Self-Censorship of Hard Information

Roi Orzach*

May 14, 2025

Abstract

Despite their potential to enhance decision-making, certain facts remain unspoken. In the model, an agent decides whether to share his verifiable information to influence the principal's decision. The agent is either biased, with different preferences from the principal, or unbiased, sharing the same preferences. Further, agents value reputation: being perceived as unbiased. When the agents cannot commit to disclosure policies, self-censorship occurs and the principal may not obtain her first-best payoff for high reputational weights by the agent. However, with commitment, the principal obtains her first-best payoff for sufficiently high reputational weights. Finally, I connect these results to academic publishing.

Keywords: Disclosure, Reputation, Self-Censorship
JEL Classification Numbers: D72, D82, D83

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

*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, Bart Lipman, Glenn Loury, Daniel Luo, Jackson Mejia, Andrea Prat, Reca Sarfati, Jean Tirole, Vod Vilfort, and especially Stephen Morris for helpful comments and discussions.

percent reported being at least somewhat likely to self-censor in publications" (Clark 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 seeks to examine the forces that influence an agent's decision to self-censor hard information when advising a principal in a setting where the agent has reputational considerations. While sharing information always results in a more efficient decision for an unbiased agent, doing so may tarnish the agent's reputation.

I will frame this question in the political-correctness paradigm introduced in Morris (2001). Here, a social scientist (he, agent) decides whether to disclose information about affirmative action with a university dean (she, principal). The social scientist may be biased, preferring a different affirmative action policy than the dean, or unbiased, sharing the dean's goals. Further, both types of social scientists have a preference for being perceived as unbiased. Each type of social scientist conducts an experiment or collects data, resulting in information about the relative benefits of affirmative action. Importantly, I assume that this information need not be generated with equal probabilities by the different types of social scientists. 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 whether 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. 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 and the Journal of Development Economics, allow editors to commit to publication outcomes based on PAPs alone.

To understand this tradeoff, I first analyze the case absent commitment. The

[1]The primary motivation behind PAPs is preventing p-hacking. In contrast, I treat p-hacking and other forms of information generation as exogenous and focus solely on communication, as is standard in the reputation literature.

agent, regardless of his type, must self-censor his information if two conditions are met: (i) he sufficiently values his 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. Therefore, given that (i) holds, he prefers to self-censor and preserve his reputation. Finally, I show that the principal receives her first-best payoff (defined as her payoff had the information been directly available to her) if and only if the agents sufficiently value reputation *and* the biased agent is strictly more likely to generate information (Proposition 1).

In contrast, if the agent commits to a disclosure rule, then the principal obtains her first-best payoff whenever the agents sufficiently value reputation (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 agents value reputation, these off-path beliefs ensure that the agents choose 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 generated information is not observed by the principal, it may be necessary for both types of agents to pool on not providing information. 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 analyzes soft information, such as Morris (2001). Qualitatively, political-correctness concerns are exacerbated in models of soft information, because an agent's equilibrium message is necessarily a function of both his type and his information about the state of the world. Naïvely, hard information should eliminate these concerns since hard information about the state is not informative about an agent's type. However, in line with empirical evidence, my results show that political-correctness concerns remain with hard information.

Within this literature, my results bare qualitative resemblance to those studying design tools to mitigate political-correctness concerns for a decision-making agent. For instance, Prat (2005) and McClellan and Rappoport (2024) show that decreasing transparency and letting the agent send a pre-play message, respectively, help restore efficiency. Most related is Bénabou et al. (2020) who empirically tests similar qualitative forces to those in my model by varying whether public donations are elicited via direct elicitation or a multiple-price list. Reassuredly, both analyses produce similar results: for low image concerns, the principal prefers decisions made after the realization of nature (direct elicitation, respectively no commitment) and for high image concerns the principal prefers decisions made before the realization of nature (multiple price lists, respectively with commitment). However, these papers all assume that the agent chooses the action, rather than communicating verifiable information to a principal who then chooses the action.

The literature on verifiable information typically assumes that the agent lacks commitment and always prefers a higher action 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 action (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 action should be taken.[2]

Finally, the descriptive literature on scientific publishing policies has also discussed the importance of commitment through PAPS, see Banerjee et al. (2020) for a review. Here, the primary tradeoff considered is that PAPs may discourage p-hacking, but researchers incur up-front costs thinking through all possible contingencies. Similarly, Coffman and Niederle (2015) argue that these costs 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]Zhang (2024) is the closest analog to Morris (2001) in the disclosure literature. However, Zhang (2024) assumes that the unbiased agent always benevolently discloses, shifting the analysis to that of a biased agent with directional preferences. In contrast, my analysis allows for general utility functions, finds that the unbiased agent may self censor, and also analyzes the environment with commitment.

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 $u^p(a, \omega)$. The relative weight on each agent's image payoff, denoted by $\lambda \geq 0$, will be referred to as the strength of reputational incentives. The utilities of the biased and unbiased agents are:

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

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

where $a(\cdot), \pi(\cdot)$ and $\phi(\cdot)$ correspond to the action chosen by the principal, disclosure rule chosen by the agent, and the reputational update by the principal about the agent's type, respectively, as discussed further below.[3]

Information: I allow the biased and unbiased agents to, potentially, utilize different data-generating processes. Here, I assume that an unbiased (respectively, biased) agent observes the realization of ω with probability $q^u \in (0, 1)$ (respectively, q^b) and observes n with probability $1 - q^u$ (respectively, $1 - q^b$), where n denotes that the agent observed no information.[4] Finally, let $\Omega^c := \Omega \cup n$.

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

Commitment: In the analysis, I vary the agent's commitment power. I say

[3]The continuity of $u^p(a, \omega)$ and the compactness of A ensure the existence of an optimal action. If multiple actions maximize the principal's utility, I assume that ties are broken in favor of the action the biased agent prefers to guarantee equilibrium existence.

[4]I discuss this assumption in depth after detailing the model in the discussion portion. Information acquisition asymmetries have been explored in the historical literature on reputational concerns, including Scharfstein and Stein (1990). Furthermore, given the generality of the payoffs, this model is equivalent to one with a latent state θ, where ω serves as a signal observed by the agent.

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 unobservable. This definition is equivalent to saying the agent lacks commitment if π is chosen after observing ω. In the discussion section below, I discuss when the commitment assumption is reasonable.

Reputation and Preferences: Denote the principal's posterior beliefs that the agent is unbiased by $\phi : \Omega^c \to [0, 1]$. The preferences in Equations (1) and (2) state that both types of agent have a reputational payoff based on their perceived type and a non-image payoff based on the action chosen. Importantly the equilibrium inference $\phi(\cdot)$ will condition on the agent's choice of $\pi(\cdot)$ for all ω when the agent has commitment, whereas without commitment it conditions on the realization of $\pi(\omega)$ alone. Throughout, I place no restrictions on off-path beliefs for $\phi(\cdot)$.

I examine the Perfect Bayesian Equilibria of this game, hereafter *equilibria*. When agents value reputation, multiple equilibria generally emerge. Therefore, some results focus on the principal's preferred equilibrium.

2.1 Discussion:

This framework asks how political-correctness concerns affect the disclosure of hard information with and without commitment. I first comment on the information generation assumption then the commitment assumption.

Information generation: I assume 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 or p-hacking. Microfounding the information acquisition stage would naturally lead to differential data-generating processes for the two types of agents. For instance, one microfoundation would endogenize q^u, q^b by allowing the unbiased (respectively, biased) agent to choose q^u (respectively q^b) at a cost $c^u(q^u)$ (respectively, $c^b(q^b)$).[5] Further, differential data-generating processes are discussed in qualitative discussions of self-censorship. For instance, Loury (1994) notes, "people genuinely committed to justice did not be-

[5]A related microfoundation is provided by Inderst and Ottaviani (2012), who show that a financial advisor earning large commissions—and thus potentially biased—has stronger incentives to acquire information. Additional microfoundations 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), resulting in differential information discovery.

come entangled in arcane technical arguments about the effects of economic boycotts (of South Africa)." Returning to the motivating vignette 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.

Commitment: I assume that the agent can commit to a disclosure policy. However, a researcher is hardly ever bound to a disclosure policy on their own. In contrast, a third-party who received access to the data (such as the research sponsor, journal, or professional organization) may be needed to invoke commitment. For example, if this third party can (i) forbid publications of results which are not included in the PAP and (ii) require all results included in the PAP to be included in the analysis, then the PAP can be viewed as commitment to a disclosure policy.[6]

A different interpretation of the model is that the agent is an academic journal. The journal wants to influence an action by the 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, which would naturally differ for biased and unbiased journals. Under this interpretation, the recent policies enacted by the Journal of Development Economics and the Journal of Political Economy Microeconomics where papers can be approved based on the PAP before the results are known can be viewed as commitment.

3 Results

This section analyzes how political-correctness concerns impact disclosure policies. In Section 3.1, I conduct the analysis without commitment. In Section 3.2, I consider the analysis when the agent commits to a disclosure policy. To simplify the exposition and focus on the case of interest, I introduce the following assumption on the equilibrium behavior absent reputational concerns.

Assumption 1. *There exists an $\epsilon > 0$ such that for any π^u, π^b which result in the principal obtaining her first-best payoff there is a subset $\Omega' \subset \Omega$ for which (i) $P(\Omega') > 0$ and (ii) $v^b(a(\omega), \omega) + \epsilon < v^b(a(n), \omega)$ for all $\omega \in \Omega'$.*

[6]Within the example of Morris (2001), a researcher studying the effects of admission to an elite college may wish to disaggregate the effects by race. However, if admission overwhelmingly favors a certain race, the researcher may prefer to self-censor. In this context, the commitment assumption is equivalent to stating that the researcher can include the disaggregated effect if and only if they are included in the PAP.

This assumption introduces conflict by stating that absent reputational considerations, the biased agent does not choose the principal's preferred policy.[7] I note that this assumption is met in many settings (c.f. Dye, 1985; Morris, 2001; Kamenica and Gentzkow, 2011), such as if $u^p(a,\omega)$ is single-peaked at ω and $v^b(a,\omega)$ is strictly monotone in a or single-peaked at $\omega + b$.

3.1 Disclosure:

This subsection considers the environment where the agent lacks commitment. 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 finite cutoff $\lambda^1(q^u, q^b)$ such that: if $\lambda > \lambda^1(q^u, q^b)$, then $a = 0$ with probability one in any equilibrium.*

Lemma 1 implies that when $\lambda > \lambda^1(q^u, q^b)$, neither agent discloses $\omega = 1$. This implication may be surprising because when $\omega = 1$, 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 has a strictly larger non-image incentive to disclose ω than the unbiased agent. If sharing $\omega = 1$ was on path, then because $q^u < q^b$, the reputation from disclosure must be strictly worse than the prior. However, for high enough reputational weights the agents will self-censor to avoid this reputational loss. □

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.

[7] Notably, absent such an assumption self-censorship may still arise. For instance, let $u^p(a,\omega) = -(a-\omega)^2$ and $v^b = -(a-\alpha\omega)^2$ with $\alpha > 1$. Using an identical argument to Example 1 below, one can show that for $q^u < q^b$ and sufficiently high λ self-censorship occurs with probability one.

Example 2: Suppose ω admits a probability density function, f_ω, with convex support Ω. The biased agent's non-image payoff is $v^b(a, \omega) = 2a$, whereas the principal's utility is $u^p(a, \omega) = -|a - \omega|$. Finally, p remains the probability that the agent is unbiased, $a \in A = \Omega$, and $a(n)$ the action the principal takes absent any information.

Lemma 2. *In the environment described above, for any f_ω, p there exists a positive Lebesgue-measure set of (q^u, q^b), and a cutoff $\lambda^2(q^u, q^b)$ such that: if $\lambda > \lambda^2(q^u, q^b)$, then in any equilibrium any $\omega > a(n)$ is self-censored. Finally, there always exists an equilibrium where the unbiased agent always reveals ω if $\omega \leq a(n)$.*

This lemma states that when $\lambda > \lambda^2(q^u, q^b)$ any information which would result in a higher action taken than $a(n)$ is self-censored. Further, $\omega > a(n)$ occurs with strictly positive probability, implying the principal does not obtain her first-best payoff. The intuition for the self-censorship is identical to Lemma 1: the biased agent has a strictly greater incentive to disclose $\omega > a(n)$, implying a strict reputational gain from self-censorship, resulting in self-censorship for high reputational weights.

However, this proposition also states that in this environment, self-censorship does not occur for all realizations. Here, as $a(n)$ must lie in the interior of the support of ω, then for $\omega < a(n)$, the unbiased agent has a strictly greater incentive to disclose 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 action or (ii) mix between revelation and concealment.[8]

I now present a characterization of when the principal obtains her first-best payoff. To juxtapose with the historical literature (c.f. Morris, 2001), I also discuss when communication is informative: defined as when there exists an equilibrium where the principal's payoff strictly exceeds her payoff when $\pi^u(\omega) = \pi^b(\omega) = n \ \forall \ \omega$.

Proposition 1. *The following are true:*

1. *The principal obtains her first-best payoff in her preferred equilibrium if and only if $q^u > q^b$ and λ exceeds a finite threshold $\lambda^*(q^u, q^b)$.*

2. *Communication is informative if $q^u \geq q^b$ for any $\lambda \geq 0$. Specifically, there always exists an equilibrium where $\pi^u(\omega) = \omega$. In contrast, when $q^u < q^b$, there exist preferences and a finite cutoff $\lambda^{**}(q^u, q^b)$, where communication is uninformative if $\lambda > \lambda^{**}(q^u, q^b)$.*

[8]The formal proof uses a fixed-point argument to construct disclosure strategies for the biased agent when $\omega < a(n)$ and both agents when $\omega > a(n)$ to prove the existence of such an equilibrium.

To gain intuition behind the first result, note that if λ is low, then by Assumption 1, the biased agent does not fully disclose. Hence, assume that λ 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. In contrast, if $q^u \leq q^b$, the principal cannot obtain her first-best payoff. If she did, then both types must fully disclose, implying self-censorship has a (weak) reputational benefit. However, Assumption 1 ensures that the biased type does not fully disclose, deriving a contradiction.

To gain intuition behind the second result, if $q^u \geq q^b$ and the unbiased agent is conjectured to fully disclose, then disclosure has a weak reputational benefit. This benefit combined with being unbiased implies the unbiased agent indeed fully discloses. Given that the unbiased agent full discloses independently of the strategy of the biased agent, one can use a fixed-point argument to prove the existence of a strategy for the biased agent which will be consistent with an equilibrium. Finally, if $q^u < q^b$, then as seen in Lemma 1 and Lemma 2 whether communication is informative or not for high values of λ depends on the shape of preferences.[9]

These results show that the presence of hard information alone does not eliminate political-correctness concerns. Further, these concerns can be sufficiently large to result in no information transmission (Lemma 1). Finally, whether or not hard information is transmitted depends on which type of agent is more likely to be informed (Proposition 1). In contrast, 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 has commitment.[10] With commitment, the principal's beliefs about an agent's type are not only a function of the realization of the information shared with the principal, but also the chosen disclosure policy. 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.

[9] When λ is low, then the unbiased agent will fully disclose implying communication is informative.

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

Lemma 3. *There always exists an equilibrium where the unbiased agent commits to fully revealing his information.*

This result is in contrast to (i) those in the previous section absent commitment and (ii) the results with soft information: communication is uninformative if reputation is sufficiently valuable(c.f. Morris, 2001). 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 the biased agent prefers 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 utilizes a mixed strategy. Here, with probability p he chooses full disclosure and with probability $1-p$ he chooses his preferred policy when $\lambda = 0$. 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.[11] Finally, when the reputational incentives get sufficiently small, each type of agent chooses 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.

[11]This equilibrium is a lower bound for the principal's payoffs. The complete proof resembles the inscrutability principle (c.f Myerson, 1983). Here, I consider a mechanism where the agent privately reports his type and the mechanism determines a disclosure policy. Given the results in this literature, unsurprisingly, there may exist a fully-pooling equilibrium that the principal prefers to the semi-pooling equilibrium in the text.

This subsection proves that commitment effectively eliminates political-correctness for the unbiased agent. Further, if the reputational incentives are high enough, then the principal can always obtain her first-best payoff.

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 both maintain a reputation for being unbiased and to influence an action. Further, I contrast the equilibrium disclosure polices when the disclosure decision is made before observing the information (agent has commitment) to when this decision is made after observing the information (agent lacks commitment). If the agents did not value reputation, commitment has no effect on the incentives of the unbiased agent, because he always prefers to truthfully reveal his information. However, commitment allows the biased agent to more selectively reveal his information (c.f. Kamenica and Gentzkow, 2011). These results suggest that commitment hurts the principal when reputational incentives are low.

However, when reputational incentives are high, I show that commitment benefits the principal. When the agent can commit, the principal can conjecture that the unbiased agent fully discloses and that any other disclosure policy is chosen by only the biased agent. Therefore, when the reputational weight is high, both agents will fully disclose. Crucial to this argument is that selective disclosure is both verifiable by the principal and off path, ensuring that such deviations are met with pessimistic reputational beliefs. In contrast, without commitment, selective disclosure can not be verified and is met with the same reputation as an agent who did not discover information. Further, if the unbiased agent is weakly less likely to discover information than the biased agent, then the principal's payoff remains bounded away from her first-best payoff for high reputational weights. These results imply that commitment weakly improves the principal's payoff when reputation is valuable and strictly so if the information is more likely to be generated by a biased agent.

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). Additionally, one could 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.

Bénabou, Roland, Armin Falk, Luca Henkel, and Jean Tirole, "Eliciting moral preferences: Theory and experiment," *Princeton University*, 2020.

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.

Inderst, Roman and Marco Ottaviani, "How (not) to pay for advice: A framework for consumer financial protection," *Journal of Financial Economics*, 2012, *105* (2), 393–411.

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.

Myerson, Roger B, "Mechanism design by an informed principal," *Econometrica: Journal of the Econometric Society*, 1983, pp. 1767–1797.

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

Scharfstein, David S and Jeremy C Stein, "Herd behavior and investment," *The American economic review*, 1990, pp. 465–479.

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

Denote $\tilde{\pi}^u(\omega) = P(\pi^u(\omega) = \omega)$, $\tilde{\pi}^b(\omega) = P(\pi^b(\omega) = \omega)$. Further, denote $\delta(\omega) = \mathbf{1}_{\tilde{\pi}^u(\omega)+\tilde{\pi}^b(\omega)>0}$, namely whether ω is disclosed on path.

Proof of Lemma 1. Note that if q^u and q^b are sufficiently low, then given that $\mu < 1/2$, $a(n) = 0$. Further $a(0) = 0$. Hence, it suffices to show that if $\delta(1) = 1$, then $\phi(1) < \phi(n) - \epsilon$, for $\epsilon > 0$, which would derive a contradiction for $\lambda > 2/\epsilon$.

If $\delta(1) = 1$ and $q^u < q^b$, then, as argued in the text, $\tilde{\pi}^b(1) = 1$ implying:

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

Therefore, it suffices to show that $\phi(n) \geq p$. Suppose by contradiction $\phi(n) < p$. By Bayes rule, then $\delta(0) = 1$ and $\phi(0) > p$. Therefore, $\phi(0) > \phi(n)$ and $a(0) = a(n)$, implying $\tilde{\pi}^u(0) = \tilde{\pi}^b(0) = 1$. As $q^u < q^b$ then $\phi(0) < p$, deriving a contradiction. □

Proof of Lemma 2. Fix (q^u, q^b) 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 because $q^u < q^b$ and the biased agent has a strictly larger incentive to disclose $\omega > a(n)$.

Now it suffices to show (ii) holds for a sufficiently large λ. For any $\omega < a(n)$, if $\delta(\omega) = 1$, then $|\phi(\omega) - \phi(n)| \leq \frac{c}{\lambda}$, where c is an exogenous constant determined by the maximum loss between two actions for any two states. Therefore, by Bayes' rule $\phi(n) \geq p$ for λ sufficiently high, proving (ii).

I now show that there always exists an equilibrium where $\tilde{\pi}^u(\omega) = 1$ if $\omega < a(n)$. To do so, I define a best-reply for each ω given two conjectured values for $a(n), \phi(n)$. If the resulting $a(n), \phi(n)$ from this best-reply are equal to the conjecture, then the strategy will constitute an equilibrium. Note that, in general, fixing $\omega, a(n), \phi(n)$, multiple strategies may constitute an equilibrium, but the construction below will select a particular best-reply resulting in an equilibrium with the desired properties.

If $\omega \leq a(n)$, set $\tilde{\pi}^u(\omega) = 1$. Further, $\tilde{\pi}^b(\omega)$ is uniquely determined by $a(n), \phi(n)$ because an increase in $\tilde{\pi}^b(\omega)$ results in a strictly lower incentive to reveal as $\phi(\omega)$ would decrease.

If $\omega > a(n)$, then the best reply is chosen as follows. If neither agent prefers to deviate when $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1$, then $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1$ is chosen. If there exists a profitable deviation, this deviation must come from the unbiased agent. In this case we construct a best-reply as follows. $\tilde{\pi}^u(\omega) = 0$ and $\tilde{\pi}^b(\omega) \in [0, 1]$. Given $\tilde{\pi}^u(\omega) = 0$, as argued in the paragraph above, $\tilde{\pi}^b(\omega)$ is uniquely determined. It suffices to check that the unbiased agent indeed prefers to self-censor. However, the unbiased agent preferred to self-censor when $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1$ (i.e., $\phi(\omega) > 0$), implying they would also prefer to self-censor when $\tilde{\pi}^u(\omega) = 0$ (i.e., $\phi(\omega) = 0$).

These best-replies determine an equilibrium if the resulting $a(n), \phi(n)$ are consistent conjectures in equilibrium. Hence, an equilibrium exists if there exists a fixed-point of the function 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 and continuous utility functions), then there exists a fixed-point determining an equilibrium with the desired properties. □

Proof of Proposition 1. Let $S := \{\pi^u, \pi^b | \text{P obtains first-best payoff}\}$. S is non-empty as it includes $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1$. Fix $\pi^u, \pi^b \in S$, then with probability 1 for any ω where $a(\omega) \neq a(n)$, $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1$. Therefore, for such ω, $\phi(\omega) = \frac{pq^u}{(1-p)q^u + pq^b}$. If $q^u \leq q^b$, then $\phi(\omega) \leq p \leq \phi(n)$. Finally, Assumption 1 implies the biased agent has a profitable deviation if $\lambda = 0$, and $\phi(\omega) \leq \phi(n)$ implies a deviation exists for $\lambda > 0$.

If $q^u > q^b$, then for any $\pi^u, \pi^b \in S$, $\phi(\omega) > \phi(n)$ and as the principal obtains her first-best payoff the unbiased agent has no deviation temptation. Therefore it suffices to analyze the biased agent. Fixing $\pi^u, \pi^b \in S$, the non-image benefit of deviating is independent of λ and the reputational loss of deviation is linear in λ, implying the existence of a threshold. This threshold is bounded away from zero by Assumption 1 and strictly finite as the reputational cost from concealment is unbounded as $\lambda \to \infty$. Taking the infimum of these thresholds over S completes the proof.

Next, I show the existence of an equilibrium where $\tilde{\pi}^u(\omega) = 1 \forall \omega$ if $q^u \geq q^b$. In this equilibrium, $\phi(\omega) \geq \phi(n) \forall \omega$, 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. □

Proof of Lemma 3. Conjecture beliefs by the principal that the unbiased agent chooses to fully reveal and that any other disclosure policy if chosen is chosen by the biased agent. Given such conjectures, the unbiased agent fully reveals.

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 continuously decreasing (due to the reputation considerations) in $\tilde{p}^b$. If neither 1 nor 0 are consistent conjectures, then the decreasing best response implies that an intermediate value must be a consistent conjecture, proving the existence of an equilibrium. □

Proof of Proposition 2. I first prove the principal's payoff in her preferred equilibrium is monotone in 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 or both chosen policies result in identical payoffs for the principal, then this equilibrium will remain an equilibrium for higher values of λ. Thus, let us consider when two policies that result in different payoffs are chosen on path for a fixed λ. There are two cases:

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. 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 is, without loss of generality, 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 to policy 2 by the principal, 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.

Given monotonicity, to prove the existence of $\bar{\lambda}^c$ it suffices to show there exists

a sufficiently high λ above which the principal obtains her first-best payoff. To construct such an equilibrium, conjecture that on-path $\tilde{\pi}^u(\omega) = \tilde{\pi}^b(\omega) = 1 \forall \omega$ and the principal assumes any off-path disclosure policy is chosen by the biased agent. For this conjecture 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 linear with respect to λ, resulting in the cutoff $\bar{\lambda}^c$.

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 with positive probability. For any such policy, one agent must have a different disclosure policy which provides an improvement in their non-image payoff by at least $\delta > 0$ (by Assumption 1). Thus, for $\lambda < \delta$, one of the types of agents would prefer to deviate to their preferred disclosure policy absent reputation. $\square$