Signaling and Discrimination in Collaborative Projects

Paula Onuchic ⓡ Debraj Ray[†]

November 2021

Abstract. We study collaborative work in pairs. Each individual draws an idea from a distribution that depends on unobserved ability. Potential collaborators choose to combine their ideas, or work separately. They are motivated not just by intrinsic project value, but also signaling payoffs, which depend on public assessments of individual ability. In equilibrium, collaboration strategies both justify and are justified by those assessments. When partners are symmetric, equilibria with symmetric collaborative strategies are often *fragile*, in a sense made precise in the paper. In such cases, the public ascribes higher credit to one of the partners based on payoff-irrelevant "identities." In asymmetric equilibria, favored identities receive a higher payoff conditional on collaborating, but may receive lower overall expected payoff relative to their disfavored counterparts, Furthermore, individuals of the disfavored identity are ex-post more likely to achieve extreme reputations. Finally, we study a simple policy based on certified random order that Pareto-improves the equilibria of our model.

1. Introduction

Research is increasingly conducted in teams. In economics, co-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 groups. Obviously, collaboration can be beneficial, as it allows workers to fruitfully combine complementary skills. However, by its very nature, teamwork obscures individual contributions, compromising an individual's ability to build reputation. This gives rise to a 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 evaluation.

We build a theory that incorporates both these aspects of collaboration. At its heart are *public perceptions* of individual ability implied by collaborative work, based not only on pre-existing priors but also on conjectures about what circumstances led to the observed collaboration. In turn, collaboration decisions themselves are endogenously

[†]Onuchic: Nuffield College, Oxford University, p.onuchic@nyu.edu; Ray: New York University and University of Warwick, debraj.ray@nyu.edu. Ray acknowledges funding under NSF grant SES-1851758. We thank Daghan Carlos Akkar, Axel Anderson, Yingni Guo, Sam Kapon, Navin Kartik, Anja Prummer, Mauricio Ribeiro, Ludvig Sinander, and Joshua Weiss for useful comments.

[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.

determined by public perceptions. The main novelty in our theory is that it incorporates this circularity in a model of collaboration with reputational concerns.

In our setting, people choose to collaborate in pre-matched pairs. Each person has one of two types, good or bad. There is a common public prior on each agent's type. (Whether or not an agent knows her own type will be irrelevant for our theory, as it often is in real life research environments.) When a pair meets, their initial interaction is a discussion of project ideas. Each person draws an idea from a distribution that depends on their type, with good-type individuals more likely to draw better ideas than bad-type individuals. Both people see their own idea and the idea of their potential partner, and choose to work together when collaboration is beneficial to both parties, and separately otherwise. In making this decision, each person seeks to maximize a combination of the *intrinsic value* and the *reputational value* of the project.

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

Section 4 characterizes equilibrium collaboration decisions. These resolve the tradeoff between *intrinsic value*, which always improves with collaboration, and *reputational value*, which is garbled in joint work. Controlling for joint project value and perceived collaboration strategies, the reputational payoff from collaboration is pinned down, while the reputational payoff from working alone rises with value of the agent's own idea. This observation 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 the payoff-relevant characteristics of the partners. For instance, for someone with an established reputation (that is, a high prior on their ability), the difference in signaling value between collaborative and individual work is very small. Such a person is almost always willing to combine ideas with a partner, and collaborates often in equilibrium. A less tested individual might seek out projects on his own, or work with other rookies, but this pattern could be non-monotonic in the prior on his own ability — there is a

rich theory of collaboration to be developed that links collaborative decisions to priors. That said, we pursue here 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, it helps to imagine that the two individuals are symmetric — the public has the same prior on them. However, suppose that each agent has a distinct payoff-irrelevant *identity*, one that is salient in the public eye. Now conjecture that the public is "biased" and allocates reputational value in favor of one identity. That is, the public thinks the favored identity contributes better ideas to collaborative outcomes, so assigning higher credit to it than to the dis-favored identity. The individuals will react to this bias with distinct collaboration strategies. Incentivized by the credit allocation, the favored identity is relatively more willing to collaborate than his dis-favored counterpart. Specifically, a favored person shares better ideas with a disfavored person than the latter is willing to do, were those ideas her own. So, at least to some degree, this reaction actually confirms the initial bias. Given the described collaboration strategies, the public should indeed "rationally" allocate reputational value the way they do.

These echo effects across a bias in public perception and optimal collaboration strategies can lead to multiple equilibria, some of them asymmetric even if the underlying collaborating agents are symmetric in all payoff-relevant characteristics. Sections 5 and 6 study these discriminatory forces. They are are more than mere theoretical abstractions. For instance, Sarsons, Gërxhani, Reuben, and Schram (2021) use data on academic economists to argue that the public responds to joint work between women and men by attributing more credit to men.[2] In related research, Ductor, Goyal and Prummer (2021) document homophily in coauthorship networks, as well as gender disparities in collaboration patterns in economic research. From the perspective of our model, these empirical observations are two sides of the same coin.[3]

Theorem 2 shows that symmetric equilibria always exist, and are unique either for large or small reputational concerns. But Theorem 3 establishes a set of sufficient conditions for that equilibrium to be *fragile* with respect to public perceptions.[4] By Proposition 2, there is always some non-fragile equilibrium. So under the conditions of Theorem 3, a non-fragile asymmetric equilibrium exists where the public perception depends on each person's payoff-irrelevant identity. That is, it features discrimination.

The main condition required by Theorem 3 is that agents sufficiently value the reputational aspect of their output, relative to their intrinsic production value. This means

[2]Specifically, 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.

[3]Ong, Chan, Torgler, and Yang (2018) also document that the decision to form coauthorships responds to expected credit assignment. Specifically, they compare coauthorship behavior between authors with surname initials earlier in the alphabet, who receive more credit, and those with later initials.

[4]We define fragility in Section 5. Informally, an equilibrium is fragile when small biases in public perceptions are confirmed, and enhanced, by individuals' behavioral responses.

that discrimination is a particularly credible threat in professions such as academia, where much of the productive value of research does not monetarily accrue to the researcher. Rather, the researcher is rewarded for signals of their underlying quality — for example, by receiving promotions and prizes that are explicitly conditioned on the perceived creativity and relevance of their past work, which is taken as a measure of the expected quality of their future work. Conversely, our model predicts that, if there are two symmetric agents who sufficiently value the intrinsic value of their work, then a unique collaborative equilibrium exists, one that is symmetric and non-fragile.

The asymmetric treatment of symmetric individuals brings to mind the enormous literature on statistical discrimination. Our approach falls under this general rubric, but the differences are noteworthy. In our setting, the equal-treatment or non-discriminatory outcome is eliminated by the fragility argument. In contrast, models of statistical discrimination usually display two equilibria (each conceivably non-fragile), one with discriminatory treatment and the other without. The question of whether the non-discriminatory equilibrium is fragile is normally not invoked, though we note that Gu and Norman (2020) take a related approach, in a different setting and with a different stability notion; see Section 2 for a more detailed comparison.

A second notable difference is that we study discrimination in a novel context, that of team formation between people with potentially different identities. Importantly, the interaction between agents is not mediated by firms, as in the usual labor market context. In this context, statistical discrimination yields new results regarding agents' payoffs in discriminatory equilibria, as well as some testable empirical implications.

In an asymmetric equilibrium ascribed to symmetric partners, one of the identities is favored, or perceived as contributing better ideas to a collaboration — thereby receiving higher collaborative credit. Perhaps counterintuitively, that favorable treatment does not necessarily map into better *overall* payoffs to the favored individual, relative to the dis-favored one. Propositions 3 and 4 in Section 7 argue that, in an asymmetric equilibrium, the expected intrinsic payoff to the dis-favored person is higher than that to the favored person. The very fact that the favored identity contributes better ideas to collaborations implies a relative gain in intrinsic payoff for the disfavored identity.

This result speaks directly to the collaborative setting in which our model is embedded. Unlike the usual labor market context, agents in our model directly transfer value to each other when they share ideas. This feature implies that relative favoritism in terms of credit assignment to an agent is invariably connected to a relative loss *to that same agent* in terms of intrinsic collaborative value.

A particularly sharp corollary applies when reputational utility is linear. Then the ex-ante reputational payoff is fixed irrespective of the collaboration structure — an implication of Bayes' plausibility. Therefore overall payoffs move in tandem with intrinsic payoffs alone, so that the disfavored person is ex-ante better off, even though she receives less credit conditional on collaboration. This is in stark contrast with the

traditional statistical discrimination literature, which generally finds that either discrimination does not affect the favored group, or unequivocally helps it.

If reputational utility is not linear, then Bayes' plausibility notwithstanding, expected signaling payoffs will depend on collaboration structure. Section 7.3 studies the *distribution* of posteriors induced in asymmetric equilibria. We posit a "target posterior" that an agent wishes to attain – for example, a reputation level that would induce a promotion. Proposition 5 argues that in an asymmetric equilibrium, the disfavored identity is more likely to reach such a target if it is extreme – either high or low. Conversely, the favored identity is more adept at reaching intermediate target posteriors.

This observation aligns two recent empirical observations that are seemingly at odds with each other. Sarsons, Gërxhani, Reuben, and Schram (2021) find that, conditional on a cross-gender academic collaboration, the probability of tenure increases more for the male rather than the female coauthor. So, conditional on collaboration, the favored identity is better off. In contrast, Card, DellaVigna, Funk and Iriberri (2021), studying the election of Fellows to the Econometric Society, argue that the female-male "gap became positive (though not statistically significant) from 1980 to 2010, and in the past decade has become large and highly significant, with over a 100% increase in the probability of selection for female authors relative to males with similar publications and citations." Our Proposition 5 states that a high target reputation (presumably needed for election to the Econometric Society) is relatively more likely to be reached by a member of the disfavored identity.

2. Related Literature

We embed a theory of discrimination in a novel context, that of team formation with reputational concerns. Collaboration garbles the reputational signal, but enhances the intrinsic value of the project. We investigate the fragility of outcomes that involve equal treatment across identities. Our results shows that discrimination is a particularly real threat when agents strongly value reputation, relative to intrinsic project value. To the best of our knowledge, ours is the first paper to propose a characterization of stability of symmetric/asymmetric outcomes in the context of collaboration.

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

In the specific context of statistical discrimination, Gu and Norman (2020) also use stability to select asymmetric equilibria in a model with multiple equilibria. They study a

search-theoretic model of the labor market, where workers sort into high-tech and low-tech sectors. They show (numerically) that the introduction of a payoff-irrelevant gender characteristic can render the symmetric equilibrium unstable, and generate gender-based sorting into the two occupations. Both the model and the forces that make for instability are entirely different from those we explore, but we mention this paper as an exception to the general approach to statistical discrimination taken in the literature.

Our results imply that an identity that is discriminated against *conditional on collaboration* may actually be better-off *overall* relative to a favored identity; see our discussions of Sarsons, Gërxhani, Reuben, and Schram (2021) and Card, DellaVigna, Funk and Iriberri (2021) in this context. But beyond that, these observations could be relevant in a larger setting in which agents choose identity. They can explain why individuals would choose to express an identity that is dis-favored along some dimension (collaborative output, in our setting), without relying on the assumption that they receive some inherent value from being their "true self" (Akerlof and Kranton 2000, Akerlof and Rayo 2020).

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

Our paper also relates to the literature on incentive provision in teams, following Holmström (1982). Winter (2004) connects team production and discrimination, arguing that 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 worker and firm reputations, and shows that working in teams creates incentives for both junior and senior team members to exert costly effort.

Our model employs a general payoff function defined on reputation and intrinsic project value that allows for nonlinear returns to reputation. In some models, that functional form can be derived from a larger game. For instance, in a dynamic setting, as in Anderson and Smith (2010) who study the possible failure of assortative matching in reputation-based models, these payoff functions would emerge as endogenous value functions. At each stage of their model, however, the collaboration decisions play no

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

role, and posterior updates are symmetric by assumption when partners are symmetric. In contrast, our main questions concern the collaboration choices of agents and the public's conjectures about collaborative patterns.[6]

Finally, as mentioned, we intersect with the theoretical literature on discrimination, especially that on statistical discrimination, stemming from Myrdal (1944), Phelps (1972) and Arrow (1973).[7] For instance, suppose an employer holds distinct beliefs about the quality of potential hires based on payoff-irrelevant identities (see, e.g., Coate and Loury 1993). In turn, these differences in perceptions incentivize different identities to make different pre-market investments in human capital, confirming their initial bias. Within that literature, our model is related to Moro and Norman (2004), who study 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.

3. Model

Two individuals have the opportunity to collaborate on a project. They bring ideas to the table. Each agent sees both ideas, and chooses whether to collaborate or work alone. If both prefer to work together, collaboration occurs. If either would rather not collaborate, then both work alone, with no plagiarism of ideas.

Ideas are generated by a distribution that depends on individual ability, which is either 0 (bad) or 1 (good). There is a public prior that a person is good, shared also by her potential partner. The individual values both the project at hand, and her *reputation*, which is the updated public belief on her ability. What the individual herself knows about her ability will turn out to be irrelevant, so we presume nothing.[8]

Each person's idea is drawn from a distribution with strictly positive densities $g(\cdot, 1)$ for the good type and $g(\cdot, 0)$ for the bad type, both with full support on $\mathbb{R}_+$. We assume that both densities have bounded derivatives on any compact set, and more substantively, that the likelihood ratio

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

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

[7]Fang and Moro (2011) survey this literature. More recent contributions include Peski and Szentes (2013), Bohren, Imas and Rosenberg (2019), Bohren, Haggag, Imas and Pope (2021) and Bardhi, Guo and Strulovici (2020).

[8]Because individual and public perceptions will generally diverge along a dynamic path, this irrelevance is particularly useful for potential dynamic extensions of our model.

With both ideas revealed — perhaps in initial discussion — agents decide whether to collaborate. Each person seeks to maximize a combination of the project's *intrinsic value*, implied by the ideas, and its *signaling value*, which determines her reputation.[9]

3.1. Intrinsic and Signaling Payoffs. Denote the public priors on the pair by p and q. We will use these letters as their names as well. Suppose that p has idea x, and q has idea y. If p and q collaborate, then the joint project has intrinsic value

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

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.[10] For any $z > 0$, 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., $f(w, \iota_z(w)) \equiv z$. We assume that ι_z has a bounded derivative on $[0, z]$.

If p and q work separately, then the Bayesian posteriors on their ability are given by

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

where $g(w, r) \equiv rg(w, 1) + (1 - r)g(w, 0)$ for $w \in \mathbb{R}_+$ and $r \in (0, 1)$. By the likelihood ratio assumption, $b_p(x)$ and $b_q(y)$ are increasing in x and y, and by our technical assumptions on g, they have derivatives bounded above and below by positive numbers on any compact set.

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

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

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

Such a set induces a probability distribution on combinations of x and y that could have led to the collaborative outcome z. The public update averages equation (2)

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

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

across every pair (x, y) in the conjectured collaboration set, using this distribution:

$$\beta_p = \mathbb{E}\left[b_p(x)|(x,y) \in C(z,p,q)\right] \quad \text{and} \quad \beta_q = \mathbb{E}\left[b_q(y)|(x,y) \in C(z,p,q)\right].$$

3.2. Overall Payoff. Each agent separably values both intrinsic and signaling payoffs outcomes from joint or solo projects. If a project has intrinsic value w and yields Bayesian posterior b, the overall payoff is

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

where $\alpha > 0$ weights project value w, and u, assumed to be smooth with positive but bounded derivative, is defined on individual reputation.

We make three remarks about this payoff structure. First, separability aside, the linearity of payoff in w is not an additional restriction provided we leave the joint production function f unrestricted. Nonlinearities in intrinsic payoff can be handled by a simple redefinition of variables. Second, the notation α is only useful because we will be interested in the case of "small" intrinsic value ($\alpha \to 0$). It is then more convenient to move α as a parameter rather than to shift u upward.

Finally, turning to reputational payoffs u, we sometimes mention the case in which u is linear. But the potential generality is useful for 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 in which some minimal value of the posterior is *necessary* for retaining a job, such as a position in a research university that prides itself on retaining "stars."

3.3. Equilibrium Definition. An *equilibrium collaboration set* is a collaboration set C used by the public to calculate posterior beliefs when agents collaborate; one which correctly describes their actual collaboration. That is, every pair of ideas in C is indeed consistent with collaborative choices. Formally, $(x, y) \in C(z, p, q)$ if and only if

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

2. Given $C(z, p, q)$, both p and q agree to collaborate when ideas are (x, y):

$$\alpha x + u(b_p(x)) \leqslant \alpha z + u(\beta_p) \text{ and } \alpha y + u(b_q(y)) \leqslant \alpha z + u(\beta_q).$$

Note that an equilibrium is defined for a particular joint outcome z — an equilibrium collaboration set is a subset of $\{(x, y) : f(x, y) = z\}$. While our analysis focuses on a single joint outcome z, it alludes to a "grand equilibrium," whereby for each z, the collaboration set is consistent with the equilibrium definition above.[11]

[11]That is, z is the intrinsic payoff of all joint projects in the locus $\{(x, y) : f(x, y) = z\}$. It is the observational unit from the public's perspective in case collaboration occurs. For example, suppose that

An equilibrium combines specific sets of ideas $\mathcal{X}$ for p and $\mathcal{Y}$ for q, with each pair of ideas generating z. We write this compactly using the notation $C(z,p,q) = \mathcal{X} \times_z \mathcal{Y}$.

4. Equilibrium

A principal goal in this section is to lead up to

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

The theorem both asserts existence and characterizes equilibria. In each nonempty equilibrium, persons p and q have thresholds $\bar{x}$ and $\bar{y}$ respectively, such that they are willing to collaborate if and only if their drawn ideas satisfty $x \leqslant \bar{x}$ and $y \leqslant \bar{y}$. Because all ideas in $C(z,p,q)$ "add up" to z, person p agrees to collaborate whenever q draws an idea $y \geqslant \underline{y}$, where $f(\bar{x},\underline{y}) = z$, and similarly for person q.

Intuitively, each agent faces the a tradeoff between intrinsic payoff gains in collaboration and potential reputational losses. As for intrinsic payoff, collaboration is always beneficial: the collaborative payoff $z = f(x,y)$ is greater than both x and y, the respective intrinsic payoffs from solo work to agents p and q. On the other hand, collaboration could be associated with a loss in reputational payoff. Specifically, if agent p draws a particularly high-quality idea x, then the posterior update from solo work — where the public sees x as strong *evidence* of high ability — is larger than the update from collaborative work, where the public only sees the joint outcome, and potentially misunderstands p's contribution.

In Section 4.1, we describe this characterization in more detail, and Section 4.2 sketches the proof of Theorem 1.

4.1. 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:

$$\alpha(z-x) \geqslant u(b_p(x)) - u(\beta_p) \quad \text{and} \quad \alpha(z-y) \geqslant u(b_q(x)) - u(\beta_q). \tag{3}$$

On the left-hand sides 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, given a conjectured collaboration set C (and implied β_p and β_q), there are unique $\bar{x}$ and $\bar{y}$ that satisfy both conditions in (3) with equality. Under that conjecture, p agrees to collaborate if and only if $x \leqslant \bar{x}$, where we resolve indifference by collaboration. The

the academic community regards all "well-published papers" as a single observational category. Then C would be defined as all the pairs of ideas that would lead to a "well-published paper" and such that both p and q agree to collaborate.

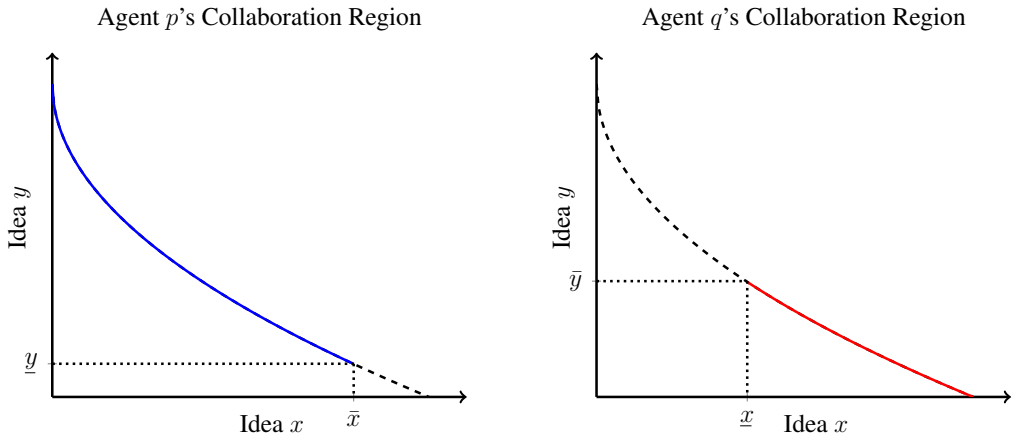

Figure 1. Collaboration regions for agents p and q. The curves display all combinations of ideas x and y that yield a project z. The left panel shows combinations (in blue) such that p agrees to collaborate. The right panel shows combinations (in red) such that q agrees to collaborate.

same is true of q and threshold idea $\bar{y}$.[12] Figure 1 displays these collaboration regions. As discussed, each region — when nonempty — 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). So $C(z,p,q) = [\underline{x},\bar{x}] \times_z [\underline{y},\bar{y}]$.

The updates β_p and β_q can be rewritten to account for this specific equilibrium shape of C. Recall that $\iota_z(w)$ maps each $w \in [0,z]$ to the partner's idea $\iota_z(w)$ on the isoquant for 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 $g(x,p)g(\iota_z(x),q)\,|\iota_z'(x)|$,[13] 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 is further adjusted to $\gamma_z(x)/[\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})]$ (when $\bar{x} > \underline{x}$). It follows that

$$\beta_p = \begin{cases} \dfrac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \displaystyle\int_{\underline{x}}^{\bar{x}} b_p(x)\gamma_z(x)dx & \text{if } \bar{x} > \underline{x} \\ b_p(\bar{x}) & \text{if } \bar{x} = \underline{x} \end{cases} \tag{4}$$

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

[13] 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.

Similarly, define ω_z, which is 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 (4),

$$\beta_q \equiv \begin{cases} \dfrac{1}{\Omega_z(\bar{y})-\Omega_z(\underline{y})}\displaystyle\int_{\underline{y}}^{\bar{y}} b_q(y)\omega_z(y)dy & \text{if } \bar{y} > \underline{y} \\ b_q(\bar{y}) & \text{if } \bar{y} = \underline{y}, \end{cases} \tag{5}$$

We summarize the discussion above as:

Proposition 1. *Under any non-empty equilibrium, there exist* $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ *with* $0 < \underline{x} < \bar{x} < z$ *and* $0 < \underline{y} < \bar{y} < z$*, such that*

$$\alpha(z - \bar{x}) = u(b_p(\bar{x})) - u(\beta_p), \tag{6}$$

$$\alpha(z - \bar{y}) = u(b_q(\bar{y})) - u(\beta_q), \text{ and} \tag{7}$$

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

where β_p *and* β_q *are given by* (4) *and* (5)*.*

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* $0 < \underline{x} < \bar{x} < z$, $0 < \underline{y} < \bar{y} < z$ *as well as* (6) *and* (7)*, then* $C(z,p,q)$ *is an equilibrium.*

Theorem 1 claims that a nonempty equilibrium set always exists. We remark here on the additional possibility that there may be an empty equilibrium. Suppose that p and q refuse 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 exist.[14]

Also note that Theorem 1 holds under the assumption that $\alpha > 0$. If instead $\alpha = 0$, 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 $\alpha > 0$.

[14] 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 set, and will have nothing else to say about the empty case for the rest of the paper.

4.2. Existence. We now sketch the proof of Theorem 1; details are 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 (6) and (7):

$$u(b_p(\bar{x})) + \alpha(\bar{x} - z) = u(\beta_p) \text{ and } u(b_q(\bar{y})) + \alpha(\bar{y} - z) = u(\beta_q), \tag{8}$$

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 (4) and (5). 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 (4) or (5) 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 (4) and (5), 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 $0 < \underline{y} < \bar{y} < z$.

5. Fragile Equilibria

Theorem 1 guarantees the existence of nonempty equilibria. A central theme of our paper revolves around the possibility that there could be many such equilibria. We explore the possibility that some of them feature the asymmetric treatment of individuals who are similar in all payoff-relevant characteristics, and that symmetric treatment is "fragile," in a sense that we now make precise.

Consider an equilibrium at (z, p, q). Temporarily assume that $p = q$, so that both players are equal in their payoff-relevant characteristics. However, suppose that the individuals can be differentiated by some payoff-irrelevant identity, such as race, gender or nationality. To keep track of these identities, we continue to refer to individuals by their "names" p and q. Further, suppose that the equilibrium in question is also symmetric, with common public update β in the event of collaboration.

Now suppose that the public sees individual identities as salient for some reason and "slightly reallocates" credit, say in favor of p: $\beta_p > \beta > \beta_q$. That is, upon seeing a collaboration, the public attributes a relatively better posterior to p and a relatively worse posterior to 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 perceived imbalance.

Understanding that they will benefit from a more generous public update, p is now more willing to collaborate with q. Conversely, q is *less* open to collaborating with p. So in response to this bias, we have $\bar{x} > \bar{y}$ — person p shares ideas of higher quality than q does. Equivalently, $\underline{x} > \underline{y}$. Observe that to some degree, this now asymmetric

sharing behavior confirms the public's initial bias. Furthermore, if these behavioral responses lead to new collaboration sets that "overshoot" the original bias, they may destabilize the symmetric outcome.

We formalize this verbal discussion using the map Θ introduced in the proof of Theorem 1. We do so generally, not presuming that $p = q$. Remember that Θ starts with collaborative updates (β_p, β_q), and then uses (8) 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 set.[15] We will say that an equilibrium at (p, q, z), with associated collaborative update vector (β_p, β_q), is *p-fragile* if there is $\zeta > 0$ and $\delta > 0$ such that for every $\epsilon \in (0, \delta)$,

$$\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) \geqslant \beta_p + (1 + \zeta)\epsilon \text{ and } \Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) \leqslant \beta_q - (1 + \zeta)\epsilon, \tag{9}$$

where the subscripts on Θ refer to its component functions. We analogously define q-fragility. In this definition, we perturb public perceptions by reallocating updates across identities. Locally, we think of an increase in credit for some identity as matched by a symmetric decrease for the other identity. Further, the requirement of $\zeta > 0$ in the definition ensures that small perturbations not only locally amplify but that they do so at some minimal geometric rate.[16]

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

Given Observation 1, we simply refer to an equilibrium as *fragile*, knowing that p- and q-fragility are equivalent. In contrast, if an equilibrium were p-, but not q-fragile, it could be fragile if the public is biased toward's p's identity, but not fragile if those identities were switched.

While fragility is a mathematical concept, we can only condition its use 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 non-fragile with respect to public perceptions.

[15] 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.

[16] 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 (9). But this would have forced us to confront technical yet non-generic situations which create complications of little conceptual interest in the present setting. The gap between the two definitions is analogous to that between a strictly increasing differentiable function, and a differentiable function with a strictly positive first derivative. We ignore such issues.

Proposition 2. *For every (z, p, q), a nonempty and non-fragile equilibrium exists.*

Proposition 2 is silent on the question of whether symmetric equilibria are fragile in symmetric settings. Indeed, we shall see that they often are. 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 such theories, unequal treatment is one equilibrium, but there could be an equally robust equilibrium with equal treatment. Our approach is different, in that it explicitly interrogates the fragility of the equal-treatment outcome.[17] 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.

6. Equilibrium with Symmetric Partners

With the above discussion in mind, we now study the case of symmetric players, who possess identical priors ($p = q$). We investigate whether such symmetric agent-pairs might nevertheless be pushed into robust asymmetric interactions that rely on payoff-irrelevant identities.

6.1. Symmetric Equilibrium.

Theorem 2. *Suppose that $p = q = r$. Then there exists a nonempty symmetric equilibrium, with $\beta_p = \beta_q$ and $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{x}, \bar{x}]$ for some $\underline{x} < \bar{x} \leqslant z$.*

Additionally for all α small or large enough, there is just one symmetric equilibrium.

We remark on the existence strategy, and details are provided in the Appendix. We adapt the map Θ from Theorem 1, imposing $\beta_p = \beta_q = \beta$ throughout. Call it Θ^S; see Figure 2. For any β over a certain domain $[\underline{\beta}, \bar{\beta}]$ (see proof), Θ^S defines $\bar{x}$ using (6) as we did for Theorem 1, then $\underline{x}$ via the isoquant $\iota_z(\bar{x})$, and finally β' by β_p or β_q as in (4) or (5). Two features require mention. First, Θ^S *imposes* symmetry. By mapping $\bar{x}$ to $\underline{x}$ directly, it presumes that the other player is behaving in identical fashion. Second, the map satisfies appropriate end-point conditions: β' lies above β 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.

Theorem 2 also asserts uniqueness of symmetric equilibrium when reputational concerns are either large (a principal case of interest to us), or small. This statement leaves open the question of whether a symmetric equilibrium is *always* unique. Certainly, in all the examples that we have explored, this has always been the case. At the same time, Theorem 2 is stated without any assumption on the curvature of utility or production

[17] As discussed earlier, Gu and Norman (2020) make some progress in examining equilibrium stability in a different setting with statistical discrimination.

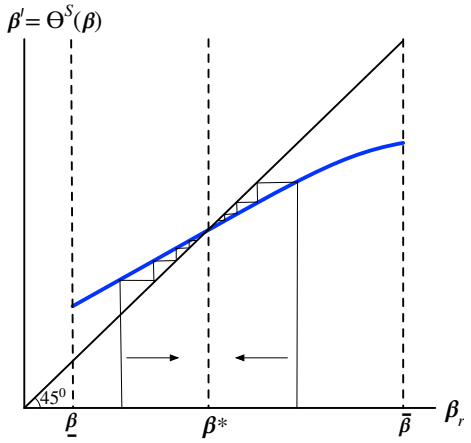

Figure 2. Symmetric Equilibrium Via the Map Θ^S. This figure depicts a situation with a unique symmetric equilibrium. Perturbations of the fixed point lead back iteratively to the equilibrium as shown, yet, as argued in the text, the symmetric equilibrium could be fragile.

functions, so we do not rule out the possibility of multiplicity for some relatively esoteric combinations of preferences and technology. That we do obtain uniqueness under general conditions for small or large α is of some interest, and for this reason we are comfortable with viewing uniqueness (under symmetry) as the leading case.

Additionally, the uniqueness of symmetric equilibrium shows that fragility is a subtle concept. Uniqueness, coupled with the end-point conditions that Θ^S satisfies, suggests that the symmetric equilibrium cannot be fragile. After all, under Θ^S, a perturbation of the equilibrium update must then lead to a sequence of update ratios that "converge back" to the equilibrium ratio. Figure 2 underscores this argument by displaying perturbations of β to the right and left of the unique fixed point β^* of Θ^S, which is the unique symmetric equilibrium. It is obvious that any pseudo-dynamic that follows upon an iterative application of Θ^S points back at β^*.

But this apparent robustness of symmetric equilibrium is misleading. Because Θ^S imposes symmetry across agents, 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, and in fact, it can render every symmetric equilibrium fragile, even when the equilibrium is unique.

6.2. Fragility of Symmetry. Loosely speaking, an equilibrium is fragile when individual responses to some initial perturbation or bias is sufficient to "more than justify" that

initial bias. This statement decomposes into two parts: (i) the effect of the bias on individual responses; and (ii) the effect of those responses on "subsequent" perceptions. Fragility requires either or both these effects to be large.

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

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

The following observation formalizes this discussion.

Observation 2. *A symmetric equilibrium $[\underline{x}, \bar{x}] \times_z [\underline{x}, \bar{x}]$ ascribed to symmetric partners with $p = q = r$ is fragile if and only if*

$$\alpha + u'(b_r(\bar{x}))b_r'(\bar{x}) < u'(\beta_r)\left[\frac{\partial \beta_r}{\partial \bar{x}} - \iota_z'(\bar{x})\frac{\partial \beta_r}{\partial \underline{x}}\right]. \tag{10}$$

where β_r is the Bayes' update from prior r conditional on collaboration.

The terms on the left-hand side of (10) need to be small for effect (i) to be strong, as discussed above. Effect (ii) is represented on the right-hand side of (10). The first term captures the direct effect due to the person's change in their own willingness to collaborate, and the second term mirrors the cross-effect due to the change in the other person's willingness to collaborate, which is mediated by the effect of $\bar{x}$ on $\underline{x}$.

Observation 2 compares these effects precisely, as described in (10). A fairly straightfoward implication is that if the intrinsic value of projects is sufficiently important (α is large enough), then the marginal reaction to biases is dampened, thereby generating symmetric equilibria that are non-fragile.

Corollary 1. *There exists $\bar{\alpha} > 0$ such that if $\alpha > \bar{\alpha}$, no symmetric equilibrium is fragile.*

We omit the proof, as all it requires are minor technical verifications that all the derivatives in (10) are bounded above even as $\bar{x}$ and $\underline{x}$ respond endogenously to α. However, we are more interested in the converse of this statement —- that there is a corresponding threshold of α *below* which a symmetric equilibrium is fragile. This requires deeper exploration, because as α goes to 0, both the right- and left-hand sides of (10)

converge to the same value. To uncover the deeper structure for small α, it will be useful to define curvature elasticities, using the notation

$$\text{Curv}(w,h) \equiv \frac{h''(w)w}{h'(w)}$$

for any twice differentiable function with $h'(w) > 0$. Two particular functions will occupy our attention under any symmetric situation (z,p,q) with $p = q = r$: the function ι_z that maps out the "isoquant" for z, and the function $u \circ b_r$ that translates the Bayes' posterior following some individual signal x into payoff $u(b_r(x))$.

Theorem 3. *Fix a symmetric situation (z,p,q), with $p = q = r$. Suppose that*

$$\text{Curv}(e_z, u \circ b_r) - \frac{1}{2}\text{Curv}(e_z, \iota_z) < -\frac{e_z}{z - e_z}, \tag{11}$$

where e_z is the "equal input" for z; i.e., $f(e_z, e_z) = z$. Then there is $\underline{\alpha} > 0$ such that for every $\alpha < \underline{\alpha}$, the symmetric equilibrium (unique by Theorem 2) is fragile.

The Theorem provides a necessary and sufficient condition, *depending only on the primitives of the model*, for the unique symmetric equilibrium to be fragile for small α; i.e., when reputational concerns are strong. This is in sharp contrast to the assertion in Corollary 1 for large α, where the symmetric equilibrium (again unique by Theorem 2) is *not* fragile.

To unpack and understand condition (11), consider the case where both u and f are linear, that is, the Bayesian posterior on an individual is her reputational payoff, and ideas are additively combined to generate the joint product ($f(x,y) = x + y$). Then a little algebra, along with the fact that $z - e_z = e_z$, shows that (11) reduces to

$$\frac{b_r''(e_z)e_z}{b_r'(e_z)} < -1. \tag{12}$$

This is a primitive condition on the curvature of the Bayesian update that depends on the distribution of ideas for good and bad types. It yields condition (10) — the responsiveness of the update from joint work under a change in collaboration thresholds exceeds the responsiveness of the corresponding update from solo work. Interestingly, condition (12) must *always* holds for an open interval of z-values.

Observation 3. *There is an open set J such that for every $z \in J$, (12) holds.*

For example, if ideas are exponentially distributed for both good type and bad type agents, then condition (12) is satisfied on $J = [\underline{z}, \infty)$, for some $\underline{z} \in \mathbb{R}_+$. The same is true when ideas are distributed according to two Weibull distributions with a common shape parameter, or according to two log-normal distributions.

Observation 3 generates the following corollary to Theorem 3:

Corollary 2. *Suppose u is linear, f linearly additive, and agents are symmetric ($p = q = r$).*

Then for every z in some open set, there is $\underline{\alpha} > 0$ such that if $\alpha < \underline{\alpha}$, there is a unique symmetric equilibrium, which is fragile.

The same condition is sufficient for fragility under other specifications on the production function and reputational payoff; certainly, linearity is not needed. This is registered in the corollary below.[18]

Corollary 3. *Suppose u is concave, f is convex, that $f_{12} \geqslant 0$, and that agents are symmetric ($p = q = r$).*

Then for every z in some open set, there is a threshold $\underline{\alpha} > 0$ such that if $\alpha < \underline{\alpha}$, there is a unique symmetric equilibrium, which is fragile.

The concavity of u helps to increase the responsiveness of reputational payoffs under collaboration, relative to that of reputational payoffs under solo work, evaluated at the threshold $\bar{x}$. That is because reputational payoff β_r is always smaller than the solo payoff $b_r(\bar{x})$ at the threshold $\bar{x}$, so concavity imparts a higher marginal utility to the former variable, accentuating the possibility of fragility. This is further reinforced by the convexity of f and its complementarity in individual ideas. Then $\text{Curv}(e_z, \iota_z) > 0$ and $z \geq 2e_z$, making it still more plausible that (11) will hold, resulting in fragility of the symmetric equilibrium for small α.

The central takeaway: if reputational concerns are uppermost, there is a real danger that symmetric players will not be treated symmetrically when their payoff-irrelevant identities are visible to the public, and those identities are associated with a salient history of unequal treatment. Of course, given Proposition 2, we know that there are other non-fragile equilibria. But these must involve asymmetric treatment.

7. Payoff Implications of Asymmetric Equilibria

Section 5 introduced the notion of fragility, and Section 6 described conditions under which symmetric equilibria across are fragile. When that happens Proposition 2 assures us that other non-fragile *asymmetric* equilibria exist. If society can distinguish between agents using otherwise-irrelevant identities, functionally identical individuals settle into such equilibria.

Because those asymmetries are supported by the existence of payoff-irrelevant identities, each identity will collaborate for distinct sets of ideas. One of the identities

[18] To see that the Corollary holds, note that, if $f_{11} \geqslant f_{12} \geqslant 0$, then $(z - e_z) \geqslant e_z$ and $\iota''_z(e_z) \leqslant 0$.

will be *favored*, in the sense that it will be perceived by the public as (stochastically) contributing better ideas to the collaboration, compared to the other identity.[19]

In this section, we discuss the notion of favored identities, and the payoff implications of favoritism. 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.

7.1. Intrinsic Gains from Collaboration. Conditioning on z, agent p's payoffs are:

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

where R_p stands for reputational payoff and I_p stands for intrinsic payoff. That is,

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

where $u_p^*(x) = u(\beta_p)$ if $x \in [\underline{x}, \bar{x}]$, and $u_p^*(x) = u(b_p(x))$ if $x \notin [\underline{x}, \bar{x}]$. Similarly,

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

Analogous expressions hold for person q, using the thresholds $\{\underline{y}, \bar{y}\}$ and density ω.

This particular formulation of payoffs takes an *ex-interim* stance: it supposes that z is known but not x and y, so that we are calculating the expected payoff conditional on a particular z. From a more global perspective, one should think that there is an "equilibrium collaboration set" ascribed to each z, and to assess payoffs from an *ex-ante* perspective, one could integrate payoffs across all z in a straightforward way.

We now compare collaboration gains across agents and across equilibria, beginning with the symmetric case in which agents have the same prior: $p = q = r$. Proposition 3 shows that, in an asymmetric equilibrium with symmetric agents, the favored identity receives *lower* intrinsic payoff than the dis-favored identity.

Proposition 3. *Consider a pair of agents with common prior $p = q = r$, and an asymmetric equilibrium at z with p enjoying the favored identity. Then*

$$I_p(z) < I_q(z),$$

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

We will discuss this result below, but we first note that its essence extends beyond the symmetric setting. To see this, we extend the concept of favoritism to allow for cases where agents are not functionally identical. That will require us to conceive of agents as being rela tively favored or dis-favored *across equilibria*. Specifically,

[19] In fact, if an asymmetric equilibrium $[\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ is ascribed to symmetric players, then one identity (say p's) must be favored, with $\underline{x} > \underline{y}$, $\bar{x} > \bar{y}$ and $\beta_p > \beta_q$. That is, the comparison is unambiguous: p is viewed as stochastically contributing the better ideas to the collaborative outcome.

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 p receives a higher collaborative update in equilibrium 1 relative to 2, while the opposite is true of q. That is, $\beta_p(z,1) > \beta_p(z,2)$ and $\beta_q(z,1) < \beta_q(z,2)$.

An extreme and rather unambiguous situation occurs when p's *worst* idea under collaboration is viewed as better than q's *best* idea; that is, when $\underline{x} > \bar{y}$. Say then that p is *super-favored*. (This case will require separate treatment.)

Finally, observe that with asymmetric agents, individuals have different "baseline payoffs": the bracketed term describing I_p in (15) is a person-specific constant. We therefore compare intrinsic payoff *gains* by netting these terms out, defining:

$$\Delta_p^I(z) \equiv \alpha \int_{\underline{x}}^{\bar{x}} (z-x)\gamma_z(x)dx,$$

with a parallel definition for q.

Proposition 4. (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 3 and 4 together make the point that persons or identities favored by the public in the perception of their joint 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 is of course "justified" in equilibrium, with p contributing the better ideas incentivized by the public bias. But it is precisely for this reason that the favored individual loses out on the intrinsic gains from collaboration.

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 do these observations have for the *overall* effects of favoritism?

7.2. Overall Gains With Linear Reputational Payoff. When reputational payoff is linear, that payoff must be equal, in expectation, across all equilibria. This is a consequence of the martingale property of Bayesian updates. It implies the following corollary to Proposition 3:

Corollary 4. *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,*

$$\alpha z + u(\beta_p) > \alpha z + u(\beta_q)$$

so that p is relatively better off conditional on collaboration (as a trivial consequence of $\beta_p > \beta_q$), but

$$\Pi_p(z) = R_p(z) + I_p(z) < \Pi_q(z) = R_q(z) + I_q(z)$$

so that q receives the higher unconditional expected payoff.

In Figure 3, Corollary 4 is illustrated by numerical calculations when both the good and the bad distributions of ideas are exponential. 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.[20] To the best of our knowledge, the observation that the payoff ordering may be *reversed* across the reputational and the overall perspectives is a novel contribution.

Specifically, discrimination in our model stems from a collaborative interaction which depends on willing participation. 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 effect is that conditional on collaboration, signaling value favors the favored identity (by construction). But the dis-favored identity becomes less willing to collaborate, which negatively affects the payoff to the favored identity. By Bayes' plausibility, when reputational payoffs are linear, the first effect *must* be dominated by the second.

7.3. Symmetric and Asymmetric Distributions of Posteriors. Corollary 4 must be qualified when reputational payoffs are nonlinear. While our observations 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.

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

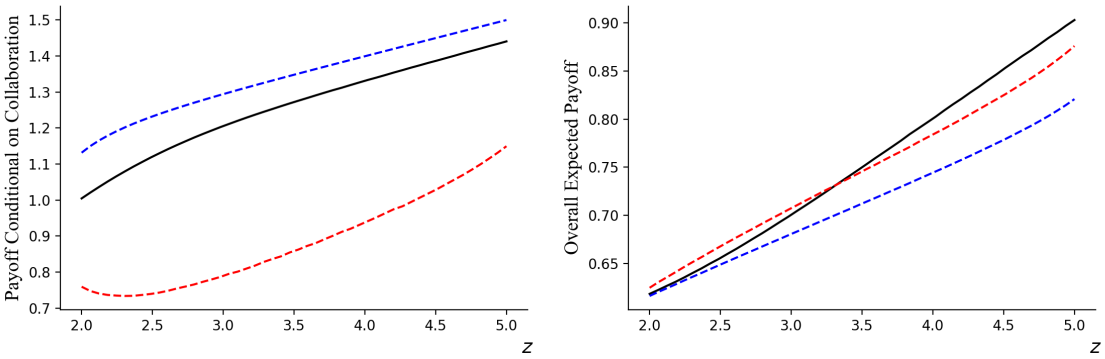

Figure 3. Payoffs in Symmetric and Asymmetric Equilibria When $g(\cdot, 1)$ and $g(\cdot, 0)$ Are Exponential. The solid black lines indicate equal payoffs in the symmetric equilibrium. The blue and red dashed lines in each panel represent payoffs to the favored and disfavored identities in an asymmetric equilibrium. The favored identity is better off conditional on collaboration (left panel), but is worse off in terms of overall payoffs (right panel).

Given z, let $P_p(t, z)$ and $P_q(t, z)$ be the probabilities that the induced posterior on agents p and q are strictly larger than some "target posterior" t.[21] Think about this target posterior as some desired threshold that is relevant for career advancement — for instance, it may be that the individual will receiver a promotion or award if the observer's posterior exceeds t.

Is this what agents care for when they build reputation? 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 5 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 5. *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,*

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

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

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

[21] 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.

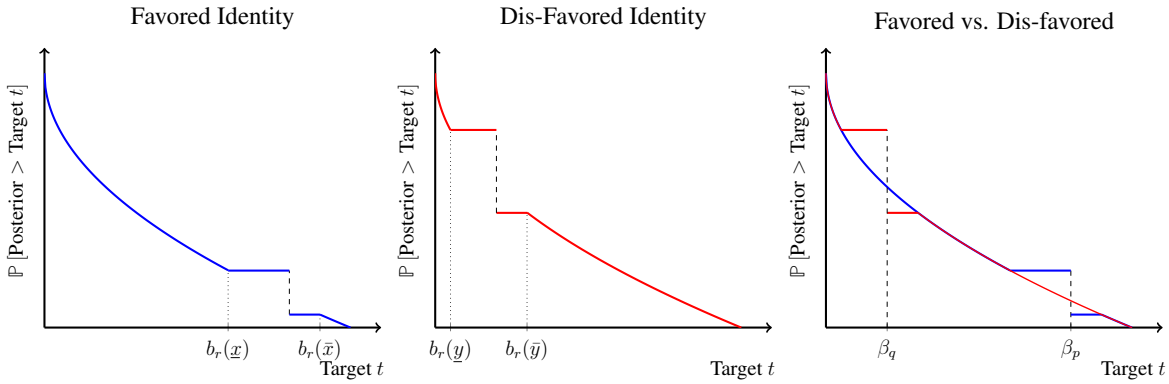

Figure 4. Distribution of Posteriors in Asymmetric Equilibria. The first panel plots this distribution for the favored identity; the second for the dis-favored identity. The third panel combines the two.

Naturally, in an equilibrium with symmetric collaboration, both agents have the same distribution of posteriors and so reach any target t with the same probability. But matters are different when collaboration is asymmetric. Figure 4 displays the distribution of posteriors for an asymmetric collaboration structure. On the horizontal axes, we plot various target thresholds for the posterior. On the vertical axes, we plot the probability that an agent's posterior will be larger than some target posterior t. Remember that when agents collaborate, the public sees only z and is unable to tell which combinations of x and y generated that outcome. So all the possible ideas that could lead to equilibrium collaboration are "garbled" to make up the expected update of the observer. That leads to the pictured flat regions and discontinuities in the posterior distributions. The first panel plots this distribution for the favored identity; the second for the dis-favored identity. The third panel combines the two.

For targets below β_q but close to it — specifically, when t lies between $b_r(\underline{y})$ and β_q — the dis-favored identity q is more likely to reach the target than her favored counterpart. For if the favored identity p has idea $\bar{x}$ (or slightly lower), then the pair collaborates and the public's update on q is β_q. If, conversely, q were to have the same idea $\bar{x}$, the agents would work separately, and the public's update on p would be $b_p(\bar{y})$, which is smaller than β_q. Similar arguments illustrate the other differences in the posterior distributions across p and q.

7.4. Majority Identities and Favoritism. The payoff gains and losses reported in Sections 7.2 and 7.3 hold fixed the identity of the partner. But there is also the question of partner matching. Specifically, consider a population version of the model, with all agents symmetric: $p = q$, but divided into two payoff-irrelevant identities of disparate sizes. Each agent is randomly paired to one potential collaborator. Following this pairing, our model proceeds as before.

Suppose that the symmetric equilibrium is fragile, so that two matched agents of different identities engage in an asymmetric equilibrium in which the majority identity, indexed by p, is favored. If two agents of the same identity meet, they play the symmetric equilibrium. Then the ex-ante payoff A for each identity is given by

$$(16)\quad A_p = \sigma \int_z \Pi_p(z) + (1-\sigma) \int_z \Pi(z) \quad \text{and} \quad A_q = \sigma \int_z \Pi(z) + (1-\sigma) \int_z \Pi_q(z),$$

where $\sigma \in (01/2)$ is the size of the minority identity, $\Pi_p(z)$ and $\Pi_q(z)$ are the expected payoffs on cross-identity matches (recall (13)) and $\Pi(z)$ is the expected payoff in symmetric equilibrium. Now consider the exponential example displayed in the right panel of Figure 3, which depicts symmetric and asymmetric ex-interim payoffs 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.

We are now potentially confronted by yet another reversal in payoffs, stemming from the fact that the minority identity faces a larger share of cross-identity matches relative to the majority identity. It is still true that Propositions 3 and 4 continue to hold for each cross-identity match, so that the dis-favored 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 6. *Consider the symmetric matching model. Suppose that expected ex-ante payoff under the symmetric equilibrium dominates expected payoffs to the dis-favored minority; that is, $\int_z \Pi(z) > \int_z \Pi_q(z)$. Then for all σ small, $A_p > A_q$, even though under each match and each z, we have $I_q(z) > I_q(z)$, as in Proposition 3.*

Under the conditions of Proposition 6, the selection of asymmetric equilibria across identities generates a disincentive for cross-identity collaboration. Symmetric equilibria played within identities have the opposite effect. In such situations, individuals seek out and collaborate disproportionally with others of their own identity. Our model therefore predicts a possible basis for collaborative homophily, though the analysis here only scratches the surface.

In summary, discrimination and favoritism have complex implications. Ex post, conditional on collaboration, the favored identity must gain — this is true by definition. Taking a step back, and conditioning only on the potential partner and not the realization of collaboration, the term “favored identity" could become a misnomer as their intrinsic expected payoffs from collaboration are lower, and could serve to outweigh the collaborative gain, as in Corollary 4. 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.

8. Efficiency and Authorship Ordering

We end with some remarks on the efficiency of equilibrium outcomes, as opposed to the distribution of payoffs across identities. In our model there is a tension between the value of collaboration and the private desire to signal, and that results in inefficient collaboration decisions. We discuss this issue, as also a possible partial solution via the use of certified random order (Ray ⓡ Robson 2018).

8.1. Inefficiency. For any $z > 0$, consider the subspace of all idea pairs that would yield z as collaborative output were they to be pursued jointly. We say that an equilibrium is *inefficient at* z if there is some other collaborative arrangement of the form $\mathcal{X} \times_z \mathcal{Y}$ such that both players receive a higher expected payoff conditional on z.

When u is linear or concave, full collaboration is socially optimal, both for society as a whole and even restricted to the small collective of our two agents. But the desire to signal ruptures optimal collaboration. To describe this, remember that the equilibrium payoff to p is

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

for any $z > 0$, where $u_p^*(x) = u(\beta_p)$ if $x \in [\underline{x}, \bar{x}]$, and $u_p^*(x) = u(b_p(x))$ if $x \notin [\underline{x}, \bar{x}]$. A parallel expression holds for q. The first term is the expected payoff from reputation. The second term is an individual-specific baseline constant, unaffected by equilibrium strategy. The third term represents the expected intrinsic gains from collaboration. All expectations are taken over individual ideas, conditional on z.

Suppose that u is linear. Then expected reputational payoff is just the expected posterior starting from a prior of p. By Bayes plausibility, this term must equal the prior, and so also becomes an individual constant. In short, *all* the private and social gains from pairwise interaction come from the intrinsic value of collaboration. 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 full collaboration is precluded in equilibrium due to a lack of commitment. Suppose that an agent has an excellent idea and their partner has a bad idea. From that ex-post perspective, the agent with the good idea understands that the intrinsic gain from collaboration may not overcome the loss of signaling value. Therefore, while collaboration is valuable in terms of its intrinsic payoff, it will not always happen.

When u is not concave, full collaboration will generally not be desirable from the joint perspective of the two agents. The local strict convexity of u in some regions

might lead them to prefer individual updates in reputation, which makes solo research more valuable. It is still true, though, that equilibria will generally fail to generate the socially optimal level of collaboration. Equilibrium and optimality conditions are distinct, barring non-generic coincidences, so the argument above works for any reputational utility function.

8.2. Merit-Based and Random Order in Collaboration. We now explore the intuition that policies that help to disentangle each person's contributions to a joint project would make 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 in the sequence of their *ordinal* contribution to the joint project. That "merit order" has the immediate impact of reducing the extent of informational garbling. Say p is the lead author. Now the observer additionally knows that contributions lie in the set $M^p(z) = C(z, p, q) \cap \{(x, y)|x \geqslant y\}$. Might that spur more collaboration?

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

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

(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 7. *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 set in the benchmark equilibrium under consideration. There is at least one person for whom the upper collaboration threshold (say $\bar{x}$) exceeds the lower threshold ($\underline{y}$) of his partner. Imagine adding to these thresholds an additional sliver of idea combinations (x, y) such that $x > \bar{x}$ and $y < \underline{y}$, demarcating these with merit order. (One can do the same with the mirror thresholds $\bar{y}$ and $\underline{x}$, assuming $\bar{y} > \underline{x}$.) Just as in the benchmark model, there will be limits to collaboration: at some idea *strictly* smaller than z, the lead author would rather go solo; simply inspect (6). 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 ordinally communicating such information *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 speculation in the first place.

9. Conclusion

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

We develop this framework, and use it (among other things) to argue that robust equilibria often necessitate discrimination, wherein the public attributes greater credit for collaborative work to individuals who belong to certain favored identities. We view these theoretical predictions as a natural accompaniment to empirical evidence regarding collaborative work in academic research, which shows that more credit is assigned to men for work produced in mixed-gender teams. 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 (with women suffering relative to men), even if partners are observationally the same in payoff-relevant attributes.

At the same time, we develop the idea that such favored individuals might be worse off because others with productive ideas may set the bar higher for collaboration, knowing that they will receive less than their fair share of credit. The net effect, while still leading to better outcomes for the publicly favored conditional on collaboration as in the paragraph above, can lead to other less favorable conclusions in an unconditional assessment. For instance, as discussed in the main text, Card, DellaVigna, Funk and Iriberri (2021) have argued that the female-male gap in elections to the Fellowship of the Econometric Society is actually positive over 1980–2010, and in the last decade, this positive coefficient has become sizable and significant. Our theory predicts possibility this even as it predicts biased updates conditional on collaboration.

There are three directions that we see as natural extensions of our current model and plan on exploring in future research. First, in our baseline model, the public's posteriors on agent types are always calculated according to Bayes' Rule. This is the case both when individuals work alone, so that the Bayesian update follows directly from the observed outcome, but also when ideas are combined, in which case the Bayesian calculation also relies on the conjectured collaboration set. However, this Bayesian assumption is not essential, and the model can be easily extended to accommodate other updating rules that rely on the observed project outcomes and the public's collaboration conjecture. With such an extension, we can explore the relation between updating behavior (specifically, behavioral distortions of that behavior away from Bayes) and the structure of equilibrium collaboration and discrimination.

Second, because our model speaks directly to empirical observations on academic collaboration and other team-based projects, it can be adapted to the empirical estimation of a model based on our framework. That estimated model would permit us to evaluate different policies — for example, the merit-based ordering policy we propose in Section 8. It could also serve to identify the direction of equilibrium selection when the equilibrium set is multi-valued, as it typically is in our setting.

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

References

Akerlof, George, and Rachel Kranton. (2000) “Economics and Identity.” *Quarterly Journal of Economics*, **115**: 715-753.

Akerlof, Robert, and Luis Rayo. (2020) “Narratives and the Economics of the Family,” *working paper*.

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

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.

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.

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

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

Card, David, DellaVigna, Stefano, Funk, Patricia, Iriberri, Nagore. (2021) “Gender Differences in Peer Recognition by Economists.” *NBER working paper* w28942.

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

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*.

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.

Levy, Gilat (2007) “Decision Making in Committees: Transparency, Reputation, and Voting Rules.” *American Economic Review*, **97**:150-168.

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

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

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

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

Ong, D., Chan, H. F., Torgler, B., and Yang, Y. A. (2018) “Collaboration Incentives: Endogenous Selection into Single and Coauthorships by Surname Initial in Economics and Management.” *Journal of Economic Behavior & Organization*, **147**: 41-57.

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

Pęski, Marcin, and Balázs Szentes. (2013) “Spontaneous Discrimination.” *American Economic Review* **103**: 2412-36.

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

Ray, Debraj, Baland, Jean-Marie, and Olivier Dagnelie (2007), “Inequality and Inefficiency in Joint Projects," *Economic Journal* **117**, 922–935

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.” *American Economic Review 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**.

Visser, Bauke and Otto H. Swank. (2007) "On Committees of Experts." *Quarterly Journal of Economics*, **122**: 337-372.

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

10. Appendix: Proofs

10.1. Proof of Proposition 1. If (3) 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 (3) holds with equality, giving us (6) and (7) via the transformations (4) and (5).

If (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 $y = \underline{y}$, (7) holds with ">," whereas at $y = z$, (7) 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 $\{\underline{x}, \bar{x}, \underline{y}, \bar{y}\}$ with $0 \leqslant \underline{x} \leqslant \bar{x} \leqslant z$ and $0 \leqslant \underline{y} \leqslant \bar{y} \leqslant z$, satisfying (6) and (7). Suppose that the public forms the beliefs $C(z, p, q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. 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 (6) holds and the right-hand side of (6) is increasing in x. The same argument holds for q, and therefore we have an equilibrium. ■

10.2. 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

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

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

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

and then β_p' and β_q' by the resulting collaborative updates as defined in (4) and (5).

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

$$\underline{x}^* = \iota_z(\bar{y}^*) < \bar{x}^* \text{ and } \underline{y}^* = \iota_z(\bar{x}^*) < \bar{y}^*. \tag{19}$$

To prove (19), it will suffice to show that $\underline{x}^* < \bar{x}^*$ and $\underline{y}^* < \bar{y}^*$. Suppose not, then (say) $\underline{x}^* = \bar{x}^*$. So by the formula for collaborative updates, $\beta_p^* = b_p(\bar{x}^*)$. At the same time, (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 the definition of the function β_q in (5), 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 (19) implies $\beta_p^* < b_p(z)$ and $\beta_q^* < b_q(z)$. Therefore, by (17), $\bar{x}^* < z$ and $\bar{y}^* < z$. Using (19), 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) and the just-established (19), along with Proposition 1. ∎

10.3. Proof of Observation 1. Suppose that an equilibrium is p-fragile. Then, recalling (9), there is $\zeta > 0$ and $\delta > 0$ such that for every $\epsilon \in (0, \delta)$,

$$\Theta_p(\beta_p + \epsilon, \beta_q - \epsilon) \geqslant \beta_p + (1+\zeta)\epsilon \text{ and } \Theta_q(\beta_p + \epsilon, \beta_q - \epsilon) \leqslant \beta_q - (1+\zeta)\epsilon, \tag{20}$$

where Θ_p and Θ_q are the component maps of Θ. Using the fact that $(\beta_p, \beta_q) = \Theta(\beta_p, \beta_q)$ in equilibrium, (9) is equivalent to

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

Recalling the construction of Θ around the equilibrium point (see (17) and (18)), noting that $\underline{x} = \iota_z(\bar{y})$ and $\underline{y} = \iota_z(\bar{x})$ at any equilibrium point, and given that f is continuously differentiable, it follows that Θ is continuously differentiable. Therefore (21) is equivalent to

$$\frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_p} - \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_q} > 1 \text{ and } \frac{\partial \Theta_q(\beta_p, \beta_q)}{\partial \beta_q} - \frac{\partial \Theta_q(\beta_p, \beta_q)}{\partial \beta_p} > 1, \tag{22}$$

where these derivatives are evaluated at the equilibrium (β_p, β_q). But (23) is entirely symmetric across p and q, and so must also be equivalent to q-fragility. ∎

10.4. Proof of Proposition 2. Fix (z, p, q). We first observe that the set of equilibria is compact, and so consequently 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 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

$$\text{(23)} \quad \text{(a)} \quad \Theta_p(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon) > \bar{\beta}_p + \epsilon \quad \text{and} \quad \text{(b)} \quad \Theta_q(\bar{\beta}_p + \epsilon, \underline{\beta}_q - \epsilon) < \underline{\beta}_q - \epsilon,$$

Given Step 1, (23a) implies $B_1(\underline{\beta}_q - \epsilon) > \bar{\beta}_p + \epsilon$. Using this inequality along with (23b) 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, and consequently can be associated with an equilibrium, as in the proof of Theorem 1. But that contradicts the definition of $\underline{\beta}_q$. ■

In what follows, for any z and for any pair of thresholds $\underline{x} < \bar{x}$, write the collaborative update β explicitly as a function of those thresholds $\underline{x}$ and $\bar{x}$, in line with (4):

$$\beta_r(\underline{x}, \bar{x}) = \frac{1}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})} \int_{\underline{x}}^{\bar{x}} b_r(x)\gamma_z(x)dx. \tag{24}$$

Lemma 1. *The function $\beta_r(\bar{x}, \underline{x})$ satisfies*

$$\lim_{\bar{x}\downarrow e_z, \underline{x}\uparrow e_z} \frac{\partial \beta_r(\underline{x}, \bar{x})}{\partial \bar{x}} = \lim_{\bar{x}\downarrow e_z, \underline{x}\uparrow e_z} \frac{\partial \beta_r(\underline{x}, \bar{x})}{\partial \underline{x}} = \frac{b_r'(e_z)}{2}. \tag{25}$$

and

$$\lim_{\bar{x}\downarrow e_z, \underline{x}\uparrow e_z} \frac{\partial^2 \beta_r(\underline{x}, \bar{x})}{\partial \bar{x}^2} = \lim_{\bar{x}\downarrow e_z, \underline{x}\uparrow e_z} \frac{\partial^2 \beta_r(\underline{x}, \bar{x})}{\partial \underline{x}^2} = \frac{b_r''(e_z)}{3} + \frac{b_r'(e_z)\gamma_z'(e_z)}{6\gamma_z(e_z)}. \tag{26}$$

Proof. It is easy to compute from (24) that for any $\bar{x} > \underline{x}$,

$$\frac{\partial \beta_r}{\partial \bar{x}}(\underline{x}, \bar{x}) = \frac{[b_r(\bar{x}) - \beta_r(\underline{x}, \bar{x})]\,\gamma_z(\bar{x})}{\Gamma_z(\bar{x}) - \Gamma_z(\underline{x})}. \tag{27}$$

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

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

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

To establish (26), differentiate (27) with respect to $\bar{x}$ to see that:

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

where we invoke (27) again. Using L'Hospital's Rule once more, we have

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

which implies that

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

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

10.5. Proof of Theorem 2. Fix r and $z > 0$. For any $\beta_r \in [0, b_r(z)]$, to be interpreted as an update ratio, define $\bar{x}$ by (6), restated here as

$$u(b_r(\bar{x})) - \alpha[z - \bar{x}] = u(\beta_r). \tag{28}$$

and then define $\underline{x}$ by

$$\underline{x} = \iota_z(\bar{x}). \tag{29}$$

Let $\underline{\beta} \in [0, b_r(z))$ be 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 (28) 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.

Restricting attention to the domain $[\underline{\beta}, b_r(z)]$, define a map $\Theta^S(\beta_r)$ as follows. Define $\bar{x}$ and $\underline{x}$ by (28) and (29), and then $\beta_r' = \Theta^S(\beta_r)$ according to (4). 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 (28) and (29). It is immediate that $(\underline{x}^*, \bar{x}^*)$ is a symmetric equilibrium.

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

$$\frac{d\beta_r'}{d\beta_r} = \left[\frac{\partial \beta_r'}{\partial \underline{x}}\frac{d\underline{x}}{d\bar{x}} + \frac{\partial \beta_r'}{\partial \bar{x}}\right]\frac{d\bar{x}}{d\beta_r} = \left[\frac{\partial \beta_r'}{\partial \underline{x}}\iota_z'(\bar{x}) + \frac{\partial \beta_r'}{\partial \bar{x}}\right]\frac{u'(\beta_r)}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha} \tag{30}$$

where the second equality follows easily from (28). By assumption, $\iota_z'(\bar{x})$ and $u'(b_r(\bar{x}))$ are bounded. It is easy to check by direct computation (use, e.g., (27)) that the partial derivatives of β_r' are bounded above. It follows that for all α large enough, the right hand side of (30) must be strictly smaller than 1, no matter which fixed point of Θ^S we pick. It follows that there can be just one fixed point, which completes the proof for large α.

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

$$e_z - \epsilon \leq \underline{x}(\alpha) < \bar{x}(\alpha) \leq e_z + \epsilon \tag{31}$$

for every pair of symmetric equilibrium thresholds, where remember that e_z is the unique value such that $f(e_z, e_z) = z$. We already know that $\underline{x}(\alpha) < \bar{x}(\alpha)$, so if (31) is false; then there exists $\epsilon > 0$ and $\alpha \to 0$ such that for every α, there is some symmetric equilibrium threshold $\bar{x}(\alpha)$ with $\bar{x}(\alpha) \geqslant e_z + \epsilon$.[22] Moreover, $\underline{x}(\alpha) \leqslant e_z$. In particular,

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

given that u and b_r are strictly increasing, there is $\delta > 0$ such that

$$u(b_r(\bar{x}(\alpha))) - u(\beta_r(\alpha)) \geqslant \delta \tag{32}$$

for all n, where $\beta_r(\alpha)$ is the corresponding equilibrium update under α conditional on collaboration. At the same time, using the equilibrium condition (6), we see that $u(b_r(\bar{x}(\alpha))) - u(\beta_r(\alpha)) \to 0$ as $\alpha \to 0$, but that contradicts (32). So the claim is true, and both $\bar{x}$ and $\underline{x}$ converge to e_z along any sequence of symmetric equilibria as $\alpha \to 0$.

To complete the proof of uniqueness for small α, use (25) of Lemma 1 in equation (30), along with the facts that $(\bar{x}, \underline{x}) \to (e_z, e_z)$, $\iota_z'(e_z) = -1$ and $u'(e_z)b_r'(e_z)$ strictly positive to conclude that the right hand side of (30) converges to 0 as $\alpha \to 0$, no matter which sequence of fixed points of Θ^S we pick. It follows that there can be just one fixed point. ∎

10.6. Proof of Observation 2. We already know that fragility is equivalent to (23). In a symmetric equilibrium, the two inequalities are identical and equivalent to

$$\frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_p} - \frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_q} > 1, \tag{33}$$

evaluated at $p = q = r$. We use the definition of the mapping Θ in section 5 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:

$$\frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_p} = \left[\frac{\partial \beta_p'}{\partial \bar{x}}\right]\left[\frac{d\bar{x}}{d\beta_p}\right] = \left[\frac{\partial \beta_r'}{\partial \bar{x}}\right]\frac{u'(\beta_r)}{u'(b_r(\bar{x}))b_r'(\bar{x}) + \alpha} \tag{34}$$

and

$$\frac{\partial \Theta_p(\beta_p, \beta_q)}{\partial \beta_q} = \left[\frac{\partial \beta_p'}{\partial \underline{x}}\right]\left[\frac{\partial \underline{x}}{\partial \bar{y}}\right]\left[\frac{d\bar{y}}{d\beta_q}\right] = \left[\frac{\partial \beta_r'}{\partial \underline{x}}\right]\frac{u'(\beta_r)\iota_z'(\bar{y})}{u'(b_r(\bar{y}))b_r'(\bar{y}) + \alpha} \tag{35}$$

Combining (33), (34) and (35), using symmetry to note that $\bar{x} = \bar{y}$ and $\gamma_z(\bar{x}) = \gamma_z(\underline{x})$,[23] and rearranging terms, we obtain (10), as desired. ∎

By Theorem 2, for small α there is a unique symmetric equilibrium. So there is a corresponding collection $\{\bar{x}(\alpha), \underline{x}(\alpha)\}$ of uniquely defined equilibrium thresholds, along with equilibrium collaborative updates $\beta_r(\underline{x}(\alpha), \underline{\lambda}(\alpha))$, satisfying

$$u(b_r(\bar{x}(\alpha))) + \alpha[\bar{x}(\alpha) - z] = u(\beta_r(\underline{x}(\alpha), \bar{x}(\alpha)), \text{ and } \underline{x}(\alpha) = \iota_z(\bar{x}(a)). \tag{36}$$

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

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

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

Proof. From (36), we have

$$u'(b_r(\bar{x}(\alpha)))b_r'(\bar{x}(\alpha))\bar{x}'(\alpha) + [\bar{x}(\alpha) - z] + \alpha[\bar{x}'(\alpha)] = u'(\beta_r(\bar{x}(\alpha), \underline{x}(\alpha)))\frac{d\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{d\alpha}. \tag{37}$$

Now observe that

$$\begin{aligned}\frac{d\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{d\alpha} &= \frac{\partial\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{\partial\bar{x}}\bar{x}'(\alpha) + \frac{\partial\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{\partial\underline{x}}\underline{x}'(\alpha) \\ &= \bar{x}'(\alpha)\left[\frac{\partial\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{\partial\bar{x}} + \frac{\partial\beta_r(\underline{x}(\alpha), \bar{x}(\alpha))}{\partial\underline{x}}\iota_z'(\bar{x}(\alpha))\right]. \qquad (38)\end{aligned}$$

By (31) in the proof of Theorem 2, we know that $\underline{x}(\alpha)$ and $\bar{x}(\alpha)$ both converge to e_z as $\alpha \to 0$. Invoking equation (25) of Lemma 1, and using the fact that $\iota_z'(\bar{x}(\alpha)) \to -1$ as $\alpha \to 0$, we see that the term in the square brackets in (38) vanishes as $\alpha \to 0$. Using this information and combining (37) with (38), we see that

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

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

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

■

10.7. Proof of Theorem 3. Recall condition (10) for fragility, slightly rewritten as:

$$\alpha < u'(\beta_r)\frac{\partial\beta_r}{\partial\bar{x}}(\underline{x}(\alpha), \bar{x}(\alpha)) - u'(\beta_r)\iota_z'(\bar{x})\frac{\partial\beta_r}{\partial\underline{x}}(\underline{x}(\alpha), \bar{x}(\alpha)) - u'(b_r(\bar{x}(\alpha)))b_r'(\bar{x}(\alpha)). \tag{39}$$

Using Lemma 1, it is easy to see that both the left-hand side and the right-hand side of condition (39) approach 0 as $\alpha \to 0$. And so, in order to evaluate whether (39) holds when α is close to 0, we must evaluate the derivatives of the left- and right-hand sides of (39) as $\alpha \to 0$. The left hand side obviously has derivative equal to 1. As for the right-hand side, we differentiate to get (arguments omitted for ease in writing):

$$\begin{aligned}\frac{\partial \text{ RHS}}{\partial\alpha} &= u''(\beta_r)\left[\frac{\partial\beta_r}{\partial\bar{x}}\bar{x}'(\alpha) + \frac{\partial\beta_r}{\partial\underline{x}}\underline{x}'(\alpha)\right]\frac{\partial\beta_r}{\partial\bar{x}} - u''(\beta_r)\iota_z'(\bar{x})\left[\frac{\partial\beta_r}{\partial\bar{x}}\bar{x}'(\alpha) + \frac{\partial\beta_r}{\partial\underline{x}}\underline{x}'(\alpha)\right]\frac{\partial\beta_r}{\partial\underline{x}} \\ &\quad -u''(b_r(\bar{x}))\left[b_r'(\bar{x})\right]^2\bar{x}'(\alpha) - \iota_z''(\bar{x})u'(\beta_r)\frac{\partial\beta_r}{\partial\underline{x}}\bar{x}'(\alpha) \\ &\quad +u'(\beta_r)\left[\frac{\partial^2\beta_r}{\partial\bar{x}^2}\bar{x}'(\alpha) + \frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}}\underline{x}'(\alpha)\right] - u'(\beta_r)\iota'(\bar{x})\left[\frac{\partial^2\beta_r}{\partial\bar{x}\partial\underline{x}}\bar{x}'(\alpha) + \frac{\partial^2\beta_r}{\partial\underline{x}^2}\underline{x}'(\alpha)\right] \\ &\quad -u'(b_r(\bar{x}))b_r''(\bar{x})\bar{x}'(\alpha).\end{aligned} \tag{40}$$

Equation (25) of Lemma 1 implies that the first two terms in (40) approach 0 as $\alpha \to 0$. Equation (26) of Lemma 1, along with the fact that $\lim_{\alpha\to 0} \frac{d\underline{x}}{d\alpha} = \lim_{\alpha\to 0} \iota_z'(\bar{x}(\alpha))\frac{d\bar{x}}{d\alpha} = -\lim_{\alpha\to 0}\frac{d\bar{x}}{d\alpha}$, imply that in the limit as $\alpha \to 0$, the fifth and sixth terms cancel each other out. Applying these cancelations, we get:

$$\lim_{\alpha\to 0}\frac{\partial \text{ RHS}}{\partial \alpha} = \lim_{\alpha\to 0}\left[-u''(b_r(\bar{x}))\left[b_r'(\bar{x})\right]^2\bar{x}'(\alpha) - \iota_z''(\bar{x})u'(\beta_r)\frac{\partial \beta_r}{\partial \underline{x}}\frac{\partial \bar{x}}{\partial \alpha} - u'(b_r(\bar{x}))b_r''(\bar{x})\bar{x}'(\alpha)\right].$$

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

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

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

$$\frac{u''(b_r(e_z))}{u'(b_r(e_z))}b_r'(e_z)e_z + \frac{1}{2}\iota''(e_z)e_z + \frac{b_r''(e_z)}{b_r'(e_z)}e_z < -\frac{e_z}{z-e_z}. \tag{41}$$

It is easy to show by direct computation that Curv $(e_z, u\circ b_r) = \frac{u''(b_r(e_z))}{u'(b_r(e_z))}b_r'(e_z)e + \frac{b_r''(e_z)}{b_r'(e_z)}e_z$ and that Curv $(e_z, \iota_z) = -\iota''(e_z)e_z$ (the latter because $\iota_z'(e_z) = -1$). Making these substitutions in (41), we obtain (11). ∎

10.8. Proof of Proposition 3. 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.

10.9. Proof of Proposition 4. Part (i). Subtracting the intrinsic gains of q from those of p, it is easy to see that

$$\Delta_p^I(z) - \Delta_q^I(z) = \int_{\underline{x}}^{\bar{x}}[\iota_z(x) - x]\gamma_z(x)dx. \tag{42}$$

Because p is unambiguously favored, $\iota_z(x) < x$ for all $x \in [\underline{x}, \bar{x}]$, so by (42), $\Delta_p - \Delta_q < 0$.

Part (ii). Because p is favored in equilibrium 1 over 2, and q is dis-favored, it must be — using (6) and (7) — 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 (42) for each equilibrium j, indexing $\Delta_p^I(z)$ and $\Delta_q^I(z)$ by j. Then

$$\delta_j \equiv \Delta_{p,j}^I(z) - \Delta_{q,j}^I(z) = \int_{\underline{x}_j}^{\bar{x}_j}[\iota_z(x) - x]\gamma_z(x)dx. \tag{43}$$

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, but p is favored in 1 over 2, we have

$$\underline{x}_2 < \underline{x}_1 \le e(x) \le \bar{x}_2 < \bar{x}_1. \tag{44}$$

Using (43), 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 (44)), and that $\iota_z(x) < x$ for $x \in [\bar{x}_2, \bar{x}_1)$ (an implication of the third and fourth inequalities in (44)). ■

10.10. Proof of Proposition 5. Consider an equilibrium collaboration set $C = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$. Then

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

and

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

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. ■

10.11. Proof of Proposition 6. See main text. ■

10.12. Proof of Proposition 7. Suppose $C(z,p,q) = [\underline{x}, \bar{x}] \times_z [\underline{y}, \bar{y}]$ is an equilibrium collaboration set of the original model with no authorship ordering. Augment the equilibrium collaboration set as follows. Define x° by the smallest solution in x (but exceeding $\bar{x}$) to

$$u(b_p(x)) + \alpha[x - z] = u(\beta_p(\bar{x}, x)). \tag{45}$$

The left hand side of (45) is strictly smaller than the right hand side at $x = \bar{x}$, because $\beta(\bar{x}, \bar{x}) > \beta(\underline{x}, \bar{x}) = u(b_p(\bar{x})) + \alpha[\bar{x} - z]$ by the equilibrium condition for $\bar{x}$. The opposite inequality holds when $x = z$. Using the continuity of b_p and β_p and the intermediate value theorem, we see that x° is well-defined, and $\bar{x} < x^\circ < z$.

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

$$u(b_q(\underline{y})) + \alpha[\underline{y} - z] \leqslant u(\beta_q(y, \underline{y})). \tag{46}$$

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

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

(48) $0 < y_* < \underline{y},$ and $u(b_q(\underline{y})) + \alpha[\underline{y} - z] < u(\beta_p(y, \underline{y}))$ for all $y_* < y \leqslant \underline{y},$

To prove this claim, note that $x^* \leqslant x^\circ < z$. Moreover, both x° and $\iota_z(y_\circ)$ strictly exceed $\bar{x}$, the latter because $y_\circ < \underline{y}$ and $\bar{x} = \iota_z(\underline{y})$. So $x^* = \min\{x^\circ, \iota_z(y_\circ)\} > \bar{x}$. Additionally, given the definition of x°, and because "$<$" holds at $x = \bar{x}$, the second inequality in (47) must hold.

Turning now to (48), 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 (46) is strictly increasing in y, while the left hand side is constant in y. So if "$\leqslant$" holds in (48) at $y = y^*$, it must do so strictly for $y_* < y \leqslant \underline{y}$. 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,

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

and

$M^q(z, p, q) \equiv \{(x, y) | f(x, y) = z,$ with $x_* \leqslant x < \underline{x}$ 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 (47) and (48), 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. ■