Self-Censorship of Hard Information

Roi Orzach*

November 25, 2024

Abstract

Despite their potential to enhance decision-making, certain facts remain unspoken. In the model, an agent receives verifiable information about an unknown state and decides whether to share such information before the principal makes her decision. The agent is either biased and has different preferences than the principal or unbiased and has the same preferences. Further, all agents aim to build a reputation for being unbiased. When the agent cannot commit to a disclosure policy, self-censorship occurs and the principal may not obtain her first-best payoff for any reputational weight by the agent. In contrast, when the agent commits to a disclosure policy, the principal's payoff increases with the reputational weight, and equals her first-best payoff when reputation is sufficiently valuable. Finally, I connect these results to academic publishing practices.

1 Introduction

In various contexts, certain facts remain unspoken or carry significant reputational consequences when spoken. For example, when deciding whether to publish data about a lack of racial disparities in police use of high-level force, economist Roland Fryer alleges that Harvard faculty advised, "Don't publish this. You'll ruin your career" (Weiss, 2024). This self-censorship is pervasive in academic publishing. In recent surveys, nearly two-thirds of psychology professors reported that "some empirically supported conclusions cannot be mentioned without punishment" and "91 percent reported being at least somewhat likely to self-censor in publications" (Clark

*Department of Economics, MIT, 50 Memorial Drive, Cambridge MA 02142 (e-mail: orzach@mit.edu). I thank Charles Angelucci, Ian Ball, Alessandro Bonatti, Glenn Ellison, Roland Fryer, Ying Gao, Robert Gibbons, Daniel Luo, Jackson Mejia, and especially Stephen Morris for helpful comments and discussions.

et al., 2023). Even though this information is verifiable (i.e., it is empirically supported rather than mere opinion), sharing it often raises the question: "What kind of individual would share this?"

The model developed in this paper aims to determine whether people will self-censor hard information when advising a principal. In the model, an agent decides whether to share hard evidence conditional on its realization. While sharing such information always results in a more efficient decision for an unbiased agent, doing so may tarnish the agent's reputation.

Let's cast this question in the political-correctness paradigm introduced in Morris (2001). Here, a social scientist (he, agent) decides whether to share information about affirmative action with a university dean (she, principal). The social scientist may be biased, preferring a different ideal affirmative action policy than the principal, or unbiased, sharing the principal's goal of maximizing student welfare. Further, both types of social scientists have a preference for being perceived as unbiased. Each social scientist conducts an experiment or collects data, resulting in information about the relative benefits of affirmative action. Importantly, I assume that this information may be generated asymmetrically between the two types of agents. Finally, in contrast to Morris (2001), I assume that this information is verifiable (cf. Milgrom, 1981; Grossman, 1981), implying the social scientist cannot fabricate information.

I consider two environments: absent commitment and with commitment. Absent commitment, the agent first observes the data and then decides if to disclose the information. With commitment, the agent pre-registers his study with the dean and chooses a disclosure policy before observing, or even obtaining, the data. The ability to commit to a disclosure policy is common in various areas of academia and medicine. For instance, one-third of economists conducting randomized control trials submit Pre-Analysis Plans (PAPs) detailing which outcomes will appear in the publication (Banerjee et al., 2020).[1] Further, various journals, such as the Journal of Political Economy Microeconomics, allow editors to commit to publication outcomes based on PAPs alone.

To understand this tradeoff, I first analyze the case absent commitment. The agent, regardless of his type, must self-censor his information if two conditions are met:

[1]The primary motivation behind PAPs is to prevent p-hacking, as opposed to combating self-censorship. In contrast, I focus squarely on information disclosure and treat information generation as exogenous.

(i) he values reputation, and (ii) the biased agent, absent reputational considerations, is more likely to disclose such information. If (ii) holds, the agent receives a better reputational payoff following self-censorship. Further, given that (i) holds, he prefers to self-censor and preserve his reputation. Finally, I show that there are environments where in any equilibrium and for any reputational weight, the principal never obtains the payoff she would obtain had the information been available to her directly, defined as her first-best payoff (Proposition 1).

In contrast, if the agent commits to a disclosure rule, then the principal obtains her first-best payoff whenever reputation is sufficiently valuable to the agent (Proposition 2). In this equilibrium, both the biased and unbiased agent are conjectured to commit to fully revealing their information. Further, the principal assumes that any off-path disclosure policy is chosen by the biased agent. When the agent values his reputation, these off-path beliefs ensure that the agent chooses the fully-revealing policy.

The unifying intuition is that whenever reputation is sufficiently valuable, the two types of agent must pool (or choose sufficiently similar strategies) when evaluated with respect to the principal's information. In the case of soft information, this pooling implies that the agents must make the same action recommendation independent of the state (c.f. Morris, 2001). For hard information absent commitment, because whether the agent acquired information is not observed by the principal, it may be necessary for both types of agents to pool on not providing information to the principal. However, for hard information with commitment, both agents can pool on full revelation.

1.1 Literature

These results broadly speak to three literatures: political correctness, disclosure, and scientific-publishing policies. The political-correctness literature typically focuses on soft information, such as Morris (2001). Qualitatively, political-correctness concerns are exacerbated in models of soft information, because, in equilibrium, an agent's message is necessarily a function of both his type and his information about the state of the world. One might think that hard information should eliminate these concerns since revealing that the state of the world takes a certain value need not conflate preferences and information about the state of the world. However, in line with empirical evidence, my results show that political-correctness concerns remain

with hard information. My results comparing commitment and non-commitment bare qualitative resemblance to the literature designing decision-making environments for the agent when he has political-correctness concerns. For instance, Prat (2005) shows that transparency exacerbates political-correctness concerns, ultimately harming the principal's payoff. Further, McClellan and Rappoport (2024) shows that the agent benefits from "taking stands," which help disambiguate preferences and information.

The literature on verifiable information disclosure typically assumes that the agent always prefers a higher decision by the principal (cf. Milgrom, 1981; Grossman, 1981; Dye, 1985). Further, in the disclosure literature studying reputation (cf. Bar-Isaac, 2003; Board and Meyer-ter Vehn, 2013; Shadmehr and Bernhardt, 2015; Zhang, 2024), an agent's (respectively, firm's) reputation is useful insofar as it encourages the principal (respectively, consumer) to take a higher decision (respectively, purchase more). In contrast, I focus on an environment, similar to Morris (2001), where the unbiased agent does not have directional preferences and is instead perfectly aligned with the principal over which decision should be taken.

Finally, the literature on scientific publishing policies has also discussed the importance of commitment. Here, the primary tradeoff considered is that such precommitment may discourage p-hacking, but incurs a large up-front burden for the researchers to think through all possible contingencies (Banerjee et al., 2020). Similarly, Coffman and Niederle (2015) argue that these costs associated with PAPs are unwarranted if replications can find flaws in the original studies. My results are complementary to this intuition, because I show that PAPs have a strong benefit when the original studies would have been self-censored (and thus unable to be replicated) absent the PAPs.

2 Model

There exist two players: a principal (she) and an agent (he). The principal has a continuous utility function $u^p(a, \omega)$ that depends on her action $a \in A$, and the state of the world $\omega \in \Omega$, where A, Ω are compact subsets of the real line. The agent and principal share a common prior belief that ω has cumulative distribution function F_ω. There are two types of agent: biased and unbiased, where an agent is unbiased with probability $p \in (0, 1)$. An agent's payoff is the sum of a non-image payoff and a reputational payoff. An unbiased agent's non-image payoff is identical to that of

the principal. Meanwhile, a biased agent's non-image payoff is a continuous function $v^b(a, \omega)$, which need not equal $u^p(a, \omega)$. Further, each type of agent has a preference for being viewed as unbiased by the principal. I delay a complete description of the utilities until after describing the disclosure policies.

Information: I allow the biased and unbiased agents to, potentially, utilize different data-generating processes. Here, I assume that the unbiased agent observes the realization of ω with probability $q^u \in (0, 1)$ and observes n with probability $1 - q^u$, where n denotes that the agent observed no information.[2] Similarly, the biased agent observes ω with probability $q^b \in (0, 1)$ and observes n with probability $1-q^b$. Finally, let $\Omega^c := \Omega \cup n$.

Disclosure: Let a disclosure policy be a map $\pi : \Omega^c \to \Omega^c$, where (i) for all ω, $\pi(\omega) \in \{\omega, n\}$ and (ii) $\pi(n) = n$. These restrictions imply that fabrication by the agent is impossible (i.e., the information is "hard").

Commitment: In the analysis, I vary the commitment power of the agent. I say the agent has commitment, if his choice of π is observable. In contrast, I say the agent lacks commitment if his choice of π is private. This definition is equivalent to saying the agent lacks commitment if his choice of π is made after observing ω.

Reputation and Preferences: Denote the principal's posterior beliefs that the agent is unbiased by $\phi : \Omega^c \to [0, 1]$. The utilities of the biased and unbiased agents are:

$$u^u(\pi) = \mathbf{E}\Big(u^p\big(a(\pi(\omega)\big) + \lambda \cdot \phi\big(\pi(\omega)\big)\Big) \tag{1}$$

$$u^b(\pi) = \mathbf{E}\Big(v^b\big(a(\pi(\omega)\big) + \lambda \cdot \phi\big(\pi(\omega)\big)\Big), \tag{2}$$

with $\lambda \geq 0$. I will refer to λ as the strength of reputational incentives. These preferences state that both types of agent have an image payoff based on their reputation and a non-image payoff based on the decision chosen. Importantly, with commitment, $\phi(\cdot)$ will depend on the observed disclosure policy, while absent commitment it will not. Throughout, I place no restrictions on off-path beliefs for $\phi(\cdot)$.

Assumptions: Throughout, I impose one assumption on the preferences.

Assumption 1. *If $\lambda = 0$, the agent lacks commitment, and the unbiased agent fully reveals, then there exists a positive measure of realizations of ω that the biased agent*

[2] I discuss this assumption in depth after detailing the model in the discussion portion. Further, I note that given the generality of the payoffs, this model is equivalent to one with a latent state θ, for which ω is a signal, and the agent observes ω.

strictly would prefer to conceal.

This assumption states that absent reputational considerations, the biased agent chooses a different disclosure policy than what is optimal for the principal. Without this assumption, there is no conflict between the biased agent and the principal, resulting in a trivial analysis. I note that this assumption is met in many settings (c.f. Dye, 1985; Morris, 2001; Kamenica and Gentzkow, 2011).

I analyze Perfect Bayesian Equilibria (PBE) of this game, henceforth *equilibria*.

Discussion: This framework asks how political-correctness concerns affect the disclosure of hard information with and without commitment. I make one non-standard assumption: the biased and unbiased agents have exogenous and potentially different data-generating processes. The exogeneity assumption is standard in models of hard information (c.f. Dye, 1985). Further, this paper focuses on censorship for given information realizations, as opposed to endogenous data generation, which would require modeling data fabrication, p-hacking, and influence activities.

One micro-foundation for the differential data-generating processes is to endogenize information acquisition similarly to Aghion and Tirole (1997). Here, each agent pays a cost $c(e)$ which with probability e allows that agent to observe hard information about ω.[3] Further, differential data-generating processes is observed in qualitative discussions of self-censorship. For instance, Loury (1994) notes, "people genuinely committed to justice did not become entangled in arcane technical arguments about the effects of economic boycotts (of South Africa)." Returning to the motivating example of Professor Roland Fryer, one might think that biased social scientists have a greater interest to show that there are no racial disparities in policing, resulting in differential information acquisition.[4]

Finally, I assume that the agent can commit to a disclosure policy. In reality, a researcher is not forced to stick to their PAPs, and a third party who received access to the data (such as the research sponsor, journal, or professional organization) may be needed to invoke such commitment. A different interpretation of the model is that the agent is an academic journal. The journal wants to influence a decision by the

[3]Additional micro-foundations could result from noting that different types of individuals may consume different news sources (c.f. Gentzkow and Shapiro, 2010) and retain different information (c.f. Angelucci and Prat, 2024), potentially resulting in differential information discovery.

[4]Roland's critics claim that the regressions he chose to run and publish may not have been truly capturing the marginal treatment effect. Again, capturing these incentives necessitates a model of p-hacking which is beyond the scope of this paper.

public or government (e.g., the principal) and further wants to maintain a reputation for being unbiased. Here, the realization of the state ω is determined by the research submitted to a journal, which would naturally differ for biased and unbiased journals. In this interpretation of the model, commitment can be viewed as the recent policy enacted by the Journal of Political Economy Microeconomics, where "authors can submit their prospective empirical projects and have approved for publication before the data is collected and the results are known."

3 Results

This section analyzes how political-correctness concerns impact disclosure policies. In subsection 1, I conduct the analysis of the disclosure game without commitment power. In subsection 2, I consider the analysis when the agent can commit to a disclosure policy.

3.1 Disclosure:

This subsection considers the environment when the agent lacks commitment power. I first consider two examples which showcase how political-correctness concerns can stifle informative communication. I next characterize when the principal obtains her first-best payoff in equilibrium.

Example 1: Let $\omega \in \{0, 1\}$, where $\mathbf{E}(\omega) = \mu < 1/2$. The biased agent's non-image payoff is $v^b(a, \omega) = 2a$ and the principal's utility is $\mathbf{1}_{a=\omega}$, where $a \in \{0, 1\}$. Finally, recall, p denotes the probability an agent is unbiased.

Lemma 1. *In the environment described above, for any (μ, p), there exists a positive Lebesgue-measure set of (q^u, q^b) and a cutoff λ^* such that: if $\lambda > \lambda^*$, then $a = 0$ with probability one in any equilibrium.*

The implication of this lemma is that the principal receives no information which alters her decision away from her prior. This result may be surprising because when the agent receives information that $\omega = 1$, the biased agent, unbiased agent, and principal all prefer $a = 1$, yet this information is self-censored resulting in $a = 0$.

Let us now build intuition behind the result. When $\omega = 1$, the biased agent's has a strictly larger non-image incentive to share $\omega = 1$ than the unbiased agent. If,

additionally, $q^u < q^b$, then the reputation from revelation must be strictly worse than the prior. As a result, if sharing $\omega = 1$ was on-path it must incur a strict reputational loss. However, for high enough reputational weights the players will self-censor to avoid this reputational loss. $\square$

I now consider a continuous variant of the setting considered in Example 1 and showcase that the principal never obtains her first-best payoff, but there always exists an equilibrium with partial communication.

Example 2: Suppose ω admits a probability density function, f_ω. The biased agent's non-image payoff is $v^b(a, \omega) = 2a$, whereas the principal's utility is $u^p(a, \omega) = -(a - \omega)^2$. Finally, p remains the probability that the agent is unbiased.

Lemma 2. *Fix f_ω and recall $a(n)$ corresponds to the decision the principal takes absent any information. There exists a positive Lebesgue-measure set of (q^u, q^b, p), and a cutoff λ^* such that: if $\lambda > \lambda^*$, then in any equilibrium $\omega > a(n)$ is self-censored. Further, for any (q^u, q^b, p, λ), there exists an equilibrium where the unbiased agent fully reveals ω if $\omega \leq a(n)$.*

This lemma states that there exist information structures such that any information which would result in a higher decision taken than $a(n)$ (i.e., the decision taken following censorship) is self-censored. The intuition is identical to Lemma 1: the biased agent has a strictly greater incentive to share such information, implying a strict reputational gain from self-censorship, resulting in self-censorship for high reputational weights.

However, this proposition also states that in the continuous environment, self-censorship does not occur for all realizations. Here, for any $a(n)$, there exist realizations $\omega < a(n)$ which would motivate the principal to take a lower decision. For these states, the unbiased agent has a strictly greater incentive to share ω. As a result, the unbiased agents will truthfully reveal such information and the biased agent will either (i) not value reputation highly enough and conceal the information to get a higher decision or (ii) mix between revelation and concealment. $\square$

I now present a characterization of when the principal obtains her first-best payoff.

Proposition 1. *The following are true:*

1. *If $q^u \leq q^b$, then for any $\lambda \geq 0$ the principal does not obtain her first-best payoff*

in any equilibrium. However, if $q^u = q^b$, then for any $\lambda \geq 0$ there always exists an equilibrium where the unbiased agent fully reveals his information.

2. *If $q^u > q^b$, then there exists a threshold $\lambda^*(q^u, q^b)$, such that the principal obtains her first-best payoff if and only if $\lambda > \lambda^*(q^u, q^b)$, where λ corresponds to the relative weight of the political-correctness concerns.*

To gain intuition behind this proposition first note that if λ is not sufficiently high, then by Assumption 1, the biased agent would prefer to selectively disclose. Hence, assume that the reputational weight is high. If $q^u > q^b$, then in the conjectured equilibrium where all parties truthfully disclose, censorship has a strict reputational cost. As a result, both types will indeed fully disclose.

However, if $q^u < q^b$, then in the conjectured equilibrium where all types fully disclose, revelation involves a strict reputational loss. As a result, for high enough reputational weights, the principal cannot obtain her first-best payoff. Further, given Example 1, the principal need not obtain any information from the agent.

Finally, if $q^u = q^b$, then the principal can always conjecture an equilibrium where the unbiased agent truthfully reveals his information. As a result, revelation improves the unbiased agent's reputation and results in a more efficient decision. However, the biased agent will then have an incentive to self-censor: if he did not self-censor, then by Assumption 1 censorship has a strict non-image payoff improvement and no reputational loss for the biased agent.

These results show that the presence of hard information alone does not mitigate political correctness incentives. Further, these concerns can be sufficiently large to result in no information transmission in equilibrium (Example 1). Finally, whether or not information is transmitted depends on which type of agent is more likely to be informed (Proposition 1). With soft information (c.f. Morris, 2001), the analysis is qualitatively unchanged if the agents are informed with differential probabilities because each party can falsely claim to be informed.

3.2 Commitment:

In this subsection, I assume that the agent commits to a disclosure policy.[5] With commitment, the principal's beliefs about an agent's type are a function of both the

[5]The proofs of the results are qualitatively identical if instead the agent could commit to any experiment, instead of simply a disclosure policy, such as in Kamenica and Gentzkow (2011).

disclosure policy and the realization of the information shared with the principal. As a result, there always exists an equilibrium where the principal conjectures that the unbiased agent fully discloses his information and that any other disclosure policy only belongs to the biased agent. This intuition is formalized in the following lemma.

Lemma 3. *For any preferences and information distributions, there always exists an equilibrium where the unbiased agent commits to fully revealing his information.*

This lemma is in contrast to (i) the results in the previous section absent commitment and (ii) the analysis with soft information, such as Morris (2001), where if reputation is sufficiently valuable communication is uninformative. Further, this lemma places a lower bound on the principal's payoff in her preferred equilibrium. The subsequent proposition further clarifies the equilibrium behavior in the principal's preferred equilibrium.

Proposition 2. *The principal's payoff in her preferred equilibrium is weakly increasing in λ. Further, there exist two thresholds $0 < \underline{\lambda}^c < \bar{\lambda}^c$ such that:*

1. *For $\lambda \leq \underline{\lambda}^c$ the unbiased agent fully discloses his information and the biased agent chooses his privately optimal disclosure policy with commitment.*

2. *For $\lambda \in \left(\underline{\lambda}^c, \bar{\lambda}^c\right)$, the principal's payoff is strictly greater than the payoff she receives when $\lambda \leq \underline{\lambda}^c$, but is strictly less than her payoff when $\lambda \geq \bar{\lambda}^c$.*

3. *For $\lambda \geq \bar{\lambda}^c$, the principal obtains her first-best payoff.*

The intuition for this proposition is that the principal's first-best payoff is obtained if and only if the biased agent would prefer to fully disclose his information and maintain a positive reputation as opposed to deviating and selectively disclosing, resulting in the worst possible reputation. The inequality corresponding to this deviation results in the cutoff $\bar{\lambda}^c$. When λ is slightly below $\bar{\lambda}^c$, the following is an equilibrium: the unbiased agent continues to fully disclose with probability one, whereas the biased agent fully discloses with probability $p \in (0, 1)$. A lower value of p increases the reputation following full disclosure, therefore for λ close enough to $\bar{\lambda}^c$, there exists a unique p that makes the biased agent indifferent between full disclosure and the optimal disclosure-policy absent reputation.[6] Finally, when the reputational

[6]This equilibrium is a lower bound for the principal's payoffs. There may there may exist a fully-pooling equilibrium such that (i) neither agent would prefer to deviate from this policy and (ii) the principal prefers this policy to the one outlined in the text.

incentives get sufficiently small, each type of agent must choose his privately-optimal policy. For the unbiased agent, this is full disclosure and Assumption 1 ensures that the biased agent chooses a partial-disclosure policy.

This subsection proves that when there is commitment in the disclosure policy, political-correctness concerns by the agents helps the principal. Further, if political-correctness concerns are high enough all information is fully revealed.

Discussion: The juxtaposition of the results in subsection 1 and subsection 2 suggests that pre-commitment is especially valuable in situations where (i) political-correctness concerns are salient and (ii) the biased and unbiased agents have differing data generating processes. If (i) fails, then the unbiased agent will truthfully reveal, and the biased agent may use commitment power to selectively reveal and persuade the principal. In contrast, if (ii) fails and the players have similar data generating processes (or the unbiased agent is more likely to discover information) then the players can pool on revealing information. However, if the biased agent is more likely to discover information, then as the agents must pool, they must pool on the action ex-ante more likely to be chosen by the unbiased agent, which is concealment.

4 Conclusion

This paper presents a model in which a possibly biased agent decides whether to disclose information about an unknown state of the world, aiming to maintain a reputation for being unbiased and to influence a decision. I show that absent the ability to commit to a revelation policy, the principal's payoffs need not equal her first-best payoff. In contrast, when the agents are required to commit to a revelation policy, the principal's payoff is increasing in the agents' reputational concerns and achieves the first-best for a finite value of the reputational concerns. These two results suggest that the principal should require the agents to commit to their revelation policy whenever the reputational considerations are high. In contrast, when the reputational concerns are low, the biased agent will more selectively conceal with commitment, whereas the unbiased agent's incentives are unchanged.

This analysis suggests revisiting seminal papers with reputational preferences is valuable. For instance, in this paper, informational discovery was fixed and exogenous. However, these reputational preferences may discourage information discovery, which generates new insights on the optimality of delegation for information discovery (cf.

Aghion and Tirole, 1997). Similarly, in this model the principal could not delegate the decision right to the agent. One could instead consider how these reputational preferences will lead to different optimal delegation sets (cf. Szalay, 2005; Alonso and Matouschek, 2008). Finally, one can consider the effects of reputation on other forms of communication such as costly communication (cf. Dewatripont and Tirole, 2005) or hierarchical communication (cf. Garicano, 2000).

References

Aghion, Philippe and Jean Tirole, "Formal and real authority in organizations," *Journal of political economy*, 1997, *105* (1), 1–29.

Alonso, Ricardo and Niko Matouschek, "Optimal delegation," *The Review of Economic Studies*, 2008, *75* (1), 259–293.

Angelucci, Charles and Andrea Prat, "Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News," *American Economic Review*, 2024, *114* (4), 887–925.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann, "In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics," Technical Report, National Bureau of Economic Research 2020.

Bar-Isaac, Heski, "Reputation and Survival: learning in a dynamic signalling model," *The Review of Economic Studies*, 2003, *70* (2), 231–251.

Board, Simon and Moritz Meyer ter Vehn, "Reputation for quality," *Econometrica*, 2013, *81* (6), 2381–2462.

Clark, Cory J, Lee Jussim, Komi Frey, Sean T Stevens, Musa Al-Gharbi, Karl Aquino, J Michael Bailey, Nicole Barbaro, Roy F Baumeister, April Bleske-Rechek et al., "Prosocial motives underlie scientific censorship by scientists: A perspective and research agenda," *Proceedings of the National Academy of Sciences*, 2023, *120* (48), e2301642120.

Coffman, Lucas C and Muriel Niederle, "Pre-analysis plans have limited upside, especially where replications are feasible," *Journal of Economic Perspectives*, 2015, *29* (3), 81–98.

Dewatripont, Mathias and Jean Tirole, "Modes of communication," *Journal of political economy*, 2005, *113* (6), 1217–1238.

Dye, Ronald A, "Disclosure of nonproprietary information," *Journal of accounting research*, 1985, pp. 123–145.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of political economy*, 2000, *108* (5), 874–904.

Gentzkow, Matthew and Jesse M Shapiro, "What drives media slant? Evidence from US daily newspapers," *Econometrica*, 2010, *78* (1), 35–71.

Grossman, Sanford J, "The informational role of warranties and private disclosure about product quality," *The Journal of law and Economics*, 1981, *24* (3), 461–483.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Loury, Glenn C, "Self-censorship in public discourse: A theory of "political correctness" and related phenomena," *Rationality and Society*, 1994, *6* (4), 428–461.

McClellan, Andrew and Daniel Rappoport, "Signaling Good Faith by Taking Stands," *Available at SSRN 4510675*, 2024.

Milgrom, Paul R, "Good news and bad news: Representation theorems and applications," *The Bell Journal of Economics*, 1981, pp. 380–391.

Morris, Stephen, "Political correctness," *Journal of political Economy*, 2001, *109* (2), 231–265.

Prat, Andrea, "The wrong kind of transparency," *American Economic Review*, 2005, *95* (3), 862–877.

Shadmehr, Mehdi and Dan Bernhardt, "State censorship," *American Economic Journal: Microeconomics*, 2015, *7* (2), 280–307.

Szalay, Dezsö, "The economics of clear advice and extreme options," *The Review of Economic Studies*, 2005, *72* (4), 1173–1198.

Weiss, Bari, "Weekend Listening: From McDonald's Drive-through to Star Harvard Professor," *The Free Press*, February 2024. Accessed: 2024-08-04.

Zhang, Wenhao, "Strategic disclosure with reputational concerns," *Journal of Mathematical Economics*, 2024, *111*, 102945.

Appendix

Proof of Lemma 1. Note that if q^u and q^b are sufficiently low, then given the prior, $a(n) = 0$ in any equilibrium. Hence, it suffices to show that $\phi(1) < \phi(n) - \epsilon$ for a strictly positive ϵ.

If $q^u < q^b$, then in any equilibrium,

$$\phi(1) \leq \frac{pq^u}{pq^u + (1-p)q^b} < p - \epsilon, \tag{3}$$

for $\epsilon > 0$ because, as argued in the text, the biased agent must disclose $\omega = 1$ with probability one if $\omega = 1$ is revealed on path.

Therefore it suffices to show that $\phi(n) \geq p$. Suppose by contradiction $\phi(n) \leq p$. By Bayes rule, then $\omega = 0$ is disclosed on-path and $\phi(0) > p$. Therefore, $\phi(0) > \phi(n)$ and $a(0) = a(n)$, implying $\omega = 0$ is disclosed with probability one by both the biased and unbiased agent. As $q^u < q^b$ then $\phi(0) < p$, deriving a contradiction. □

Proof of Lemma 2. Fix (q^u, q^b, p) where $q^u < q^b$; I first show that $\omega > a(n)$ are self-censored for sufficiently high λ. As in Lemma 1, it suffices to show (i) $\phi(\omega) < p - \epsilon$ for $\epsilon > 0$ and (ii) $\phi(n) \geq p$. (i) holds as $q^u < q^b$ and the biased agent has a strictly larger incentive to share $\omega > a(n)$.

Now it suffices to show (ii) holds for a sufficiently large λ. To see why this occurs, note that for any $\omega < a(n)$, if ω is revealed on path, then $\phi(\omega) - \phi(n) \leq \frac{c}{\lambda}$, where c is an exogenous constant determined by the maximum loss between two decisions for any two states. Therefore, by Bayes' rule $\phi(n) > p$ for λ sufficiently high. As a result, there must exist a exists a λ^* above which these states must be self-censored.

I now show that there always exists an equilibrium where the unbiased agent fully reveals ω if $\omega < a(n)$. I claim that given two values $a(n), \phi(n)$ there exists a best-reply constructed as follows: (1) if $\omega \leq a(n)$, then the unbiased agent fully reveals and the biased agent reveals with probability $p(\omega) \in [0, 1]$. $p(\omega)$ is unique as the biased agent has a strictly lower incentive to reveal as more biased agents choose to reveal. (2) If $\omega > a(n)$, then either (a) both types fully reveal, (b) only the biased agent reveals, or (c) neither type reveals. If (a) is such that neither type prefers to deviate, then (a) is chosen. If (a) fails, and (b) holds, then (b) is chosen. Finally, if both (a) and (b) fail, then (c) must hold and is thus chosen. These best-replies determine an equilibrium if and only if the resulting $a(n), \phi(n)$ are consistent conjectures in equilibrium. However, this best-reply results in a fixed-point problem mapping conjectures of $a(n), \phi(n)$ into resulting actions and beliefs given these conjectures. As this mapping is smooth (given the continuous state space), then there exists a fixed-point and hence an equilibrium with the desired properties. □

Proof of Proposition 1. I begin by proving that if $q^u \leq q^b$ the principal never obtains her first-best payoff. Let us proceed by contradiction. That $q^u \leq q^b$ implies that for all ω, $\phi(\omega) \geq \phi(n)$. Further, Assumption 1 implies that when $\lambda = 0$, the biased agent has a strict incentive to conceal. Further, the weak reputational benefit from concealment implies the biased agent has a profitable deviation, resulting in a contradiction.

Next, if $q^u = q^b$ I show the existence of an equilibrium where the unbiased type fully reveals. If this occurs, then for all ω, $\phi(\omega) \geq \phi(n)$, implying the unbiased agent indeed reveals. Further, using an identical fixed point argument to Lemma 2, there exists a strategy for the biased agent that is consistent with an equilibrium. Finally, this fixed point argument need not rely on a continuous state space as the strategy of the biased agent is continuous with respect to the conjectures given that the unbiased agent always reveals.

I now prove statement 2. Note the principal obtains her first best payoff if and only if neither agent has a benefit from deviating from fully revealing their information. As $q^u > q^b$, $\phi(n) < \phi(\omega)$ for all ω, implying that the unbiased agent never has an incentive to conceal. Let us now consider the incentives of the biased agent. In this conjectured equilibrium, $a(n), a(\omega), \phi(n)$ are independent of λ and $\phi(\omega)$ is independent of both λ

and ω. Therefore, this conjectured equilibrium is an equilibrium if and only if

$$\max_{\omega}\{v^b(a(n),\omega) - v^b(a(\omega),\omega)\} \leq \lambda\big(\phi(\omega) - \phi(n)\big), \tag{4}$$

resulting in the cutoff described in the proposition. □

Proof of Lemma 3. Suppose the unbiased agent chooses to fully reveal and conjecture that any other disclosure policy leads to the principal assigning probability one to the biased agent. Given such beliefs, the unbiased agent, irrespective of the strategy of the biased agent has a strict incentive to disclose.

Given such beliefs, the biased agent's strategy must place positive probability on only (i) fully revealing or (ii) choosing his preferred revelation policy absent reputational forces. Denote the conjectured probability the biased agent fully reveals by $\tilde{p}^b$. An equilibrium exists if and only if the best response to a given conjecture is equal to the conjecture itself. However, the best response is decreasing (due to the reputation considerations) and continuous in $\tilde{p}^b$ implying a fixed point exists, proving the existence of an equilibrium. □

Proof of Proposition 2. The principal obtains her first-best payoff if and only if both types of agent commit to full disclosure (or an outcome-equivalent disclosure rule). Further, this equilibrium is supported by off-path beliefs that any other disclosure rule is chosen by the biased agent. In this conjectured equilibrium, the unbiased agent has no gain from deviating. Further, the biased agent's optimal deviation is to his optimal revelation policy absent reputation. Therefore, the non-image gain from deviation is independent of λ, and the reputational loss is monotone and continuous with respect to λ, resulting in the cutoff in the proposition $\bar{\lambda}^c$.

Next, I prove monotonicity of the principal's payoff in her preferred equilibrium with respect to λ. As there are only two types of agent, it suffices to consider equilibria where at-most two disclosure policies are chosen on path. If for a fixed value λ, only one disclosure policy is chosen on path, then this equilibrium will remain an equilibrium for higher values of λ. Thus, let us consider when two different policies that result in different payoffs are chosen on path for a fixed λ. There are two cases of interest:

Case 1: The unbiased agent mixes. Let policy 1 be the one that the principal prefers. There are three subcases. (i) The biased agent chooses policy 1 with probability 1. Then one can construct an alternative equilibrium where policy 1 is chosen

by both agents with probability 1. That neither agent deviated when policy 1 had a worse reputation, implies neither party will deviate under the alternative equilibrium. (ii) The biased agent chooses policy 2 with probability 1. This cannot occur as the unbiased agent gets a better non-image payoff and reputation under policy 1. (iii) Both agents mix. As policy 1 confers a greater non-image utility to the unbiased agent, policy 2 must confer a better reputation. For the biased agent to mix, the biased agent must also prefer policy 1. As a result, both players choosing policy 1 with probability 1 is an equilibrium if and only if neither party prefers to deviate to a policy which was off path in the original equilibrium. However, now the reputation following policy 1 is larger, implying that such deviations are deterred in the new equilibrium.

Case 2: The unbiased agent selects policy 1 with probability 1. Again there are two subcases, as we assumed that two policies are chosen on path. (a) If the biased agent chooses policy 2 with probability 1, then policy 1 must be full revelation and policy 2 the biased agent's preferred policy absent reputation. Now following an increase in λ, the biased agent will either continue selecting policy 2 or mix between the two, proving the result. (b) The biased agent mixes between policy 2 and policy 1. Therefore, policy 2 must be the policy the biased agent prefers absent reputational considerations. Note that policy 1 must be strictly preferred by the principal than policy 1, else by Lemma 1, the principal could have the unbiased agent fully disclose. However, now following an increase in λ, there are two subcases. (i) If policy 1 is the principal's preferred policy (e.g., full disclosure) then the biased agent following an increase in λ chooses this policy more and the unbiased agent has no deviation temptation. (ii) If this policy is not the principal's preferred policy, then upon increasing λ one can maintain the probability that the biased agent chooses policy 2, by changing policy 1 to be the appropriate mix between the old policy 1 and fully disclosure. To check if this is an equilibrium, one must check the deviation temptation for the unbiased agent. However, that the unbiased agent did not deviate for the lower value of λ, implies he will not in this new candidate equilibrium because the reputational cost from doing so is strictly higher and the new on-path policy provides a strictly greater non-image payoff.

Finally, I must prove the existence of $\underline{\lambda}^c$. As the principal's payoff is monotone in λ, it suffices to show that there exists a threshold for which no pooling or partially pooling equilibrium exists. Suppose that both the biased and unbiased agent choose

the same disclosure policy. For any such policy, one agent must have a different disclosure policy which provides a strict improvement in their non-image payoff (by Assumption 1) by a fixed constant ϵ. Thus, for $\lambda < \epsilon$, even for the maximal possible difference in reputation, one of the types of agent's would prefer to deviate to their preferred disclosure policy absent reputation.

□