymmetric equilibrium, the disfavored agent is more likely to reach extreme target posteriors, either very large or very small. Conversely, the favored agent is more likely to reach intermediate targets. One possible interpretation is that the distribution of career outcomes of disfavored agents (induced by the reputation distribution) is more risky than that of favored agents. Figure 2 illustrates this by displaying the distribution of posteriors for an asymmetric equilibrium. The horizontal axes plot various target thresholds for the posterior, and the vertical axes the probability that an agent's posterior will be larger than some target t. Recall that when agents collaborate, the public sees only z and is unable to tell

[19] This novel payoff result could be relevant in a larger setting in which agents choose identity. They can explain why individuals would choose to express an identity that is disfavored along some dimension (collaborative output, in our setting), without relying on the assumption that they receive some inherent value from being their "true self" (Akerlof and Kranton 2000; Akerlof and Rayo 2020).

[20] In Appendix B, we discuss the possibility of yet another payoff reversal across favored and disfavored identities in a larger game in which partners are randomly matched at a prior stage. The reversal occurs because a minority identity faces a larger share of cross-identity matches relative to a majority identity.

which combinations of x and y generated z. So all the possible ideas that could lead to collaboration are "garbled" to make up the expected update of the observer. That leads to the pictured flat regions and discontinuities in the posterior distributions.

The first panel plots this distribution for the favored identity; the second for the disfavored identity. The third panel combines the two. For targets below β_q but close to it—specifically, when $t \in \left[b_r(\underline{y}), \beta_q\right)$—the disfavored identity q is more likely to reach the target than her favored counterpart. For if p has idea $\bar{x}$, then the pair collaborates and the public's update on q is β_q. If, conversely, q were to have the same idea $\bar{x}$, the agents work separately, and the public's update on p is $b_p(\bar{y}) < \beta_q$. Similar arguments illustrate the other differences in the posterior distributions across p and q.

These differences in posterior distributions across favored and disfavored agents suggest a possible impact of discrimination in credit assignment on the career dynamics of agents. Suppose we think of the target posterior t as a level of reputation that an individual must attain in order to stay in the profession for one more period. In that case, if t is large, the discrimination dynamics would be "self-correcting"—more disfavored agents would achieve the threshold necessary to stay in the profession, and would work for another period (and have another chance to establish their reputation). If instead t is less ambitious, the resulting dynamics would be "spiraling": agents with the disfavored identity would be disproportionally excluded from the profession and have less future opportunities to reveal their underlying quality.[21]

These observations also speak to a point recently made by Bohren, Hull, and Imas (2022) regarding the measurement of discrimination. They argue that if—controlling for observables—members of identity A are more likely to reach a certain good outcome than members of identity B, that should not necessarily be taken as evidence of discrimination against group B. Rather, that comparison would measure what they call "direct discrimination," but may not account for the entire trajectory that should contextualize this comparison, or what they call "total discrimination" in the environment.[22] A local observation may or may not be indicative of a more global diagnosis.

This is precisely what our model also notes, but in the cross section (of targets) rather than on trajectories. Suppose that a researcher were to measure the probability that agents with the favored and disfavored identity reach some career outcome induced by a target posterior t. Proposition 7 tells us that if that target is sufficiently ambitious, then the researcher would find that agents with the *disfavored* identity are *more* likely to reach it. (The opposite conclusion would be reached if the target t is a bit less ambitious.) Our model thus illustrates a mechanism through which the initial local measurement may not reflect the "true" underlying discrimination within a system, while at the same time it points to the pattern that overall discrimination takes. The fact that the higher moments of success are affected by "true discrimination" means that measurement at different targets or thresholds have different meanings.

[21] Of course, this informal description only hints at a possible career dynamic that's implied by the repetition of our stage-game equilibrium. A proper analysis of the dynamic game would be necessary in order to understand the implications of credit assignment on career trajectories. In a different context, Bardhi, Guo, and Strulovici (2020) also argue that early career discrimination may be "spiraling" or "self-correcting," depending on the characteristics of the learning environment through which an employer learns their skill-types of employees.

[22] They refer to the difference between these two objects as "systemic discrimination."

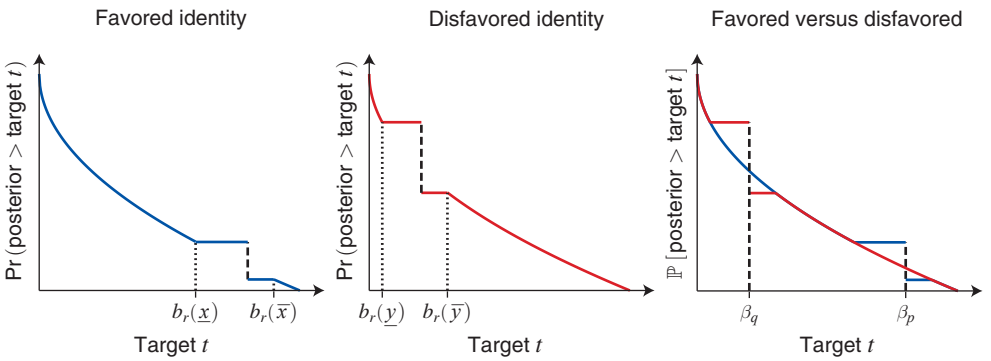

FIGURE 2. DISTRIBUTION OF POSTERIORS IN ASYMMETRIC EQUILIBRIA

Notes: The first panel plots this distribution for the favored identity; the second for the disfavored identity. The third panel combines the two.

Relatedly, Proposition 7 also speaks to two recent empirical observations. Sarsons et al. (2021) find that, conditional on cross-gender academic collaboration, the probability of tenure increases more for male rather than female authors. In contrast, Card et al. (2022, p.1937), studying the election of fellows to the Econometric Society, argue that the female-male "gap became positive (though not statistically significant) from 1980 to 2010, and in the past decade has become large and highly significant, with over a 100 percent increase in the probability of selection for female authors relative to males with similar publications and citations." Proposition 7 states that, in an equilibrium where women are discriminated against in terms of credit assignment (as documented by the first fact), a high target reputation (presumably needed for election to the Econometric Society) is relatively more likely to be reached by a member of the disfavored identity. We are, of course, fully aware of recent initiatives to correct gender imbalances in the profession (certainly in the last decade) and do not intend to entirely attribute this empirical finding to the forces of our model. Rather, we view the juxtaposition of the findings in Sarsons et al. (2021) and in Card et al. (2022) as an illustration of the varying career patterns possibly implied by discrimination.

D. *Welfare in Symmetric and Asymmetric Equilibria*

Propositions 5–7 compare the distribution of payoffs across favored and disfavored agents within an asymmetric equilibrium. A separate question concerns the comparison of aggregate welfare across symmetric and asymmetric equilibria, when both exist. When the production function is linearly additive and the reputational payoff is linear, we can measure aggregate welfare in each of these equilibria by the probability that the two partners choose to collaborate. We record this formally as:

OBSERVATION 1: *If* $f(x,y) = x + y$ *and* $u(b) = b$*, then the probability of collaboration* (*for a given* z) *is a measure of ex interim aggregate welfare.*

Aggregate Welfare in the Exponential Model.—Suppose that ideas are exponentially distributed for both agent types. That is, let $g(w,1) = \lambda_1 e^{-\lambda_1 w}$ and $g(w,0) = \lambda_0 e^{-\lambda_0 w}$, with $\lambda_0 > \lambda_1$. Then $g(w,1)$ dominates $g(w,0)$ in the likelihood ratio order.

Numerical computation shows that when α is small enough and z large enough, there is a unique symmetric equilibrium, which is fragile. (This is consistent with the discussion in Section IV.) For those same parameters, asymmetric equilibria exist.[23] When both equilibria coexist, we can compare the probability of collaboration, given by $\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})$, across equilibria. The results are displayed in the left panel of Figure 3. It shows that for low values of z, the probability of collaboration is larger in the asymmetric equilibrium than in the symmetric one, but that the order is reversed for higher values of the joint product z.[24] By Observation 1, these collaboration probabilities map directly into comparisons of aggregate welfare. When z is small, asymmetric equilibria entail higher welfare, with the comparison overturned at larger values.[25]

The right panel of Figure 3 uses a different measure of collaboration: the "size" of the equilibrium collaboration set ($\bar{x} - \underline{x} = \bar{y} - \underline{y}$). It is less accurate for the measure of aggregate welfare but is still of intrinsic interest. The panel shows that the symmetric equilibrium always displays a "larger set" of ideas for which collaboration occurs, relative to the asymmetric equilibrium. This is intriguing because as already noted, the collaboration probabilities do switch rank across equilibria with z. Briefly, the probability of collaboration may be larger in the asymmetric equilibrium, because the distribution Γ_z of ideas conditional on the joint outcome z assigns higher probability to more unequal points in the isoquant $f(x,y) = z$ than to points where x and y are close to each other; thereby weighting asymmetric collaboration sets relatively more than symmetric ones. We return to this issue below.

Beyond the Exponential Model.—For any $z > 0$, consider branches of symmetric and asymmetric equilibrium collaboration sets—$\left[\underline{x}_s(\alpha), \bar{x}_s(\alpha)\right] \times_z \left[\underline{y}_s(\alpha), \bar{y}_s(\alpha)\right]$ and $\left[\underline{x}_a(\alpha), \bar{x}_a(\alpha)\right] \times_z \left[\underline{y}_a(\alpha), \bar{y}_a(\alpha)\right]$ respectively—indexed by $\alpha \downarrow 0$. Their limits all involve singleton sets of the form $\{x\} \times_z \{y\}$, with $f(x,y) = z$. That's because direct value is given no weight in the limit, so that by an unraveling argument, each person's contributions will need to be perfectly identified under collaboration.

Say that an asymmetric branch is *distinct* if it does not merge with the symmetric branch in the limit; that is, it converges to some $\{x\} \times_z \{y\}$ with $x \neq y$.

PROPOSITION 8: *Let $f(x,y) = x + y$ and $u(b) = b$. Consider a distinct asymmetric equilibrium branch. Let $\hat{x}$ the common limit of $\underline{x}_a(\alpha)$ and $\bar{x}_a(\alpha)$ as $\alpha \to 0$. Then, if*

$$b'(\hat{x}) + b'(z - \hat{x}) > 2b'(z/2), \tag{16}$$

[23] When the symmetric equilibrium is not fragile—which occurs either when z is sufficiently small or α is sufficiently large—there are no asymmetric equilibria.

[24] The parametric region $z \in [0, 1.6]$ is not depicted in Figure 3, because asymmetric equilibria do not exist for those values of z.

[25] In a binary collaborative setting, Tumlinson (2012) also delineates some conditions under which discriminatory equilibria are Pareto dominate symmetric equilibria, when both exist.

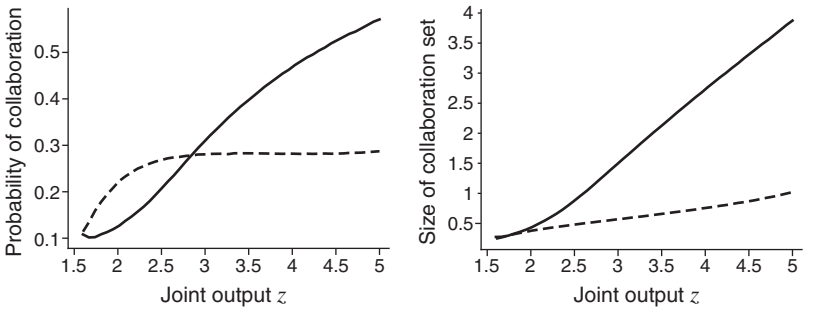

FIGURE 3

Notes: Probability of collaboration $(\Gamma_z(\bar{x}) - \Gamma_z(\underline{x}))$ and size of collaboration set $(\bar{x} - \underline{x})$ in symmetric (solid lines) and asymmetric equilibria (dashed lines). The parameters of the model are $g(w,1) = e^{-w}$, $g(w,0) = 4e^{-4w}$, $p = q = 0.5$, and $\alpha = 0.1$.

the collaboration set is larger under symmetric equilibrium for all α small enough; *i.e.*, *there is $\bar{\alpha} > 0$ such that $\alpha < \bar{\alpha}$ implies $\bar{x}_s(\alpha) - \underline{x}_s(\alpha) > \bar{x}_a(\alpha) - \underline{x}_a(\alpha)$.*

Condition (16) holds in the exponential model, in the parameter region considered in Figure 3. Indeed, the following stronger condition holds:

$$b'(z+\epsilon) + b'(z-\epsilon) > 2b'(z/2), \text{ for every } \epsilon \in (0, z/2]. \tag{17}$$

As in the discussion above, we must qualify this assertion. While of intrinsic interest, the "size" of the collaboration set does not fully pin down the probability of collaboration, as we saw in the exponential model. The latter also depends on the distribution Γ_z of ideas conditional on the joint outcome being z. That distribution is symmetric around the equal idea e_z, but exhibits two canonical shapes: (i) Γ_z is S-shaped and (ii) Γ_z is reverse-S-shaped. Case (i) indicates that, given an outcome z, it is more likely that both agents had "similar" ideas, and case (ii) indicates that, given an outcome z, it is more likely that agents contributed differently. (The uniform distribution would lie neutrally in between.) In the former case, the size of the collaboration set maps unambiguously to the probability of collaboration.

COROLLARY 3 (to Proposition 8): *Suppose $f(x,y) = x + y$ and $u(b) = b$, and consider collections of symmetric and asymmetric equilibria as in Proposition 8. Then if (17) holds and Γ_z is S-shaped, there exists some $\bar{\alpha}$ such that $\alpha < \bar{\alpha}$ implies that the probability of collaboration in the symmetric equilibrium is larger than that in the asymmetric equilibrium.*

VI. Efficiency and Authorship Ordering

We end with some remarks on the efficiency of equilibrium outcomes, as opposed to the distribution of payoffs across identities. There is a tension between the value of collaboration and the private desire to build reputation, and that results in inefficient collaboration decisions. We discuss this issue, as also a partial resolution of it.

A. *Inefficiency*

An equilibrium is *inefficient at* $z > 0$ if there is some other collaborative arrangement of the form $\mathcal{X} \times_z \mathcal{Y}$ such that both players receive a higher expected payoff conditional on z. For instance, when u is linear or concave, anything short of full collaboration is inefficient. To see why, recall the equilibrium payoff to p:

$$\begin{aligned}\Pi_p(z) &= R_p(z) + D_p(z)\\ &= \int_0^z v_p(x)\gamma_z(x)dx + \alpha\int_0^z x\gamma_z(x)dx + \alpha\int_{\underline{x}}^{\bar{x}}(z-x)\gamma_z(x)dx\end{aligned}$$

for any $z > 0$, where $v_p(x) = u(\beta_p)$ if $x \in [\underline{x},\bar{x}]$, and $v_p(x) = u\big(b_p(x)\big)$ if $x \notin [\underline{x},\bar{x}]$. A parallel expression holds for q. The first term is the expected payoff from reputation. The second term is an individual-specific baseline constant, unaffected by equilibrium strategy. The third term represents the expected direct gains from collaboration. All expectations are taken over individual ideas, conditional on z. Suppose that u is linear. Then expected reputational payoff is just the expected posterior starting from a prior of p. All the private and social gains from pairwise interaction come from the direct value of collaboration.

The same is true a fortiori when reputational utility is concave. In that case, collaboration is additionally useful because it creates a reduction in the spread of Bayes' updates; that contraction is mean preserving by Bayes plausibility and therefore unrestrained collaboration is again welcomed. In summary, full collaboration is unequivocally valuable with weakly concave reputational utility.

But full collaboration is precluded in equilibrium due to lack of commitment. Suppose that an agent has an excellent idea and their partner has a bad idea. From that ex post perspective, the agent with the good idea understands that the direct gain from collaboration may not overcome the loss of reputational value. Therefore, while collaboration is valuable in terms of its direct payoff, it will not always happen.

When u is not concave, full collaboration will generally not be desirable from the joint perspective of the two agents. The local strict convexity of u in some regions might lead them to prefer individual updates in reputation, which makes solo research more valuable. It is still true, though, that equilibria will generally be inefficient. Equilibrium and optimality conditions are distinct, barring nongeneric coincidences, so the argument above works for any reputational utility function.

B. *Merit-Based and Random Order in Collaboration*

We now explore the intuition that policies that help to disentangle each person's contributions to a joint project would make for greater collaboration, and efficiency. Obviously, a policy that states that "p contributed x, q contributed y" would be first best in theory, but alas, only in theory. Such a policy would be blind to the fact that such statements are hard, if not impossible, to make in practice; see the discussion in Section VIC of Ray ⊙ Robson (2018). One policy, standard in the

publishing process of many scientific fields, is to arrange authors in the sequence of their *ordinal* contribution to the joint project. That "merit order" has the immediate impact of reducing the extent of informational garbling. Say p is the lead author. Now the observer additionally knows that contributions lie in the set $M^p(z) = C(z) \cap \{(x,y) \mid x \geq y\}$. Might that spur more collaboration?

Certainly, holding fixed the collaboration set from our baseline model, p would willingly reveal this additional information. But q might not want to. The problem is most severe when q's idea is just short of the equal input e_z, where a decision to go solo would yield (approximately) $u(b_q(e_z)) + \alpha e(z)$, while a collaborative decision would generate a payoff of $\hat{\beta}_q + \alpha z$, where $\hat{\beta}_q$ is calculated from $M^p(z)$. That may or may not be enough for q to participate—it is certainly not as attractive a prospect as in our benchmark model, because $\hat{\beta}_q < \beta_q$. Merit order solves one problem at the potential cost of creating another.

Fortunately, it is possible to have one's cake and eat it too. Consider an arrangement in which merit order is not revealed unless the contributions are disparate enough. With relatively egalitarian ideas, let authors randomize their name order in a way that signals that merit order is *not* being used; this could be done, for instance, by using a particular symbol as proposed in Ray ⓣ Robson (2018). Under this convention, the absence of a symbol would signify the use of merit order. Following this line of reasoning, a *merit-augmented equilibrium* at z is defined by three disjoint collections $R(z)$, $M^p(z)$, and $M^q(z)$ of (x,y) pairs, to be respectively interpreted as zones for which random order, merit order favoring p, and merit order favoring q are employed, such that

(i) For every $(x,y) \in R(z) \cup M^p(z) \cup M^q(z), f(x,y) = z$.

(ii) $x > y$ for all $(x,y) \in M^p(z)$ and $x < y$ for all $(x,y) \in M^q(z)$.

(iii) For $C \in \{R(z), M^p(z), M^q(z)\}$, we have $(x,y) \in C$ if and only if $\alpha x + u(b_r(x)) \leq \alpha z + u(\beta_r(z,C))$ for $r = p,q$, where $\beta_r(z,C)$ is the public update ratio conditional on observing z and one of the three specific collaboration sets.

PROPOSITION 9: *For each equilibrium of our baseline model, there is a merit-augmented equilibrium that strictly Pareto dominates it* (*both ex interim and ex post*).

To illustrate, let C be the equilibrium set in the benchmark equilibrium under consideration. There is at least one person for whom the upper collaboration threshold (say $\bar{x}$) exceeds the lower threshold ($\underline{y}$) of his partner. Imagine adding to these thresholds an additional sliver of idea combinations (x,y) such that $x > \bar{x}$ and $y < \underline{y}$, demarcating these with merit order. (One can do the same with the mirror thresholds $\bar{y}$ and $\underline{x}$, assuming $\bar{y} > \underline{x}$.) Just as in the benchmark model, there will be limits to collaboration: at some idea *strictly* smaller than z, the lead author would rather go solo; simply inspect (5). So the new equilibrium with its combination of merit and random order will still fall sort of complete efficiency, but it will improve on the old one.

Might the merit-augmented equilibrium be fragile as in the benchmark model? We do not formally develop a definition of fragility for this expanded equilibrium concept. But the very existence of equilibrium zones that are "merit augmented" discourages—perhaps without entirely eliminating—public speculation on who contributed more. Now the authors themselves have a language to ordinally communicate such information *of their own volition*. If they choose the set R, they make it clear to the public that merit differences are not severe enough to be pointed out. If they choose M^p and M^q, that removes some of the need for speculation in the first place.

VII. Conclusion

We propose a model of collaborative work in pairs, in which individuals choose to combine ideas, or work alone based on the direct and reputational values of their projects. Our model captures two important aspects of collaboration: the direct gains derived from combining people's complementary skills, coupled with the potential reputational losses that arise from intertwined contributions, thereby compromising each individual's ability to build reputation.

Among other things, we argue that robust equilibria often display discrimination: the public attributes greater credit for collaborative work to individuals who belong to certain favored identities. We view these theoretical predictions as a natural accompaniment to empirical evidence regarding collaborative work in academic research, which shows that greater credit is assigned to men for work produced in mixed-gender teams. Most prominently, Sarsons (2017) and Sarsons et al. (2021) study gender differences in recognition for group work. Using two experiments, as well as observational data on academic production in economics, they argue that credit attribution for joint work depends on gender (with women suffering relative to men), even if partners are observationally the same in payoff-relevant attributes.

But a fuller consideration of the welfare implications of discriminatory equilibria yields more nuanced findings. We compare favored and disfavored identities, both within a single asymmetric equilibrium and across equilibria. It is certainly the case—by definition—that a favored individual receives greater credit conditional on collaboration. But that generates endogenous reactions in collaborative strategies that work against the favored individual, who is unable to participate in the best ideas from cross-identity matches. This latter effect reduces direct payoffs to the favored individual. We then turn to higher moments of the reputation distribution, arguing that discrimination against a certain identity creates varied reflections along the cross-section of career outcomes. For example, individuals with the disfavored identity may be relatively *more* likely to attain very ambitious career "targets," while the opposite is true of intermediate career targets. Finally, we compare aggregate welfare across symmetric and asymmetric equilibria, and consider policies that might extend the scope of symmetric cooperation.

There are three directions that we see as natural extensions of our current model and plan on exploring in future research. First, in our baseline model, the public's posteriors on agent types are always calculated according to Bayes' rule. However, this Bayesian assumption is not essential, and the model can accommodate other updating rules that rely on the observed project outcomes and the public's

conjectured collaboration structure. Second, because our model speaks directly to empirical observations on academic collaboration and other team-based projects, it can be adapted to the empirical estimation of a model based on our framework. That estimated model would permit us to evaluate different policies—for example, the random-name-order/merit-based-order policy we propose in Section VI. It could also serve to identify the nature of equilibrium selection when the equilibrium set is multivalued, as it typically is in our setting.

Finally, our simple model uses random matching and may be interpreted as describing a single step in the evolution of an entire career dynamic. That makes it a good base on which other empirically relevant extensions can be constructed, such as prematch considerations and a fuller account of career dynamics. We do not mean to suggest that such an extension would be immediate or fully amenable to analytical treatment, situated as it is in a complex interactive system. But we do believe that the model constructed here represents a useful first step.

Appendix A. Conditional Idea Distributions

Define the conditional density that p has idea x, under the presumption that p and q *always collaborate* on joint project z, as

$$\gamma_z(x) \equiv \frac{g(x,p)g\big(\iota_z(x),q\big)\big|\iota_z'(x)\big|}{\int_0^z g\big(x',p\big)g\big(\iota_z(x'),q\big)\big|\iota_z'(x')\big|dx'} \tag{A.1}$$

$$\text{with associated cumulative distribution function } \Gamma_z \text{ on } [0,z],$$

where recall that $\iota_z(w)$ maps $w \in [0,z]$ to the partner's idea $\iota_z(w)$ on the isoquant for z. That is, knowing z, the density of x is given by $g(x,p)g\big(\iota_z(x),q\big)\big|\iota_z'(x)\big|$,[26] normalized by the term in the denominator of (A.1) to account for conditioning on z. Similarly, define ω_z, which is the counterpart of γ_z for person q:

$$\omega_z(y) \equiv \frac{g(y,q)g\big(\iota_z(y),p\big)\big|\iota_z'(y)\big|}{\int_0^z g\big(y',q\big)g\big(\iota_z(y'),p\big)\big|\iota_z'(y')\big|dy'}$$

$$\text{with associated cumulative distribution function } \Omega_z \text{ defined on } [0,z].$$

Note that γ_z and ω_z are model primitives and not endogenous. If p and q collaborate only on $C(z) = [\underline{x},\bar{x}] \times_z [\underline{y},\bar{y}]$, then the conditional density of x is further adjusted to $\gamma_z(x)/\big[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})\big]$, as we do when taking the conditional expectations in (3). Writing these out, we have

$$\beta_p = \begin{cases} \dfrac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})}\displaystyle\int_{\underline{x}}^{\bar{x}} b_p(x)\gamma_z(x)dx, & \text{if } \bar{x} > \underline{x}; \\ b_p(\bar{x}), & \text{if } \bar{x} = \underline{x}; \end{cases} \tag{A.2}$$

[26] The density of the partner's idea at $\iota_z(x)$ is given by $g\big(\iota_z(x),q\big)\big|\iota_z'(x)\big|$, a standard transformation using change of variable.

and

$$\beta_q \equiv \begin{cases} \dfrac{1}{\Omega_z(\bar{y}) - \Omega_z(\underline{y})} \displaystyle\int_{\underline{y}}^{\bar{y}} b_q(y)\omega_z(y)dy, & \text{if } \bar{y} > \underline{y}; \\ b_q(\bar{y}), & \text{if } \bar{y} = \underline{y}. \end{cases} \tag{A.3}$$

Appendix B. More on Favored and Disfavored Identities

8.1. *Favored and Disfavored Identities for Asymmetric Individuals*

Our notions of favoritism in the main text extend to cases where agents are not functionally identical. One particularly extreme equilibrium situation occurs when p's *worst* idea under collaboration is viewed as better than q's *best* idea, that is, when $\underline{x} > \bar{y}$. Say then that p is *super favored* in that equilibrium. Less drastically, consider two distinct equilibria 1 and 2, and persons p and q, where p may not equal q. Say that p—or p's identity—is *relatively favored* (and q *relatively disfavored*) in equilibrium 1 relative to 2 if p receives a higher collaborative update in equilibrium 1 relative to 2, while the opposite is true of q. That is, $\beta_p(z,1) > \beta_p(z,2)$ and $\beta_q(z,1) < \beta_q(z,2)$.

It should be noted that with asymmetric agents, individuals have different "baseline payoffs": the bracketed term describing D_p in (15) is a person-specific constant. We therefore compare direct payoff *gains* by netting these terms out, defining

$$\Delta_p(z) \equiv \alpha \int_{\underline{x}}^{\bar{x}} (z - x)\gamma_z(x)dx \quad \text{and} \quad \Delta_q(z) \equiv \alpha \int_{\underline{y}}^{\bar{y}} (z - y)\omega_z(y)dy.$$

We now state the following extension of Proposition 5 (for a proof, see Appendix C):

PROPOSITION A.1:

(i) *If p is super favored in some equilibrium with joint output z conditional on collaboration, then $\Delta_p(z) < \Delta_q(z)$. That is, p obtains a lower direct payoff gain than q in that equilibrium, relative to always working alone.*

(ii) *If p is relatively favored (and q relatively disfavored) in equilibrium 1 over 2, and there are no super favored individuals in either equilibrium, then $\Delta_p(z,1) - \Delta_p(z,2) < \Delta_q(z,1) - \Delta_q(z,2)$: q's gain in direct payoff in moving from equilibrium 2 to 1 is larger than p's gain.*

8.2. *Majority Identities and Favoritism*

The payoff gains and losses reported in Section V hold fixed partner identity. Consider now a population version of the model, with all agents symmetric but divided into two payoff-irrelevant identities of disparate sizes. Each agent is randomly paired to one potential collaborator. Following this pairing, our model proceeds as before.

Suppose that the symmetric equilibrium is fragile, so that two matched agents of different identities engage in an asymmetric equilibrium in which the majority

identity, indexed by p, is favored. If two agents of the same identity meet, they play the symmetric equilibrium. Then the ex ante payoff A for each identity is given by

$$A_p = \sigma \int_z \Pi_p(z) + (1-\sigma)\int_z \Pi(z) \quad \text{and} \tag{A.4}$$

$$A_q = \sigma \int_z \Pi(z) + (1-\sigma)\int_z \Pi_q(z),$$

where $\sigma \in (0, 1/2)$ is the size of the minority identity, $\Pi_p(z)$ and $\Pi_q(z)$ are the expected payoffs on cross-identity matches (recall (13)) and $\Pi(z)$ are the expected payoff in symmetric equilibrium.

We are now potentially confronted by yet another reversal in payoffs, stemming from the fact that the minority identity faces a larger share of cross-identity matches relative to the majority identity. Proposition 5 continues to hold for each cross-identity match, so that the disfavored identity benefits from larger direct payoffs, conditional on each encounter. Nevertheless, the ex ante payoff to the minority identity could be lower by the fact that the symmetric equilibrium has payoffs that Pareto dominate those from asymmetric equilibria. Indeed, the smaller the disfavored minority, the more likely it is that a second payoff reversal could occur from this ex ante perspective.

PROPOSITION A.2: *Consider the symmetric matching model. Suppose that expected ex ante payoff under the symmetric equilibrium dominates expected payoffs to the disfavored minority; that is,* $\int_z \Pi(z) > \int_z \Pi_q(z)$. *Then for all* σ *small,* $A_p > A_q$, *even though under each match and each* z, *we have* $D_q(z) > D_p(z)$, *as in Proposition 5.*

The proof follows from the discussion above and is omitted.

Under the conditions of Proposition A.2, the play of asymmetric equilibria across identities creates a disincentive for cross-identity collaboration. Symmetric equilibria played within identities have the opposite effect. In such situations, and under the conditions described in the proposition, individuals (or at least disfavored individuals) will attempt to seek out others of their own identity. Our model describes a possible basis for collaborative homophily, though the analysis here only scratches the surface.

Appendix C. Proofs

8.3. *Proof of Proposition 1*

We defer the proof of existence to Step 2. Step 1 characterizes all equilibria.

Step 1 (Characterization): We claim that in any equilibrium, there exist $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ with $0 < \underline{x} < \bar{x} < z$ and $0 < \underline{y} < \bar{y} < z$, such that

$$\alpha(z - \bar{x}) = u\big(b_p(\bar{x})\big) - u(\beta_p), \tag{A.5}$$

$$\alpha(z - \bar{y}) = u\big(b_q(\bar{y})\big) - u(\beta_q), \text{ and} \tag{A.6}$$

$$C(z) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}], \tag{A.7}$$

where β_p and β_q are given by (A.2) and (A.3).

To prove this claim, note that if (4) holds for some x and y, then it also does for all $x' < x$ and $y' < y$. So the collaboration set of p is of the form $[0, \bar{x}]$, and that for q is of the form $[0, \bar{y}]$, for some $\bar{x}$ and $\bar{y}$ in $[0, z]$. Define $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = 1\iota_z(\bar{x})$; then it must be that $C(z) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. Because $C(z)$ is nonempty, $0 \leq \underline{x} \leq \bar{x} \leq z$ and $0 \leq \underline{y} \leq \bar{y} \leq z$. In turn, given $\underline{x}$ and $\underline{y}$, the upper bounds $\bar{x}$ and $\bar{y}$ are determined by indifference between collaboration and working alone, so that (4) holds with equality, giving us (5) and (6) via the transformations (A.2) and (A.3).

Next, we claim that $\underline{x} < \bar{x}$ (and likewise that $\underline{y} < \bar{y}$). For suppose this is false; then $\underline{x} = \bar{x}$, and because $\underline{y} = \iota_z(\bar{x})$ and $\bar{y} = \iota_z(\underline{x})$, it must be that $\underline{y} = \bar{y}$ as well. Recalling (A.2) and (A.3), we must conclude that $\beta_p = b_p(\bar{x})$ and $\beta_q = b_q(\bar{y})$, so that the right-hand sides of both (A.5) and (A.6) are 0. But given $\alpha > 0$, the left-hand sides of at least one of these equations must be strictly positive, a contradiction.

We note next that $\bar{x} < z$ (and likewise that $\bar{y} < z$). Suppose not; then $\bar{x} = z$. But at this threshold, collaborative output is the same as solo output, while by (1) and $\underline{x} < \bar{x}$, the signaling update is *strictly* smaller, a contradiction.

For the converse, take any $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ with $0 \leq \underline{x} \leq \bar{x} \leq z$ and $0 \leq \underline{y} \leq \bar{y} \leq z$, satisfying (5) and (6). Suppose that the public forms the beliefs $C(z) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. Then p will be happy to collaborate if $x < \bar{x}$ and unwilling to collaborate if $x > \bar{x}$, by virtue of that fact that (5) holds and the right-hand side of (5) is increasing in x. The same argument holds for q, and therefore we have an equilibrium.

Step 2 (Existence): Now we complete the proof of Proposition 1 by arguing that a nonempty equilibrium exists. Fix p, q and z. Let $\mathbf{B} \equiv [b_p(0), b_p(z)] \times [b_q(0), b_q(z)]$. Define a mapping $\Theta : \mathbf{B} \to \mathbf{B}$ as follows. For $(\beta_p, \beta_q) \in \mathbf{B}$, let $\bar{x}$ and $\bar{y}$ solve

$$u\big(b_p(\bar{x})\big) - \alpha[z - \bar{x}] = u(\beta_p) \quad \text{and} \quad u\big(b_q(\bar{y})\big) - \alpha[z - \bar{y}] = u(\beta_q). \tag{A.8}$$

Next, define $\underline{x}$ and $\underline{y}$ by

$$\underline{x} = \min\{\bar{x}, \iota_z(\bar{y})\} \quad \text{and} \quad \underline{y} = \min\{\bar{y}, \iota_z(\bar{x})\}, \tag{A.9}$$

and then β_p' and β_q' by the resulting collaborative updates as defined in (A.2) and (A.3).

Note that $(\beta_p', \beta_q') \in \mathbf{B}$. Denote by Θ this map from (β_p, β_q) to (β_p', β_q'). It is easy to see that Θ is continuous. By Brouwer's fixed point theorem, it has a fixed

point (β_p^*, β_q^*). Let $(\bar{x}^*, \bar{y}^*, \underline{x}^*, \underline{y}^*)$ be the corresponding values generated by (A.8) and (A.9). We claim that all these values lie strictly between 0 and z, and that

$$\underline{x}^* = \iota_z(\bar{y}^*) < \bar{x}^* \quad \text{and} \quad \underline{y}^* = \iota_z(\bar{x}^*) < \bar{y}^*. \tag{A.10}$$

To prove (A.10), it will suffice to show that $\underline{x}^* < \bar{x}^*$ and $\underline{y}^* < \bar{y}^*$. Suppose not, then (say) $\underline{x}^* = \bar{x}^*$. So by the formula for collaborative updates, $\beta_p^* = b_p(\bar{x}^*)$. At the same time, (A.8) implies that $b_p(\bar{x}^*) > \beta_p^*$ whenever $\bar{x}^* < z$, so the previous equality must imply that $x^* = z$. Therefore by (A.9), $\underline{y}^* = \min\{\bar{y}^*, \iota_z(\bar{x}^*)\} = 0$. Using the definition of the function β_q in (A.3), this implies $\beta_q^* < b_q(z)$, and therefore (A.8) implies $\bar{y}^* \in (0, z)$. But then, using (A.9) again, $\underline{x}^* = \min\{\bar{x}^*, \iota_z(\bar{y}^*)\} = \min\{z, \iota_z(\bar{y}^*)\} = \iota_z(\bar{y}^*) \in (0, z)$. At the same time, $x^* = z$, as we have already deduced. Together, these assertions contradict $\underline{x}^* = \bar{x}^*$.

To prove the rest of the claim, observe that (A.10) implies $\beta_p^* < b_p(z)$ and $\beta_q^* < b_q(z)$. Therefore, by (A.8), $\bar{x}^* < z$ and $\bar{y}^* < z$. Using (A.10), that implies $\underline{x}^* > 0$ and $\underline{y}^* > 0$.

It only remains to check that $(\bar{x}^*, \bar{y}^*, \underline{x}^*, \underline{y}^*)$ is an equilibrium. This is immediate using (A.8) and the just-established (A.10), along with Step 1. ∎

8.4. *Proof of Proposition 2*

Fix (z, p, q). Each equilibrium can be identified with a collection $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$; see Step 1 in the proof of Proposition 1. The set of all such equilibrium collections is compact—because the equilibrium correspondence is continuous—and so therefore is the set of equilibrium updates conditional on collaboration. Fix some agent, say q. Let $\underline{\beta}_q$ be the minimum value of equilibrium updates for her, over all equilibria at (z, p, q). Also let $\bar{\beta}_p$ be the maximum value of equilibrium updates for the other agent, p. Recall the mapping Θ with component functions Θ_p and Θ_q, introduced in Step 2 of the proof of Proposition 1.

Step 0: Θ_q is decreasing in its first argument and increasing in its second, and the opposite is true of Θ_p.

This is immediate from the definition of Θ. Next, for each $\beta_q \in [b_q(0), \underline{\beta}_q)$, let $B_1(\beta_q)$ be the largest value of β_p such that

$$\Theta_p(\beta_p, \beta_q) = \beta_p,$$

which is well defined by Step 0, and let

$$B_2(\beta_q) = \Theta_q(B_1(\beta_q), \beta_q).$$

Step 1: For all $\beta_q \in [b_q(0), \underline{\beta}_q)$ and $\beta_p \geq B_1(\beta_q)$,

$$\Theta_p(\beta_p, \beta_q) \leq \beta_p.$$

That follows from the definition of B_1 and the fact that $\Theta_p\big(b_p(z),\beta_q\big) \leq b_p(z)$.

Step 2: $B_2\big(\beta_q\big)$ is nondecreasing.

To verify this, let $\beta_q, \beta_q' \in \big[b_q(0), \underline{\beta}_q\big)$, with $\beta_q' > \beta_q$. By Step 0,

$$\Theta_p\big(\beta_p, \beta_q'\big) \leq \Theta_p\big(\beta_p, \beta_q\big).$$

And so for all $\beta_p \geq B_1\big(\beta_q\big)$, using Step 1,

$$\Theta_p\big(\beta_p, \beta_q'\big) \leq \Theta_p\big(\beta_p, \beta_q\big) \leq \beta_p.$$

But that just means $B_1\big(\beta_q'\big) \leq B_1\big(\beta_q\big)$. By Step 0 again, $B_2\big(\beta_q'\big) \geq B_2\big(\beta_q\big)$.

Step 3: $B_2\big(b_q(0)\big) > b_q(0)$.

By (A.8), $\Theta_q\big(\beta_p, b_q(0)\big) > b_q(0)$ for all $\beta_p \in \big[b_p(0), b_p(z)\big]$, and so $B_2\big(b_q(0)\big) > b_q(0)$.

Step 4: If an equilibrium with update $\underline{\beta}_q$ for q is fragile, then $B_2\big(\underline{\beta}_q - \epsilon\big) < \underline{\beta}_q - \epsilon$ for some $\epsilon > 0$.

If an equilibrium with updates $\big(\bar{\beta}_p, \underline{\beta}_q\big)$ is fragile, then by (8), there is $\epsilon > 0$ such that

$$\text{(A.11)} \qquad \text{(a)}\ \ \Theta_p\big(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon\big) > \bar{\beta}_p + \epsilon \ \text{ and}$$

$$\text{(b)}\ \ \Theta_q\big(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon\big) < \underline{\beta}_q - \epsilon.$$

Given Step 1, (A.11a) implies $B_1\big(\underline{\beta}_q - \epsilon\big) > \bar{\beta}_p + \epsilon$. Using this inequality along with (A.11b) and Step 0, we have $B_2\big(\underline{\beta}_q - \epsilon\big) < \underline{\beta}_q - \epsilon$.

To complete the proof, we claim that any equilibrium with updates $\big(\bar{\beta}_p, \underline{\beta}_q\big)$ is not fragile. For suppose it were fragile. Then Step 4 applies, and together with Steps 2 and 3, implies that there is $\beta_q \in \big(b_q(0), \underline{\beta}_q - \epsilon\big)$ such that

$$B_2\big(\beta_q\big) = \beta_q.$$

But then $(B_1(\beta_q), \beta_q)$ is a fixed point of the map Θ, and consequently can be associated with an equilibrium, as in the proof of Proposition 1. But that contradicts the definition of $\underline{\beta}_q$. ∎

8.5. *An Auxiliary Result*

In what follows, for any z and for any pair of thresholds $\underline{x} < \bar{x}$, and any prior $r \in (0,1)$, write the collaborative update β explicitly as a function of those thresholds $\underline{x}$ and $\bar{x}$, in line with (A.2):

$$\beta_r(\underline{x},\bar{x}) = \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})}\int_{\underline{x}}^{\bar{x}} b_r(x)\gamma_z(x)dx. \tag{A.12}$$

LEMMA 1: *For any* $x \in (0,z)$*:*

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial\beta_r(\underline{x},\bar{x})}{\partial\bar{x}} = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial\beta_r(\underline{x},\bar{x})}{\partial\underline{x}} = \frac{b_r'(x)}{2}, \tag{A.13}$$

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r(\underline{x},\bar{x})}{\partial\bar{x}^2} = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r(\underline{x},\bar{x})}{\partial\underline{x}^2} = \frac{b_r''(x)}{3} + \frac{b_r'(x)\gamma_z'(x)}{6\gamma_z(x)}, \tag{A.14}$$

and

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r(\underline{x},\bar{x})}{\partial\bar{x}\partial\underline{x}} = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r(\underline{x},\bar{x})}{\partial\underline{x}\partial\bar{x}} = \frac{b_r''(x)}{6} - \frac{b_r'(x)\gamma_z'(x)}{6\gamma_z(x)}. \tag{A.15}$$

PROOF:

It is easy to compute from (A.12) that for any $\bar{x} > \underline{x}$,

$$\frac{\partial\beta_r}{\partial\bar{x}}(\underline{x},\bar{x}) = \frac{\left[b_r(\bar{x}) - \beta_r(\underline{x},\bar{x})\right]\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})}. \tag{A.16}$$

To calculate the limit as $\bar{x}\downarrow x$ and $\underline{x}\uparrow x$, we use L'Hôpital's rule to see that

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial\beta_r}{\partial\bar{x}}(\underline{x},\bar{x}) = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\left\{\frac{\left[b_r'(\bar{x}) - \frac{\partial\beta_r}{\partial\bar{x}}(\underline{x},\bar{x})\right]\gamma_z(\bar{x}) + \left[b_r(\bar{x}) - \beta_r(\underline{x},\bar{x})\right]\gamma_z'(\bar{x})}{\gamma_z(\bar{x})}\right\},$$

Now $b_r(\bar{x}) - \beta_r(\underline{x},\bar{x}) \to 0$ as the limit above is taken, while $\gamma_z'(\bar{x})$ is bounded. Using this information in the equation above, we conclude that the required limit of $\frac{\partial\beta_r}{\partial\bar{x}}(\underline{x},\bar{x})$ equals $b_r'(x)/2$. The same steps can be used to show that $\frac{\partial\beta_r(x,x)}{\partial\underline{x}} = b_r'(x)/2$.

To establish (A.14), differentiate (A.16) with respect to $\bar{x}$ to see that

$$\begin{aligned}\frac{\partial^2\beta_r}{\partial\bar{x}^2} &= \frac{\left[b_r'(\bar{x}) - \frac{\partial\beta_r}{\partial\bar{x}}\right]\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} - \frac{\left[b_r(\bar{x}) - \beta_r\right]\gamma_z(\bar{x})^2}{\left[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})\right]^2} + \frac{\left[b_r(\bar{x}) - \beta_r\right]\gamma_z'(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \\ &= \frac{\left[b_r'(\bar{x}) - 2\frac{\partial\beta_r}{\partial\bar{x}}\right]\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} + \frac{\gamma_z'(\bar{x})}{\gamma_z(\bar{x})}\frac{\partial\beta_r}{\partial\bar{x}},\end{aligned}$$

where we invoke (A.16) again. Using L'Hôpital's rule once more, we have

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r}{\partial\bar{x}^2} = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\left\{\frac{\left[b_r''(\bar{x}) - 2\frac{\partial^2\beta_r}{\partial\bar{x}^2}\right]\gamma_z(\bar{x}) + \left[b_r'(\bar{x}) - 2\frac{\partial\beta_r}{\partial\bar{x}}\right]\gamma_z'(\bar{x})}{\gamma_z(\bar{x})} + \frac{\gamma_z'(\bar{x})}{\gamma_z(\bar{x})}\frac{\partial\beta_r}{\partial\bar{x}}\right\},$$

which implies that

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r}{\partial\bar{x}^2} = \frac{b_r''(x)}{3} + \frac{b_r'(x)\gamma_z'(x)}{6\gamma_z(x)}$$

as claimed. The same steps show that $\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r}{\partial\underline{x}^2} = \frac{b_r''(x)}{3} + \frac{b_r'(x)\gamma_z'(x)}{6\gamma_z(x)}$.

To establish (A.15), differentiate (A.16) with respect to $\underline{x}$ to see that

$$\frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}} = \frac{-\frac{\partial\beta_r}{\partial\underline{x}}\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} + \frac{\left[b_r(\bar{x}) - \beta_r\right]\gamma_z(\bar{x})\gamma_z(\underline{x})}{\left[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})\right]^2} = \frac{\frac{\partial\beta_r}{\partial\bar{x}}\gamma_z(\underline{x}) - \frac{\partial\beta_r}{\partial\underline{x}}\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})},$$

where we invoke (A.16) again. Using L'Hôpital's rule once more, we have

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}} = \lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\left[\frac{-\frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}}\gamma_z(\underline{x}) - \frac{\partial\beta_r}{\partial\bar{x}}\gamma_z'(\underline{x}) + \frac{\partial^2\beta_r}{\partial\underline{x}^2}\gamma_z(\bar{x})}{\gamma_z(\underline{x})}\right],$$

which implies that

$$\lim_{\underline{x}\uparrow x,\bar{x}\downarrow x}\frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}} = \frac{b_r''(x)}{6} - \frac{b_r'(x)\gamma_z'(x)}{6\gamma_z(x)}$$

as claimed. ∎

8.6. *Proof of Proposition 3*

Fix $z > 0$ and $p = q$. We drop the common subscripts $p = q$ on b_p, b_q, β_p and β_q. For any $B \in \left[0, b(z)\right]$, define $\bar{x}$ by (5), restated here as

$$u\left(b(\bar{x})\right) - \alpha(z - \bar{x}) = u(B). \tag{A.17}$$

and then define $\underline{x}$ by

$$\underline{x} = \iota_z(\bar{x}). \tag{A.18}$$

Let $\underline{B} \in [0, b(z))$ be the smallest value of B such that $\underline{x} \leq \bar{x}$. This threshold is well defined. For when $B \to b(z)$, it is evident from (A.17) that $\bar{x} \to z$ as well, but then $\underline{x} = \iota_z(\bar{x})$ must be close to 0 and therefore below $\bar{x}$. Moreover, for all $B > \underline{B}$, it is also true that $\underline{x} < \bar{x}$, because $\bar{x}$ is increasing in B and $\underline{x}$ is decreasing.

Restricting attention to the domain $[\underline{B}, b(z)]$, define a map $\Theta_s(B)$ as follows. Define $\bar{x}$ and $\underline{x}$ by (A.17) and (A.18), and then $\Theta_s(B) = B' = \beta(\underline{x}, \bar{x})$ according to (A.12). Two end-point conditions are to be noted. First, for $B = \underline{B}$, $b(\bar{x})$ is strictly larger than B. If $\underline{B} > 0$, it must also be that $\underline{x} = \bar{x}$, and so $B' = \Theta_s(B) > B$. If $B = \underline{B} = 0$, then certainly the same inequality $B' = \Theta_s(B) > B$ holds a fortiori. Second, for $B = b(z)$, $\bar{x} = z$ while $\underline{x} = 0$, so $\Theta_s(b(z)) < b(z)$. Finally, Θ_s is continuous, so there must be some $B^* \in (\underline{B}, b(z))$ with $\Theta_s(B^*) = B^*$. Define the accompanying values $\bar{x}^*$ and $\underline{x}^*$ from (A.17) and (A.18). It is immediate that $(\underline{x}^*, \bar{x}^*)$ is a symmetric equilibrium.

We now prove uniqueness. Recalling $B' = \Theta_s(B)$ and evaluating the derivative $\frac{d\Theta_s(B)}{dB} = \frac{dB'}{dB}$ at any symmetric fixed point with accompanying thresholds $\underline{x}$ and $\bar{x}$, we have

$$\begin{aligned} \frac{d\Theta_s(B)}{dB} = \frac{dB'}{dB} &= \left[\frac{\partial \beta}{\partial \underline{x}} \frac{d\underline{x}}{d\bar{x}} + \frac{\partial \beta}{\partial \bar{x}}\right] \frac{d\bar{x}}{dB} \\ &= \left[\frac{\partial \beta}{\partial \underline{x}} \iota_z'(\bar{x}) + \frac{\partial \beta}{\partial \bar{x}}\right] \frac{u'(\beta(\underline{x}, \bar{x}))}{u'(b(\bar{x}))b'(\bar{x}) + \alpha}, \end{aligned} \tag{A.19}$$

where the second equality follows easily from (A.17). By assumption, $u'(b(\bar{x}))$ is bounded, and because f has derivatives bounded above and below by positive numbers, $\iota_z'(\bar{x})$ is also bounded. Then it is easy to check by direct computation (use, e.g., (A.16)) that the partial derivatives of β are bounded above. It follows that for all α large enough, the right-hand side of (A.19) must be strictly smaller than 1, no matter which fixed point of Θ_s we pick. It follows that there can be just one fixed point, which completes the proof for large α.

Now take α small. We claim that for each $\epsilon > 0$, there is $\alpha(\epsilon)$ such that

$$e_z - \epsilon \leq \underline{x}(\alpha) < \bar{x}(\alpha) \leq e_z + \epsilon \tag{A.20}$$

for every pair of symmetric equilibrium thresholds $\{\underline{x}(\alpha), \bar{x}(\alpha)\}$ indexed by $\alpha \in (0, \alpha(\epsilon))$. We already know that $\underline{x}(\alpha) < \bar{x}(\alpha)$, so if (A.20) is false, then there exists $\epsilon > 0$ and a sequence $\{\alpha\}$ with $\alpha \to 0$ such that for every α, there is some symmetric equilibrium threshold $\bar{x}(\alpha)$ with $\bar{x}(\alpha) \geq e_z + \epsilon$.[27] Moreover, $\underline{x}(\alpha) \leq e_z$. In particular, given that u and b are strictly increasing, there is $\delta > 0$ such that

$$u(b(\bar{x}(\alpha))) - u(\beta(\underline{x}(\alpha), \bar{x}(\alpha))) \geq \delta \tag{A.21}$$

[27] This assertion is without loss. For if $\underline{x}(\alpha) \leq e_z - \epsilon$ instead, then using $\underline{x}(\alpha) = \iota_z(\bar{x}(\alpha))$, there is $\epsilon' > 0$ with $\bar{x}(\alpha) \geq e_z + \epsilon'$.

for all n. At the same time, using (5), we see that $u\big(b\big(\bar{x}(\alpha)\big)\big) - u\big(\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)\big) \to 0$ as $\alpha \to 0$, but that contradicts (A.21). So both $\bar{x}$ and $\underline{x}$ converge to e_z along any sequence of symmetric equilibria as $\alpha \to 0$, which establishes (A.20).

To complete the proof of uniqueness for small α, use (A.13) of Lemma 1 in (A.19), along with $(\bar{x},\underline{x}) \to (e_z,e_z)$, $\iota_z'(e_z) = -1$ and $u'(e_z)b'(e_z)$ strictly positive to conclude that the right-hand side of (A.19) converges to 0 as $\alpha \to 0$, no matter which sequence of fixed points of Θ_s we pick. It follows that there can be just one fixed point. ∎

8.7. *Proof of Proposition 4*

Using the fact that $(\beta_p,\beta_q) = \Theta(\beta_p,\beta_q)$ in equilibrium, (8) is equivalent to

$$\frac{\Theta_p(\beta_p+\epsilon,\beta_q-\epsilon) - \Theta_p(\beta_p,\beta_q)}{\epsilon} \geq 1+\zeta \text{ and} \tag{A.22}$$

$$\frac{\Theta_q(\beta_p+\epsilon,\beta_q-\epsilon) - \Theta_q(\beta_p,\beta_q)}{-\epsilon} \geq 1+\zeta.$$

Recalling the construction of Θ around the equilibrium ((A.8) and (A.9)), noting that $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = \iota_z(\bar{x})$ at any equilibrium, and noting that f is twice differentiable, it follows that Θ is continuously differentiable. So (A.22) is equivalent to

$$\frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_p} - \frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_q} > 1 \quad \text{and} \quad \frac{\partial\Theta_q(\beta_p,\beta_q)}{\partial\beta_q} - \frac{\partial\Theta_q(\beta_p,\beta_q)}{\partial\beta_p} > 1,$$

where these derivatives are evaluated at the equilibrium updates (β_p,β_q). In a symmetric equilibrium, these two inequalities are identical and equivalent to

$$\frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_p} - \frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_q} > 1, \tag{A.23}$$

evaluated at $\beta_p = \beta_q$. In the equations below, we drop the subscripts p and q when referring to the common value of the agents' prior. Wherever endogenous variables such as $\underline{x}$ and $\bar{x}$ appear, they are taken to refer to the symmetric equilibrium in question. We have

$$\frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_p} = \left[\frac{\partial\beta(\underline{x},\bar{x})}{\partial\bar{x}}\right]\left[\frac{d\bar{x}}{d\beta_p}\right] = \left[\frac{\partial\beta(\underline{x},\bar{x})}{\partial\bar{x}}\right]\frac{u'\big(\beta(\underline{x},\bar{x})\big)}{u'\big(b(\bar{x})\big)b'(\bar{x}) + \alpha} \tag{A.24}$$

and

$$\frac{\partial\Theta_p(\beta_p,\beta_q)}{\partial\beta_q} = \left[\frac{\partial\beta(\underline{x},\bar{x})}{\partial\underline{x}}\right]\left[\frac{\partial\underline{x}}{\partial\bar{y}}\right]\left[\frac{d\bar{y}}{d\beta_q}\right] = \left[\frac{\partial\beta(\underline{x},\bar{x})}{\partial\underline{x}}\right]\frac{u'\big(\beta(\underline{x},\bar{x})\big)\iota_z'(\bar{y})}{u'\big(b(\bar{y})\big)b'(\bar{y}) + \alpha}. \tag{A.25}$$

Combining (A.23), (A.24), and (A.25), using symmetry to note that $\bar{x} = \bar{y}$ and $\gamma_z(\bar{x}) = \gamma_z(\underline{x})$,[28] and rearranging terms, we obtain (9) as desired. ∎

[28] If $p = q$, $\big[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})\big]\gamma_z(\bar{x}) = g(\bar{x},p)g\big(\iota_z(\bar{x}),p\big) = g(\bar{x},p)g(\underline{x},p) = g(\underline{x},p)g\big(\iota_z(\underline{x}),p\big) = \big[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})\big]\gamma_z(\underline{x})$.

8.8. *Equivalence of (9) and (10) in Section IV*

Suppose $f(x,y) = x + y$ and $g(x,p) = g(x,q) = \lambda e^{-\lambda x}$, for some $\lambda > 0$. Then inequalities (9) and (10) are equivalent.

PROOF:

First, we show that Γ_z as defined in (A.1) is the uniform distribution over $[0,z]$. By definition, Γ_z is supported on $[0,z]$. Take some $x \in [0,1]$; we have

$$\gamma_z(x) = \frac{\left(\lambda e^{-\lambda x}\right)\left(\lambda e^{-\lambda(z-x)}\right)}{\int_0^z \left(\lambda e^{-\lambda x'}\right)\left(\lambda e^{-\lambda(z-x')}\right)dx'} = \frac{1}{z},$$

where the first equality uses $\iota_z(x) = z - x$ and $\iota_z'(x) = -1$. Given this distribution, it follows that

$$\beta(\underline{x},\bar{x}) = \frac{\int_{\underline{x}}^{\bar{x}} b(x)dx}{\bar{x} - \underline{x}},$$

so that

$$\frac{\partial\beta(\underline{x},\bar{x})}{\partial\bar{x}} = \frac{b(\bar{x}) - \beta(\underline{x},\bar{x})}{\bar{x} - \underline{x}} \quad \text{and} \quad \frac{\partial\beta(\underline{x},\bar{x})}{\partial\underline{x}} = \frac{\beta(\underline{x},\bar{x}) - b(\underline{x})}{\bar{x} - \underline{x}}.$$

Combining this observation with $\iota_z'(x) = -1$, we have

$$\frac{\partial\beta(\underline{x},\bar{x})}{\partial\bar{x}} - \iota_z'(\bar{x})\frac{\partial\beta(\underline{x},\bar{x})}{\partial\underline{x}} = \frac{b(\bar{x}) - b(\underline{x})}{\bar{x} - \underline{x}} = \frac{\int_{\underline{x}}^{\bar{x}} b'(x)dx}{\bar{x} - \underline{x}},$$

which completes the proof of the claim. ∎

8.9. *An Auxiliary Result (Toward the Proof of Corollary 1)*

By Proposition 3, for small α there is a unique symmetric equilibrium. So, continuing to suppress the common subscripts p and q, there is a collection $\{\bar{x}(\alpha),\underline{x}(\alpha)\}$ of uniquely defined equilibrium thresholds, along with equilibrium collaborative updates $\beta(\underline{x}(\alpha),\bar{x}(\alpha))$, satisfying

$$\text{(A.26)}\quad u\left(b\left(\bar{x}(\alpha)\right)\right) + \alpha\left[\bar{x}(\alpha) - z\right] = u\left(\beta\left(\underline{x}(\alpha),\bar{x}(\alpha)\right)\right) \quad \text{and} \quad \underline{x}(\alpha) = \iota_z\left(\bar{x}(a)\right).$$

LEMMA 2: *The functions $\bar{x}(\alpha)$ and $\underline{x}(\alpha)$ have the property that*

$$\lim_{\alpha\to 0}\bar{x}'(\alpha) = -\lim_{\alpha\to 0}\underline{x}'(\alpha) = \frac{z - e_z}{u'\left(b\left(e_z\right)\right)b'\left(e_z\right)}.$$

PROOF:

From (A.26), we have

$$u'\big(b\big(\bar{x}(\alpha)\big)\big)b'\big(\bar{x}(\alpha)\big)\bar{x}'(\alpha) + \big[\bar{x}(\alpha) - z\big] + \alpha\big[\bar{x}'(\alpha)\big] = u'\Big(\beta\big(\bar{x}(\alpha),\underline{x}(\alpha)\big)\Big)\frac{d\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{d\alpha}. \tag{A.27}$$

Now observe that

$$\begin{aligned}\frac{d\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{d\alpha} &= \frac{\partial\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{\partial\bar{x}}\bar{x}'(\alpha) + \frac{\partial\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{\partial\underline{x}}\underline{x}'(\alpha)\\ &= \bar{x}'(\alpha)\left[\frac{\partial\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{\partial\bar{x}} + \frac{\partial\beta\big(\underline{x}(\alpha),\bar{x}(\alpha)\big)}{\partial\underline{x}}\iota_z'\big(\bar{x}(\alpha)\big)\right].\end{aligned} \tag{A.28}$$

By (A.20) in the proof of Proposition 3, we know that $\underline{x}(\alpha)$ and $\bar{x}(\alpha)$ both converge to e_z as $\alpha \to 0$. Invoking equation (A.13) of Lemma 1, and using the fact that $\iota_z'\big(\bar{x}(\alpha)\big) \to -1$ as $\alpha \to 0$, we see that the term in the square brackets in (A.28) vanishes as $\alpha \to 0$. Using this information and combining (A.27) with (A.28), we see that

$$\lim_{\alpha\to 0}\bar{x}'(\alpha) = \frac{z - e_z}{u'(b(e_z)b'(e_z)}.$$

Again using $\underline{x}'(\alpha) = \iota_z'\big(\bar{x}(\alpha)\big)\bar{x}'(\alpha)$ and $\lim_{\alpha\to 0}\iota_z'\big(\bar{x}(\alpha)\big) = -1$, we also have

$$\lim_{\alpha\to 0}\underline{x}'(\alpha) = -\frac{z - e_z}{u'\big(b(e_z)\big)b'(e_z)}.$$

which completes the proof. ∎

8.10. *Proof of Corollary 1*

Recall condition (9) for fragility, slightly rewritten as

$$\alpha < u'(\beta)\frac{\partial\beta}{\partial\bar{x}}\big(\underline{x}(\alpha),\bar{x}(\alpha)\big) - u'(\beta)\iota_z'(\bar{x})\frac{\partial\beta}{\partial\underline{x}}\big(\underline{x}(\alpha),\bar{x}(\alpha)\big) - u'\big(b\big(\bar{x}(\alpha)\big)\big)b'\big(\bar{x}(\alpha)\big). \tag{A.29}$$

Using Lemma 1, it is easy to see that both the left-hand side and the right-hand side of condition (A.29) approach 0 as $\alpha \to 0$. And so, in order to evaluate whether (A.29) holds when α is close to 0, we must evaluate the derivatives of the left- and right-hand sides of (A.29) as $\alpha \to 0$. The left-hand side obviously has derivative

equal to 1. As for the right-hand side, we differentiate to get (arguments omitted for ease in writing):

$$\begin{aligned}\frac{\partial \text{RHS}}{\partial \alpha} = {} & u''(\beta)\left[\frac{\partial \beta}{\partial \bar{x}}\bar{x}'(\alpha) + \frac{\partial \beta}{\partial \underline{x}}\underline{x}'(\alpha)\right]\frac{\partial \beta}{\partial \bar{x}} - u''(\beta)\iota_z'(\bar{x})\left[\frac{\partial \beta}{\partial \bar{x}}\bar{x}'(\alpha) + \frac{\partial \beta}{\partial \underline{x}}\underline{x}'(\alpha)\right]\frac{\partial \beta}{\partial \underline{x}} \\ & - u''\big(b(\bar{x})\big)\big[b'(\bar{x})\big]^2\bar{x}'(\alpha) - \iota_z''(\bar{x})u'(\beta)\frac{\partial \beta}{\partial \underline{x}}\bar{x}'(\alpha) \\ & + u'(\beta)\left[\frac{\partial^2 \beta}{\partial \bar{x}^2}\bar{x}'(\alpha) + \frac{\partial^2 \beta}{\partial \bar{x}\partial \underline{x}}\underline{x}'(\alpha)\right] \\ & - u'(\beta)\iota'(\bar{x})\left[\frac{\partial^2 \beta}{\partial \bar{x}\partial \underline{x}}\bar{x}'(\alpha) + \frac{\partial^2 \beta}{\partial \underline{x}^2}\underline{x}'(\alpha)\right] - u'\big(b(\bar{x})\big)b''(\bar{x})\bar{x}'(\alpha).\end{aligned}$$

Equation (A.13) of Lemma 1 implies that the first two terms on the right-hand side above approach 0 as $\alpha \to 0$. Equation (A.14) of Lemma 1, along with the fact that $\lim_{\alpha\to 0}\frac{d\underline{x}}{d\alpha} = \lim_{\alpha\to 0}\iota_z'\big(\bar{x}(\alpha)\big)\frac{d\bar{x}}{d\alpha} = -\lim_{\alpha\to 0}\frac{d\bar{x}}{d\alpha}$, imply that in the limit as $\alpha \to 0$, the fifth and sixth terms cancel each other out. Applying these cancellations, we get

$$\begin{aligned}\lim_{\alpha\to 0}\frac{\partial \text{RHS}}{\partial \alpha} = \lim_{\alpha\to 0}\Bigg\{ & -u''\big(b(\bar{x})\big)\big[b'(\bar{x})\big]^2\bar{x}'(\alpha) - \iota_z''(\bar{x})u'(\beta)\frac{\partial \beta}{\partial \underline{x}}\frac{\partial \bar{x}}{\partial \alpha} \\ & - u'\big(b(\bar{x})\big)b''(\bar{x})\bar{x}'(\alpha)\Bigg\}.\end{aligned}$$

Applying Lemmas 1 and 2 and $\big(\underline{x}(\alpha),\bar{x}(\alpha)\big) \to (e_z,e_z)$ as $\alpha \to 0$, we finally have

$$\lim_{\alpha\to 0}\frac{\partial \text{RHS}}{\partial \alpha} = -\frac{u''\big(b(e_z)\big)}{u'\big(b(e_z)\big)}b'(e_z)(z-e_z) - \frac{1}{2}\iota''(e_z)(z-e_z) - \frac{b''(e_z)}{b'(e_z)}(z-e_z).$$

The fragility condition holds for small enough α whenever the derivative above exceeds 1, or equivalently,

$$\frac{u''\big(b(e_z)\big)}{u'\big(b(e_z)\big)}b'(e_z)e_z + \frac{1}{2}\iota''(e_z)e_z + \frac{b''(e_z)}{b'(e_z)}e_z < -\frac{e_z}{z-e_z}. \tag{A.30}$$

It is easy to see that $\frac{v''(e_z)e_z}{v'(e_z)} = \frac{u''\big(b(e_z)\big)}{u'\big(b(e_z)\big)}b'(e_z)e_z + \frac{b''(e_z)}{b'(e_z)}e_z$. Making these substitutions in (A.30), we obtain (11). ∎

8.11. *Proof of the Claim Concerning (12) in Section IV*

Let v be twice differentiable and bounded on $\mathbb{R}_+$, with $v'(e) > 0$ for all e. Let $Z = \left\{z : -\frac{v''(e_z)e_z}{v'(e_z)} > 1\right\}$, where e_z is defined by $f(e_z,e_z) = z$. Then Z is an unbounded union of open intervals.

PROOF:

Clearly Z is an open set and therefore a union of open intervals if nonempty. We show that Z is unbounded. Because e_z is continuously increasing in z, with $e_z \to \infty$ as $z \to \infty$, it suffices to show that $-v''(e)e/v'(e) \geq 1$ on some unbounded set E.

Suppose on the contrary that there is $e^* \geq 0$ such that for all $e \geq e^*$, $-v''(e)e/v'(e) \leq 1$. Rearranging this inequality, we see that $v'(e) + v''(e)e \geq 0$ for $e \geq e^*$, or equivalently,

$$\frac{dv'(e)e}{de} \geq 0 \quad \text{for} \quad e \geq e^*,$$

which implies in turn that $v'(e)e \geq v'(e^*)e^* \equiv d > 0$ for $e \geq e^*$. It follows that for all $e \geq e^*$,

$$v(e) - v(e^*) = \int_{e^*}^{e} v'(x)dx \geq d\int_{e^*}^{e} \frac{1}{x}dx = \ln(e) - \ln(e^*),$$

but this contradicts the fact that v is bounded. ∎

8.12. *Proof of Propositions 5 and 6*

Both inequalities in Proposition 5 follow from the fact that $\underline{x}^p > \underline{x}^q$ and $\bar{x}^p > \bar{x}^q$ in any equilibrium in which p is favored. Proposition 6 follows from the martingale property of Bayes' updates and from Proposition 5.

8.13. *Proof of Proposition 7*

Given an equilibrium set $C = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ we have

$$P_p(t,z) = \begin{cases} 1 - \Gamma_z\big(b^{-1}(t)\big), & \text{if } t < b(\underline{x}) \text{ or } t \geq b(\bar{x}); \\ 1 - \Gamma_z(\underline{x}), & \text{if } t \in \big[b(\underline{x}), \beta_p\big); \\ 1 - \Gamma_z(\bar{x}), & \text{if } t \in \big[\beta_p, b(\bar{x})\big); \end{cases}$$

and

$$P_q(t,z) = \begin{cases} 1 - \Gamma_z\big(b^{-1}(t)\big), & \text{if } t < b(\underline{y}) \text{ or } t \geq b(\bar{y}); \\ 1 - \Gamma_z(\underline{y}), & \text{if } t \in \big[b(\underline{y}), \beta_q\big); \\ 1 - \Gamma_z(\bar{y}), & \text{if } t \in \big[\beta_q, b(\bar{y})\big). \end{cases}$$

Now note that in an asymmetric equilibrium where p is favored, either $\underline{y} < \bar{y} \leq \underline{x} < \bar{x}$ or $\underline{y} < \underline{x} < \bar{y} < \bar{x}$. In either case, the inequalities in the proposition hold. ∎

8.14. *Proof of Proposition 8*

Take any collection of equilibrium collaboration sets indexed by α, with $\{\bar{x}(\alpha), \underline{x}(\alpha)\} \to \{\bar{x}(0), \underline{x}(0)\}$ as $\alpha \to 0$, where we already know that $\bar{x}(0) = \underline{x}(0)$. From the equilibrium condition (5), we have

$$\alpha\bar{x}(\alpha) = \alpha z + \beta\big(\underline{x}(\alpha), \bar{x}(\alpha)\big) - b\big(\bar{x}(\alpha)\big), \tag{A.31}$$

while rewriting (6) using $\bar{y} = z - \underline{x}$ and $\underline{y} = z - \bar{x}$, we have

$$\alpha\underline{x}(\alpha) = b\big(z - \underline{x}(\alpha)\big) - \beta\big(z - \bar{x}(\alpha), z - \underline{x}(\alpha)\big). \tag{A.32}$$

Combining (A.31) and (A.32),

$$\begin{aligned}\alpha\big[\bar{x}(\alpha) - \underline{x}(\alpha)\big] = {} & \alpha z + \beta\big(\underline{x}(\alpha), \bar{x}(\alpha)\big) \\ & + \beta\big(z - \bar{x}(\alpha), z - \underline{x}(\alpha)\big) - b\big(\bar{x}(\alpha)\big) - b\big(z - \underline{x}(\alpha)\big).\end{aligned}$$

Differentiating both sides with respect to α, we get

$$\begin{aligned}\big(\bar{x}(\alpha) - \underline{x}(\alpha)\big) + \alpha\big(\bar{x}'(\alpha) - \underline{x}'(\alpha)\big) = {} & z + \beta_1\big(\underline{x}(\alpha), \bar{x}(\alpha)\big)\underline{x}'(\alpha) \\ & + \beta_2\big(\underline{x}(\alpha), \bar{x}(\alpha)\big)\bar{x}'(\alpha) \\ & - \beta_1\big(z - \bar{x}(\alpha), z - \underline{x}(\alpha)\big)\bar{x}'(\alpha) \\ & - \beta_2\big(z - \bar{x}(\alpha), z - \underline{x}(\alpha)\big)\underline{x}'(\alpha) \\ & - b'\big(\bar{x}(\alpha)\big)\bar{x}'(\alpha) + b'\big(z - \underline{x}(\alpha)\big)\underline{x}'(\alpha).\end{aligned}$$

Taking limits as $\alpha \to 0$ (and $\bar{x} \to \underline{x}$) and invoking Lemma 1,

$$\begin{aligned}0 = {} & z + \frac{b'\big(\bar{x}(0)\big)}{2}\big[\bar{x}'(0) + \underline{x}'(0)\big] - \frac{b'\big(z - \bar{x}(0)\big)}{2}\big[\underline{x}'(0) + \bar{x}'(0)\big] \\ & - b'\big(\bar{x}(0)\big)\bar{x}'(0) + b'\big(z - \bar{x}(0)\big)\underline{x}'(0),\end{aligned}$$

which implies in turn that

$$0 = z + \frac{b'\big(\bar{x}(0)\big)}{2}\big[\underline{x}'(0) - \bar{x}'(0)\big] + \frac{b'\big(z - \bar{x}(0)\big)}{2}\big[\underline{x}'(0) - \bar{x}'(0)\big], \tag{A.33}$$

or equivalently,

$$\bar{x}'(0) - \underline{x}'(0) = \frac{z}{\frac{b'\big(\bar{x}(0)\big)}{2} + \frac{b'\big(z - \bar{x}(0)\big)}{2}}.$$

The final assertion of the proposition holds if

$$\left[\bar{x}_s'(0) - \underline{x}_s'(0)\right] > \left[\bar{x}_a'(0) - \underline{x}_a'(0)\right],$$

or equivalently, using (A.33) and noting that $\bar{x}_s(0) = \underline{x}_s(0) = z/2$, if (16) holds. ∎

8.15. *Proof of Proposition 9*

Suppose $C(z) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ is an equilibrium collaboration set of the original model with no authorship ordering. Augment the set as follows. Define x° by the smallest solution in x (but exceeding $\bar{x}$) to

$$u\big(b_p(x)\big) + \alpha[x - z] = u\big(\beta_p(\bar{x}, x)\big). \tag{A.34}$$

The left-hand side of (A.34) is strictly smaller than the right-hand side at $x = \bar{x}$, because $\beta(\bar{x}, \bar{x}) > \beta(\underline{x}, \bar{x})$, and thus $u\big(\beta(\bar{x}, \bar{x})\big) > u\big(\beta(\underline{x}, \bar{x})\big) = u\big(b_p(\bar{x})\big) + \alpha[\bar{x} - z]$ by the equilibrium condition for $\bar{x}$. The opposite inequality holds when $x = z$. Using the continuity of b_p and β_p and the intermediate value theorem, we see that x° is well defined, and $\bar{x} < x^\circ < z$.

Next, define $y_\circ$ by the smallest nonnegative value y such that

$$u\big(b_q(\underline{y})\big) + \alpha[\underline{y} - z] \leq u\big(\beta_q(y, \underline{y})\big). \tag{A.35}$$

This is well defined because the inequality does hold—strictly—when $y = \underline{y}$. So $y_\circ < \underline{y}$.

Define $x^* = \min\{x^\circ, \iota_z(y_\circ)\}$ and $y_* = \max\{y_\circ, \iota_z(x^\circ)\}$. We claim that

$$\bar{x} < x^* < z, \quad \text{and} \quad u\big(b_p(x)\big) + \alpha[x - z] < u\big(\beta_p(\bar{x}, x)\big) \tag{A.36}$$

for all $\bar{x} \leq x < x^*$, while

$$0 < y_* < \underline{y}, \quad \text{and} \quad u\big(b_q(\underline{y})\big) + \alpha[\underline{y} - z] < u\big(\beta_p(y, \underline{y})\big) \tag{A.37}$$

for all $y_* < y \leq \underline{y}$.

To prove this claim, note that $x^* \leq x^\circ < z$. Moreover, both x° and $\iota_z(y_\circ)$ strictly exceed $\bar{x}$, the latter because $y_\circ < \underline{y}$ and $\bar{x} = \iota_z(\underline{y})$. So $x^* = \min\{x^\circ, \iota_z(y_\circ)\} > \bar{x}$. Additionally, given the definition of x°, and because "<" holds at $x = \bar{x}$, the second inequality in (A.36) must hold.

Turning now to (A.37), note that $y_* \geq \iota_z(x^\circ) > 0$, because $x^\circ < z$. Moreover, $y_\circ < \underline{y}$ as already noted, and also $\iota_z(x^\circ) < \underline{y}$ because $x^\circ > \bar{x}$. Therefore $y_* =$

$\max\{y_\circ, \iota_z(x^\circ)\} < \underline{y}$. And finally, observe that the right-hand side of (A.35) is strictly increasing in y, while the left-hand side is constant in y. So if "$\leq$" holds in (A.37) at $y = y_*$, it must do so strictly for $y_* < y \leq \underline{y}$. That completes the proof of the claim.

In an entirely parallel manner, define $y^* \in (\bar{y}, z)$ and $x_* \in (0, \underline{x})$.

Now define $R(z,p,q) = C(z,p,q)$, and additionally,

$$M^p(z,p,q) \equiv \{(x,y) \,|\, f(x,y) = z, \text{ with } \bar{x} < x \leq x^* \text{ and } y_* \leq y < \underline{y}\} \cap \{(x,y) \,|\, x > y\}$$

and

$$M^q(z,p,q) \equiv \{(x,y) \,|\, f(x,y) = z, \text{ with } x_* \leq x < \underline{x} \text{ and } \bar{y} < y \leq y^*\} \cap \{(x,y) \,|\, x < y\}.$$

At least one of M^p and M^q is nonempty. Using (A.36) and (A.37), it is easy to verify that the collection $\{R, M^p, M^q\}$ satisfies the conditions for an equilibrium at (z,p,q).

Because this equilibrium adds zones of collaboration to the old equilibrium C without disturbing any updates there, and because each individual always has the option not to collaborate, this equilibrium must strictly Pareto dominate the old equilibrium in an ex post sense, and a fortiori in the ex ante sense. ∎

8.16. *Proof of Observation 1*

We've already argued that when $u(b) = b$, the ex interim expected reputational payoff is independent of collaboration strategies. From (15), the sum of direct payoffs across agents is

$$\begin{aligned}
D_p(z) + D_q(z) &= \left[\alpha \int_0^z x\gamma_z(x)dx\right] + \alpha \int_{\underline{x}}^{\bar{x}} (z-x)\gamma_z(x)dx + \left[\alpha \int_0^z y\omega_z(y)dy\right] \\
&\quad + \alpha \int_{\underline{y}}^{\bar{y}} (z-y)\omega_z(y)dy \\
&= \alpha\left[\int_0^z x\gamma_z(x)dx + \int_0^z y\omega_z(y)dy\right] + \alpha \int_{\underline{x}}^{\bar{x}} (z-x)\gamma_z(x)dx \\
&\quad + \alpha \int_{\underline{x}}^{\bar{x}} [z - \iota(x)]\gamma_z(x)dx \\
&= \alpha\left[\int_0^z x\gamma_z(x)dx + \int_0^z y\omega_z(y)dy\right] \\
&\quad + \alpha \int_{\underline{x}}^{\bar{x}} [2z - x - \iota_z(x)]\gamma_z(x)dx \\
&= \alpha\left[\int_0^z x\gamma_z(x)dx + \int_0^z y\omega_z(y)dy\right] + \alpha z[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})]. \blacksquare
\end{aligned}$$

8.17. *Proof of Proposition A.1*

(i) Subtracting the direct gains of q from those of p,

$$\Delta_p(z) - \Delta_q(z) = \int_{\underline{x}}^{\bar{x}} \big[\iota_z(x) - x\big]\gamma_z(x)dx. \tag{A.38}$$

Because p is super favored, $\iota_z(x) < x$ for all $x \in [\underline{x},\bar{x}]$, so by (A.38), $\Delta_p - \Delta_q < 0$.

(ii) Because p is favored in equilibrium 1 over 2, and q disfavored, it follows from (5) and (6) that $\bar{x}_1 > \bar{x}_2$ and $\bar{y}_1 < \bar{y}_2$. The latter inequality means that $\underline{x}_1 > \underline{x}_2$.

Recall (A.38) for each equilibrium j, indexing $\Delta_p(z)$ and $\Delta_q(z)$ by j. Then

$$\delta_j \equiv \Delta_{p,j}(z) - \Delta_{q,j}(z) = \int_{\underline{x}_j}^{\bar{x}_j} \big[\iota_z(x) - x\big]\gamma_z(x)dx. \tag{A.39}$$

We wish to sign $\delta_1 - \delta_2$. Because no agent is unambiguously favored in any equilibrium, but p is favored in 1 over 2, we have

$$\underline{x}_2 < \underline{x}_1 \leq e_z \leq \bar{x}_2 < \bar{x}_1. \tag{A.40}$$

Using (A.39), we must conclude that

$$\begin{aligned}\delta_1 - \delta_2 &= \int_{\underline{x}_1}^{\bar{x}_1} \big[\iota_z(x) - x\big]\gamma_z(x)dx - \int_{\underline{x}_2}^{\bar{x}_2} \big[\iota_z(x) - x\big]\gamma_z(x)dx \\ &= \int_{\bar{x}_2}^{\bar{x}_1} \big[\iota_z(x) - x\big]\,\gamma_z(x)\,dx - \int_{\underline{x}_2}^{\underline{x}_1} \big[\iota_z(x) - x\big]\gamma_z(x)dx < 0,\end{aligned}$$

where the last inequality follows from the fact that $\iota_z(x) > x$ for $x \in [\underline{x}_2,\underline{x}_1)$ (an implication of the first two inequalities in (A.40)), and that $\iota_z(x) < x$ for $x \in [\bar{x}_2,\bar{x}_1)$ (an implication of the third and fourth inequalities in (A.40)). ∎

REFERENCES

Akerlof, George, and Rachel Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics* 115 (3): 715–53.

Akerlof, Robert, and Luis Rayo. 2020. "Narratives and the Economics of the Family." Unpublished.

Anderson, Axel, and Lones Smith. 2010. "Dynamic Matching and Evolving Reputations." *Review of Economic Studies* 77 (1): 3–29.

Arrow, Kenneth. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets*, edited by Orley Ashenfelter and Albert Rees, 3–33. Princeton: Princeton University Press.

Bar-Isaac, Heski. 2007. "Something to Prove: Reputation in Teams." *RAND Journal of Economics* 38 (2): 495–511.

Bardhi, Arjada, Yingni Guo, and Bruno Strulovici. 2020. "Early-Career Discrimination: Spiraling or Self-Correcting?" Unpublished.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." NBER Working Paper 25935.

Bohren, J. Aislinn, Peter Hull, and Alex Imas. 2022. "Systemic Discrimination: Theory and Measurement." NBER Working Paper 29820.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109 (10): 3395–436.
Bowles, Samuel, Glenn C. Loury, and Rajiv Sethi. 2014. "Group Inequality." *Journal of the European Economic Association* 12 (1): 129–52.
Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. 2022. "Gender Differences in Peer Recognition by Economists." *Econometrica* 90 (5): 1937–71.
Chade, Hector, and Jan Eeckhout. 2020. "Competing Teams." *Review of Economic Studies* 87 (3): 1134–73.
Chaudhuri, Shubham, and Rajiv Sethi. 2008. "Statistical discrimination with peer effects: can integration eliminate negative stereotypes?" *Review of Economic Studies* 75 (2): 579–96.
Chalioti, Evangelia. 2016. "Team Production, Endogenous Learning about Abilities and Career Concerns." *European Economic Review* 85: 229–44.
Coate, Stephen, and Glenn C. Loury. 1993. "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review* 83 (5): 1220–40.
Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer. 2021. "Gender and Collaboration." Unpublished.
Fang, Hanming, and Andrea Moro. 2011. "Theories of Statistical Discrimination and Affirmative Action: A Survey." In *Handbook of Social Economics,* Vol. 1, edited by Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, 133–200. Amsterdam: Elsevier.
Gu, Jiadong, and Peter Norman. 2020. "A Search Model of Statistical Discrimination." Unpublished.
Holmström, Bengt. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13 (1): 324–40.
Jones, Benjamin. 2021. "The Rise of Research Teams: Benefits and Costs in Economics." *Journal of Economic Perspectives* 35 (2): 191–216.
Levy, Gilat. 2007. "Decision Making in Committees: Transparency, Reputation, and Voting Rules." *American Economic Review* 97 (1): 150–68.
Lissoni, Francesco, Fabio Montobbio, and Lorenzo Zirulia. 2013. "Inventorship and Authorship as Attribution Rights: An Enquiry into the Economics of Scientific Credit." *Journal of Economic Behavior and Organization* 95: 49–69.
Mookherjee, Dilip, and Debraj Ray. 2002. "Is Equality Stable?" *American Economic Review Papers and Proceedings* 92 (2): 253–59.
Mookherjee, Dilip, and Debraj Ray. 2003. "Persistent Inequality." *Review of Economic Studies* 70 (2): 369–94.
Moro, Andrea, and Peter Norman. 2004. "A General Equilibrium Model of Statistical Discrimination." *Journal of Economic Theory* 114 (1): 1–30.
Myrdal, Gunnar. 1944. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper and Row.
Ong, David, Ho Fai Chan, Benno Torgler, and Yu Yang. 2018. "Collaboration Incentives: Endogenous Selection into Single and Coauthorships by Surname Initial in Economics and Management." *Journal of Economic Behavior & Organization* 147: 41–57.
Onuchic, Paula. 2022. "Recent Contributions to Theories of Discrimination." Unpublished.
Onuchic, Paula ⍟ Debraj Ray. 2023. "Replication Data for: Signaling and Discrimination in Collaborative Projects." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E179821V1.
Ozerturk, Saltuk, and Huseyin Yildirim. 2021. "Credit Attribution and Collaborative Work." *Journal of Economic Theory* 195: 105264.
Pęski, Marcin, and Balázs Szentes. 2013. "Spontaneous Discrimination." *American Economic Review* 103 (6): 2412–36.
Phelps, Edmund. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review* 62 (4): 659–61.
Ray, Debraj, Jean-Marie Baland, and Olivier Dagnelie. 2007. "Inequality and Inefficiency in Joint Projects." *Economic Journal* 117 (533): 922–35.
Ray, Debraj ⍟ Arthur Robson. 2018. "Certified Random: A New Order for Coauthorship." *American Economic Review* 108 (2): 489–520.
Sarsons, Heather. 2017. "Recognition for Group Work: Gender Differences in Academia." *American Economic Review Papers and Proceedings* 107 (5): 141–45.
Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender Differences in Recognition for Group Work." *Journal of Political Economy* 129 (1): 101–47.
Tumlinson, Justin. 2012. "Adverse Selection in Team Formation under Discrimination." Unpublished.
Visser, Bauke, and Otto H. Swank. 2007. "On Committees of Experts." *Quarterly Journal of Economics* 122 (1): 337–72.
Winter, Eyal. 2004. "Incentives and Discrimination." *American Economic Review* 94 (3): 764–73.