Signaling and Discrimination in Collaborative Projects

Paula Onuchic ⓡ Debraj Ray[†]

June 2021

Abstract. We propose a model of collaborative work in pairs. Each potential partner draws an idea from a distribution that depends on their unobserved ability. The partners then choose to combine their ideas, or work separately. These decisions are based on the intrinsic value of their projects, but also on signaling payoffs, which depend on the public's assessment of individual contributions to joint work. In equilibrium, collaboration strategies both justify and are justified by public assessments. When partners are symmetric, equilibria with symmetric collaborative strategies are often fragile, in a sense made precise in the paper. In such cases, asymmetric equilibria exist: upon observing a collaborative outcome, the public ascribes higher credit to one of the partners based on payoff-irrelevant "identities." Such favored identities do receive a higher payoff relative to their disfavored counterparts conditional on collaborating, but may receive lower overall expected payoff. Finally, we study a policy that sometimes (but not always) clarifies the ordinal ranking of partners' contributions, and find that such disclosures can be Pareto-improving and reduce the scope for discrimination across payoff-irrelevant identities.

1. Introduction

Research is increasingly conducted in teams. In economics, team-authored papers make up over 70% of all published research, up from 20% in 1960.[1] The prominence of teamwork extends to other academic fields, as well as non-academic work — large technology companies such as Facebook or Amazon are known for fostering collaborative environments where workers self-organize in teams. Collaborative work can be beneficial, as it allows workers to fruitfully combine their complementary skills. However, by its very nature, teamwork obscures individual contributions, compromising each worker's ability to build their reputation. In this paper, we study this fundamental tension in collaborative activity, one that pits the intrinsic gains from joint work against the difficulty of revealing person-specific ability to the lens of public perception.

Consider a setting in which agents can collaborate in pre-matched pairs. Each agent is one of two types, good or bad. There is a common prior on each agent's type, which might vary across agents. When a pair meets, their initial interaction is a discussion of project ideas. Each agent draws an idea from a distribution that depends on their type, with good agents more likely to draw better ideas than bad agents. Both people then see their own idea and the idea of their potential partner, and they choose to work together when collaboration is beneficial to both parties, and separately otherwise. In making this decision, each agent seeks to maximize a combination of the *intrinsic value* and the *reputational value* of the project.

[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.

The intrinsic value of a project depends just on the agent's own idea if the work is completed alone, or on both agents' ideas if the work is undertaken in collaboration. Reputational value comes from an observer, also referred to as the public. In the event of solo work, the public observes the project outcome and updates their prior on the individual. In the event of collaboration, the public sees the joint outcome, but 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. These conjectures and updates affect reputational value, and therefore the agents' collaboration decisions.

Those decisions could be additionally affected by the possibility that public perceptions are biased towards or against certain individual characteristics that are payoff-irrelevant, such as gender, age, religion or nationality. In that case, collaborative work will be disproportionally incentivized for an individual with a "favored identity," and discouraged for an individual with a disfavored identity. A study of such discriminatory forces is a central theme of our paper, but they are more than mere theoretical abstractions.

Empirically, Sarsons et al. (2021) use data from academic economists to argue that upon observing joint work between women and men, the public attributes more credit to male collaborators. They find that, conditional on quality and other observables, an extra unit of joint research improves the probability of tenure of a male coauthor more than that of a female coauthor. Separately, Ductor, Goyal and Prummer (2021) document homophily in coauthorship networks, as well as other disparities across genders in collaboration patterns in economics academia. From the perspective of our model, these empirical observations are two sides of the same coin.

In Section 3, we establish existence and provide a general characterization of equilibrium collaboration in our model. The collaboration decision resolves the tradeoff between intrinsic value, which benefits from collaboration, and reputational value, which is garbled in joint work. Controlling for the value of a joint project, an agent with a good idea receives higher reputational payoff by not collaborating, and thus clarifying their own contribution to the observer. This tradeoff yields the characterization in Proposition 1: in any 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 sufficiently large. Theorem 1 establishes the existence of a nonempty collaboration set with this property.

Of course, the extent of equilibrium collaboration is intimately linked to payoff-relevant characteristics of the two partners. For instance, for someone with an established reputation (that is, a high prior), the difference in signaling value between collaboration and individual work is very small. Such person is almost always willing to combine ideas with a partner, and collaborates often in equilibrium. That said, we choose to pursue a different line of inquiry, focusing on the less evident observation that equilibrium collaboration patterns may also depend on individuals' payoff-*irrelevant* characteristics, such as gender, nationality, age or race.

To build this theme, Section 4 introduces the notion of equilibrium *fragility*. While we study the concept in more generality, it helps to imagine that the two individuals are perfectly symmetric, in the sense that the public has the same prior on them. However, suppose that each agent has a distinct payoff-irrelevant *identity*, which has cultural salience for the public, perhaps because of

a history of discrimination based on those identities. Suppose the public is "slightly biased" and reallocates some reputational value in favor of one identity, call it m. That is, the public thinks individual m contributes a little more to collaborative outcomes, and so assigns a slightly higher credit to m than to the other identity, say w, upon observing a cross-identity collaboration. The individuals in question will react to the small misperception by changing their willingness to collaborate in the first place. Identity m is now more willing to collaborate than identity w, the former incentivized by the credit allocation, the latter discouraged by it. This reaction actually confirms the initial bias, at least to some degree, and if such echo effects lead to behavior that "overshoots" the initial public misperception, they may destabilize symmetric interaction. In that case, we say the original (possibly symmetric) equilibrium is *fragile*.

Mathematically, fragility is related to the behavior of the pseudo-dynamic of updates around an equilibrium point, but our view is that this can only take hold in the presence of payoff-irrelevant but salient individual characteristics that the public can condition upon. Proposition 2 reassures us that there is always *some* equilibrium that is not fragile, but it may not be the one that displays equal treatment across equal individuals. Theorem 2 in Section 5 specifically studies symmetric agents, and establishes a necessary and sufficient condition for symmetric equilibria to be fragile. Proposition 4 follows this up by asserting that for an open set of collaboration values (large for several families of idea distributions), if signaling is a sufficiently strong motive, then there is a *unique and yet fragile* symmetric equilibrium. Invoking Proposition 2, a non-fragile but asymmetric equilibrium exists, where the public perception of contributions to collaborative outcomes depends on each person's payoff-irrelevant identity.

The use of multiple equilibrium to shore up statistical discrimination is well known, and in part our approach falls under this rubric. But the differences are also noteworthy. In our setting, the equal treatment or non-discriminatory outcome is eliminated by the fragility argument. In contrast, models of statistical discrimination need only display two equilibria (each conceivably non-fragile, though the question is typically not even asked), one with discriminatory treatment and the other without. Fragility relies on amplification of the smallest differences in perception. Section 6 takes up a second theme that brings out an additional and even starker contrast.

Recall that in an asymmetric equilibrium across symmetric agents, one of the identities is *favored*, or perceived as contributing better ideas to a collaboration — and must do so in equilibrium, thereby receiving higher credit when collaboration does takes place. But, and perhaps counterintuitively, that favorable treatment does not necessarily map into better overall payoffs. Section 6 studies the payoff implications of asymmetric treatment. Propositions 5 and 6 argue that while signaling payoff conditional on collaboration is higher for the favored identity, the intrinsic gains from collaboration are, in fact, lower. The very fact of more enthusiastic participation by the favored identity implies that they collaborate more when their ideas are better. That translates into an intrinsic payoff gain for the disfavored identity.

The paradox is particularly sharp when reputational utility is linear. Then the reputational payoff is fixed irrespective of the collaboration structure — a simple yet powerful implication of Bayes' plausibility. Therefore *overall* payoffs move in perfect tandem with intrinsic payoffs. Conditional on the partner match, the disfavored type is better off. This result contrasts with the literature on statistical discrimination, which generally finds that either discrimination does not

affect the favored group, or unequivocally favors it.[2] But the discussion is more nuanced than that. Section 6.3 considers an anterior step by embedding our model into a population game with random matching of partners. It shows that these payoff rankings could be reversed when majority identities are favored; see that Section and especially Proposition 7.

These nuances are heightened when reputational utility is nonlinear. Then, Bayes' plausibility notwithstanding, signaling payoffs do depend on the collaboration structure. Section 6.4 studies the *distribution* of posteriors induced by favoritism in a special case, one in which a "target posterior" is salient. For instance, agents may get promoted if they are perceived to be good with high enough probability. Proposition 8 argues that in an asymmetric equilibrium, the dis-favored identity is more likely to reach a target posterior when that target is extreme (either high or low). Conversely, the favored identity is more likely to reach intermediate target posteriors.

Finally, in Section 7, we consider a policy intervention. In our model, signaling concerns may inefficiently impede collaboration, owing to the lack of transparency regarding individual contributions that a joint project might entail. It is natural to conjecture that a policy that clarifies individual contributions to a joint project could be Pareto-enhancing. While it is unrealistic to expect that the precise individual contribution could be fully described, it is commonplace in many academic disciplines to use authorship order to signal ordinal contributions to a project. That comes with its own share of problems, as near-equal authors would then refrain from collaboration. However, Proposition 9 argues that a policy along the lines advocated by Ray ⓡ Robson, which explicitly randomizes authorship over a sub-domain of contributions but uses merit-based ranking for suitably unequal contributions, can Pareto-improve the equilibria of our model.

Proofs of all propositions in the paper are collected together in an Appendix.

Related Literature. As already mentioned, we intersect with the literature on statistical discrimination, starting from Myrdal (1944), Phelps (1972) and Arrow (1973).[3] One strand of the literature proposes models where an employer's holds distinct beliefs about the quality of potential hires based on payoff-irrelevant identities, despite their being equal in all payoff-relevant characteristics.[4] In turn, the difference in the employer's perceptions incentivizes different identities to make different pre-market investments in human capital, which in equilibrium confirm the employer's bias. Within that literature, our model is especially related to Moro and Norman (2004), who study a model of statistical discrimination in general equilibrium. In their model, people of different identities are hired by the same firm, and in asymmetric equilibria, one identity specializes in unskilled labor, while the other provides skilled labor.

Unlike the statistical discrimination literature, our theory of discrimination does not rely on individuals of different identities choosing different levels of pre-market investments in human capital. In fact, in our asymmetric equilibria, all agents produce ideas using the same technology. Ours is a theory of discrimination based on unequal credit attribution in collaborative outcomes. More fundamentally, we actively investigate the fragility of outcomes with equal

---

[2]See Moro and Norman (2004) for a discussion.

[3]Fang and Moro (2011) survey the literature. Some more recent relevant papers are Bohren, Imas and Rosenberg (2019), Bohren, Haggag, Imas and Pope (2021) and Bardhi, Guo and Strulovici (2020).

[4]See, for example, Coate and Loury (1993).

treatment across identities. Specifically, while unequal treatment is a possible outcome in the literature on statistical discrimination, we find that asymmetric equilibria are often the only robust outcomes. That said, contributions in other contexts consider the stability of symmetric and asymmetric equilibria. Chaudhuri and Sethi (2008) and Bowles, Loury and Sethi (2014) study segregation and group inequality, and find that discriminatory equilibria can be destabilized by social integration. In a model of statistical discrimination with sectoral choice, Gu and Norman (2020) argue that symmetric outcomes may be fragile, though with a different notion of fragility.

Our paper is also related to a literature that studies career concerns, credit attribution and the dynamic of reputations in teams. Anderson and Smith (2010) study a dynamic matching model that includes evolving public Bayesian reputations. They show that, despite productive complementarity, agents' desire to build a reputation leads assortative matching to fail, in contrast to the prediction in Becker (1973). Unlike Anderson and Smith (2010), our model takes a pair of potential contributors as given and studies their potentially unequal contributions to a joint project. In our model, agents' collaboration choices and the public's conjecture about collaborative patterns are co-determined in equilibrium, creating scope for multiple and potentially asymmetric equilibria even if partners are symmetric.[5] In Anderson and Smith (2010), the distribution of posteriors generated by a particular match is exogenously determined, and, in particular, updates are symmetric when partners are equal.

Ozerturk and Yildirim (2021) studies a model of team production where the market attributes credit for success based on its inference of individual efforts. They find that endogenous credit attribution may inefficiently affect individuals' effort decisions. Different from our model, Ozerturk and Yildirim (2021) assume that each person's productivity is commonly known, in order to isolate from effects of career concerns. In their model, the credit that is attributed to each agent determines who keeps what share of the physical outcome of the project. In our model, we take an opposite approach and study the interplay of endogenous credit attribution and career concerns — the public's conjecture precisely affects the signaling value of collaborative work.

Our paper is also related to the literature on incentive provision in teams, following Holmström (1982). Winter (2004) connects team production and discrimination. More specifically, Winter (2004) argues that, from the point of view of optimal incentives, differential rewards may be unavoidable even when individuals are completely identical. Chalioti (2016) studies career concerns in teams and finds that, to manipulate the market's assessment of their type, a worker has incentives to help or even sabotage her colleagues. Bar-Isaac (2008) considers the co-evolution of the reputation of workers and firms, and shows that working in teams creates incentives for both junior and senior team members to exert costly effort. There is also a connection to symmetry-breaking in general equilibrium models with imperfect capital markets, in which ex ante symmetric agents must make different occupational choices with consequent implications for economic inequality (see, e.g., Mookherjee and Ray 2002, 2003).

---

[5]Chade and Eeckhout (2020) study a model of team formation, where the formed teams later 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 individuals' actions creates scope for multiple equilibria with distinct matching patterns.

Finally, our paper contributes to the interpretation of empirical evidence on teamwork and discrimination in academia and other collaborative environments. Most prominently, Sarsons (2017) and Sarsons, Gërxhani, Reuben and Schram (2020) 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, even if partners are observationally the same. Ductor, Goyal and Prummer (2021) document gender disparities in research output and collaboration patterns in economics. Einav and Yariv (2006) document that credit attribution and career success may be impacted even by the prominence of authors' names in published papers.[6] With this evidence in sight, many papers propose policies to clarify authorial contributions in joint work — see Jones (2021) for a thorough discussion.

2. MODEL

There are two agents, each of whom perform a task either alone or in paired collaboration, and an observer, to be interpreted as the public. Each agent's type can be good (1) or bad (0). The observer's prior that an agent is good lies strictly between 0 and 1, but can vary over agents. These priors about others are public and commonly held by each agent, though a person's opinion about herself could differ from the public prior about her. That will not matter for behavior in our model. After observing the outcome of the task and whether it was executed alone or in collaboration, priors are updated by the observer. Each agent derives payoffs from the public posterior on her, as well as from the realized quality of the task. Details follow.

2.1. **Idea-Generating Process.** Each agent has an idea w, distributed with density $g(w, 1)$ for the good type and $g(w, 0)$ for the bad type with full support on all of $\mathbb{R}_+$. We assume that the former dominates the latter in the likelihood ratio order; that is,

(1) $$\frac{g(w, 1)}{g(w, 0)} \text{ is strictly increasing in } w.$$

Define, for each $w \in \mathbb{R}_+$ and $r \in [0, 1]$, $g(w, r) \equiv rg(w, 1) + (1 - r)g(w, 0)$. Then

$$b_r(w) \equiv \frac{g(w, 1)r}{g(w, r)}$$

is the *Bayes' update* on prior r and revealed idea w, yielding a posterior on agent type. By the likelihood ratio ordering assumption, $b_r(w)$ is increasing in w for each r.

2.2. **Intrinsic and Signaling Payoffs.** Unless otherwise noted, we refer to an agent with prior r simply as "r." Suppose that p has idea x, and q has idea y. Each agent sees both ideas, and chooses whether to collaborate or work alone. If both agents prefer to work together, collaboration takes place. If either would rather not collaborate, then both work alone. In making this decision, each agent seeks to maximize a combination of the *intrinsic value* of the project and its *signaling value*. If p and q collaborate with ideas x and y, the joint project has intrinsic value

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

---

[6]Ray ⓡ Robson (2018) propose "certified random" authorship order to mitigate prominence differentials.

where f is continuously differentiable, symmetric and strictly increasing, and where $f(x,0) = x$ and $f(0,y) = y$ are, respectively, the intrinsic value of p and q's projects if they work alone.

The signaling value of a joint project is given by the expected value of the observer's posterior about the agent's type, calculated "in equilibrium." That is, if a collaboration happens, the observer sees the outcome z, but not each idea x and y separately. The observer takes as given some *collaboration correspondence*

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

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

Such a correspondence, along with z, induces a measure $\nu(x,y;z)$ on combinations of x and y that could have led to the collaborative outcome z. For each such (x,y), there are two updates on the abilities of p and q, given by $b_p(x)$ and $b_q(y)$ respectively. The expected update ratio used by the observer averages these using ν:

$$\beta_p(z) = p \int_{(x,y) \in C(z,p,q)} b_p(x) d\nu(x,y;z), \text{ and likewise}$$

$$\beta_q(z) = q \int_{(x,y) \in C(z,p,q)} b_q(y) d\nu(x,y;z).$$

If p works alone and has outcome x, the observer assigns to them the update $b_p(x)$. No information regarding the counterfactual possibility of collaboration is required to perform this update.

Each agent separably values both project outcome and her reputation, the latter measured by the posterior on ability. For any agent with prior r, project value w and Bayes' update b, payoff is

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

where $\alpha > 0$ is the weight on project value w, and u is a smooth function with $u' > 0$.

Remark 1. Separability aside, the linearity of payoff in w is not a restriction provided we leave f unrestricted. Nonlinearities in payoff can be suitably folded into the joint production function f by a simple redefinition of variables.

Remark 2. The notation α is superfluous and could be normalized to 1. But we will sometimes be interested in the case of "small" intrinsic value, so that signaling becomes the principal concern. It will then be convenient to regards α as small rather than shift u up. We refer to this as the case of *approximately pure signaling*.

Remark 3. We retain the generality of u for some, but not all our results. We will often be interested in the *linear case* in which u is the identity function. But the potential generality may be useful in exploring other applications. For instance, a strictly concave u can approximate career concerns in which some minimal value of the posterior is sufficient for retaining a job; e.g., in teaching colleges where research considerations might be secondary (subject to being minimally satisfactory). On the other hand, a strictly convex u could approximate situations call for high reputational thresholds; e.g., a research university that prides itself on retaining

"stars." For a useful discussion of how dynamic considerations can impart particular reduced-form curvature to reputation, see, e.g., Anderson and Smith (2010).

2.3. **Equilibrium Definition.** A *equilibrium* at (z, p, q) is a collection

$$C(z, p, q) = \{(x, y) \big| f(x, y) = z \text{ and } \alpha w + u(b_r(x)) \leqslant \alpha z + u(\beta_r(z)) \text{ for } (w, r) = (p, x), (q, y)\}.$$

Sometimes, when there is no danger of confusion, we will also refer to an equilibrium at (z, p, q) as just an equilibrium, or as an equilibrium collaboration set.

3. EQUILIBRIUM

3.1. **Preliminary Characterization.** Rewriting the equilibrium condition, we see that p and q with ideas x and y (with $z = f(x, y)$) each prefer to collaborate when:

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

On the left-hand sides of (2) are the intrinsic gains from collaboration, relative to working alone. For given z, these are decreasing in x and in y. On the right-hand sides are the losses in signaling value that might arise from collaboration, which are increasing in x and in y, *given* some public perception of collaboration summarized by C. So there is some $\bar{x}$, depending on z and C, such that p agrees to collaborate if and only if $x \leqslant \bar{x}$, where we resolve indifference by collaboration. Or, equivalently, p agrees to collaborate if and only if agent q's idea is $y \geqslant \underline{y}$, where $f(\bar{x}, \underline{y}) = z$. The same is true of q, similarly generating bounds $\bar{y}$ and $\underline{x}$.[7]

The two panels of Figure 1 display these collaboration regions. As discussed, each region can equivalently be described by $\bar{x}$ or by $\underline{y}$ (as far as p's decision goes), and by $\bar{y}$ or by $\underline{x}$ (as far as q's decision goes). Any equilibrium collaboration set under z must be an interval of ideas for agent p, which is associated with a corresponding interval of ideas for agent q, with each pair of ideas generating z. We write this property compactly using the notation $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$.

The updates $\beta_p(z)$ and $\beta_q(z)$ can be rewritten to account for this specific equilibrium shape of C. Let $\iota_z(w)$ map each individual's idea $w \in [0, z]$ to her partner's idea $\iota_z(w)$ on the isoquant for z; i.e., $f(w, \iota_z(w)) \equiv z$. Define the conditional density that p has idea x, under the presumption that p and q *always collaborate* on joint project z, as

$$\gamma_z(x) \equiv \frac{g(x, p)g(\iota_z(x), q) \, |\iota_z'(x)|}{\int_0^z g(x', p)g(\iota_z(x'), q) \, |\iota_z'(x')| \, dx'} \text{ with associated cdf } \Gamma_z \text{ on } [0, z].$$

That is, knowing z, the density of x is given by $g(x, p)g(\iota_z(x), q) \, |\iota_z'(x)|$,[8] normalized by the term in the denominator to account for conditioning on z. Note that γ_z is a model primitive and not endogenous. If p and q collaborate only on $[\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$, then the conditional density of x

---

[7]This is a standard solution concept in models of networks which applies naturally here; see, e.g., Jackson and Wolinsky (1996).

[8]The density of the partner's idea at $\iota_z(x)$ is given by $g(\iota_z(x), p) \, |\iota_z'(x)|$, which is a standard transformation.
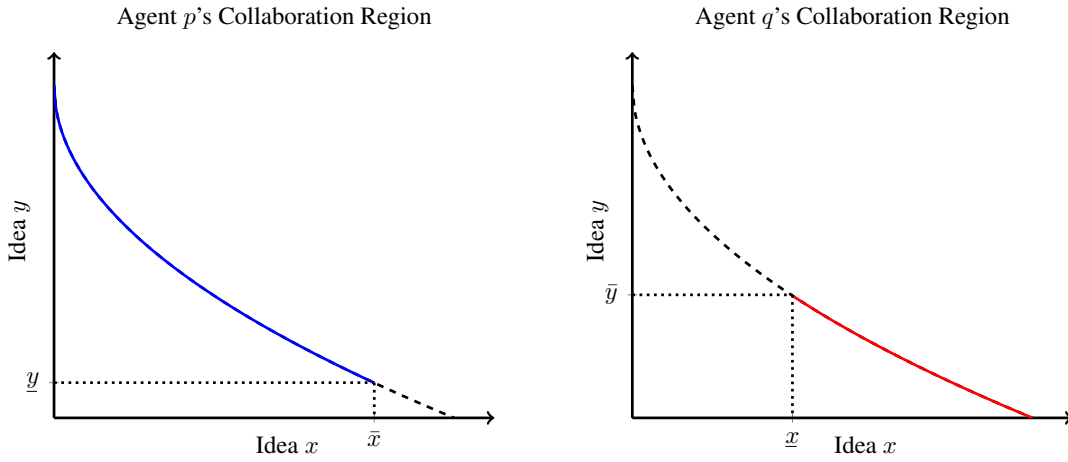
FIGURE 1. Equilibrium collaboration regions for agents p and q. The dashed curve displays all combinations of ideas x and y that yield a project z. On the left panel, in blue, are all the combinations such that p agrees to collaborate. On the right panel, in red, are all the combinations such that q agrees to collaborate.

is further adjusted to $\gamma_z(x)/[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})]$ (when $\bar{x} > \underline{x}$). It follows that

$$(3) \qquad \beta_p(z) = \begin{cases} \dfrac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \displaystyle\int_{\underline{x}}^{\bar{x}} b_p(x)\gamma_z(x)dx & \text{if } \bar{x} > \underline{x} \\ b_p(\bar{x}) & \text{if } \bar{x} = \underline{x} \end{cases}$$

Similarly, define another model primitive ω_z, the counterpart of γ for person q:

$$\omega_z(y) \equiv \frac{g(y,q)g(\iota_z(y),p)\,|\iota'_z(y)|}{\int_0^z g(y',q)g(\iota_z(y'),p)\,|\iota'_z(y')|\,dy'} \text{ with associated cdf } \Omega_z \text{ defined on } [0,z].$$

By the same logic leading to (3),

$$(4) \qquad \beta_q(z) \equiv \begin{cases} \dfrac{1}{\Omega_z(\bar{y}) - \Omega_z(\underline{y})} \displaystyle\int_{\underline{y}}^{\bar{y}} b_q(y)\omega_z(y)dy & \text{if } \bar{y} > \underline{y} \\ b_q(\bar{y}) & \text{if } \bar{y} = \underline{y}, \end{cases}$$

We summarize the discussion above as:

Proposition 1. *In any equilibrium at (z,p,q) with $C(z,p,q) \neq \varnothing$, there exist $\underline{x}, \bar{x}, \underline{y}, \bar{y}$ with $0 \leqslant \underline{x} \leqslant \bar{x} \leqslant z$ and $0 \leqslant \underline{y} \leqslant \bar{y} \leqslant z$, such that*

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

$$(6) \qquad\qquad \alpha(z - \bar{y}) = u(b_q(\bar{y})) - u(\beta_q(z)), \text{ and}$$

$$C(z,p,q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}],$$

where $\beta_p(z)$ and $\beta_q(z)$ are given by (3) and (4). If either $[\underline{x}, \bar{x}]$ or $[\underline{y}, \bar{y}]$ is nondegenerate, then both are, and in that case $0 < \underline{x} < \bar{x} < z$ and $0 < \underline{y} < \bar{y} < z$.

Conversely, if $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ for some $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ satisfying $0 \leqslant \underline{x} \leqslant \bar{x} \leqslant z$, $0 \leqslant \underline{y} \leqslant \bar{y} \leqslant z$ as well as (5) *and* (6)*, then $C(z, p, q)$ is an equilibrium at (z, p, q).*

Proofs of all propositions are in the Appendix.

Remark 4. In Theorem 1, we show that a nonempty equilibrium set always exists. An equilibrium could be empty-valued for some z, with p and q refusing to collaborate no matter what combination of ideas they have. Such an equilibrium must specify off-path beliefs in case a "surprise collaboration" is observed. However, if those beliefs assign probability 1 to any *one* combination of x and y, then both agents would be better off collaborating when ideas are x and y. With this restriction on off-path beliefs, equilibria with empty values cannot occur.[9]

Remark 5. We maintain $\alpha > 0$ throughout. If $\alpha = 0$, then signaling is the only concern, and by an unraveling argument, it is easy to see that only (and all) *singleton* sets $C(z, p, q) = \{x\} \times_z \{y\}$ with $x \in [0, z]$ and $y = i_z(x)$ are nonempty equilibria. As already mentioned, we will sometimes be interested in approximating this case, but always with $\alpha > 0$.

3.2. **Existence.** We establish the existence of a nonempty equilibrium.

Theorem 1. *For each $z > 0$ and $p, q \in (0, 1)$, a nonempty equilibrium exists: $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$, where $0 < \underline{x} < \bar{x} < z$ and $0 < \underline{y} < \bar{y} < z$.*

We provide a sketch of the proof; details are available in the Appendix. We rely on a suitable fixed point mapping on the domain of updates. Take as given p, q and $z > 0$. Define a domain $\mathbf{B} \equiv [b_p(0), b_p(z)] \times [b_q(0), b_q(z)]$. Obviously, all pairs of updates must lie in $\mathbf{B}$. Define $\Theta : \mathbf{B} \to \mathbf{B}$ as follows. For $(\beta_p, \beta_q) \in \mathbf{B}$, define $\bar{x}$ and $\bar{y}$ by (5) and (6) (with the numbers (β_p, β_q) in place of the functions $\beta_p(z)$ and $\beta_q(z)$ in those equations):

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

Next, let $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = \iota_z(\bar{x})$, and recover a new update vector (β_p', β_q') using (3) and (4). The difficulty with this construction is that as described, it is possible that $\underline{x} > \bar{x}$ or $\underline{y} > \bar{y}$, in which case (3) or (4) are not well-defined. We therefore modify the definitions of $\underline{x}$ and $\underline{y}$ by setting $\underline{x} = \min\{\bar{x}, \iota_z(\bar{y})\}$ and $\underline{y} = \min\{\bar{y}, \iota_z(\bar{x})\}$, and then proceed with (3) and (4), which are now well-defined. That yields a composite mapping $(\beta_p', \beta_q') = \Theta(\beta_p, \beta_q)$. It is continuous and has a fixed point. A non-trivial step is then to show that at any such fixed point, the corresponding values $(\bar{x}, \bar{y}, \underline{x}, \underline{y})$ will indeed satisfy $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = \iota_z(\bar{x})$. The Appendix contains the details of this argument, and additionally shows that $0 < \underline{x} < \bar{x} < z$ and $0 < \underline{y} < \bar{y} < z$.

3.3. **Collaboration and Signaling in Equilibrium.** There is a tension between the value of collaboration and the private desire to signal. To bring this out, let us write the equilibrium

---

[9]Without this constraint — that is, if off-path beliefs assign probabilities to several pairs (x, y) — then an empty-valued equilibrium might exist, though that would depend on the curvature of the "production function" f. We do establish the existence of a nonempty-valued equilibrium correspondence, and will have nothing else to say about the empty case for the rest of the paper.

payoffs to p and q from posterior reputation as follows:

$$(8) \qquad u_p^*(x) = \begin{cases} u(\beta_p(z)) & \text{if } x \in [\underline{x}, \bar{x}] \\ u(b_p(x)) & \text{if } x \notin [\underline{x}, \bar{x}] \end{cases} \text{ and } u_q^*(y) = \begin{cases} u(\beta_q(z)) & \text{if } y \in [\underline{y}, \bar{y}] \\ u(b_q(y)) & \text{if } y \notin [\underline{y}, \bar{y}] \end{cases}$$

Conditional on both players following their equilibrium strategies, player p's payoff is given by

$$\pi_p \equiv \int_0^{\underline{x}} [\alpha x + u(b_p(x))]\gamma_z(x)d + [\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})][\alpha z + u(\beta_p(z))] + \int_{\bar{x}}^z [\alpha x + u(b_p(x))]\gamma_z(x)dx$$

$$= \alpha \left[\int_0^{\underline{x}} x\gamma_z(x)dx + [\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})]z + \int_{\bar{x}}^z x\gamma_z(x)dx \right] + \int_0^z u_p^*(x)\gamma_z(x)dx$$

$$(9) \quad = \alpha \int_0^z x\gamma_z(x)dx + \int_0^z u_p^*(x)\gamma_z(x)dx + \alpha \int_{\underline{x}}^{\bar{x}} (z - x)\gamma_z(x)dx,$$

where we use (8) to obtain the last line in (9). (A parallel expression holds for individual q.) The first term is an individual-specific constant, unaffected by equilibrium strategy. The second term is the expected payoff from reputation. The third term represents the expected intrinsic gains from collaboration. All expectations are taken over individual ideas, conditional on z.

Suppose that u is linear. Then the expected reputational payoff is just the expected posterior starting from a prior of p. By the martingale property of Bayes' updates, this term must *equal* the prior, and so also becomes an individual-specific constant. Indeed, *all* the private and social gains from pairwise interaction come from the intrinsic value of collaboration. With this point made transparent, it is clear that the same is true a fortiori for the case of concave reputational utility. Collaboration is additionally useful because it creates a reduction in the spread of Bayes' updates over some range of ideas; 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 that degree of collaboration is precluded due to a lack of commitment. Under full collaboration, signaling value is transferred across realizations with high x and low y, both generating the same joint value z. By Bayes' plausibility, these transfers of signaling value wash out on average. However, suppose that an agent has an excellent idea and is matched with a partner with a bad idea. From that ex-post perspective, the agent with the good idea understands that the intrinsic gain from collaboration may not overcome the loss of signaling value. Therefore, while collaboration is always valuable in terms of its intrinsic payoff, it will not always happen. [10]

4. FRAGILE EQUILIBRIA

A central theme of our paper concerns the possibility of asymmetric treatment of individuals who are similar in all payoff-relevant characteristics but differ on other observable features that are payoff-irrelevant to the interaction at hand. These could include gender, nationality, age, or

---

[10] When u is not concave, full collaboration will generally not be socially desirable. Because utility is strictly convex in reputation over some zones, the insurance aspect of collaboration mentioned above is removed. Individual agents may be worse off ex ante under full collaboration. The convexity of utility pushes them to confront individual variation in updates, which makes solo research more valuable. It is still true, though, that equilibria will generally fail to generate the socially optimal level of collaboration.

race. We will refer to such features as *identities*. We wish to trace this asymmetric treatment to a notion of "fragility" in their symmetric interaction.

To fix ideas, consider an equilibrium at (z, p, q) with associated collaborative updates β_p and β_q. Let us (temporarily) assume that $p = q$, so that both players are equal in all payoff-relevant characteristics, and that the equilibrium in question is symmetric, with common update ratio β in the event of collaboration. We shall nevertheless continue to use the notation p and q to keep track of the payoff-irrelevant identities of each agent. Suppose that the public sees these identities as salient for some reason and "slightly reallocates" its credit, say 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. We are interested in how each player reacts to this small anticipated imbalance.

Controlling for the prior, which is the same across the two individuals, p would now be more open to cross-identity collaboration, thereby benefiting from the more generous public update. For exactly the same reasons but in reverse, q will be *less* open to cross-identity collaboration. In summary, and in response to this bias, we have $\bar{x} > \bar{y}$, and therefore $\underline{x} > \underline{y}$. To some degree, that confirms the public's initial bias. But more than that, if these echo effects lead to new collaboration zones that "overshoot" the original bias, they may destabilize the symmetric outcome. In this way, payoff-irrelevant notions of identity could precipitate an asymmetric outcome across functionally identical individuals.

We formalize this verbal discussion using the map Θ introduced in the proof of Theorem 1, and no longer presume that $p = q$. Remember that Θ starts with updates (β_p, β_q), and then uses (7) to generate best-response collaboration thresholds, $\bar{x}$ and $\bar{y}$. These thresholds are then mapped into lower bounds $\underline{x}$ and $\underline{y}$, and finally into a new update vector (β_p', β_q') consistent with the implied collaboration correspondence.[11] We will say that an equilibrium with update vector (β_p, β_q) is *p-fragile* if there is $\zeta > 0$ and $\delta > 0$ such that for every $\epsilon \in (0, \delta)$,

(10) $\qquad \Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) \geqslant \beta_p + (1 + \zeta)\epsilon$ and $\Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) \leqslant \beta_q - (1 + \zeta)\epsilon,$

where the subscripts on Θ refer to its component functions. We analogously define q-fragility.

In the above definition, we perturb public perceptions by "reallocating" updates across identities. Locally, we think of an increase in update credit for some identity as matched by a symmetric decrease for the other identity — hence our use of $+\epsilon$ and $-\epsilon$. Furthermore, the requirement of $\zeta > 0$ in the definition insures that small perturbations not only locally amplify but that they do so at some minimal geometric rate. Alternatively p-fragility could be defined by the less demanding requirement that $\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) > \beta_p + \epsilon$ and $\Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) < \beta_q - \epsilon$, instead of (10). But this would have forced us to confront technical yet non-generic situations which create complications of little conceptual interest in the present setting.[12]

The following observation is useful in both simplifying and interpreting fragility:

Observation 1. *If an equilibrium is p-fragile, then it is q-fragile.*

---

[11]For mathematical convenience, this map did not write $\underline{x}$ and $\underline{y}$ as the isoquant images of $\bar{x}$ and $\bar{y}$ respectively, but as mentioned above and shown explicitly in the proof of Theorem 1, these equalities *do* hold at every equilibrium and in a neighborhood of every equilibrium. For our purposes, that is all we need.

[12]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. We ignore such issues.

Observation 1 permits us to simply refer to an equilibrium as *fragile* with respect to public perceptions, knowing that p-fragility is equivalent to q-fragility. More importantly, it permits an unconstrained interpretation of fragility using "identities." In contrast, if an equilibrium were p-fragile but not q-fragile, it could be eliminated by a public that is biased toward's p's identity, but not by the same public if those identities were switched across p and q. Given Observation 1, the equilibrium would be fragile against *any* assignment of identities as long as the public is slightly biased in favor of one of them.

More generally, our formal concept of fragility rests on the existence of identities that are conducive to differential treatment for reasons of predisposed bias or perceptions of historical inequality. If two agents were identical in all ways that could be conceivably regarded as salient, the mapping Θ might still be fragile, but that "mathematical fragility" would not translate itself into asymmetric treatment — a salient pair of identities is needed to anchor the asymmetry.

The following proposition augments Theorem 1, establishing the existence of a non-empty equilibrium that is robust with respect to public perceptions.

Proposition 2. *There exists an equilibrium with the same properties as those in Theorem 1, which is also not fragile.*

The possibly unequal treatment of equals has received extensive attention in the literature on statistical discrimination, starting from Myrdal (1944), Phelps (1972) and Arrow (1973). In theories of statistical discrimination, unequal treatment is one equilibrium, but there could be an equally "robust" equilibrium with equal treatment. Our approach is different, in that it actively investigates the fragility of the equal-treatment outcome. In part, this is possible because the two players actively interact, and beliefs about the one must directly map into the beliefs about the other — they are not independent. It is possible that similar progress can be made in other settings of statistical discrimination.[13] For other differences, see Proposition 6 in Section 6.

5. Equilibria with Symmetric Partners

With the above discussion in mind, we now study the case of symmetric players, who possess identical priors. We investigate whether such symmetric agent-pairs might nevertheless be pushed into asymmetric yet robust interactions that rely on payoff-irrelevant identities. Assume, then, that $p = q$. It is natural to begin by looking for *symmetric* equilibria.

5.1. Symmetric Equilibria. We establish the existence of a symmetric equilibrium.

Proposition 3. *Suppose that $p = q = r$. Then there is a symmetric equilibrium at (z, r, r) in which each individual employs identical collaboration thresholds.*

The proof, provided in the Appendix, adapts the map Θ from Theorem 1, by additionally imposing $\beta_p = \beta_q$ throughout. Call it Θ^S. For any β_r over a certain domain described in the proof, Θ^S defines $\bar{x}$ using (5) as we did for Theorem 1, then $\underline{x}$ via the isoquant $\iota_z(\bar{x})$, and finally β_r' by

---

[13]In this sense, our concept is perhaps more closely related to the Battle of the Sexes, where the mixed strategy equilibrium is "unstable," though in our setting all the equilibria are pure, with strict best responses a.e.

$\beta_r(z)$ as in (3) or (4). Two features of Θ^S deserve special mention in the light of the discussion to come. First, Θ^S *imposes* symmetry. By mapping $\bar{x}$ to $\underline{x}$ directly, it presumes that the other player — who really determines $\underline{x}$ — is behaving in identical fashion. Second, the map satisfies appropriate end-point conditions: β_r' lies above β_r for low values of β_r, and below it for high values. Continuity assures us that a fixed point exists, which is then shown to be a symmetric equilibrium. In fact, *every* symmetric equilibrium is identifiable with some fixed point of Θ^S.

We remark that multiple symmetric equilibria may exist, although we view uniqueness as the leading case; see Proposition 4(i) for more on uniqueness. When u is concave, multiple symmetric equilibria — if they exist — are identically payoff ranked *both* by their collaboration zones and by their equilibrium update ratios. To see that, note that if β_r^1 and β_r^2 are update ratios for two symmetric equilibria, with $\beta_r^1 < \beta_r^2$, then $\bar{x}^1 < \bar{x}^2$ and $\underline{x}^1 > \underline{x}^2$. Therefore the latter equilibrium has more collaboration and consequently higher intrinsic value than the former. Moreover, the distribution of updates is squeezed by collaboration, leading to further payoff gains when u is strictly concave. Multiple symmetric equilibria are therefore akin to Pareto-ranked equilibria in coordination games.

We now turn to investigating the fragility of symmetric equilibria.

5.2. **Fragility of Symmetry.** Given the end-point conditions on Θ^S, there is always some symmetric equilibrium at which Θ^S intersects the 45° line "from above." That appears to suggest that there is always some non-fragile symmetric equilibrium. After all, under Θ^S, a perturbation of the equilibrium update leads to a sequence of update ratios that returns to the equilibrium ratio.

But this seeming robustness is misleading. Because Θ^S imposes symmetry across agent behavior, it fails to capture the fact that an individual's best response is generally to collaborate *more* when her partner collaborates *less*. This behavioral feature, central to our definition of fragility, is missing from the symmetric fixed point map Θ^S. It is, however, perfectly well-contained in the higher-dimensional map Θ that we used to define fragility.

The following theorem fully characterizes fragile symmetric equilibria.

Theorem 2. *A symmetric equilibrium collaboration set $[\underline{x}, \bar{x}] \times_z [\underline{x}, \bar{x}]$, with associated update ratio β_r, is fragile if and only if*

$$(11) \qquad u'(\beta_r)\frac{[b_r(\bar{x}) - \beta_r] - \iota_z'(\bar{x})[\beta_r - b_r(\underline{x})]}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha} > \frac{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})}{\gamma_z(\bar{x})}.$$

Condition (11) can be roughly interpreted as comparing as comparing the curvatures of two sets of functions. Such a comparison determines whether agents' responses to small biases in the observer's perception are sufficiently strong so as to "more than justify" the observer's initial bias and destabilize the equilibrium. To see this connection, consider the case of a linearly additive production function, so that $f(x, y) = x + y$. Then $\iota_z'(\bar{x}) = -1$, and (11) is equivalent to:

$$(12) \qquad u'(\beta_r)\frac{[b_r(\bar{x}) - b_r(\underline{x})]/[\bar{x} - \underline{x}]}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha} > \frac{[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})]/[\bar{x} - \underline{x}]}{\gamma_z(\bar{x})},$$

where we've divided by the common term $\bar{x} - \underline{x}$ on either side. On the right hand side, then, we have a measure of the curvature of Γ_z: the ratio of the "chord-slope" connecting $\Gamma_z(\underline{x})$ and $\Gamma_z(\bar{x})$ to the "point-slope" of Γ_z at $\bar{x}$. On the left, we have a similar ratio, this time applied to b_r weighted by reputational marginal utility. An additional asymmetry is that the payoff term α also appears in the comparison.

5.3. **Signaling and Fragility.** In this section, we argue that symmetric outcomes are often — though not always — fragile when signaling is approximately pure, and the utility u from reputation is weakly concave. Recall that this refers to a setting where the weight α placed on the intrinsic value of the project is small. Throughout this section, we presume that the production function is linear: $f(x + y) = x + y$. The following observation identifies a distinctive set of joint values that we shall focus on.

Observation 2. *There is an open set J such that for every joint value $z \in J$,*

(13)
$$- b_r''(e)e > b'(e),$$

where e is the unique solution to $f(e, e) = z$.

We comment on this set J in more detail below, but first we note:

Proposition 4. *Suppose that $f(x, y) = x + y$. Then for each $z \in J$ and $p = q = r \in (0, 1)$, there is $\underline{\alpha} > 0$ such that for every $\alpha < \underline{\alpha}$:*

(i) *There is a unique symmetric equilibrium collaboration set under (z, r, r).*

(ii) *If u is weakly concave, $\underline{\alpha} > 0$ can be chosen so that each such equilibrium set is fragile.*

Under (a) the linearity of f, (b) the concavity of signaling payoff u, (c) a high weight on signaling (α small), and (d) joint output lying in a certain set J (more in Remark 6 below), Proposition 4 argues that the symmetric equilibrium is *both* unique and fragile. That combination makes for a strong assertion. In this case, it is not that pairs of agents can take refuge in some other symmetric equilibrium when a particular equilibrium is fragile — there *are* no others. Yet, given Proposition 2, we know that there are indeed other non-fragile equilibria, and so they must involve asymmetric treatment. So if signaling considerations are uppermost, there is a real danger that symmetric players will not be treated symmetrically, especially when other payoff-irrelevant features of their identity are evident to the public.

Now we attempt to intuitively explain Proposition 4. First, approximately pure signaling implies that the collaboration zones under any equilibrium are small. The proof of Proposition 4 will verify this formally, but an inspection of (5) should clarify the argument. The thresholds $\underline{x}$ and $\bar{x}$ edge very close to each other under approximately pure signaling, and perforce (under symmetry) to the *equal input threshold* $e(z)$ for x, defined by $f(e(z), e(z)) = z$. Additionally, either partner is equally likely to have contributed x as she has x', where x and x' are bound by $f(x, x') = z$. In short, with f linear, γ_z must be symmetric around $e(z)$. That suggests that the curvature for Γ_z, on the right hand side of (12), must vanish as $\underline{x}$ and $\bar{x}$ approach each other, or as they each converge to $e(z)$. The proof in the Appendix formally verifies this claim.
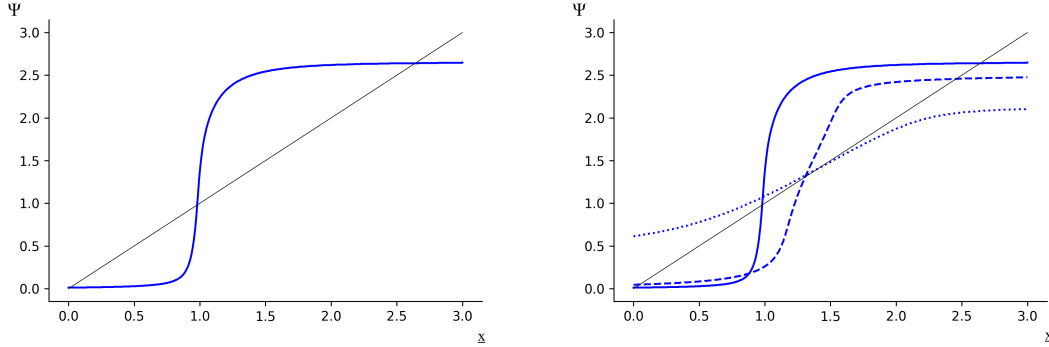
FIGURE 2. The map Ψ in the Exponential Model. The parameters in this simulation are given by $\lambda_0 = 4$, $\lambda_1 = 1$, $\alpha = .1$, and $z = 3$. On the left panel, in blue, we plot Ψ when $p = q = .5$. The thin black line is the $45°$ line. On the right panel, the blue plots are Ψ when $p = q = .5$ (solid), $p = q = .2$ (dashed) and $p = q = .05$ (dotted).

On the other hand, while the update function b_r on the left hand side of (12) is increasing, it has no particular symmetry around $e(z)$ or anywhere else. But this function is bounded above, because it equals a probability, and so must have the feature that it is "at least occasionally strictly concave," otherwise it will overshoot its bound. Indeed, it is occasionally "more than strictly concave," as described by (13). And when z happens to lie in the set J, with $e(z)$ satisfying (13), then we prove that the left hand side retains non-zero curvature even as α vanishes, thereby allowing it to increase with α at 0 and satisfying the fragility condition under approximately pure signaling. Any concavity in u only adds to this tendency.

Remark 6. If we endow the model with functional forms for the distributions of ideas, we can further characterize the set J. For example, in Section 5.4, we find that if ideas are exponentially distributed for both good type and bad type agents, then condition (13) is satisfied on a set $J = [\underline{z}, \infty)$, for some $\underline{z} \in \mathbb{R}_+$. The set J is also of the form $[\underline{z}, \infty)$ when ideas are distributed according to two Weibull distributions with a common shape parameter, or according to two log-normal distributions.

5.4. **An Example: Exponential Distribution of Ideas.** Suppose that ideas are drawn from exponential distributions with parameters λ_1 and λ_0 for good and bad types respectively, where $0 < \lambda_1 < \lambda_0$. The update ratio function for an agent with prior r and idea w is given by:

$$b_r(w) = \frac{1}{r + (1-r)\frac{g(w,0)}{g(w,1)}} = \frac{1}{r + (1-r)\frac{\lambda_0}{\lambda_1}e^{-(\lambda_0-\lambda_1)w}}.$$

Because $\lambda_0 > \lambda_1$, the distribution of ideas of the good type dominates the one of the bad type in the likelihood ratio order, and so b_r is increasing in w. Assume, moreover, that the technology that combines the ideas is additive and linear, so $f(x,y) = x + y$, and that the signaling payoff is linear — u is the identity function.

Because the fixed points of Θ must be visualized in four-dimensional space, it will be convenient to employ a different map, less powerful for the formal analysis but more useful for this exposition. This map, Ψ, composes four functions as follows:

$$\Psi(\underline{x}) = \iota_z \circ \phi_q \circ \iota_z \circ \phi_p(\underline{x}).$$

It begins, for each $\underline{x}$, by choosing $\phi_p(\underline{x})$ as some value of $\bar{x}$ for player p that satisfies (5), then finds y via the isoquant map, then chooses $\phi_q(y)$ as the value of $\bar{y}$ for player q that satisfies (6), and finally maps back to $\underline{x}$ via the isoquant map. Analogously to the fixed point map Θ, any fixed point of Ψ defines an equilibrium. Figure 2 displays some numerical simulations of Ψ in our exponential example, when both agents start with identical priors; that is, $p = q$. The parameters in this simulation are given by $\lambda_0 = 4$, $\lambda_1 = 1$, $\alpha = .1$, and $z = 3$.

In the left panel, when $p = q = .5$, Ψ crosses the $45°$ line three times, implying three different fixed points: the first crossing corresponds to an asymmetric equilibrium where q has the favored identity, the second crossing defines a symmetric equilibrium and the third crossing pertains to an asymmetric equilibrium where p has the favored identity. Observe that at the symmetric equilibrium, Ψ crosses the $45°$ line from below, with a slope that exceeds 1. Indeed, it can be shown that when Ψ is a well-defined and differentiable function, this crossing pattern is equivalent to condition (11) that notes the fragility of the symmetric equilibrium. [14]

On the right, we see that the pattern of three equilibria is maintained for a range of values of the agents' priors. However, for very low priors, the two asymmetric equilibria do not exist, and there is a unique symmetric equilibrium. Numerically simulating the exponential model, we "verify" that these are the two possible patterns — either three equilibria, or a unique symmetric equilibrium. In all cases, there is just one equilibrium that issymmetric. Figure 3 depicts the parameter areas where each of the patterns holds, confirming that, as stated in Proposition 4, the symmetric equilibrium is fragile when α is small and z is large.

6. Payoff Implications of Equilibria

Section 4 introduced the notion of fragility, and Section 5 described conditions under which symmetric equilibria across similar agents are fragile. When that happens — see condition (11) in Theorem 2 and especially Proposition 4 — Proposition 2 assures us that other non-fragile *asymmetric* equilibria exist. If society can distinguish between agents using otherwise-irrelevant identities, functionally identical individuals will settle into such equilibria. Because those asymmetries are supported by the existence of payoff-irrelevant identities, each identity will choose to collaborate for distinct sets of ideas. In particular, one of the identities will be *favored*, in the sense that it will be perceived by the public as (stochastically) contributing better ideas to the collaboration.[15]

---

[14]In general, ϕ_p and ϕ_q — and therefore Ψ — are correspondences and not functions. But for our numerical parameters in the exponential case, they are uniquely defined.

[15] In fact, if an asymmetric equilibrium $[\underline{x}, \bar{x}] \times_z [y, \bar{y}]$ is ascribed to symmetric players, then one identity (say p's) must be favored, with $\underline{x} > y$, $\bar{x} > \bar{y}$ and $\beta_p > \beta_q$. That is, the comparison is unambiguous: p is viewed as stochastically contributing the better ideas to the collaborative outcome.
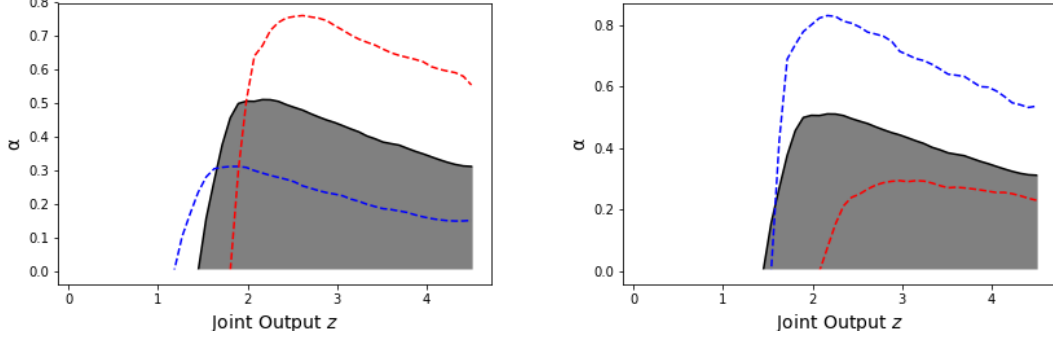
FIGURE 3. The black area in both panels depicts (α, z) combinations for which the symmetric equilibrium is fragile, when $\lambda_0 = 4$, $\lambda_1 = 1$ and $p = q = .5$. In the left panel, the red and blue dashed lines are the fragility boundaries when $p = q = .3$ and $p = q = .9$, respectively. In the right panel, the red and blue dashed lines are the fragility boundaries when $(\lambda_0, \lambda_1) = (3, 1)$ and $(\lambda_0, \lambda_1) = (4, .5)$, respectively.

In this section, we discuss the notion of favored identities, and the payoff implications of favoritism. Of course, such a favored identity benefits — almost by definition — from the signaling aspects of collaboration. But matters are more complicated when not just signaling payoffs but also the intrinsic payoffs of collaboration are taken into account.

6.1. **Intrinsic Gains from Collaboration.** In equation (9) of Section 3.3, we showed that conditioning on z, agent p's payoffs are

$$\pi_p = \int_0^z u_p^*(x)\gamma_z(x)dx + \alpha \int_0^z x\gamma_z(x)dx + \alpha \int_{\underline{x}}^{\bar{x}} (z - x)\gamma_z(x)dx.$$

The first of these terms have to do with the signaling value of collaboration, and the second and third with its intrinsic value. We focus on the latter to begin with, and write these separately as

$$(14) \qquad \pi_p^I(z) = \left[\alpha \int_0^z x\gamma_z(x)dx\right] + \alpha \int_{\underline{x}}^{\bar{x}} (z - x)\gamma_z(x)dx, \text{ and}$$

$$(15) \qquad \pi_q^I(z) = \left[\alpha \int_0^z y\omega_z(y)dy\right] + \alpha \int_{\underline{y}}^{\bar{y}} (z - y)\omega_z(y)dy,$$

where the superscript I stands for "intrinsic." Our particular formulation may look odd because it supposes that z is known but not x and y. But we could easily integrate this payoff over all z without adding or eliminating anything of substance, and so stick to this formulation.

We now compare collaboration gains across agents and across equilibria, first focusing on the symmetric case in which agents have the same prior: $p = q = r$. Proposition 5 shows that, in an asymmetric equilibrium with symmetric agents, the favored identity receives lower intrinsic payoff than the dis-favored identity.

Proposition 5. *In an asymmetric equilibrium at z ascribed to agents with a common prior $p = q = r$, where p has the favored identity,*

$$\pi_p^I(z) < \pi_q^I(z),$$

so that p, while favored, receives a lower intrinsic payoff from collaboration.

Before embarking on a discussion of this result, we note that its essence extends beyond the symmetric setting. To see this, we first need to extend the concept of favoritism to allow for cases where agents are not functionally identical. One possible approach is to say that p is *super-favored* in some equilibrium at (z, p, q) if $\underline{x} > \bar{y}$. That is, the worst possible idea that p brings to the joint collaboration is viewed as better than q's best possible idea. This is unambiguous, given that p and q have idea distributions that are always overlapping, even when $p \neq q$, but restrictive. (It will play a role below, though.)

A more robust notion is to conceive of agents as being relatively favored or dis-favored *across* equilibria. Specifically, consider two distinct equilibria, denoted 1 and 2, and two individuals p and q with distinct identities. Say that p — or p's identity — is *relatively favored* (and q *relatively dis-favored*) in equilibrium 1 relative to 2 if $\pi_p^P(z, 1) > \pi_p^P(z, 2)$ and $\pi_q^P(z, 1) < \pi_q^P(z, 2)$.

Additionally, in the asymmetric case, individuals have different "baseline payoffs": in equations (14) and (15), observe that for each agent, the bracketed term is a person-specific constant. So it be useful to compare intrinsic payoffs as *gains* by netting these terms out:

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

Proposition 6. (i) *If p is super-favored in some equilibrium with joint output z conditional on collaboration, then $\Delta_p^I(z) < \Delta_q^I(z)$. That is, p obtains a lower intrinsic 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 $\Delta_p^I(z, 1) - \Delta_p^I(z, 2) < \Delta_q^I(z, 1) - \Delta_q^I(z, 2)$: q's gain in intrinsic payoff in moving from equilibrium 2 to 1 is larger than p's gain.*

Propositions 5 and 6 together make the point that persons or identities favored by the public in the perception of their collaborative contributions are actually worse off in terms of their intrinsic payoff gains from collaboration. Being favored means that the public singles out a particular individual; that is, his *type* when $p \neq q$, or his *identity* when $p = q$, and gives him better treatment for the contribution of ideas. That very treatment needs to be justified in equilibrium, so that individual indeed contributes the better ideas, incentivized by the public bias. But it is precisely because the favored individual on average contributes better ideas than their partner, that he loses out on the intrinsic gains from collaboration.

We reiterate that none of this hinges on z being known — indeed, as already noted, z cannot be known if x and y are yet to be realized. But it does not matter. For instance, in the symmetric case, we simply integrate payoffs over z, picking out the symmetric equilibrium in case it is not fragile and replacing it with an asymmetric equilibrium that favors some given identity in

case it is fragile. The same result holds. The question then arises: what implications does this observation have for the *overall* implications of favoritism?

6.2. **Overall Gains With Linear Reputational Payoff.** When reputational payoff is linear, those payoffs must be equal across all equilibria. This is a consequence of the martingale property of Bayesian updates. That implies the following corollary to Proposition 5. To state it, let π_p^C and π_q^C denote the payoffs to p and q, conditional on collaboration taking place at z; that is, $\pi_p^C(z) = \alpha z + u(\beta_p(z))$ and $\pi_q^C(z) = \alpha z + u(\beta_q(z))$.

Corollary 1. *Suppose that reputational payoff is linear. Then in any asymmetric equilibrium at z across agents with a common prior $p = q = r$, where p has the favored identity,*

$$\pi_p^C(z) > \pi_q^C(z), \text{ but } \pi_p(z) < \pi_q(z),$$

so that p is better off conditional on collaboration, but q receives higher overall expected payoff.

In Figure 4, Corollary 1 is confirmed in simulations of the exponential example from Section 5.4. That is, the figure shows that the favored identity – in blue in either panel – is better off, relative to the dis-favored identity (in red) conditional on collaboration, but worse off in terms of overall expected payoffs.

These results contrast sharply with the literature on statistical discrimination. That literature typically finds either that discrimination does not affect the favored group — the one favored by public beliefs — or that the favored group benefits from discrimination.[16] Furthermore, to the best of our knowledge, the observation that the payoff ordering may be reversed between the reputational and the intrinsic perspectives is a novel contribution.

Specifically, and again in contrast to the traditional literature, discrimination in our model stems from a collaborative interaction which depends on willing participation by both agents. However, when public perception favors one individual's identity at the expense of another's, there are two effects on the value to the favored agent. The direct and positive effect is that conditional on collaboration, signaling value favors the favored identity (by construction). On the other hand, the dis-favored identity becomes less willing to collaborate. This has a negative effect on payoff to the favored identity. By Bayes' plausibility, when reputational payoffs are linear, the first effect *must* be dominated by the second.

6.3. **Majority Identities and Favoritism.** To summarize the discussion so far: a favored identity receives larger credit conditional on collaboration. But that credit needs to be justified in equilibrium. In that process, the favored identity loses out on the intrinsic gains from collaboration. In the special case with linear reputational payoffs, the net gain is generally negative.

It should be noted the payoff gains and losses reported in Section 6.2 hold fixed the identity of the partner. But there are also questions of matching across partners. Specifically, consider a population version of the model. There is some positive measure of symmetric agents, all with $p = q$. Each is randomly paired to one potential collaborator. Following this pairing, our model proceeds as before.

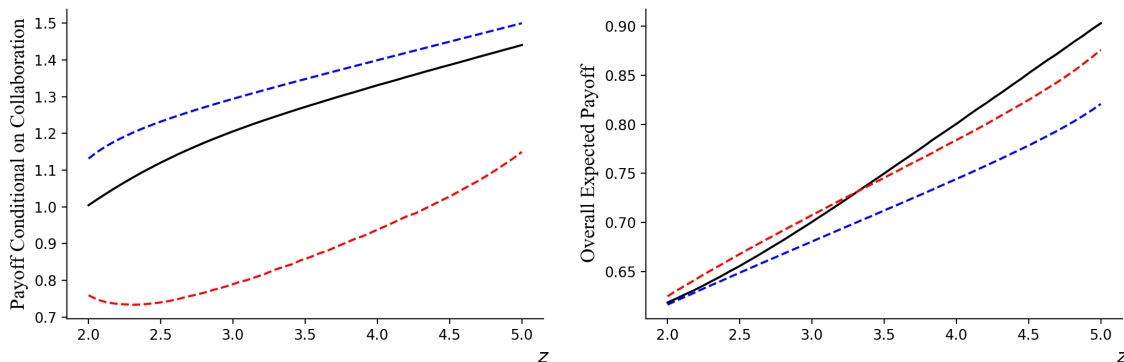[16]For a discussion, see Moro and Norman (2004).

FIGURE 4. Payoffs in Symmetric and Asymmetric Equilibria of the Exponential Model. The solid black lines indicate equal payoffs in the symmetric equilibrium. The blue and red dashed lines in each panel represent payoffs to the favored and disfavored identities in an asymmetric equilibrium. The favored identity is better off conditional on collaboration (left panel), but is worse off in terms of overall payoffs (right panel).

Suppose that this population is made up of a small fraction σ of persons of a minority identity (labeled by q) and a large fraction $(1 - \sigma)$ of a majority identity (labeled by p). Suppose, moreover, that whenever the symmetric equilibrium is fragile, two agents of different identities engage in an asymmetric equilibrium in which the majority identity is favored. Otherwise, or if two agents of the same identity meet, they play the symmetric equilibrium. Then ex-ante payoff π^a for each identity is given by

$$(16) \qquad \pi_p^a \equiv \sigma \int_z \pi_p(z) + (1 - \sigma) \int_z \pi(z) \quad \text{and} \quad \pi_q^a \equiv \sigma \int_z \pi(z) + (1 - \sigma) \int_z \pi_q(z),$$

where $\pi_p(z)$ and $\pi_q(z)$ are the expected payoffs on cross-identity matches and $\pi(z)$ is the symmetric expected payoff. In line with our earlier remarks, we use the mathematical device of conditioning on but then integrating over z. Now consider the exponential example displayed in the right panel of Figure 4, which depicts symmetric and asymmetric ex-interim payoffs, conditioning for different values of z. Observe that there is a threshold for z above which the symmetric payoff dominates *both* the asymmetric payoffs for p and q.

In any such situation, we are potentially confronted by yet another reversal in payoffs. The minority identity faces a larger share of cross-identity matches relative to the majority identity. It is true that Propositions 5 and 6 continue to hold for each cross-identity match, so that the disfavored identity benefits from intrinsic payoffs, conditional on each encounter. Nevertheless, the *ex ante* payoff to the minority identity could be lower by the fact that the symmetric equilibrium is Pareto dominant. All other things equal, the smaller the dis-favored minority, the more likely it is that a second payoff reversal could occur from this ex ante perspective. We summarize this discussion as:

Proposition 7. *Consider the symmetric random matching model with linear reputational payoffs. Suppose that expected ex-ante payoff under the symmetric equilibrium dominates expected pay-offs to the disfavored minority; that is, $\pi^a > \int_z \pi_q(z)$. Then for all σ small, $\int_z \pi_p^a(z) > \int_z \pi_q^a(z)$, even though under each match and each z $\pi_q^I(z) > \pi_q^I(z)$ as in Proposition 5.*

If, on the other hand, $\pi^a \leqslant \int_z \pi_q(z)$, then the ranking $\pi_q^I(z) > \pi_p^I(z)$ is retained ex-ante no matter what the value of σ is.

One way to interpret homophily is as an indicator that the conditions for Proposition 7 are met. To clarify, suppose that we empirically observe homophily, i.e., that people seek out and collaborate disproportionally with individuals of their own identity. That would suggest that asymmetric equilibria across identities generate a disincentive for cross-identity collaboration, while symmetric equilibria played within identities have the opposite effect.

In summary, then, discrimination and favoritism have complex implications. Ex post, conditional on collaboration, the favored identity gains. Taking a step back, and conditioning only on the potential partner and not the realization of collaborative, the term "favored identity" could become a misnomer as the their intrinsic expected payoffs from collaboration are lower. And finally, the dominance of symmetric payoffs could cause yet another payoff reversal as described in this section, when the favored identity also happens to be in the majority.

6.4. **Symmetric and Asymmetric Distributions of Posteriors.** Return finally to nonlinear reputational payoffs. While our results on intrinsic gains remain unaffected, there are additional expected gains and losses from signaling per se — Bayesian plausibility holds for posteriors, but not for the expected *utility* from those posteriors, when that utility is nonlinear. The entire distribution of posteriors is relevant in determining expected reputational payoffs.

There is a large range of possibilities when reputational utility is nonlinear. As a specific example, let us study $P_r(t, z)$: the probability that the induced posterior on agent r is strictly larger than some "target posterior" t.[17] Think about this target posterior as some desired threshold that is relevant for career advancement – for instance, the agent will get promoted if the observer's posterior is larger or equal to t.

Is this precisely what agents care for in their own reputational payoff functions? We remain agnostic on the matter. One view is that we should respect the precise functional form that agents choose to maximize. Another view is that the target posterior is an interesting by-product of agent interaction, irrespective of what it is that agents are maximizing. Proposition 8 below is comfortable with either view. It argues that, when symmetric agents collaborate in an asymmetric equilibrium, the dis-favored 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.

Proposition 8. *In an asymmetric equilibrium at z with updates (β_p, β_q) ascribed to agents with a common prior $p = q$, where p has the favored identity,*

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

---

[17]As before, this probability is computed across ideas x and y, but presuming that $f(x, y) = z$. Again, we can just as easily integrate this object across all possible z's, if desired.
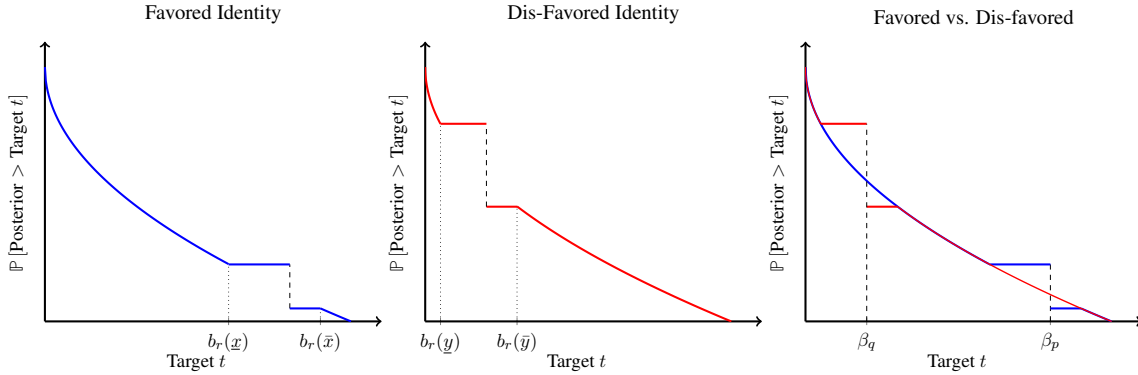
FIGURE 5. Distribution of Posteriors in Asymmetric Equilibria.

$$\text{and } P_p(t, z) \leqslant P_q(t, z), \text{ if } t < \beta_q \text{ or } t \geqslant \beta_p.$$

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

To visualize Proposition 8, refer to Figure 5, which displays the distribution of posteriors for symmetric and asymmetric collaboration structures. On the vertical axis of each panel, we plot the probability that an agent's posterior will be larger than some target posterior t. Remember that when agents' ideas are such that they collaborate, the observer sees only z and is unable to tell which x and y ideas made up that outcome. Therefore all the possible ideas that would lead to equilibrium collaboration in equilibrium are "garbled" to make up the expected update of the observer. That leads to the pictured flat regions and discontinuities in the posterior distributions.

Naturally, in an equilibrium with symmetric collaboration, both p and q have the same posterior distributions and so reach the target posterior t with the same probability, regardless of t – this is pictured in the first panel of Figure 5. On the other hand, in an asymmetric equilibrium — as in the second and third panels — the distribution of posteriors is distinct across identities.

In particular, say the asymmetric equilibrium for $p = q = r$ has collaboration regions $[\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. If the target posterior t is between $b_r(\underline{y})$ and β_q, then q is more likely to reach the target. To understand this, note that if q has idea $\bar{y}$, then agents collaborate and the public's update on q is β_q. If, conversely, p were to have idea $\bar{y}$, agents would work separately, and the public's update on p would be $b_p(\bar{y}) < \beta_q$. Similar arguments imply the other differences in the posterior distributions across p and q.

7. MERIT-BASED AND RANDOM ORDER IN COLLABORATION

Individual concerns with signaling prevent potential partners from achieving efficient full collaboration. The informational garbling that occurs with joint projects keeps agents with particularly strong ideas from collaborating with partners with relatively weak ideas. They would rather sacrifice a better joint project so as to achieve better signaling. Yet committing to collaborate is

Pareto-superior whenever reputational utilities are concave. In this section, we explore the intuition that policies that help to disentangle each person's contributions to a joint project would make agents more willing to collaborate, and lead to greater 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 VI.C of Ray ⓡ Robson (2018). One policy, standard in the publishing process of many scientific fields, is to arrange authors explicitly to reveal their ordinal contribution to a 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 $M^p(z) = C(z, p, q) \cap \{(x, y) | x \geqslant y\}$. Might that spur more collaboration?

Certainly, holding fixed the collaboration correspondence 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) $b_q(e(z)) + \alpha e(z)/q$, while a collaborative decision would generate a payoff of $\beta_q(z) + \alpha z/q$, where $\beta_a(z)$ 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. 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, p, q) 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 $(x, y) \in R(z) \cup M^p(z) \cup M^q(z)$, $f(x, y) = x$;

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

(ii) For $C \in \{R(z), M^p(z), M^q(z)\}$, we have $(x, y) \in C$ if and only if $V(x, r, b_r(x)) \leqslant V(z, r, \beta_r(z, C))$ 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 (in both the ex ante and the ex-post sense).*

To illustrate, let C be the equilibrium correspondence in the benchmark equilibrium under consideration. There is at least one person for wom 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 subjected to the same fragility critique as equilibria 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 for communicating such information, albeit just ordinally, *of their own volition*. If they choose the set R, then certainly they are making clear to the public that the merit differences are not severe enough to be pointed out. If they choose the sets M^p and M^q, that removes the need for extensive speculation in the first place.

REFERENCES

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

Arrow, Kenneth. (1973) "The Theory of Discrimination", in: O. Ashenfelter, A. Rees (Eds.), *Discrimination in Labor Markets*, Princeton University Press, Princeton, NJ, 1973, pp. 3-33.

Bagnoli, Mark, and Ted Bergstrom. (2005) "Log-Concave Probability and its Applications." *Economic Theory* **26**: 445-469.

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

Bardhi, Arjada, Yingni Guo, and Bruno Strulovici. (2020) "Early-Career Discrimination: Spiraling or Self-Correcting?" *working paper*.

Becker, Gary. (1973) "A Theory of Marriage: Part I." *Journal of Political Economy* **81**: 813-846.

Bowles, Samuel, Glenn Loury, and Rajiv Sethi. (2014) "Group Inequality." *Journal of the European Economic Association* **12**: 129-152.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. (2019) "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* **109**: 3395-3436. Harvard

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. (2019) "Inaccurate Statistical Discrimination." *NBER working paper* w25935.

Chade, Hector, and Jan Eeckhout. (2020) "Competing Teams." *Review of Economic Studies* **87**: 1134-1173.

Chaudhuri, Shubham, and Rajiv Sethi. (2008) "Statistical Discrimination with Peer Effects: Can Integration Eliminate Negative Stereotypes?" *Review of Economic Studies* **75**: 579-596.

Chalioti, Evangelia. (2016)"Team Production, Endogenous Learning about Abilities and Career Concerns." *European Economic Review* **85**: 229-244.

Coate, Stephen, and Glenn C. Loury. (1993) "Will Affirmative-Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, **83**: 1220-1240.

Ductor, Lorenzo, Sanjeev Goyal and Anja Prummer. (2021) "Gender and Collaboration." *working paper*.

Einav, Liran, and Leeat Yariv. (2006) "What's in a Surname? The Effects of Surname Initials on Academic Success." *Journal of Economic Perspectives* **20**: 175-187. Harvard

Fang, Hanming, and Andrea Moro. (2011) "Theories of Statistical Discrimination and Affirmative Action: A Survey." *Handbook of Social Economics* **1**: 133-200.

Gu, Jiadong, and Peter Norman. (2020) "A Search Model of Statistical Discrimination." *working paper*.

Holmström, Bengt. (1982) "Moral Hazard in Teams," *The Bell Journal of Economics* **13**, 324-340.

Jackson, Matthew, and Asher Wolinsky. (1996) "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* **71**: 44-74.

Jones, Benjamin. (2021) "The Rise of Research Teams: Benefits and Costs in Economics." *Journal of Economic Perspectives* **35**: 191-216.

Kambhampati, Ashwin, and Carlos Segura-Rodriguez. (2020) "The Optimal Assortativity of Teams Inside the Firm." *working paper*.

Kambhampati, Ashwin, Carlos Segura-Rodriguez, and Peng Shao. (2020) "Matching to Produce Information." *working paper*.

Mookherjee, Dilip, and Debraj Ray (2002), "Is Equality Stable?," *American Economic Review* **92** (Papers and Proceedings), 253-259.

Mookherjee, Dilip, and Debraj Ray (2003), "Persistent Inequality," *Review of Economic Studies*, 70(2), April 2003, 369-394.

Moro, Andrea, and Peter Norman. (2004) "A General Equilibrium Model of Statistical Discrimination." *Journal of Economic Theory* **114**: 1-30. Harvard

Myrdal, G. (1944), *An American Dilemma: The Negro Problem and Modern Democracy*. Harper & Row, New York.

Ozerturk, Saltuk, and Huseyin Yildirim. (2021) "Credit Attribution and Collaborative Work." *Journal of Economic Theory*, forthcoming.

Phelps, Edmund. (1972) "The Statistical Theory of Racism and Sexism", *American Economic Review* **62**: 659-66.

Ray, Debraj, ⓡ Arthur Robson. (2018) "Certified Random: A New Order for Coauthorship." *American Economic Review* **108**: 489-520.

Sarsons, Heather. (2017) "Recognition for Group Work: Gender Differences in Academia." *AER Papers & Proceedings*, **107**: 141-145.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. (2021) "Gender Differences in Recognition for Group Work." *Journal of Political Economy* **129**.

Winter, Eyal. (2004) "Incentives and Discrimination." *American Economic Review* **94**: 764-773.

8. APPENDIX: PROOFS

8.1. **Proof of Proposition 1.** As discussed, if (2) 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 $\underline{y} = \iota_z(\bar{x})$; then it must be that $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. Because $C(z, p, q)$ is nonempty, $0 \leqslant \underline{x} \leqslant \bar{x} \leqslant z$ and $0 \leqslant \underline{y} \leqslant \bar{y} \leqslant z$. 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 (2) holds with equality, giving us (5) and (6) via the transformations (3) and (4).

If $[\underline{x}, \bar{x}]$ or $[\underline{y}, \bar{y}]$ — say $[\underline{x}, \bar{x}]$ — is non-degenerate, then additionally $\bar{x} < z$. Suppose not; then $\bar{x} = z$. But at this threshold, collaborative output is the same as solo output, while by (1) and the nondegeneracy of $[\underline{x}, \bar{x}]$, the signaling update is *strictly* smaller, a contradiction. It follows from $\bar{x} < z$ and $\underline{y} = \iota_z(\bar{x})$ that $\underline{y} \in (0, z)$. Now we can check that $\bar{y} \in (\underline{y}, z)$, because at $y = \underline{y}$, (6) holds with ">," whereas at $y = z$, (6) holds with "<." In turn, $\bar{y} < z$ and $\underline{x} = \iota_z(\bar{y})$ imply $\underline{x} > 0$, and all the strict inequalities are established.

For the converse, take any collection $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ with $0 \leqslant \underline{x} \leqslant \bar{x} \leqslant z$ and $0 \leqslant \underline{y} \leqslant \bar{y} \leqslant z$, and so that (5) and (6) are satisfied. Suppose that the public forms the beliefs $\tilde{C}(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. 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. ∎

8.2. **Proof of Theorem 1.** 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

$$(17) \qquad u(b_p(\bar{x})) - \alpha[z - \bar{x}] = u(\beta_p) \text{ and } u(b_q(\bar{y})) - \alpha[z - \bar{y}] = u(\beta_q).$$

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

$$(18) \qquad \underline{x} = \min\{\bar{x}, \iota_z(\bar{y})\} \text{ and } \underline{y} = \min\{\bar{y}, \iota_z(\bar{x})\},$$

and then

$$(19) \qquad \beta'_p = \beta_p(z) \text{ and } \beta'_q = \beta_q(z),$$

where the right hand sides of these equations are defined in (3) and (4).

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 (17) and (18). We claim that all these values lie strictly between 0 and z, and that

(20) $$\underline{x}^* = \iota_z(\bar{y}^*) < \bar{x}^* \text{ and } \underline{y}^* = \iota_z(\bar{x}^*) < \bar{y}^*.$$

To prove (20), it will suffice to show that $\underline{x}^* < \bar{x}^*$ and $\underline{y}^* < \bar{y}^*$. Suppose not, then (say) $\underline{x}^* = \bar{x}^*$. Then (19) implies $\beta_p^* = b_p(\bar{x}^*)$. At the same time, (17) implies that $b_p(\bar{x}^*) > \beta_p^*$ whenever $\bar{x}^* < z$, so the previous equality must imply that $x^* = z$. Therefore by (18), $\underline{y}^* = \min\{\bar{y}^*, \iota_z(\bar{x}^*)\} = 0$. Using (19) and the definition of the function β_q in (4), this implies $\beta_q^* < b_q(z)$, and therefore (17) implies $\bar{y}^* \in (0, z)$. But then, using (18) 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 (20) implies $\beta_p^* < b_p(z)$ and $\beta_q^* < b_q(z)$. Therefore, by (17), $\bar{x}^* < z$ and $\bar{y}^* < z$. Using (20), 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 (17), (19) and the just-established (20), along with Proposition 1. ∎

8.3. **Proof of Observation 1.**
Suppose that an equilibrium is p-fragile. Then, recalling (10), there is $\zeta > 0$ and $\delta > 0$ such that for every $\epsilon \in (0, \delta)$,

(21) $$\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) \geqslant \beta_p + (1 + \zeta)\epsilon \text{ and } \Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) \leqslant \beta_q - (1 + \zeta)\epsilon.$$

Using the fact that $(\beta_p, \beta_q) = \Theta(\beta_p, \beta_q)$, (10) is equivalent to

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

Recalling the construction of Θ around the equilibrium point (see (17), (18) and (19)), recalling that $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = \iota_z(\bar{x})$ at any equilibrium point, and given our assumption that f is continuously differentiable, it follows that Θ is continuously differentiable as well. Therefore (22) is equivalent to

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

where these derivatives are evaluated at the equilibrium (β_p, β_q). But (24) is entirely symmetric across p and q, and so must also be equivalent to q-fragility. ∎

8.4. **Proof of Proposition 2.**
Fix (p, q) and a value of $z > 0$. We first observe that the set of equilibria is compact, and therefore so is the set of equilibrium updates conditional on collaboration. Fix some agent, say q, and let $\underline{\beta}_q$ be the minimum value of equilibrium updates for her, over the set of all equilibria.

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

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

and let

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

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.)

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

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

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

Step 2. B_2 is nondecreasing.

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

$$\Theta_p(\beta_p, \beta_q') \leqslant \Theta_p(\beta_p, \beta_q).$$

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

$$\Theta_p(\beta_p, \beta_q') \leqslant \Theta_p(\beta_p, \beta_q) \leqslant \beta_p$$

But that just means $B_1(\beta_q') \leqslant B_1(\beta_q)$. By Step 0 again, $B_2(\beta_q') \geqslant B_2(\beta_q)$.

Step 3. $B_2(b_q(0)) > b_q(0)$.

By (17), $\Theta_q(b_q(0), \beta_p) > b_q(0)$ for all $\beta_p \in [b_p(0), b_p(z)]$. In particular, $B_2(b_q(0)) > b_q(0)$.

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

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

(24)   (a) $\Theta_p(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon) > \bar{\beta}_p + \epsilon$ and   (b) $\Theta_q(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon) < \underline{\beta}_q - \epsilon,$

Given Step 1, (24a) implies $B_1(\underline{\beta}_q - \epsilon) > \bar{\beta}_p + \epsilon$. Using this inequality along with (24b) and Step 0, we have $B_2(\underline{\beta}_q - \epsilon) < \underline{\beta}_q - \epsilon$.

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

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

But then $(B_1(\beta_q), \beta_q)$ is a fixed-point of the map Θ, contradicting the definition of $\underline{\beta}_q$. ∎

8.5. **Proof of Proposition 3.** We adapt the fixed point mapping used in the proof of Theorem 1. Fix r and $z > 0$, and consider a map with $[0, b_r(z)]$ as domain. For any β_r in this domain, to be interpreted as an update ratio, define $\bar{x}$ by (5), restated here with minor notational change as

(25) $$u(b_r(\bar{x})) - \alpha[z - \bar{x}] = u(\beta_r).$$

and then define $\underline{x}$ by

(26) $$\underline{x} = \iota_z(\bar{x}).$$

As in the proof of Theorem 1, (26) might result in $\underline{x} > \bar{x}$, a situation that has no meaning for us. Accordingly, we set β_r to be greater than or equal to $\underline{\beta}$, where $\underline{\beta} \in [0, b_r(z))$ is the smallest value of β_r such that $\underline{x} \leqslant \bar{x}$. This threshold is well-defined because for values of β_r approaching

$b_r(z)$, it is evident from (25) that $\bar{x}$ must approach z as well, but then $\underline{x} = \iota_z(\bar{x})$ must be close to 0 and therefore below $\bar{x}$. Moreover, for all $\beta_r > \underline{\beta}$, it is also true that $\underline{x} < \bar{x}$, because $\bar{x}$ is increasing in β_r and $\underline{x}$ is decreasing.

Therefore, restrict attention to the sub-interval $[\underline{\beta}, b_r(z)]$. For $\beta_r \in [\underline{\beta}, b_r(z)]$, define $\bar{x}$ and $\underline{x}$ by (25) and (26), and then β'_r according to (3). Two end-point conditions are to be noted. First, for $\beta_r = \underline{\beta}$, $b_r(\bar{x})$ is strictly larger than β_r. If $\underline{\beta} > 0$, it must also be that $\underline{x} = \bar{x}$, and so $\beta'_r = \Theta^S(\beta_r) > \beta_r$. If $\beta_r = \underline{\beta} = 0$, then certainly the same inequality $\beta'_r = \Theta^S(\beta_r) > \beta_r$ holds a fortiori. Second, for $\beta_r = b_r(z)$, $\bar{x} = z$ while $\underline{x} = 0$, so $\Theta^S(b_r(z)) < b_r(z)$. Finally, Θ^S is continuous, so there must be some $\beta_r^* \in (\underline{\beta}, b_r(z))$ with $\Theta^S(\beta_r^*) = \beta_r^*$. Define the accompanying values $\bar{x}^*$ and $\underline{x}^*$ from (25) and (26). It is immediate that $(\underline{x}^*, \bar{x}^*)$ is a symmetric equilibrium. ■

8.6. **Proof of Theorem 2.** We already know that fragility is equivalent to (24). In a symmetric equilibrium, the two inequalities are identical and equivalent to

$$(27) \qquad \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_p} - \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_q} > 1,$$

evaluated at $p = q = r$. We use the definition of the mapping Θ in section 4 to compute these derivatives. In the equations below, we write the common value of p and q as r. Wherever endogenous variables such as $\underline{x}$ and $\bar{x}$ appear, they are taken to refer to the symmetric equilibrium in question. We have:

$$(28) \quad \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_p} = \left[\frac{\partial \beta'_p}{\partial \bar{x}}\right]\left[\frac{d\bar{x}}{d\beta_p}\right] = \left[\frac{\partial \beta'_p}{\partial \bar{x}}\right] \frac{u'(\beta_r)}{u'(b_r(\bar{x}))b'_r(\bar{x}) + \alpha} = \frac{u'(\beta_r)[b_r(\bar{x}) - \beta_r]\gamma_z(\bar{x})}{[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})][u'(b_r(\bar{x}))b'_r(\bar{x}) + \alpha]},$$

and

$$(29) \quad \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_q} = \left[\frac{\partial \beta'_p}{\partial \underline{x}}\right]\left[\frac{\partial \underline{x}}{\partial \bar{y}}\right]\left[\frac{d\bar{y}}{d\beta_q}\right] = \left[\frac{\partial \beta'_p}{\partial \underline{x}}\right] \frac{u'(\beta_r)\iota'_z(\bar{y})}{u'(b_r(\bar{y}))b'_r(\bar{y}) + \alpha} = \frac{u'(\beta_r)\iota'_z(\bar{y})[\beta_r - b_r(\underline{x})]\gamma_z(\underline{x})}{[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})][u'(b_r(\bar{y}))b'_r(\bar{y}) + \alpha]}.$$

Combining (27), (28) and (29), using symmetry to note that $\bar{x} = \bar{y}$ and $\gamma_z(\bar{x}) = \gamma_z(\underline{x})$,[18] and rearranging terms, we obtain (11), as desired. ■

8.7. **Proof of Proposition 4.** Pick any $z > 0$. We claim that for every $\epsilon > 0$, there exists $\alpha(\epsilon)$ such that if $\alpha \in (0, \alpha(\epsilon))$, then $\bar{x} - e \leqslant \epsilon$ for every symmetric equilibrium with upper threshold $\bar{x}$, where remember that e is the unique value such that $f(e, e) = z$. For suppose this assertion is false; then there exists $\epsilon > 0$ and a sequence $\alpha^n \to 0$ such that for every n, there is some symmetric equilibrium threshold $\bar{x}^n$ with $\bar{x}^n \geqslant e + \epsilon$. Moreover, $\underline{x}^n \leq e$. In particular, given that u and b_r are strictly increasing, there is $\delta > 0$ such that

$$(30) \qquad u(b_r(\bar{x}^n)) - u(\beta_r^n) \geqslant \delta$$

for all n, where β_r^n is the corresponding sequence of equilibrium updates conditional on collaboration. At the same time, we know that by the equilibrium conditions,

$$(31) \qquad u(b_r(\bar{x}^n)) + \alpha^n\left[\bar{x}^n - z\right] = u(\beta_r^n)$$

_____

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

for all n. As $n \to \infty$, $\alpha^n \to 0$, while $\bar{x}^n$ is bounded, but then (30) and (31) cannot both hold, which is a contradiction.

Part (i). Pick any $z \in J$. Then there is $\epsilon' > 0$ such that $b_r''(x) < 0$ on $[e - \epsilon', e + \epsilon']$. By the Claim above, there is $\underline{\alpha} > 0$ such that if $\alpha \in (0, \underline{\alpha})$, then $\bar{x} - e = \epsilon \leqslant \epsilon'$ for every symmetric equilibrium threshold $\bar{x}$. Recall the map Θ^S that picks out all symmetric equilibria; see the proof of Proposition 3 in the main text. Writing $\beta_r' = \Theta^S(\beta_r)$ and evaluating the derivative at a symmetric fixed point with accompanying thresholds $\underline{x}$ and $\bar{x}$, we have:

$$
\begin{aligned}
\frac{d\beta_r'}{d\beta_r} &= \left[\frac{d\beta_r'}{d\underline{x}} \frac{d\underline{x}}{d\bar{x}} + \frac{d\beta_r'}{d\bar{x}} \right] \frac{d\bar{x}}{d\beta_r} \\
&= \left[\frac{[\beta_r - b_r(\underline{x})]\gamma_z(\underline{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \iota_z'(\bar{x}) + \frac{[b_r(\bar{x}) - \beta_r]\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \right] \frac{u'(\beta_r)}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha} \\
(32) \qquad &= \left[\frac{[b_r(\bar{x}) + b_r(\underline{x}) - 2\beta_r]}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \right] \frac{u'(\beta_r)\gamma_z(\bar{x})}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha},
\end{aligned}
$$

where the last equality invokes the symmetry of γ_z, so that $\gamma_z(\bar{x}) = \gamma_z(\underline{x})$, and the assumption that f is linear, so that $\iota_z'(\bar{x}) = -1$. Now, because $b_r''(x) < 0$ on $[e - \epsilon, e + \epsilon]$ and $\bar{x} - e = \epsilon \leqslant \epsilon'$ (and by symmetry, $e - \underline{x} = \epsilon \leqslant \epsilon'$), it follows that $b_r''(x) < 0$ on $[\underline{x}, \bar{x}]$. Using this information along with the symmetry of γ_z, we must conclude that

$$
\begin{aligned}
\beta_r &= \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \int_{\underline{x}}^{\bar{x}} b_r(x)\gamma_z(x)dx \\
&= \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \left[\int_{e-\epsilon}^{e} b_r(x)\gamma_z(x)dx + \int_{e}^{e+\epsilon} b_r(x)\gamma_z(x)dx \right] \\
&= \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \int_{0}^{\epsilon} [b_r(e-\eta) + b_r(e+\eta)]\gamma_z(e+\eta)d\eta \\
(33) \qquad &> \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \int_{0}^{\epsilon} [b_r(e-\epsilon) + b_r(e+\epsilon)]\gamma_z(e+\eta)d\eta = \frac{1}{2} [b_r(\underline{x}) + b_r(\bar{x})],
\end{aligned}
$$

where the third equality uses symmetry and sets $\gamma_z(e + w) = \gamma_z(e - w)$ for each $w \in [0, \epsilon]$, and the final inequality uses the strict concavity of b_r on $[\underline{x}, \bar{x}]$. Combining (32) and (33), we must conclude that $d\beta_r'/d\beta_r < 0$ at any symmetric equilibrium. Given the end-point conditions on Θ^S, this implies that there is one and only one such equilibrium.

Part (ii). Pick any $z \in J$. Take a sequence $\{\alpha^n\}$ with $\alpha^n \to 0$. We know that for large n, there is a unique symmetric equilibrium for each n. So there is a corresponding sequence $\{\bar{x}^n, \underline{x}^n\}$ of uniquely defined equilibrium thresholds, along with equilibrium collaborative updates β_r^n, satisfying

$$
\bar{x}^n = e + \epsilon(\alpha^n), \underline{x}^n = e - \epsilon(\alpha^n), \text{ and } \beta_r^n = \beta_r(\epsilon(\alpha^n)),
$$

where $\epsilon(\alpha)$ is a function satisfying

$$
(34) \qquad u(b_r(e + \epsilon(\alpha))) + \alpha[e + \epsilon(\alpha) - z] = u(\beta_r(\epsilon(\alpha^n))),
$$

and where $\beta_r(\epsilon)$ satisfies (3), or in the present specific context:

$$
(35) \qquad \beta_r(\epsilon) = \frac{1}{\Gamma_z(e + \epsilon) - \Gamma_z(e - \epsilon)} \int_{e-\epsilon}^{e+\epsilon} b_r(x)\gamma_z(x)dx.
$$

Using (34) and (35), and the fact that $\lim_{\epsilon \to 0} \beta_r'(\epsilon) = 0$ it is easy to see that $\epsilon(\alpha)$ is differentiable and

$$(36) \qquad \epsilon'(0) = \frac{e}{u'(b_r(e))b_r'(e)}.$$

Recall the fragility condition (12) when $f(x, y) = x + y$, reproduced here using ϵ as

$$(37) \qquad u'(\beta_r(\epsilon))\frac{[b_r(e + \epsilon) - b_r(e - \epsilon)]/[2\epsilon]}{u'(b_r(e + \epsilon))b_r'(e + \epsilon) + \alpha} > \frac{[\Gamma_z(e + \epsilon) - \Gamma_z(e - \epsilon)]/[2\epsilon]}{\gamma_z(e + \epsilon)},$$

The two sides of (37) have a common limit — 1 — as α (and $\epsilon(\alpha)$) $\to 0$, and so we examine the *derivatives* of each side as $\alpha \to 0$. Denoting the right-hand side of (37) by $R(\alpha)$, it follows that

$$R'(\alpha) = \frac{\gamma_z(e + \epsilon)\left[\frac{d}{d\epsilon}\frac{\Gamma_z(e+\epsilon)-\Gamma_z(e-\epsilon)}{2\epsilon}\right]\epsilon'(\alpha) - \left[\frac{\Gamma_z(e+\epsilon)-\Gamma_z(e-\epsilon)}{2\epsilon}\right]\gamma_z'(e + \epsilon)}{\gamma_z(e + \epsilon)^2},$$

and passing to the limit as $\alpha \to 0$, we have

$$(38) \qquad R'(0) = 0,$$

where we use the fact that $\gamma_z'(e) = 0$ by the symmetry of γ_z around e, and the fact that for any C^2 function f,

$$(39) \qquad \frac{d}{d\epsilon}\left[\frac{f(e + \epsilon) - f(e - \epsilon)}{2\epsilon}\right] = 0$$

evaluated at $\epsilon = 0$. Next, denoting the left-hand side of (37) by $L(\alpha)$, we have

$$L'(\alpha) = \frac{[u'(b_r(e + \epsilon))b_r'(e + \epsilon) + \alpha]\left[u'(\beta_r(\epsilon))\frac{d}{d\epsilon}\frac{b_r(e+\epsilon)-b_r(e-\epsilon)}{2\epsilon} + \frac{b_r(e+\epsilon)-b_r(e-\epsilon)}{2\epsilon}u''(\beta_r(\epsilon))\beta_r'(\epsilon)\right]\epsilon'(\alpha)}{[u'(b_r(e + \epsilon))b_r'(e + \epsilon) + \alpha]^2}$$

$$- \frac{u'(\beta_r(\epsilon))\frac{b_r(e+\epsilon)-b_r(e-\epsilon)}{2\epsilon}\left\{[u'(b_r(e + \epsilon))b_r''(e + \epsilon) + u''(b_r(e + \epsilon))b_r'(e + \epsilon)]\epsilon'(\alpha) + 1\right\}}{[u'(b_r(e + \epsilon))b_r'(e + \epsilon) + \alpha]^2}.$$

We pass to the limit as $\alpha \to 0$ above, while keeping in mind that $\lim_{\epsilon \to 0} \beta_r'(\epsilon) = 0$ and also using (36) and (39). The entire first set of terms above converges to zero, and we obtain:

$$L'(0) = -\frac{u'(b_r(e))b_r'(e)[\{u'(b_r(e))b_r''(e) + u''(b_r(e))b_r'(e)\}\epsilon'(0) + 1]}{u'(b_r(e))^2 b_r'(e)^2}$$

$$(40) \qquad = -\frac{\{u'(b_r(e))b_r''(e) + u''(b_r(e))b_r'(e)\}\epsilon'(0) + 1}{u'(b_r(e))b_r'(e)} = -\frac{\frac{eb_r''(e)}{b_r'(e)} + \frac{eu''(b_r(e))}{u'(b_r(e))} + 1}{u'(b_r(e))b_r'(e)} > 0,$$

where the equality in the last line uses (36), and the final inequality uses $z \in J$, as well as the concavity of u. Comparing (38) and (40), we must conclude that for small α, the condition (37) is satisfied, so that the symmetric equilibrium is fragile. ∎

8.8. **Proof of Proposition 5.** Both inequalities follow immediately from the fact that $\underline{x}^p > \underline{x}^q$ and $\bar{x}^p > \bar{x}^q$ in any equilibrium where p has the favored identity.

8.9. **Proof of Proposition 6.** Part (i). As shown in the main text, we have

$$(41) \qquad \Delta_p - \Delta_q = \int_{\underline{x}}^{\bar{x}} [\iota_z(x) - x]\gamma_z(x)dx.$$

Because p is unambiguously favored, $\iota_z(x) < x$ for all $x \in [\underline{x}, \bar{x}]$, so by (41), $\Delta_p - \Delta_q < 0$.

Part (ii). Because p is favored in equilibrium 1 over 2, and q is disadvantaged, it must be — using (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 (41) for each equilibrium i, then

$$(42) \qquad \delta_i \equiv \Delta_p^i - \Delta_q^i = \int_{\underline{x}_i}^{\bar{x}_i} [\iota_z(x) - x]\gamma_z(x)dx.$$

We wish to sign $\delta_1 - \delta_2$. Recall the definition of the equal input $e(z)$. Because no agent is unambiguously favored in any equilibrium, we have

$$(43) \qquad \underline{x}_2 < \underline{x}_1 \leq e(x) \leq \bar{x}_2 < \bar{x}_1.$$

Recalling (42), we must conclude that

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

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 (43)), and that $\iota_z(x) < x$ for $x \in [\bar{x}_2, \bar{x}_1)$ (an implication of the third and fourth inequalities in (43)). ∎

8.10. **Proof of Proposition 7.** See main text. ∎

8.11. **Proof of Proposition 8.** Consider an equilibrium collaboration set $C = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. Then

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

and

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

Now note that $p = q = r$ and that, in an asymmetric equilibrium where p is favored, either $\underline{y} < \bar{y} \leqslant \underline{x} < \bar{x}$ or $\underline{y} < \underline{x} < \bar{y} < \bar{x}$. In either case, it is easy to check that the inequalities in the proposition hold. ∎

8.12. **Proof of Proposition 9.** Suppose $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ is an equilibrium collaboration set under no authorship ordering. In what follows, we extend our notation to define $\beta_p(z, \underline{x}, \bar{x})$ and $\beta_q(z, \underline{y}, \bar{y})$ to stand for the right hand sides in (3) and (4); that is, we carry the limits of the integration explicitly. Now augment the equilibrium collaboration set as follows. Define x° by the smallest solution in x (but weakly exceeding $\bar{x}$) to

$$(44) \qquad u(b_p(x)) + \alpha[x - z] = u(\beta_p(z, \bar{x}, x)).$$

The left hand side of (44) is strictly smaller than the right hand side at $x = \bar{x}$. 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

$$(45) \qquad u(b_q(y)) + \alpha[y - z] \leqslant u(\beta_q(z, y, \underline{y})).$$

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

$$(46) \qquad \bar{x} < x^* < z, \text{ and } u(b_p(x)) + \alpha[x - z] \leqslant u(\beta_p(z, \bar{x}, x)) \text{ for all } \bar{x} \leqslant x \leqslant x^*, \text{ while}$$

$$(47) \qquad 0 < y_* < \underline{y}, \text{ and } u(b_q(y)) + \alpha[y - z] \leqslant u(\beta_p(z, y, \underline{y})) \text{ for all } y_* \leqslant y \leqslant \underline{y},$$

with at least one of the two second inequalities in (46) and (47) holding with equality. To prove this claim, note that $x^* \leqslant x^\circ < z$. Moreover, both x° and $\iota_z(y_\circ) > \bar{x}$, 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 (46) must hold.

Turning now to (47), note that $y^* \geqslant \iota_z(x^\circ) > 0$, 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 (45) is strictly increasing in y, while the left hand side is constant in y. So if "$\leqslant$" holds in (47) at $y = y^*$, it must do so for $y_* \leqslant y \leqslant \underline{y}$. 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,

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

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

Note that at least one of M^p and M^q is non-empty. Using (46) and (47), it is easy to verify that the collection $\{R, M^p, M^q\}$ satisfies all 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. ∎