A Dynamic Model of Censorship

Yiman Sun*

February 19, 2023

Abstract

We study the interaction between an agent of uncertain type, whose project gives rise to both good and bad news, and an evaluator who must decide if and when to fire the agent. The agent can hide bad news from the evaluator at some cost, and will do so if this secures her a significant increase in tenure. When bad news is conclusive, censorship hurts the evaluator, the good agent, and possibly the bad agent. However, when bad news is inconclusive, censorship may benefit all those players. This is because the good agent censors bad news more aggressively than the bad agent, which improves the quality of information.

Keywords: Censorship, Information Manipulation, Learning, Dynamic Games

JEL Codes: C73, D82, D83

*CERGE-EI, a joint workplace of Charles University and the Economics Institute of the Czech Academy of Sciences. Email: yiman.sun@cerge-ei.cz

I am very grateful to V. Bhaskar and Caroline Thomas for their constant support and invaluable guidance. I would like to thank three anonymous referees, Daniel Ackerberg, Svetlana Boyarchenko, Yi Chen, Cary Deck, Tommaso Denti, Laura Doval, Ignacio Esponda, Andrew Glover, Yingni Guo, Ayça Kaya, Frédéric Koessler, Stephen Morris, Xiaosheng Mu, Mallesh Pai, Harry Di Pei, Jacopo Perego, Jérôme Renault, Larry Samuelson, Vasiliki Skreta, Andrzej Skrzypacz, Joel Sobel, Dale Stahl, Colin Stewart, Maxwell Stinchcombe, Rodrigo Velez, Thomas Wiseman, Takuro Yamashita, Muhamet Yildiz, and other seminar participants at UT Austin, Toulouse School of Economics, the 29th Stony Brook Game Theory Conference, 2018 Midwest Economic Theory Conference, 2018 Texas Economic Theory Camp, 2018 SEA Conference, and 2020 SITE Conference for their excellent suggestions and helpful discussions. I acknowledge support by the ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program), the European Research Council (Starting Grant #714693), and the Lumina Quaeruntur fellowship (LQ300852101, Challenges to Democracy) of the Czech Academy of Sciences.

1 Introduction

Individuals or firms often suppress information that negatively affects their reputation. Such censorship is widely regarded as undesirable. Nonetheless, suppressing bad news may help those individuals or firms that have good potential but suffer from a low reputation due to bad luck. For example, customers often check online reviews before buying products. However, a bad review may exist for reasons unrelated to the product's quality, e.g., damage during delivery or the customer was in a bad mood. A seller who has good products but has not established her reputation may suppress and thereby survive bad reviews thanks to Reverse SEO.[1]

This paper studies censorship in a dynamic game between an agent who seeks to remain in her job and an evaluator who wants to retain the competent agent and dismiss the incompetent one. The agent's project gives rise to both good and bad news, from which the evaluator learns the agent's competence. Good news is publicly observed and confirms the agent is competent. Bad news arises more frequently when the agent is incompetent. The agent knows her competence and can suppress bad news when it materializes, but this is costly.

In addition to online reputation management, our model can be applied to broader contexts involving information censorship. The agent may be a division manager of a firm, while the evaluator is the CEO. The firm has multiple divisions and wishes to keep only those that are potentially profitable. Consequently, the manager will lose her job if her division is deemed unprofitable and abolished. The CEO relies on internal reports to evaluate divisions and managers. However, accounting books can be "cooked," machine log files faked, and inspectors/auditors bribed,[2] all of which cost effort and money.

Basic economic intuition suggests that concealing information hurts the evaluator. Moreover, the possibility of censorship makes the evaluator suspicious about the agent's performance. In the absence of bad news, he does not know whether it is because the agent is competent or because she is censoring information. The possi-

[1]Reverse SEO is a reputation management tactic that maintains Internet reputations by suppressing or burying negative results on search engines.

[2]Internal audit only detects fraud 15% of the time (including financial statement fraud). See a report from ACFE (https://s3-us-west-2.amazonaws.com/acfepublic/2018-report-to-the-nations.pdf).

bility of censorship hurts a competent agent because she cannot prove that she did not censor anything. Thus, censorship can only benefit an incompetent agent since it helps her survive bad news. If this intuition is correct, the policy implication would be to reduce censorship by making it as costly as possible.

We contribute to the literature by showing that the above intuition is only partially correct, depending critically on whether bad news is conclusive. Indeed, when bad news only arises for the incompetent agent, censorship hinders the transmission of information, which hurts the evaluator, the competent agent, and sometimes the incompetent agent. However, when bad news also arises for the competent agent, we show that censorship may increase the welfare of all parties. To our best knowledge, this is the first paper that identifies this positive welfare effect of censorship.

We now turn to the details of our model and results. There are both good and bad news events. Good news events are conclusive and can only be obtained by good types. Bad news is generally inconclusive, arriving at higher rates for bad types than good types. We also assume that good news arrives faster than bad news, which corresponds to cases where the project has a relatively promising prospect and adverse news events are not abundant. Note that it is a "good news" model à la Keller et al. (2005), as the evaluator's belief about the agent being good declines in the absence of news.

In Section 4.1 we assume that bad news is also conclusive. We show that although the evaluator is forward-looking, the belief threshold at which he fires the agent is independent of censorship (Lemma 1). Since censorship is costly, the agent censors if and only if the reputation damage caused by bad news exceeds the cost (Lemma 3). As conclusive bad news destroys reputation completely, the reputation damage is the reputation value itself. We show there is a unique belief threshold, at which the reputation value crosses the cost of censorship (Lemma 2). The agent switches from Full-Censorship to No-Censorship from that point (Proposition 1). The evaluator and the good agent are worse off with censorship, confirming our initial intuition. The bad agent may also be worse off as she has a shorter tenure in the absence of news than she does when censorship is impossible (Proposition 2). This is because she has no creditable way to commit not to censor bad news and thus the evaluator is skeptical. As a result, her reputation declines faster.

The analysis in Section 4.2 assumes inconclusive bad news. We find an equilibrium with a similar structure where both types of agent stop censoring at certain belief thresholds, and show that it is the unique pure strategy Markov Perfect Equilibrium if bad news is sufficiently bad (Proposition 3). Our main insight arises from the fact that the good type has a greater incentive to censor bad news than the bad one. This is because the good type knows that good news may arise and secure her permanency in tenure, and she is less likely to get further bad news. It is both more beneficial and less costly for her to maintain her reputation. Thus, the belief threshold p^G at which the good type stops censoring is lower than the threshold $\hat{p}^B$ of the bad type. Consequently, the two types separate in censorship strategies when their reputation is intermediate in the interval $(p^G, \hat{p}^B)$, and bad news becomes endogenously conclusive—any bad news may only come from the bad type as the good one always censors it. Thus, censorship improves the quality of information and benefits the evaluator. Moreover, the evaluator becomes less pessimistic over time due to better information. Thus, the agent's reputation declines at a slower rate, which also benefits her (Proposition 4).

In our model, the evaluator solves a Poisson bandit problem where the news process is endogenously determined by the agent's censorship strategy. This allows us to study the relationship between experimentation and censorship. However, we find that the evaluator's experimentation strategy, i.e., the dismissal belief threshold, is not affected by censorship. This is due to our focus on the "good news" model. We discuss how the "bad news" model changes this situation in Section 5. Although the evaluator's experimentation strategy is not affected, we explore how censorship affects the evolution of the reputation in our dynamic model (Figure 1). This is a key factor that shapes the agent's incentive to censor and drives the welfare results.

Our results have policy implications. The European Union recently adopted a series of legal actions that made it easier to remove negative search results from Google, which is called the right of erasure (Article 17, 19) in the General Data Protection Regulation (GDPR). Our results suggest that whether this improves welfare depends on the nature of the information. Information such as criminal records may be considered to be conclusive. Making it easier to remove them may hurt the public. However, the right of erasure may benefit the person concerned as well as the public, if the negative search results contain only inconclusive evidence.

1.1 Related literature

We contribute to the literature with new welfare insights into costly censorship. Most relatedly, Hauser (2021) considers a similar dynamic environment where a firm invests in its quality and censors bad news to maintain its reputation. He studies only conclusive bad news and finds that cheap censorship hurts the firm by crowding out investment. We also find that censorship can hurt the censor when bad news is conclusive. However, our result is unrelated to investment but driven solely by the evaluator's skepticism about the absence of news. Unlike Hauser (2021), we also allow for inconclusive bad news. Remarkably, we find that censorship may benefit both the informed and uninformed parties[3] because different types of agent censor bad news differently, which is only possible when bad news is inconclusive.

Shadmehr and Bernhardt (2015) study political censorship in a one-shot game, and also find that censorship may backfire due to skepticism following no news. They assume that news from different types of ruler is distinguishable, and citizens observe the differences, while we assume the evaluator cannot tell where bad news comes from (unless it is conclusive). We both explore the equilibrium meaning of "no news," but only our model explores that of "bad news." Thus, our results on inconclusive bad news have no counterpart in their paper. We also differ from Shadmehr and Bernhardt (2015) in that we emphasize the welfare effect on the evaluator, whereas they, like other works on political censorship,[4] do not provide a framework to study citizens' welfare or presume censorship affects them negatively.

Smirnov and Starkov (2021) study censorship on product reviews, but consider costless censorship and assume a seller faces both Bayesian and naive buyers. The model and results are quite different from ours. They show that bad news exists even if censorship is costless, because Bayesian buyers treat bad reviews as good news.

Our paper is related to the literature on information disclosure and reputation.[5] Board and Meyer-ter Vehn (2013) study how information shapes a firm's reputation

[3]Hauser (2021) focuses only on the informed party and assumes the uninformed one non-strategic.

[4]This literature focuses on the role of media outlets. See Besley and Prat (2006), Eraslan and Ozerturk (2017), Gehlbach and Sonin (2014), Guriev and Treisman (2018), Kolotilin et al. (2019), Egorov et al. (2009), and Lorentzen (2014). Also, Edmond (2013) and Redlicki (2017) study a signal-jamming model in global games.

[5]Dye (2017), Daughety and Reinganum (2018), and Kartik et al. (2017) study disclosure of verifiable information in other applications where concealing information is costly.

and investment in a dynamic model. Their paper provides a workhorse model for our analysis of the agent. Marinovic et al. (2018)[6] extend the model and assume a firm can certify its quality perfectly. They show that the firm suffers from "over-certification traps": it must certify frequently because of buyers' expectations. Our model with conclusive bad news leads to analogous "over-censorship traps": the agent suffers a rapidly declining reputation because the evaluator expects censorship. Both traps are driven by the lack of commitment to reduce certification/censorship so that "no news" is interpreted unfavorably in equilibrium. We additionally analyze the interpretation of news arrivals when bad news is inconclusive, which has no reflection in their model as they assume certification perfectly reveals quality.

Similarly to our setup, Ekmekci et al. (2020) consider a dynamic game where a principal wants to terminate a bad agent, but the agent can manipulate information to stay in the relationship at a cost. Unlike us, they study the fabrication of good news, which only the bad agent can produce. Kuvalekar and Lipnowski (2020) study a dynamic game between a firm and a worker with stopping decisions, but consider observable actions that control symmetric information about the quality of the match. In contrast, our results rely on asymmetric information and unobservable censorship.

Bhaskar and Thomas (2019) and Kovbasyuk and Spagnolo (2021) both show that a rating system that erases public records can be efficient. This is similar to our finding that censorship can be beneficial. However, the reasons behind this efficiency entirely differ from those found in our study. Bhaskar and Thomas (2019) study how erasing records sustains punishments and trust in a repeated game. Kovbasyuk and Spagnolo (2021) show how hiding information can enhance market transactions and generate information externalities for future buyers.

Our idea that inconclusive news becomes conclusive through separation in strategies appears in other applications. Povel and Strobl (2019) analyze a principal-agent model with manipulable performance reports, and show that inducing the agent to fabricate a positive report can make a noisy report more informative. In contrast to our approach, they study the optimal contract design. Bar-Isaac (2003) studies the survival of a firm in a dynamic signaling model where good and bad firms separate

[6]Other papers use a similar framework to study reputation, but also consider the firm's exit decision (Board and Meyer-ter Vehn (2021)), the promotion of reputation (Hauser (2017)), and the design of optimal monitoring (Varas et al. (2020)).

in trading behaviors. In contrast, we show that different types of agent separate in how they manipulate information, and focus on the welfare consequences.

Finally, this paper relates to two-armed Poisson bandit models in continuous time, e.g., Presman (1991), Keller et al. (2005), Keller and Rady (2015), Thomas (2016). We borrow results from this literature to solve the evaluator's problem.

2 The model

Two risk neutral players, an agent (she) and an evaluator (he), play a game in continuous time $t \in [0, \infty)$. The common discount rate is $\rho > 0$.

The agent owns a project. At time $t = 0$, nature chooses the type θ of the project from the set $\Theta = \{G, B\}$. We assume that the agent observes nature's choice, while the evaluator does not. Let $p_0 \in (0, 1)$ denote the probability that nature chooses G. Thus, the project's type is the agent's private information or her type.

The agent enjoys a flow payoff $w > 0$ that is independent of her type while she stays in her job, and has no payoffs after she is dismissed by the evaluator.[7] Only the type G agent can succeed in her project.[8] A success is publicly observable and arrives at each jumping time of a Poisson process $\mathcal{S} = \{S_t\}_{t\geq 0}$ with an arrival rate $\gamma > 0$. The first success reveals that the agent is type G.

A success yields a lump-sum payoff $k > 0$ to the evaluator. He can choose a time $t \geq 0$ to irreversibly dismiss the agent. After the dismissal, the game ends and he receives his outside option normalized to m. We assume $h := \gamma k > m > 0$; he prefers the type G agent to the outside option, and the outside option to the type B agent.

The agent's project may produce adverse news events. The news arrival rate is type dependent. The type θ project generates a piece of news at each jumping time of a Poisson process $\mathcal{N}^\theta = \{N_t^\theta\}_{t\geq 0}$ with an arrival rate β^θ; $\beta^B > \beta^G \geq 0$. The success

[7] Fixed hourly payment is common for wage and salary workers, according to the US Bureau of Labor Statistics; see Kuvalekar and Lipnowski (2020). Besides the fixed payment, our main results still hold if we include an incentive payment conditional on a success on the project. This would increase the type G agent's incentive to maintain her reputation.

[8] This assumption is analytically convenient. However, it would be sufficient to assume that the type B agent succeeds at a lower rate, because the relevant driving force is that the type G has a higher rate of success and can thus stay in her job longer.

process $\mathcal{S}$ and the news process $\mathcal{N}^G$ are independent, conditional on the type G. We call such news bad news since it happens more often to the type B agent, although it has no impact on the payoff. We also call a success good news. We assume $\gamma > \beta^B$; good news arrives faster than bad news from the type B agent. This implies that beliefs about the agent being type G drift down in the absence of news. We will discuss this assumption in Section 5.

The agent observes the bad news process. When a piece of bad news arrives, she can choose to incur a lump-sum cost $c > 0$ to censor it. We assume $c < \bar{c} := w/(\rho + \beta^B)$ so that the cost is not prohibitively high.[9] The evaluator observes bad news if and only if the agent does not censor it, but he cannot distinguish whether it comes from $\mathcal{N}^G$ or $\mathcal{N}^B$ unless $\beta^G = 0$. Let $X_t^\theta \in \{1, 0\}$ denote the censoring decision of the type θ agent at time $t \geq 0$ when a piece of bad news arrives at time t; $X_t^\theta = 1$ denotes censoring. We say that a piece of bad news is revealed to the evaluator if it is not censored (i.e., $X_t^\theta = 0$). Good news is always revealed.

Histories and Strategies – A history of the type θ agent at time t is denoted by h_t^θ. It consists of a finite sequence of news realizations and her censoring decisions up to t. Her strategy $\boldsymbol{x}^\theta = \{x_t^\theta\}_{t\geq 0}$ is predictable with respect to the associated filtration, where $x_t^\theta \in [0, 1]$ is the censoring probability, conditional on bad news arriving at time t. Intuitively, x_t^θ depends on the history h_{t-}^θ prior to t.[10]

The evaluator only observes the public history—the revealed news. Both his strategy and his conjecture of the agent's strategy depend on the public history. A public history $\tilde{h}_t$ at time t consists of a finite sequence of revealed news realizations up to time t. The evaluator's strategy is a stopping time T with respect to the filtration generated by the public history. Let $\tilde{\boldsymbol{x}}^\theta = \{\tilde{x}_t^\theta\}_{t\geq 0}$ be a predictable process with respect to the same filtration, where $\tilde{x}_t^\theta \in [0, 1]$. T is the time at which the agent is dismissed. $\tilde{\boldsymbol{x}}^\theta$ is the public conjectured strategy of the type θ agent.

Payoffs – When the type θ agent is not dismissed (i.e., $T \geq t$), she enjoys a flow benefit of $w\,\mathrm{d}t$ but pays a cost of $c\,X_t^\theta\,\mathrm{d}N_t^\theta$. Thus, the cost is paid when a piece of bad news arrives and she chooses to censor it. Given T, $\boldsymbol{x}_t^\theta = \{x_\nu^\theta\}_{\nu\geq t}$, and a private

[9] If the censorship cost is very high, the agent will not censor news in equilibrium. Assuming $c < \bar{c}$ ensures that is not the case: $c > \bar{c}$ implies the type B agent prefers being dismissed by revealing bad news to staying in the job forever but at a cost of censoring all bad news.

[10] We use $y_{t-} := \lim_{s\uparrow t} y_s$ to denote the left limit of a process $\boldsymbol{y}$ at time t and $y_{0-} := y_0$.

history h^{θ}_{t-}, the type θ agent's expected discounted payoff at time t is

$$v_{\theta}^{T,\boldsymbol{x}^{\theta}}(h^{\theta}_{t-}) = \mathbb{E}\left[\int_t^T \rho\, e^{-\rho(\nu-t)}(w\,\mathrm{d}\nu - c\; X^{\theta}_{\nu}\,\mathrm{d}N^{\theta}_{\nu})\middle|\, h^{\theta}_{t-}\right], \tag{1}$$

where the expectation is taken over news processes, her censoring strategy, and the evaluator's stopping time. Note that conditioning on h^{θ}_{t-} means that the payoff is evaluated before the agent knows whether a piece of bad news arrives at time t.

The evaluator's continuation payoff after the agent is dismissed is m. Until then he enjoys payoffs from the type G agent's successes. Given T, $\tilde{\boldsymbol{x}}^{\theta}_t = \{\tilde{x}^{\theta}_{\nu}\}_{\nu\geq t}$, and a public history $\tilde{h}_{t-}$, the evaluator's expected discounted payoff at time t is

$$u^{T,\tilde{\boldsymbol{x}}}(\tilde{h}_{t-}) = \mathbb{E}\left[\int_t^T \rho\; e^{-\rho(\nu-t)}\; \mathbb{1}_{\theta=G}\, k\,\mathrm{d}S_{\nu} + e^{-\rho(T-t)}\; m\middle|\, \tilde{h}_{t-}\right], \tag{2}$$

where $\tilde{\boldsymbol{x}} := \{\tilde{\boldsymbol{x}}^{G}, \tilde{\boldsymbol{x}}^{B}\}$, and the expectation is taken over news processes, his stopping time, and the public conjectured strategy of the agent.

Beliefs and Equilibria – The public belief at time t about the agent being type G is denoted by $p_t := \mathbb{P}[\theta = G|\tilde{h}_t]$, where the probability measure is induced by the public conjectured strategies $\tilde{\boldsymbol{x}}$. This can be interpreted as the agent's reputation.

When a success arrives, the public belief jumps up to 1. If a piece of bad news is revealed at time t, the public belief jumps from $p_{t\text{-}}$ to

$$J(p_{t\text{-}},\, \tilde{x}^{G}_t,\, \tilde{x}^{B}_t) := \frac{p_{t\text{-}}\,\beta^{G}\,(1-\tilde{x}^{G}_t)}{p_{t\text{-}}\,\beta^{G}\,(1-\tilde{x}^{G}_t)\;+\;(1-p_{t\text{-}})\,\beta^{B}\,(1-\tilde{x}^{B}_t)}. \tag{JUMP}$$

In the absence of censorship, the belief jumps from $p_{t\text{-}} \in (0,1)$ to $j(p_{t\text{-}}) := J(p_{t\text{-}}, 0, 0) < p_{t\text{-}}$ as $\beta^{B} > \beta^{G}$; bad news is detrimental to reputation in the absence of censorship. When the conjectured strategy is $\tilde{x}^{G}_t = \tilde{x}^{B}_t = 1$, (JUMP) is not well-defined. We assume $J(1,1,1) = 1$ as $p = 1$ is an absorbing state, and $J(p,1,1) = 0$ for $p < 1$.[11]

If no news is revealed over the time interval $[t, t+\mathrm{d}t)$, the public belief evolves

[11] This is a natural restriction when $\beta^{G} = 0$. Moreover, making such an assumption does not reduce the set of equilibria; any equilibrium with an off-path belief that supports censorship as an equilibrium strategy for both types can also be supported with the worst belief. In particular, Lemma 3 does not make use of this assumption.

continuously according to the following ordinary differential equation,

$$\dot{p}_t = d(p_t, \tilde{x}_t^G, \tilde{x}_t^B) := -p_t(1-p_t)\left[\gamma + (1-\tilde{x}_t^G)\beta^G - (1-\tilde{x}_t^B)\beta^B\right]. \qquad \text{(DRIFT)}$$

Note that $\gamma > \beta^B$ implies $d(p_t, \tilde{x}_t^G, \tilde{x}_t^B) < 0$ for any $\tilde{x}_t^G$, $\tilde{x}_t^B$, and $p_t \in (0,1)$. The belief declines in the absence of news, regardless of the public conjectured strategy.

We study Markov Perfect Equilibria (MPE), which are Perfect Bayesian Equilibria (PBE) with equilibrium strategies being Markovian with respect to the state variable—the left limit of the public belief p_{t-}. Note that $p_{t-} = p_t$ for almost all t.

A PBE consists of strategies of each type θ of agent $\boldsymbol{x}^\theta = \{x_t^\theta\}_{t\geq 0}$ and the evaluator T, a public belief $\boldsymbol{p} = \{p_t\}_{t\geq 0}$, and a public conjectured strategy $\tilde{\boldsymbol{x}}^\theta = \{\tilde{x}_t^\theta\}_{t\geq 0}$ such that conditional on $T \geq t$,

1. For each h_{t-}^θ, $\boldsymbol{x}^\theta$ is optimal for the type θ agent, given $\boldsymbol{p}$ and T;
2. For each $\tilde{h}_{t-}$, T is is optimal for the evaluator, given $\boldsymbol{p}$ and $\tilde{\boldsymbol{x}}$;
3. $\tilde{\boldsymbol{x}}^\theta = \mathbb{E}[\boldsymbol{x}^\theta|\tilde{h}]$; $\boldsymbol{p}$ is updated according to (JUMP) and (DRIFT).

We write Markov strategies as functions of the public belief: for any public history starting at time t with a public belief p_t,[12] $x_t^\theta = \boldsymbol{x}^\theta(p_t)$, $\tilde{x}_t^\theta = \tilde{\boldsymbol{x}}^\theta(p_t)$, and $T = T(\Sigma) := \inf\{\nu \geq t : p_\nu \in \Sigma\}$, where Σ is a closed Borel subset of $[0,1]$. Thus, an MPE is a PBE, where the agent's strategy $\boldsymbol{x}^\theta(p)$ is a best response to the evaluator's strategy $T(\Sigma)$, $T(\Sigma)$ is a best response to the public conjectured strategy $\tilde{\boldsymbol{x}}(p)$, and $\boldsymbol{x}^\theta(p) = \tilde{\boldsymbol{x}}^\theta(p)$. We also write (1) and (2) as functions of the initial belief p, and their value functions $V_\theta^T(p) = \sup_{\boldsymbol{x}^\theta} v_\theta^{T,\boldsymbol{x}^\theta}(p)$ and $U^{\tilde{\boldsymbol{x}}}(p) = \sup_T u^{T,\tilde{\boldsymbol{x}}}(p)$.

To ensure (DRIFT) with an initial condition $p_0 \in (0,1)$ admits a well-defined solution, we restrict $\tilde{\boldsymbol{x}}^\theta(p)$ to be admissible: $\tilde{\boldsymbol{x}}^\theta(p)$ is *admissible* if there exists a finite number of cutoffs q_i^θ with $0 \leq q_1^\theta < \cdots < q_n^\theta \leq 1$, such that it is Lipschitz-continuous on any interval $[0, q_1^\theta), \cdots, (q_i^\theta, q_{i+1}^\theta), \cdots, (q_n^\theta, 1]$ and it is left-continuous at any interior cutoff $q_i^\theta \in (0,1)$. For a given Markov $\tilde{\boldsymbol{x}}$, we rewrite $d^{\tilde{\boldsymbol{x}}}(p) := d(p, \tilde{\boldsymbol{x}}^G(p), \tilde{\boldsymbol{x}}^B(p))$ and $J^{\tilde{\boldsymbol{x}}}(p) := J(p, \tilde{\boldsymbol{x}}^G(p), \tilde{\boldsymbol{x}}^B(p))$. By restricting to admissible $\tilde{\boldsymbol{x}}^\theta(p)$ for both θ, $d^{\tilde{\boldsymbol{x}}}(p)$ inherits the same admissible property. This approach follows Board and Meyer-ter

[12]The definition of a Markov strategy at a belief p_t is immaterial if no history induces p_t.

Vehn (2013), who prove that there exists a unique solution to (DRIFT).[13]

Lastly, we say the evaluator's strategy is a *cutoff strategy* with $p \in (0,1)$, denoted by T_p, if it is a Markov strategy $T(\Sigma)$ with $\Sigma = [0,p]$, and the type θ agent's strategy is a *cutoff strategy* with $p^\theta \in (0,1)$ if it is a Markov strategy $\boldsymbol{x}^\theta(p) = \mathbb{1}_{p \in (p^\theta, 1)}$.

3 Preliminary analysis

The evaluator and two benchmarks – Censorship may be too costly and infeasible under certain circumstances due to institutional or technological constraints. Without censorship, the evaluator faces a two-armed bandit problem with exogenous Poisson news processes. We call this benchmark the No Censorship Benchmark (NCB). It serves as a welfare comparison to our main analysis where costly censorship is possible. Another benchmark—the Full Censorship Benchmark (FCB)—is that the agent censors all bad news so that only good news is ever revealed. If censorship is costless, censoring all bad news is trivially an equilibrium. From the evaluator's perspective, he then faces a two-armed bandit problem with only good news.

Thomas (2016) and Keller et al. (2005)[14] solve the evaluator's problem in the NCB and the FCB, respectively. In both benchmarks they show the evaluator's optimal policies are the same cutoff strategy T_{p^*} with $p^* := \rho\, m / (\rho\, h + \gamma(h-m))$. The evaluator retains the agent until p^* even when it is myopically suboptimal ($p^* < m/h$), indicating an option value of retaining the agent. But why the same cutoff? Consider the evaluator's decision at the belief margin when he is indifferent. Censorship affects his information but not his payoff. His decision depends on the probabilities of three events: (a) no news, (b) bad news, and (c) good news. He dismisses the agent if and only if (a) or (b) happens, as the belief drifts or jumps down below the indifference margin. His option value comes from (c); only good news changes his decision of dismissing the agent. While censorship changes the relative probabilities of (a) and (b), it does not change their total probability. Thus, censorship does not change his trade-off at the margin, so the optimal belief cutoff is the same. In fact, the cutoff

[13] Our assumption $\gamma > \beta^B$ implies $d^{\tilde{\boldsymbol{x}}}(p) < 0$, which corresponds to a special case in Board and Meyer-ter Vehn (2013) so that a unique local solution exists by the Picard-Lindelöf theorem and a unique global solution can be concatenated by local solutions.

[14] See Proposition 7 in Thomas (2016) and Proposition 3.1 in Keller et al. (2005).

remains the same in equilibrium in our main model as long as the belief drifts down in the absence of news and jumps down when bad news arrives.

Thomas (2016) and Keller et al. (2005) also characterize the evaluator's value functions, denoted by $U^{\mathbf{0}}(p)$ in the NCB and $U^{\mathbf{1}}(p)$ in the FCB. Both functions are continuously differentiable; they are equal to m for $p \in [0, p^*]$ and strictly increasing and strictly convex for $p \in [p^*, 1]$.[15] Lemma 1 summarizes the comparison of value functions under different censorship policies and its implications on strategies. All proofs are in the Appendix.

Lemma 1. *Fix any $\tilde{\boldsymbol{x}}$. (a) $U^{\tilde{\boldsymbol{x}}}(p) \geq U^{\mathbf{1}}(p)$ and it is strictly dominated to dismiss the agent when $p > p^*$; (b) assume $\beta^G = 0$, $U^{\tilde{\boldsymbol{x}}}(p) \leq U^{\mathbf{0}}(p)$ and it is strictly dominated to retain the agent when $p < p^*$.*

It is easy to see that the FCB is the worst information for the evaluator, because the information generated by it is a garbling of that generated by any conjectured strategy $\tilde{\boldsymbol{x}}$ à la Blackwell (1953); the FCB pools the no-news event with the event where the uncensored bad news (by $\tilde{\boldsymbol{x}}$) has arrived. In the proof, we show the evaluator's optimal policy in the FCB can be implemented by ignoring bad news, which is feasible for any strategy $\tilde{\boldsymbol{x}}$ and gives him the same payoff as in the FCB, regardless of $\tilde{\boldsymbol{x}}$. Thus, facing any strategy $\tilde{\boldsymbol{x}}$, the evaluator can do at least as well as in the FCB by simply ignoring bad news. Therefore, for any strategy $\tilde{\boldsymbol{x}}$, if the evaluator follows the optimal policy in the FCB and thus retains the agent at a belief $p > p^*$, he obtains a payoff higher than $U^{\mathbf{1}}(p) > m$, so dismissing the agent is strictly dominated.

One may attempt to show the NCB is the best information via a similar information garbling argument. This is true with conclusive bad news. As news is conclusive, it is optimal to dismiss the agent when bad news arrives and retain her when good news arrives. We prove that, using any policy that respects this rule, the evaluator obtains a higher payoff in the NCB than she does when facing any strategy $\tilde{\boldsymbol{x}}$, since bad news arrives earlier in the NCB and thus he dismisses the type B agent earlier. This logic applies to any strategy that retains the agent at a belief $p < p^*$. Since such a strategy gives the evaluator a payoff strictly less than m in the NCB, it gives him even less payoff when facing any other strategy $\tilde{\boldsymbol{x}}$; the strategy is strictly dominated.

[15] Smooth at p^* reflects the regularity of the stopping boundary. Convexity exhibits the value of information.

Censorship hurts the evaluator in models with conclusive bad news. When bad news is inconclusive, a strategy $\tilde{\boldsymbol{x}}$ pools a no-news event with a bad-news event from each type of agent according to $\tilde{\boldsymbol{x}}^\theta$. If $\tilde{\boldsymbol{x}}^G \neq \tilde{\boldsymbol{x}}^B$, the information generated by the NCB and $\tilde{\boldsymbol{x}}$ is not Blackwell-comparable; one is not sufficient for another. In Section 4.2, we will show that in such a case, censorship can improve the evaluator's information.

The agent – Turning to analysis of the agent, we start with some observations that hold in any MPE.[16] First, the agent never censors news if her reputation is degenerate ($p = 0$ or 1), as the belief remains unchanged and censorship is costly. Second, the agent is dismissed immediately at $p = 0$ and retained forever at $p = 1$, giving her a lower bound 0 and an upper bound w of her value function. Last, bad news never brings the belief $p < 1$ to 1 (i.e., $J^{\tilde{\boldsymbol{x}}}(p) < 1$). Otherwise, this means that only the type B censors news. However, revealing the news is her dominant strategy to achieve the highest continuation value and save the censorship cost. The public conjecture on censorship is correct in equilibrium, so this scenario cannot happen.

Fixing a public conjectured strategy $\tilde{\boldsymbol{x}}$ and the evaluator's strategy T_q with $q \in (0, 1)$, we analyze the agent's value of her reputation. Clearly, the type θ agent's value function $V_\theta^{T_q}(p) = 0$ for $p \leq q$, as (DRIFT) implies $T_q = 0$ with probability one for any $\tilde{\boldsymbol{x}}$. For $p > q$, we show that $V_\theta^{T_q}(p)$ satisfies the following properties.

Lemma 2. *Fix the evaluator's strategy T_q and the public conjectured strategy $\tilde{\boldsymbol{x}}$. (a) The agent's value function $V_\theta^{T_q}(p)$ is continuous in $p \in [0, 1)$; (b) there exists a unique $q^\theta \in (q, 1)$ such that $V_\theta^{T_q}(p) > \rho\, c$ if and only if $p > q^\theta$.*

The agent's value of her reputation $V_\theta^{T_q}(p)$ exhibits the single-crossing property: there is a unique cutoff belief $q^\theta \in (q, 1)$, her value function crosses the censorship cost from below at q^θ. Thus, comparing to the censorship cost, the agent's value of her reputation is higher if and only if her reputation is higher than the cutoff q^θ.

However, the agent's censorship incentive is not determined by the absolute reputation value $V_\theta^{T_q}(p)$, but by the relative reputation damage caused by bad news $\Delta_\theta^{T_q}(p) := V_\theta^{T_q}(p) - V_\theta^{T_q}(J^{\tilde{\boldsymbol{x}}}(p))$ (unless bad news is conclusive so they are the same). To see this,

[16] The type G agent's analysis is meaningful only when $\beta^G > 0$. Otherwise, she is a passive player.

we write her value function $V_\theta^{T_q}(p_t)$ by truncating (1) at the first news arrival:

$$\sup_{\boldsymbol{x}^\theta} \int_t^{T_q(\varnothing)} y^\theta(\nu;t) \left\{ \rho\, w + \gamma^\theta V_\theta^{T_q}(1) + \beta^\theta \left[x_\nu^\theta (\Delta_\theta^{T_q}(p_\nu) - \rho\, c) + V_\theta^{T_q}(J^{\tilde{\boldsymbol{x}}}(p_\nu)) \right] \right\} \mathrm{d}\nu,$$

where $\gamma^G := \gamma$, $\gamma^B := 0$, $y^\theta(\nu;t) := e^{-(\rho+\beta^\theta+\gamma^\theta)(\nu-t)}$ is the effective discount factor, and $T_q(\varnothing)$ is the time of dismissal when no news arrives.[17] Clearly, the following strategy that maximizes the integrand is an optimal strategy: it is optimal to censor bad news if the reputation damage caused by it exceeds the censorship cost.[18]

$$\boldsymbol{x}^\theta(p) = \begin{cases} 1, & \text{if } \Delta_\theta^{T_q}(p) > \rho\, c, \\ 0, & \text{if } \Delta_\theta^{T_q}(p) < \rho\, c. \end{cases} \tag{3}$$

Moreover, in any MPE where the evaluator uses a cutoff strategy, any optimal strategy of the agent must satisfy (3). This is partially driven by the continuity assumption from admissibility so that a pointwise optimization is necessary. Thus, we have the following result.

Lemma 3. *(T_q, $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}, \boldsymbol{p}$) is an MPE if and only if $\boldsymbol{x} = \tilde{\boldsymbol{x}}$, T_q is a best response to $\tilde{\boldsymbol{x}}$, (3) holds for all $p > q$, and $\boldsymbol{p}$ is updated according to (JUMP) and (DRIFT).*

Note that the agent's strategy $\boldsymbol{x}^\theta(p)$ has no impact on her payoff for $p \leq q$. Without loss of generality, we assume $\boldsymbol{x}^\theta(p) = 0$ for $p \leq q$ in any MPE where the evaluator uses a strategy T_q so that (3) also holds for $p \leq q$ (i.e., $\Delta_\theta^{T_q}(p) = 0 < \rho\, c$).

Last, we consider the agent's payoff function $\Pi_\theta(p)$ under the NCB. This serves as the welfare comparison and provides the key variables p^G and p^B. In the NCB, the public conjectured strategy $\tilde{\boldsymbol{x}}$ is zero and the evaluator's strategy is T_{p^*}. We show that (a) the type G has a higher payoff as she can succeed and has less bad news and (b) both types have a higher payoff for a higher reputation as they can stay longer in the job, and (c) find the beliefs at which their payoffs equal the censorship cost.

Lemma 4. *(a) $\Pi_G(p) > \Pi_B(p)$ for $p \in (p^*, 1)$; (b) $\Pi_\theta(p)$ is continuous in $p < 1$ and strictly increasing in $p \in (p^*, 1)$; (c) there is a unique $p^\theta \in (p^*, 1)$ such that*

[17] We use $\varnothing$ to denote the history when no news arrives.

[18] If $\Delta_\theta^{T_q}(p) = \rho\, c$, $\boldsymbol{x}^\theta(p) \in [0,1]$ is only regulated by the admissibility of the conjectured strategy in equilibrium.

$\Pi_\theta(p^\theta) = \rho\ c$, *where* $p^G < p^B$.

4 Main results

4.1 Conclusive bad news

This section assumes conclusive bad news ($\beta^G = 0$). The type G agent is a passive player and revealed bad news concludes that the agent's type is B. The analysis is simplified by two facts related to this assumption. First, the NCB gives the evaluator the best information. Lemma 1 pins down the unique candidate equilibrium strategy T_{p^*} of the evaluator. Second, the agent's reputation is zero after bad news is revealed. The reputation damage caused by bad news $\Delta_B^{T_{p^*}}$ is her reputation value $V_B^{T_{p^*}}$. Thus, the censorship incentive is determined by the reputation value and the censorship cost (Lemma 3). Since the reputation value exhibits the single-crossing property (Lemma 2), the agent's equilibrium strategy must be a cutoff strategy.

Moreover, the type B agent's cutoff strategy is uniquely determined by p^B defined in Lemma 4. At the belief where the type B agent stops censoring (call it q^B), her reputation value must be equal to the censorship cost (Lemma 2). In the MPE, the public conjecture is that the agent stops censoring when the belief is below q^B, and that is her optimal strategy as the conjecture is correct in equilibrium. Thus, below q^B, her value function $V_B^{T_{p^*}}$ actually coincides with her payoff function Π_B in the NCB. Lemma 4 shows that $\Pi_B(p)$ is monotone and $\Pi_B(p) = \rho\ c$ only at p^B. Thus, if the public conjecture is that the agent stops censoring at a point $q^B > p^B$ before the belief reaches p^B, at q^B her reputation value is actually higher than the censoring cost so she should not stop censoring. If the public conjecture is that the agent stops censoring at a point $q^B < p^B$ after the belief reaches p^B, at q^B her reputation value is actually lower than the censoring cost so she should stop censoring earlier. Thus, she must stop censoring at p^B. The result is formalized below.[19]

Proposition 1. *Assume* $\beta^G = 0$. *There exists a unique MPE. In the MPE, the evaluator's strategy is* T_{p^*} *and the type* B *agent's strategy is* $\boldsymbol{x}^B = \mathbb{1}_{p \in (p^B, 1)}$.

Figure (1a) illustrates the public belief evolution in the absence of news in the MPE

[19] One can show that this is the unique PBE, as the belief always drifts down and the trajectory $t \mapsto p_t(\varnothing)$ induces a homeomorphism between some time interval $[0, \bar{t}]$ and $[p_0, \bar{p}]$ for some $\bar{p}$.

and the NCB. In both cases, the evaluator dismisses the agent whenever the belief is below p^*. The censorship policy has two phases in the MPE. At the beginning when the public belief is higher than p^B, the agent censors all bad news. We call this the Full-Censorship period, the length of which is denoted by s_1. When the belief falls below p^B, the agent stops censoring. We call this the No-Censorship period, the length of which is denoted by s_2. If no bad news arrives, the agent can stay in the job until s_2 at which point the belief reaches the dismissal threshold p^*. But if bad news arrives before s_2, the agent is also dismissed. In the NCB, the maximal duration the agent can stay in her job is $\bar{s}$.

We notice two things in Figure (1a). First, the public belief drifts down faster in the Full-Censorship period. In fact, the greater the intensity of censorship, the faster the public belief declines in the absence of news, i.e., $|d(p, \tilde{x}^G, \tilde{x}^B)|$ is increasing $\tilde{x}^B$. Although the evaluator's strategy remains the cutoff strategy T_{p^*}, the drifting rate determines how long it takes for the belief to reach p^*. Thus, it takes less time for the belief to reach p^* in the MPE than in the NCB. Second, the agent's cutoff p^B is determined by the indifference condition that equalizes the reputation value and the censorship cost. Thus, the larger the cost, the earlier the agent stops censoring, and thus the higher the p^B.

Welfare: Equilibrium versus NCB – We now compare the welfare consequences of censorship between the MPE and the NCB. In particular, the welfare is measured by the expected payoff. Since the type B agent's equilibrium strategy is to stop censoring when $p \leq p^B$, nothing is different between these two scenarios if the prior belief is low. Thus, we assume that we start with a high prior belief $p_0 > p^B$.

The evaluator is worse off in the MPE, as the NCB is the best information that the evaluator can have (Lemma 1). The fact that the public belief drifts down faster in the MPE implies that the type G agent is also worse off, since, in order to stay in her job, she now has to succeed within a shorter period of time before the belief drifts below the evaluator's dismissal cutoff belief. Only the type B agent's comparison is non-trivial. The results are given below.

Proposition 2. *Assume $\beta^G = 0$ and $p_0 > p^*$. (a) The evaluator and the type G agent have lower payoffs in the MPE than in the NCB. (b) There exist $\mathfrak{c}_1, \mathfrak{c}_2 \in [0, \bar{c})$, $\mathfrak{c}_1 < \mathfrak{c}_2$, such that the type B agent has a higher (resp. lower) payoff in the MPE than*

in the NCB if $c \in (0, \mathfrak{c}_1)$ *(resp. if* $c \in (\mathfrak{c}_1, \mathfrak{c}_2)$*), and has the same payoff if* $c \geq \mathfrak{c}_2$.

The type B agent may also be worse off in the MPE. Censorship gives her a higher expected flow payoff in the Full-Censorship period as she censors and survives bad news. It is the positive effect for her, due to censorship. The longer the Full-Censorship period is, the larger the benefit she enjoys. However, if bad news does not arrive, the agent cannot prove that. Unable to prove or commit that she was not censoring makes the evaluator skeptical about the absence of news—the belief drifts down faster, regardless of what she actually does. When the Full-Censorship period is short, the positive effect is dominated by the negative effect so that she is also worse off in the MPE. If possible, the agent would want to commit to less censorship.

We now show the implications of the above on the censoring cost. The cost reflects the difficulty in suppressing evidence and the strength of anti-censorship institutions. This determines when the agent is willing to give up censorship and the length of the Full-Censorship period. The higher the cost, the shorter the Full-Censorship period. If the cost is high enough, the Full-Censorship period disappears. Then, the type B agent has the same payoff in the MPE and the NCB. If the cost is low, it means that the Full-Censorship period is long. If it is long enough, the type B agent is better off in the MPE. However, if the cost is moderate, the Full-Censorship period exists but is short, and the type B agent is worse off in the MPE. In fact, we can show that her payoff in the MPE is single-peaked at some point between $\mathfrak{c}_1$ and $\mathfrak{c}_2$.

According to the above results, the right of erasure seems a bad idea. However, our result depends crucially on the assumption that bad news is indeed conclusive. GDPR states that the right of erasure does not apply to some circumstances (Article 23), mostly for public interest reasons. Policy makers apparently have taken the welfare implications into consideration. In the next section, we show that allowing people to remove search results can be a good idea if bad news is inconclusive.

4.2 Inconclusive bad news

This section assumes inconclusive bad news ($\beta^G > 0$). In the NCB, bad news as a signal is imperfect. After bad news arrives at a belief $p \in (0, 1)$, the evaluator updates his belief, in the natural direction, to $j(p) = J(p, 0, 0) \in (0, p)$; it is more likely but not certain that the agent's type is B. Decisions based on bad news may cause both

type I and type II errors.

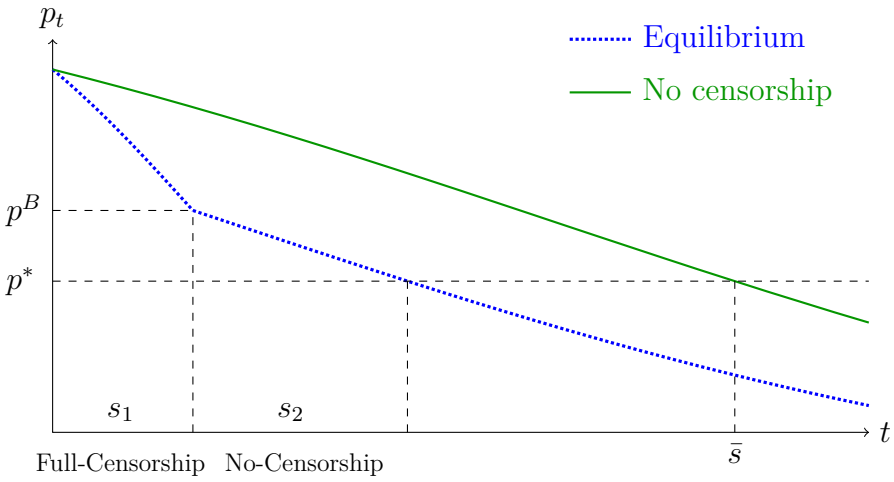
When censorship is possible, inconclusive bad news complicates the analysis but brings new insights. The reputation damage caused by bad news determines the censorship incentive, while the public conjecture on censorship determines how detrimental bad news is to reputation and the corresponding damage; the conjecture agrees with the equilibrium strategies. Unlike Section 4.1, the reputation damage caused by bad news is, in general, not the reputation value. The single-crossing property of the reputation value gives us the cutoff structure of the equilibrium in Section 4.1. However, the reputation damage may not have this property. In fact, it is hard to obtain a general property for the reputation damage, as different conjectures on censorship may be sustained by corresponding strategies, which results in multiple consistent interpretations of bad news and the resulting reputation damage.

We exploit the single-crossing property of the reputation value and construct an equilibrium similar to that in Section 4.1, and show that it is the unique pure strategy MPE if bad news is sufficiently bad. We now state the result formally.

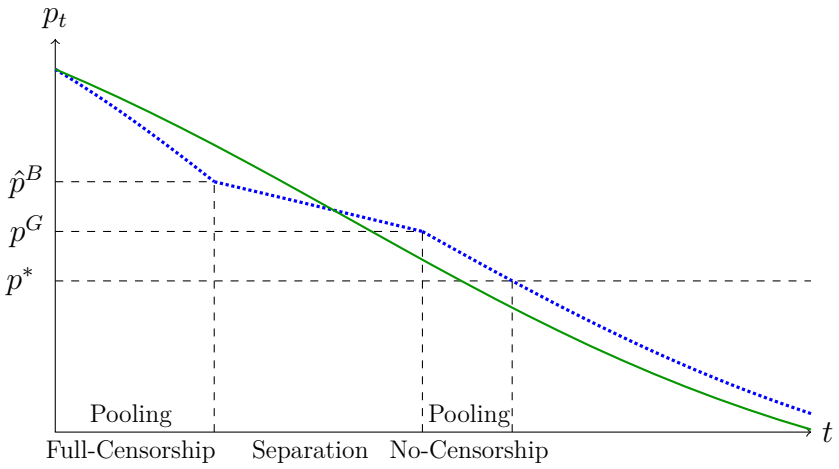
Proposition 3. *Assume $\beta^G > 0$. There exists an MPE, where the evaluator's strategy is T_{p^*}, the type G agent's strategy is $\boldsymbol{x}^G = \mathbb{1}_{p\in(p^G,1)}$, and the type B agent's strategy is $\boldsymbol{x}^B = \mathbb{1}_{p\in(\hat{p}^B,1)}$. Moreover, if $j(p_0) \leq p^*$, it is the unique pure strategy MPE.*

Recall that p^G is defined by $\Pi_G(p^G) = \rho\ c$ in Lemma 4. $\hat{p}^B$ is defined in an auxiliary problem in Lemma 5, where $\hat{\Pi}_B(\hat{p}^B) = \rho\ c$. Figure (1b) depicts the belief evolution in the absence of news in the MPE and the NCB. The evaluator's strategy is the same cutoff strategy T_{p^*} as before. The belief p^G (resp. $\hat{p}^B$) is the cutoff belief at which the type G (resp. B) agent stops censoring. There are three phases in equilibrium. When the public belief is higher than $\hat{p}^B$, both types of agent censor bad news for sure. We call this the Pooling-Full-Censorship period. When the public belief is in between p^G and $\hat{p}^B$, the type G agent censors bad news, but the type B agent does not. We call this the Separation period. When the public belief is lower than p^G, both types stop censoring. We call this the Pooling-No-Censorship period.

We emphasize three features of this MPE. First, in the Separation period, bad news endogenously becomes conclusive as it can only come from the type B agent. This will give us the main welfare result for the evaluator. In the Pooling-Full-Censorship



(a) Conclusive bad news ($\beta^G = 0$)



(b) Inconclusive bad news ($\beta^G = 2$)

Figure 1: The belief drifting process $(\gamma, \beta^B, p_0, p^*, w, c, \rho) = (5, 3.5, 0.75, 0.4, 10, 1, 6)$

period, bad news does not exist. Without loss of generality, it is assumed that off-path bad news brings the public belief to 0 (see footnote 11). Moreover, in both periods, the reputation damage caused by bad news $\Delta_\theta^{T_{p^*}}$ is the reputation value $V_\theta^{T_{p^*}}$, as the reputation is completely destroyed after bad news. This allows us to use the familiar properties of the value function to analyze the censorship incentive. In the Pooling-No-Censorship period, no one censors, so bad news remains inconclusive. At a belief p, the agent is dismissed immediately after bad news if and only if the belief after bad news $j(p)$ is below p^*. If $j(p^G) \leq p^*$ (or β^G is low),[20] the agent is always

[20]The working paper Sun (2021) shows that $j(p^G) \leq p^*$ if and only if $\beta^G \leq \bar{\beta}$, and $j(\hat{p}^B) \leq p^*$ if

dismissed right after bad news in the Pooling-No-Censorship period, since the belief in this period is bounded above by p^G and $j(p)$ is monotone. However, if $j(p^G) > p^*$ (or β^G is high), there is a belief threshold, $j^{-1}(p^*)$, in between p^* and p^G such that the agent is dismissed right after bad news only when the belief is below $j^{-1}(p^*)$; she survives a piece of bad news when the belief is above $j^{-1}(p^*)$ but below p^G.

Second, in the MPE, the belief drifting rate is fastest in the Pooling-Full-Censorship period and slowest in the Separation period. In the Pooling-No-Censorship period and in the NCB, it is intermediate. This is because the drifting rate is increasing in the censorship intensity of the type B agent, but decreasing in that of the type G agent, i.e., $|d(p, \tilde{x}^G, \tilde{x}^B)|$ is increasing in $\tilde{x}^B$ and decreasing in $\tilde{x}^G$. We now provide some intuition. At a belief p_t, consider all possible events that may happen in the time interval $[t, t + \mathrm{d}t)$. Take the NCB as a baseline. When good news arrives, p_t jumps up to 1; when bad news arrives, p_t jumps down to $j(p_t)$; in the absence of news, p_t drifts down to $p_t + d^{\tilde{x}}(p_t)\,\mathrm{d}t$. Beliefs are a martingale. The expectation of $p_{t+\mathrm{d}t}$—an average of 1, $j(p_t)$ and $p_t + d^{\tilde{x}}(p_t)\,\mathrm{d}t$ weighted by their probabilities—remains p_t. In the Pooling-Full-Censorship period, both types of agent censor bad news, so it never arrives. To compensate, $p_t + d^{\tilde{x}}(p_t)\,\mathrm{d}t$ must become smaller so the drifting rate is faster. In the Separation period, bad news becomes conclusive, so the belief jumps down to 0 instead of $j(p_t)$. To compensate, $p_t + d^{\tilde{x}}(p_t)\,\mathrm{d}t$ must become larger so the drifting rate is slower. The slower drifting rate in the Separation period is important for our main welfare result for the agent.

Third, both types of agent use a cutoff strategy, and it is the type B who gives up censoring first ($\hat{p}^B > p^G$). The cutoff beliefs $\hat{p}^B$ and p^G are determined by the indifference condition that equalizes the agent's value of her reputation and the censorship cost for each type. The reason they can be used to construct the cutoff structure of the censorship strategy is as follows. Above the cutoff beliefs, the reputation value and the reputation damage caused by bad news coincide because of the first feature of this MPE. Thus, both are larger than the cost and censorship is optimal. Below the cutoff beliefs, the reputation damage caused by bad news is smaller than the reputation value, and the latter is smaller than the censorship cost. Thus, not censoring is optimal. Hence, not only the reputation value but also the reputation damage single

and only if $\beta^G \leq \underline{\beta}$, where $\underline{\beta}, \bar{\beta} \in (0, \beta^B)$ are independent of p_0 and $\underline{\beta} < \bar{\beta}$.

crosses the censorship cost from blow.[21] Thus, the agent uses a cutoff strategy. The cutoff belief for the type G agent is smaller than that for the type B because the type G agent has a higher value of her reputation. First, the type G agent has a chance to succeed in her project, so she stays longer in her job on average. Second, although the censorship cost is the same for both types, the fact that bad news arises more often for the type B agent makes it more costly for her to use the same censorship policy. Thus, maintaining reputation is both more beneficial and less costly for the type G agent.[22]

The above equilibrium may not be the unique MPE. It is the equilibrium where bad news is most heavily censored. In that MPE, the type G agent censors bad news whenever the belief is above p^G. We can show that censorship does not exist if the belief is below p^G in any pure strategy MPE. However, if the belief is high, whether censorship exists depends on the public conjecture of the censorship strategy, since it determines how much reputation value is left if the agent chooses not to censor bad news. Multiple conjectures may be consistent. Fixing the type G agent's cutoff strategy with p^G, we can further show that the type B agent does not censor bad news if the public belief is below $\hat{p}^B$ in any pure strategy MPE, while in our MPE the type B agent censors bad news whenever the belief is above $\hat{p}^B$. In this sense, we construct an equilibrium that supports censorship as much as possible.

To see other possibilities, suppose that the bad news arrival rate β^G of the type G agent is high and close to the arrival rate β^B of the type B agent. Thus, if the reputation is high, a consistent conjecture on censorship may well be that no one censors. This is because the reputation and its value do not change much when bad news is revealed, if β^G is close to β^B and if no one is expected to censor news. Bad news is a weak signal. We can show that in any pure strategy MPE, there are belief cutoffs $\tilde{p}^\theta \in (p^*, 1)$, $\tilde{p}^G < \tilde{p}^B$, such that the agent's strategy satisfies the following: (1) $\boldsymbol{x}^B = \boldsymbol{x}^G = 0$ for $p \in (p^*, \tilde{p}^G)$; (2) $\boldsymbol{x}^B = 0 \leq \boldsymbol{x}^G \in \{0, 1\}$ for

[21] Note that the reputation value is continuous in the belief, but the reputation damage may be discontinuous at the cutoff belief.

[22] The two types have different incentives to maintain reputation, but they value the job equally if maintaining reputation is not needed. If we allow the agent to signal herself by burning money (Milgrom and Roberts (1986)) at the beginning of the game, the unique separating equilibrium is that the type G agent burns money that equals w—the value of staying in the job forever without censoring news. In this equilibrium, both types have zero payoffs and are indifferent between burning money and not. This equilibrium does not exist if the agent is slightly financially constrained.

$p \in (\tilde{p}^G, \tilde{p}^B)$; (3) $\boldsymbol{x}^B = \boldsymbol{x}^G \in \{0,1\}$ for $p \in (\tilde{p}^B, 1)$. The cutoff $\tilde{p}^\theta$ is defined by the belief at which the agent's value function equals the censorship cost, $V_\theta^{T_{p^*}}(\tilde{p}^\theta) = \rho\ c$ (Lemma 2), and $\tilde{p}^G < \tilde{p}^B$ follows the fact that $V_B^{T_{p^*}}(p) < V_G^{T_{p^*}}(p)$ for $p \in (p^*, 1)$.[23] Moreover, $\tilde{p}^G = p^G$ as $V_G^{T_{p^*}}(p^G) = \Pi_G(p^G) = \rho\ c$. When $p < \tilde{p}^\theta$, $\boldsymbol{x}^\theta(p) = 0$ as $\Delta_\theta^{T_{p^*}}(p) \leq V_\theta^{T_{p^*}}(p) < \rho\ c$. When $p \in (\tilde{p}^B, 1)$, $\boldsymbol{x}^B(p) \neq \boldsymbol{x}^G(p)$ cannot be supported in a pure strategy MPE. First, $\boldsymbol{x}^B(p) = 1 > \boldsymbol{x}^G(p) = 0$ is ruled out, as in that case $J^{\tilde{\boldsymbol{x}}}(p) = 1$. Second, $\boldsymbol{x}^B(p) = 0 < \boldsymbol{x}^G(p) = 1$ is impossible, since in that case $J^{\tilde{\boldsymbol{x}}}(p) = 0$ and $\Delta_B^{T_{p^*}}(p) = V_B^{T_{p^*}}(p) > \rho\ c$. The MPE in Proposition 3 captures the situation where censorship is expected to happen very often. This may happen when customers care greatly about product reviews and believe the seller will do her best to maintain her reputation, including suppressing bad reviews as much as possible.[24]

When $\beta^G = \beta^B$, all the above possibilities, including the strategies in Proposition 3, can be supported in equilibrium. We now construct a pure strategy MPE from an arbitrary admissible pure Markov strategy $\boldsymbol{x}(p)$, where the type G's equilibrium strategy is $\boldsymbol{x}^G(p) := \boldsymbol{x}(p)\, \mathbb{1}_{p \in (p^G, 1)}$. In any pure strategy MPE, no one censors at low beliefs $p \leq p^G$. We first consider the type B's strategy $\boldsymbol{x}_0^B(p) := 0$ and the type G's strategy $\boldsymbol{x}_0^G(p) := \boldsymbol{x}^G(p)$. Given the conjecture $\tilde{\boldsymbol{x}}_0^\theta(p) := \boldsymbol{x}_0^\theta(p)$, we can find a $\tilde{p}_{\boldsymbol{x}}^B \in (p^G, 1)$ such that $V_B^{T_{p^*}}(\tilde{p}_{\boldsymbol{x}}^B) = \rho\ c$ (Lemma 2). Note that $\tilde{p}_{\boldsymbol{x}}^B$ does not depend on $\tilde{\boldsymbol{x}}_0^\theta(p)$ for $p \in (\tilde{p}_{\boldsymbol{x}}^B, 1)$, because the belief only goes down or goes up to 1, and thus the value function $V_B^{T_{p^*}}(p')$ at a belief $p' < 1$ does not depend on $\tilde{\boldsymbol{x}}_0^\theta(p)$ for $p \in (p', 1)$. We now keep the type G's strategy $\boldsymbol{x}^G(p)$ unchanged and change the type B's strategy to $\boldsymbol{x}^B(p) := \boldsymbol{x}(p)\, \mathbb{1}_{p \in (\tilde{p}_{\boldsymbol{x}}^B, 1)}$. According to our argument, given the new conjecture $\tilde{\boldsymbol{x}}^\theta(p) := \boldsymbol{x}^\theta(p)$, we still have $V_B^{T_{p^*}}(\tilde{p}_{\boldsymbol{x}}^B) = \rho\ c$. We can verify that $\boldsymbol{x}^\theta(p)$, $\tilde{\boldsymbol{x}}^\theta(p)$ and the evaluator's strategy T_{p^*} constitute an equilibrium. When the belief is above $\tilde{p}_{\boldsymbol{x}}^B$, censorship from both types can be supported in equilibrium if it is expected, since "bad news" can be expected to be a strong signal and the reputation damage is high. Similarly, censorship from the type G, but not from the type B, can be supported in

[23] In any pure strategy MPE, the evaluator's strategy is T_{p^*} (Lemma 6) and the type G's value function is higher. The latter is due to (1) $V_B^{T_{p^*}}(p) < V_G^{T_{p^*}}(p)$ for some interval $p \in (p^*, \tilde{p})$ because of Lemma 4 and $V_\theta^{T_{p^*}}(p) = \Pi_\theta(p)$ for $p \in (p^*, \tilde{p})$, and (2) at a belief $p \in (p^*, 1)$, the type G can mimic the type B's equilibrium strategy and obtain a higher payoff, if $V_B^{T_{p^*}}(p') \leq V_G^{T_{p^*}}(p')$ for $p' < p$. This result still holds when $\beta^G = \beta^B$, since only the type G can succeed.

[24] 94% of consumers said a bad review has convinced them to avoid a business, according to Online Reviews Statistics and Trends by ReviewTrackers (https://www.reviewtrackers.com/reports/online-reviews-survey/).

equilibrium at intermediate beliefs $p \in (p^G, \tilde{p}^B_{\boldsymbol{x}})$, since the censorship cost is higher than the type B's value and lower than that of type G. But if censorship is not expected, "bad news" is uninformative ($j(p) = p$ as $\beta^G = \beta^B$) and causes no damage. In particular, there is an MPE without any censorship.

However, if bad news is a strong signal, i.e., $j(p_0) \leq p^*$ (or β^G is small),[25] we have a unique pure strategy MPE as in Proposition 3. The above assumption implies that the agent will be dismissed right after bad news, for all possible conjectured strategies in a pure strategy MPE. Thus, the reputation damage caused by bad news can always be represented by the reputation value. Using the same logic in Section 4.1, the censorship strategy must be a cutoff strategy and the cutoff belief is also unique.

Welfare: Equilibrium versus NCB – We now compare the welfare consequence of censorship between the MPE of Proposition 3 and the NCB, under the assumption that the game starts from the Separation period. We find the evaluator has a higher payoff in the MPE. If we additionally assume $j(p_0) \leq p^*$, we find that both types of agent also have higher payoffs in the MPE.

Proposition 4. *Assume $\beta^G > 0$ and $p_0 \in (p^G, \hat{p}^B]$. (a) The evaluator has a higher payoff in the MPE of Proposition 3 than in the NCB. (b) If $j(p_0) \leq p^*$, both types of agent have higher payoffs in the MPE of Proposition 3 than in the NCB.*

Note that the set $(p^G, \hat{p}^B]$ is non-empty (Lemma 5), and the welfare comparison is strict. Thus, the above results still hold if the prior p_0 is slightly above $\hat{p}^B$ due to the continuity of the value functions. We restrict attention to $p_0 \in (p^G, \hat{p}^B]$ to make a sharp exposition on the underlying economic forces. For the result of the agent, we also assume $j(p_0) \leq p^*$. The set of priors $P_0 := (p^G, \hat{p}^B] \cap (0, j^{-1}(p^*)]$ that satisfies both assumptions is non-empty if and only if $j(p^G) < p^*$. The latter is equivalent to $\beta^G < \bar{\beta}$, where $\bar{\beta} \in (0, \beta^B)$ is independent of the prior (see footnote 20). Thus, the set of priors required for the second result is non-empty if and only if $\beta^G < \bar{\beta}$. Moreover, $P_0 = (p^G, \hat{p}^B]$ if and only if $j(\hat{p}^B) \leq p^*$. The latter is equivalent to $\beta^G \leq \underline{\beta}$, where $\underline{\beta} \in (0, \bar{\beta})$ is independent of the prior. Thus, if $\beta^G \leq \underline{\beta}$, the assumption $j(p_0) \leq p^*$ is automatically satisfied for any $p_0 \in (p^G, \hat{p}^B]$ and can be dropped.

[25] Equivalently, $\beta^G \leq \beta^B \Omega(p_0)/\Omega(p^*)$, where $\Omega(p) := (1-p)/p$ and p^* is independent of β^G and β^B. Under this assumption, p_0 can be above or below $\hat{p}^B$; the set $(\hat{p}^B, j^{-1}(p^*)]$ is non-empty if $\beta^G < \underline{\beta}$ (see footnote 20).

We now provide some intuition for the results. As mentioned in the first feature of the MPE, in the Separation period, revealed bad news indicates that the agent is of type B. This improves the evaluator's information quality. During this period, he will never wrongly dismiss a type G agent, and the agent he ever dismissed must be a type B agent. The endogenously conclusive signal helps him avoid making type I and type II errors based on bad news. This is why his payoff is higher in the MPE when the game starts from the Separation period (i.e., $p_0 \in (p^G, \hat{p}^B]$).

As explained in the second feature of the MPE, the public belief drifts down slower in the Separation period than in the NCB. Since the dismissal threshold belief is the same in both situations, in equilibrium the agent stays longer in her job in the absence of news than she does in the NCB. For the type B agent, when the game starts from the Separation period, she does not censor bad news in both the MPE and the NCB. The assumption $j(p_0) \leq p^*$ ensures that the posterior belief is driven below the evaluator's dismissal threshold after revealed bad news even in the NCB. Thus, the revealed bad news leads to the dismissal of the agent in both situations. Thus, her payoff is the same in both situations when bad news arrives. However, if bad news does not arrive, she stays longer in her job in the MPE than in the NCB. Thus, the type B agent is strictly better off in equilibrium. The type G agent is also strictly better off in equilibrium. First, censoring bad news in the Separation period gives her a higher expected flow payoff in the MPE. We have seen the same effect when bad news is conclusive. In addition, she also stays longer in her job in the MPE. This effect is reversed when bad news is conclusive, since then the belief drifts down faster, not slower, in the absence of news. Therefore, both effects imply that the type G agent has a strictly higher payoff in the MPE than in the NCB.

Although our results extend if the prior p_0 is slightly above $\hat{p}^B$, the welfare comparison can go either way when p_0 is far above $\hat{p}^B$. When the game starts from the Pooling-Full-Censorship period, the evaluator never observes bad news in the MPE. Compared with the NCB, he first has worse information in the Pooling-Full-Censorship period and then better information in the Separation period. The welfare comparison for him depends on the net effect from the earlier inferior information and the later superior information. With a higher prior p_0, greater disadvantage arising from inferior information is accumulated. It is possible that the earlier inferior information dominates when p_0 is far above $\hat{p}^B$, so the evaluator has a lower payoff in the MPE than

in the NCB. Worse information in the Pooling-Full-Censorship period also implies that the evaluator's belief drifts down faster than in the NCB. This faster decrease in reputation hurts the agent. When p_0 is far above $\hat{p}^B$, the cost of a faster decrease in reputation in the Pooling-Full-Censorship period may outweigh the benefit of a slower decrease in reputation during the Separation period. In this case, both types of agent have lower payoffs in the MPE than in the NCB.

Coming back to the right of erasure, what is the implication if censorship is made easier? Reducing the censorship cost increases the agent's censorship incentive. If the cost is relatively high, neither type censors bad news. However, with a lower cost, two types with an intermediate reputation may separate. Our result implies that making it easier to remove search results may improve welfare when the relevant party has an intermediate reputation, especially if it used to be too difficult to do. From this viewpoint, the right of erasure and Reverse SEO seem to be a good idea, especially in cases where individuals and small businesses have not yet established high reputations.

5 Discussion: the bad news model

We have studied the "good news" model ($\gamma > \beta^B$), in which the reputation declines in the absence of news, regardless of the public conjectured strategy. This implies that censorship does not affect the evaluator's experimentation strategy so his dismissal threshold belief p^* remains unchanged.

Sun (2021) studies the "bad news" model ($\gamma < \beta^B$), in which the direction of the drift of beliefs depends on the public conjectured strategy. The reputation drifts up in the NCB but can drift down in equilibrium, as the consistent conjecture is that the agent censors news until the dismissal threshold.[26] This changes the evaluator's trade-off at the belief margin when he is indifferent. When the reputation drifts up, the "default" action is to retain the agent as she will become more optimistic. However, when the reputation drifts down due to censorship, the "default" action is to dismiss the agent as she will become more pessimistic. Sun (2021) shows that, with conclusive bad news, the dismissal threshold becomes higher in equilibrium than in the NCB.

[26]The dismissal threshold acts as a reflecting barrier on the reputation process as both players use mixed strategies at the threshold.

With inconclusive bad news, it becomes higher if the censorship cost is low but lower if the cost is intermediate. Thus, censorship affects the evaluator's experimentation strategy only in the "bad news" model.

Nevertheless, our welfare results still hold. With conclusive bad news, all parties can be worse off due to censorship. With inconclusive bad news, the type G agent censors news more aggressively than the type B agent. Separation may occur if the censorship cost is intermediate, in which case the evaluator is better off. However, if the cost is small, both types censor too much bad news, limiting the information available to the evaluator to the point that the evaluator essentially has the same information as in the FCB. This lack of information hurts the evaluator.

Appendices

A Proofs for Section 3

Proof of Lemma 1. (a) In the FCB, bad news is never revealed and thus the optimal strategy T_{p^*} can be implemented as follows: fixing an initial belief $p > p^*$, the evaluator dismisses the agent at time $\tau^*(p) > 0$ if no good news is revealed by $\tau^*(p)$ and retains her forever otherwise, where $\tau^*(p) := T_{p^*}(\varnothing)$ is the time it takes for the belief to drift from p to p^* in the absence of news. We also use $\tau^*(p)$ to denote this policy.

Thus, $U^{\mathbf{1}}(p) := u^{T_{p^*},\mathbf{1}}(p) = u^{\tau^*(p),\mathbf{1}}(p)$. Since $\tau^*(p)$ is a feasible policy for an arbitrary $\tilde{\boldsymbol{x}}$, we have $U^{\tilde{\boldsymbol{x}}}(p) := \sup_T u^{T,\tilde{\boldsymbol{x}}}(p) \geq u^{\tau^*(p),\tilde{\boldsymbol{x}}}(p)$. Moreover, fixing the policy $\tau^*(p)$, $u^{\tau^*(p),\tilde{\boldsymbol{x}}}(p)$ is independent of $\tilde{\boldsymbol{x}}$. This is because $\tau^*(p)$ simply ignores the bad news, which is independent of the arrival of good news and payoff-irrelevant to the evaluator.

Thus, $U^{\tilde{\boldsymbol{x}}}(p) \geq u^{\tau^*(p),\tilde{\boldsymbol{x}}}(p) = u^{\tau^*(p),\mathbf{1}}(p) = U^{\mathbf{1}}(p)$. For any $\tilde{\boldsymbol{x}}$, firing the agent at a belief $p > p^*$ gives the evaluator a continuation payoff of m, but following the policy $\tau^*(p)$ the continuation payoff is $U^{\mathbf{1}}(p) > m$. Thus, it is strictly dominated to dismiss the agent when $p > p^*$.

(b) Assume $\beta^G = 0$ so that $\tilde{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}^B\}$. Both good and bad news are conclusive.

Consider the set $\mathbb{T}$ of the evaluator's strategies (i.e., stopping times) with the following restrictions: firing the agent immediately when bad news is revealed, and retaining the agent forever when good news is revealed. For $T \in \mathbb{T}$, we use $T(\varnothing)$ to denote the time of dismissal when no news arrives. Note that any $T \in \mathbb{T}$ is a feasible strategy for any agent's strategy $\tilde{\boldsymbol{x}}$. As both types of news are conclusive, we can focus on the set $\mathbb{T}$ of the evaluator's strategies. That is, $U^{\tilde{\boldsymbol{x}}}(p) = \sup_{T\in\mathbb{T}} u^{T,\tilde{\boldsymbol{x}}}(p)$.

Given the agent's strategy $\tilde{\boldsymbol{x}}$, let $\tau_G(\tilde{\boldsymbol{x}})$ (resp. $\tau_B(\tilde{\boldsymbol{x}})$) be the random time at which the first revealed good (resp. bad) news arrives conditional on the agent's type being G (resp. B). By definition, for any $\tilde{\boldsymbol{x}}$, $\tau_G(\tilde{\boldsymbol{x}}) = \tau_G(\mathbf{0})$ and $\tau_B(\tilde{\boldsymbol{x}}) \geq \tau_B(\mathbf{0})$.

Thus, for a given strategy $\tilde{\boldsymbol{x}}$ of the agent, an evaluator's strategy $T \in \mathbb{T}$ dismisses the type θ agent at random time $T_\theta(\tilde{\boldsymbol{x}})$, where $T_B(\tilde{\boldsymbol{x}}) = \min\{T(\varnothing), \tau_B(\tilde{\boldsymbol{x}})\}$ and

$$T_G(\tilde{\boldsymbol{x}}) = \begin{cases} T(\varnothing), & \text{if } \ \tau_G(\tilde{\boldsymbol{x}}) \geq T(\varnothing), \\ \infty, & \text{if } \tau_G(\tilde{\boldsymbol{x}}) < T(\varnothing). \end{cases}$$

Thus, for any $T \in \mathbb{T}$, $T_G(\tilde{\boldsymbol{x}}) = T_G(\mathbf{0})$ and $T_B(\tilde{\boldsymbol{x}}) \geq T_B(\mathbf{0})$, which implies $u^{T,\tilde{\boldsymbol{x}}}(p) \leq u^{T,\mathbf{0}}(p)$ as $h > m$. Thus, $u^{T,\tilde{\boldsymbol{x}}}(p) \leq U^{\mathbf{0}}(p)$ for any $T \in \mathbb{T}$, and $U^{\tilde{\boldsymbol{x}}}(p) \leq U^{\mathbf{0}}(p)$.

Lastly, we now show that it is strictly dominated to retain the agent when $p < p^*$. Without loss of generality, we only consider the set $\mathbb{T}$ of strategies, since any strategy that is not in $\mathbb{T}$ is strictly dominated by a strategy in $\mathbb{T}$. For any $\tilde{\boldsymbol{x}}$ and $p < p^*$, consider an evaluator's strategy $T \in \mathbb{T}$ that retains the agent at the belief p. We have $u^{T,\tilde{\boldsymbol{x}}}(p) \leq u^{T,\mathbf{0}}(p)$. Note that in the NCB, retaining the agent at a belief $p < p^*$ gives the evaluator a payoff strictly less than m. Thus, $u^{T,\tilde{\boldsymbol{x}}}(p) \leq u^{T,\mathbf{0}}(p) < m$. □

Proof of Lemma 2. (a) The proof uses Lemma 1 and Lemma 3 in Board and Meyer-ter Vehn (2013). As in Lemma 1 in Board and Meyer-ter Vehn (2013), $t \mapsto V_\theta^{T_q}(p_t(\varnothing))$ is Lipschitz-continuous, given that the value function is bounded within $[0, w]$. As in Lemma 3 in Board and Meyer-ter Vehn (2013), $V_\theta^{T_q}(p)$ is continuous in $p \in [0, 1)$. Continuity in $(q, 1)$ is because $d^{\tilde{\boldsymbol{x}}}(p) < 0$ implies that the trajectory $t \mapsto p_t(\varnothing)$ always induces a homeomorphism between some time interval $[0, \bar{t}]$ and $[p_0, \bar{p}]$ for some $\bar{p}$, and $t \mapsto V_\theta^{T_q}(p_t(\varnothing))$ is Lipschitz-continuous. Continuity in $[0, q)$ is because $V_\theta^{T_q}(p) = 0$ for $p \leq q$. Last, continuity at q is because $\lim_{p\downarrow q} T_q(\varnothing) = 0$ almost surely.

(b) First, $\liminf_{p\uparrow 1} V_\theta^{T_q}(p) > \rho\ c$. Consider a strictly suboptimal strategy—censoring all bad news until good news arrives. When good news arrives, the agent's continuation value is $V_\theta^{T_q}(1) = w$. The payoff $\hat{V}_\theta(p)$ under this strategy is

$$\begin{aligned}\hat{V}_\theta(p) &= \int_0^{T_q(\varnothing)} e^{-(\rho+\gamma^\theta)t}\left[\rho(w-\beta^\theta c)+\gamma^\theta w\right]\mathrm{d}t \\ &= \frac{\rho(w-\beta^\theta c)+\gamma^\theta w}{\rho+\gamma^\theta}\left(1-e^{-(\rho+\gamma)T_q(\varnothing)}\right),\end{aligned}$$

where $T_q(\varnothing)$ is the time that the public belief drifts from p to q. Clearly, $\lim_{p\uparrow 1} T_q(\varnothing) = +\infty$ and $\lim_{p\uparrow 1}\hat{V}_\theta(p) > \rho c$, since $c < \bar{c}$. Thus, $\liminf_{p\uparrow 1} V_\theta^{T_q}(p) \geq \lim_{p\uparrow 1}\hat{V}_\theta(p) > \rho\ c$.

Second, let $q^\theta = \inf\{p \in [0,1] : V_\theta^{T_q}(p) = \rho\ c\}$. Given $V_\theta^{T_q}(p)$ is continuous in $[0,1)$, $V_\theta^{T_q}(p) = 0$ for $p \leq q$, and $\liminf_{p\uparrow 1} V_\theta^{T_q}(p) > \rho\ c$, $q^\theta \in (q,1)$ is attained and $V_\theta^{T_q}(p) < \rho\ c$ for $p < q^\theta$. We now show $V_\theta^{T_q}(p) > \rho\ c$ for $p > q^\theta$ and the result follows.

Let $p > q^\theta$ and τ be the time that the belief drifts from p to q^θ according to (DRIFT) and the public conjectured strategy $\tilde{\boldsymbol{x}}$. Consider the following feasible strategy: before time τ, censoring all bad news until good news arrives, and resuming the optimal strategy at the public belief q^θ starting at time τ if no news arrives by then. The payoff $\tilde{V}_\theta(p)$ under this strategy is

$$\begin{aligned}\tilde{V}_\theta(p) =& \int_0^{\tau} e^{-(\rho+\gamma^\theta)t}\left[\rho(w-\beta^\theta c)+\gamma^\theta w\right]\mathrm{d}t + e^{-(\rho+\gamma^\theta)}V_\theta^{T_q}(q^\theta) \\ =& \frac{\rho(w-\beta^\theta c)+\gamma^\theta w}{\rho+\gamma^\theta}\left(1-e^{-(\rho+\gamma)\tau}\right) + e^{-(\rho+\gamma^\theta)\tau}V_\theta^{T_q}(q^\theta) \\ >& \rho\ c\left(1-e^{-(\rho+\gamma)\tau}\right) + e^{-(\rho+\gamma^\theta)\tau}V_\theta^{T_q}(q^\theta) = \rho\ c,\end{aligned}$$

where the inequality comes from $c < \bar{c}$. Thus, $V_\theta^{T_q}(p) \geq \tilde{V}_\theta(p) > \rho\ c$ for $p > q^\theta$. □

Proof of Lemma 3. The "if" part is obvious. We now prove the "only if" part by following Lemma 1 in Board and Meyer-ter Vehn (2013). We already have that $t \mapsto V_\theta^{T_q}(p_t(\varnothing))$ is Lipschitz-continuous and $V_\theta^{T_q}(p)$ is continuous in $p \in [0,1)$. Moreover, $t \mapsto \boldsymbol{x}^\theta(p_t(\varnothing))$ is right-continuous, given the admissibility of $\tilde{\boldsymbol{x}}^\theta(p)$ and $\tilde{\boldsymbol{x}}^\theta(p) = \boldsymbol{x}^\theta(p)$ in equilibrium.

Suppose at some p_0, $\Delta_\theta^{T_q}(p_0) > \rho\ c$ but $\boldsymbol{x}^\theta(p_0) < 1$. $\boldsymbol{x}^\theta(p_0) < 1$ implies that $J^{\tilde{\boldsymbol{x}}}(p_0)$ is defined on-path as $\boldsymbol{x}^\theta(p_0) = \tilde{\boldsymbol{x}}^\theta(p_0) < 1$ in equilibrium. The admissibility of $\tilde{\boldsymbol{x}}^\theta(p)$

implies that $t \mapsto J^{\tilde{\boldsymbol{x}}}(p_t(\varnothing))$ is defined on-path and right-continuous on some time interval. Together with continuity of $V_\theta^{T_q}(p)$, we have that $t \mapsto V_\theta^{T_q}(J^{\tilde{\boldsymbol{x}}}(p_t(\varnothing)))$ is right-continuous. Thus, $t \mapsto \Delta_\theta^{T_q}(p_t(\varnothing))$ and $t \mapsto \boldsymbol{x}^\theta(p_t(\varnothing))$ are right-continuous. We have $\Delta_\theta^{T_q}(p_t(\varnothing)) > \rho\, c$ and $\boldsymbol{x}^\theta(p_t(\varnothing)) < 1$ for all t on some time interval $[0, \delta]$. Setting $\boldsymbol{x}^\theta(p) = 1$ for all $p \in \{p_t(\varnothing) : t \in [0, \delta]\}$ can strictly improve the agent's payoff.

Suppose at some p_0, $\Delta_\theta^{T_q}(p_0) < \rho\, c$ but $\boldsymbol{x}^\theta(p_0) > 0$. This means that $V_\theta^{T_q}(p_0) \leq \Delta_\theta^{T_q}(p_0) < \rho\, c$. Right-continuity of $t \mapsto V_\theta^{T_q}(p_t(\varnothing))$ and $t \mapsto \boldsymbol{x}^\theta(p_t(\varnothing))$ imply that we have $V_\theta^{T_q}(p_t(\varnothing)) < \rho\, c$ and $\boldsymbol{x}^\theta(p_t(\varnothing)) > 0$ for all t on some time interval $[0, \delta]$. Thus, $\Delta_\theta^{T_q}(p_t(\varnothing)) \leq V_\theta^{T_q}(p_t(\varnothing)) < \rho\, c$ on $[0, \delta]$. Setting $\boldsymbol{x}^\theta(p) = 0$ for all $p \in \{p_t(\varnothing) : t \in [0, \delta]\}$ can strictly improve the agent's payoff. □

Proof of Lemma 4. In the NCB, the public conjectured strategy $\tilde{\boldsymbol{x}}$ is zero. Given an initial belief $p \in (p^*, 1)$, let $p_t(p)$ be the unique solution to $\dot{p}_t = d(p_t, 0, 0)$, $\tau(p) = T_{p^*}(\varnothing)$ be the time that the belief drifts from p to p^* in the absence of news, $k_t(p) = \min\{k \in \mathbb{N} : j^k(p_t(p)) \leq p^*\}$ be the minimal number of bad news jumps after which the belief falls below p^* at time t, and $\mathbb{P}_{\theta,p}[T_{p^*} \leq t]$ be the probability that type θ agent is dismissed by time $t > 0$. The agent is dismissed by t if and only if there are at least $k_t(p)$ pieces of bad news revealed and no good news is revealed by t. Note that $k_t(p) = 0$ for $t \geq \tau(p)$. Thus, for any $t > 0$,

$$\mathbb{P}_{\theta,p}[T_{p^*} \leq t] = \sum_{n \geq k_t(p)} \frac{e^{-\beta^\theta t}\left(\beta^\theta t\right)^n}{n!} e^{-\gamma^\theta t}.$$

(a) Fix $p \in (p^*, 1)$, $\Pi_\theta(p) = \mathbb{E}\left[\int_0^{T_{p^*}} e^{-\rho t} \rho\, w\, \mathrm{d}t\right]$. As the agent has a constant flow payoff when she is not dismissed, her discounted payoff is strictly increasing in the time she stays in the job. We now show that the type G stays strictly longer than the type B in the usual stochastic dominance order and the result follows.

For any $t \geq \tau(p)$, $k_t(p) = 0$, we have $\mathbb{P}_{G,p}[T_{p^*} \leq t] = e^{-\gamma t} < 1 = \mathbb{P}_{B,p}[T_{p^*} \leq t]$. For $t < \tau(p)$, $k_t(p) \geq 1$; for any $n \geq k_t(p)$, $\gamma > \beta^B > \beta^G$ implies $e^{-\beta^G t}\left(\beta^G t\right)^n e^{-\gamma t}/n! < e^{-\beta^B t}\left(\beta^B t\right)^n/n!$. Thus, $\mathbb{P}_G[T_{p^*} \leq t] < \mathbb{P}_B[T_{p^*} \leq t]$ for any $t > 0$ and the result follows.

(b) Continuity follows from the same proof as in Lemma 2. Fixing $\theta \in \Theta$, we now show that $\Pi_\theta(p)$ is strictly increasing in $p \in (p^*, 1)$ by showing that the agent stays

strictly longer (in the usual stochastic dominance order) with a higher initial belief.

According to (DRIFT) and (JUMP), for any $p^* < p' < p < 1$ and $t > 0$, we have $p_t(p) > p_t(p')$, $\tau(p) > \tau(p')$ and $k_t(p) \geq k_t(p')$; for $t \in [\tau(p'), \tau(p))$, $k_t(p) > 0 = k_t(p')$. Thus, for any $p^* < p' < p < 1$ and $t > 0$, $\mathbb{P}_{\theta,p}[T_{p^*} \leq t] \leq \mathbb{P}_{\theta,p'}[T_{p^*} \leq t]$ and the inequality is strict for $t \in [\tau(p'), \tau(p))$. The result follows.

(c) Obviously, $\Pi_\theta(p) = 0$ for $p \leq p^*$. We now show that $\lim_{p\uparrow 1} \Pi_\theta(p) > \rho\ c$ so that there is a unique $p^\theta \in (p^*, 1)$ such that $\Pi_\theta(\theta) = \rho\ c$. $\Pi_B(p)$ can be written by truncating itself at the first news arrival

$$\int_0^{T_{p^*}(\varnothing)} e^{-(\rho+\beta^B)t} \left\{\rho\ w + \beta^\theta \Pi_B(j(p_\nu))\right\} \mathrm{d}t$$
$$\geq \int_0^{T_{p^*}(\varnothing)} e^{-(\rho+\beta^B)t} \rho\ w\, \mathrm{d}t = \frac{\rho\ w}{\rho + \beta^B} \left(1 - e^{-(\rho+\beta^B)T_{p^*}(\varnothing)}\right) \to \frac{\rho\ w}{\rho + \beta^B}$$

when $p \to 1$, as $T_{p^*}(\varnothing) \to \infty$. Note that $c < \bar{c}$ implies that $\rho w/(\rho + \beta^B) > \rho c$. Thus, $\lim_{p\uparrow 1} \Pi_B(p) > \rho\ c$. $\lim_{p\uparrow 1} \Pi_G(p) > \rho\ c$ follows as $\Pi_G(p) > \Pi_B(p)$ and $\Pi_G(p)$ is a bounded monotone function that converges. Lastly, $p^G < p^B$ as $\Pi_G(p) > \Pi_B(p)$. □

B Proofs for Section 4.1

Proof of Proposition 1. Lemma 1 implies that the unique candidate equilibrium strategy of the evaluator is T_{p^*}. Given T_{p^*}, we have $V_B^{T_{p^*}}(J^{\tilde{\boldsymbol{x}}}(p)) = 0$ for $p < 1$, since $J^{\tilde{\boldsymbol{x}}}(p) = 0$. According to Lemma 2 and 3, the type B's best response is a cutoff strategy, with the cutoff denoted by q^B. We now (a) show $q^B = p^B$ (uniqueness) and (b) verify its optimality (existence), and (c) verify the evaluator's best response is indeed T_{p^*}.

(a) Since the type B's best response $\boldsymbol{x}^B$ is a cutoff strategy with $q^B \in (p^*, 1)$ for any public conjectured strategy $\tilde{\boldsymbol{x}}^B$, $\tilde{\boldsymbol{x}}^B$ must agree with $\boldsymbol{x}^B$ in equilibrium with the same cutoff q^B. Moreover, given $\tilde{\boldsymbol{x}}^B$, we must have $V_B^{T_{p^*}}(q^B) = \rho\ c$ by Lemma 2. For $p \leq q^B$, we have $\tilde{\boldsymbol{x}}^B(p) = \boldsymbol{x}^B(p) = 0$ and thus $V_B^{T_{p^*}}(p) = \Pi_B(p)$. By Lemma 4, $\Pi_B(p) = \rho\ c$ if and only if $p = p^B$. $V_B^{T_{p^*}}(q^B) = \rho\ c$ implies $q^B = p^B$.

(b) Given $\tilde{\boldsymbol{x}}^{\boldsymbol{B}} = \mathbb{1}_{p\in(p^B,1)}$, we now verify the optimality of $\boldsymbol{x}^B = \mathbb{1}_{p\in(p^B,1)}$. By Lemma 2, the type B's value function single crosses $\rho\ c$ from below at some be-

lief q^B so that $V_B^{T_{p^*}}(q^B) = \rho\ c$. Thus, given Lemma 3 and $j(p_0) \leq p^*$, it is optimal to censor bad news if and only if $p > q^B$. The optimality of $\boldsymbol{x}^B = \mathbb{1}_{p\in(p^B,1)}$ is verified by showing the belief q^B is p^B. Given $\tilde{\boldsymbol{x}}^{\boldsymbol{B}} = \mathbb{1}_{p\in(p^B,1)}$, we have $V_B^{T_{p^*}}(p) \geq \Pi_B(p)$ for $p \leq p^B$. As $\Pi_B(p)$ is monotone and $\Pi_B(p^B) = \rho\ c$, we have $q^B \leq p^B$. Suppose $q^B < p^B$, then $V_B^{T_{p^*}}(p) = \Pi_B(p)$ for $p \leq q^B$ and $\Pi_B(q^B) < \Pi_B(p^B) = \rho\ c$. This means $V_B^{T_{p^*}}(q^B) < \rho\ c$, which is a contradiction.

(c) Given $\tilde{\boldsymbol{x}}^{\boldsymbol{B}} = \mathbb{1}_{p\in(p^B,1)}$, there is no censorship for $p \leq p^B$. Thomas (2016) solves the same problem in the NCB and finds there exists a unique best response T_{p^*}. Moreover, $p \leq p^B$ never enters the region $(p^B, 1)$. Thus, T_{p^*} is the unique best response when $p \leq p^B$. For $p > p^B$, note that, with probability one, p enters the region $[0, p^B] \cup \{1\}$ by some bounded time. With part (a) of Lemma 1, the optimality of T_{p^*} follows. □

Proof of Proposition 2. The part for the evaluator and the type G agent is obvious. We now prove the part for the type B agent. We (a) compute some key variables, (b) show the type B has a higher payoff in the MPE if and only if $s_1 > s_1^*$ for some s_1^*, and (c) prove the result.

(a) Note that $\bar{s}$ is the time it takes for the belief to drift from p_0 to p^* in the NCB. According to (DRIFT) and $\tilde{\boldsymbol{x}}^B(p) = 0$, $\bar{s} = \ln\left[\Omega(p^*)/\Omega(p_0)\right]/(\gamma - \beta^B)$, where $\Omega(p) := (1-p)/p$.

In the MPE, p^B is defined in Lemma 4 by $\Pi_B(p^B) = \rho\, c$, where

$$\Pi_B(p^B) = \int_0^{s_2} e^{-(\rho+\beta^B)t}\rho\ w\,\mathrm{d}t = \frac{\rho\ w}{\rho+\beta^B}\left(1 - e^{-(\rho+\beta^B)s_2}\right),$$

and s_2 is the time it takes for the belief to drift from p^B to p^*. $\Pi_B(p^B) = \rho\ c$ gives

$$s_2 = \frac{1}{\rho+\beta^B}\ln\left[\frac{w}{w - (\rho+\beta^B)\,c}\right]. \tag{4}$$

According to (DRIFT) and $\tilde{\boldsymbol{x}}^B(p) = 0$ for $p \leq p^B$, $s_2 = \ln\left[\Omega(p^*)/\Omega(p^B)\right]/(\gamma - \beta^B)$. Together,

$$p^B = \frac{p^*}{p^* + (1-p^*)e^{-(\gamma-\beta^B)s_2}}. \tag{5}$$

According to (DRIFT) and $\tilde{\boldsymbol{x}}^B(p) = 1$ for $p > p^B$, the time s_1 it takes for the belief

to drift from p_0 to p^B is given by

$$s_1 = \frac{1}{\gamma} \ln \left[\frac{p_0}{1-p_0} \frac{1-p^B}{p^B} \right] = \frac{\gamma - \beta^B}{\gamma} \left(\bar{s} - s_2 \right), \tag{6}$$

where the second equality comes from the definition of $\bar{s}$ and s_2.

(b) When $p_0 > p^B$, the expected payoff of the type B agent in the MPE is

$$\left(w - \beta^B c \right) \left(1 - e^{-\rho s_1} \right) + e^{-\rho s_1} \rho c.$$

In the NCB, her expected payoff is $\rho w \left(1 - e^{-(\rho + \beta^B)\bar{s}} \right) / (\rho + \beta^B)$. Thus, she has a higher payoff in the MPE if and only if

$$\left(w - \beta^B c \right) \left(1 - e^{-\rho s_1} \right) + e^{-\rho s_1} \rho c > \frac{\rho w}{\rho + \beta^B} \left(1 - e^{-(\rho + \beta^B)\bar{s}} \right).$$

Note that we have $w - \beta^B c > \rho w \left(1 - e^{-(\rho + \beta^B)\bar{s}} \right) / (\rho + \beta^B) > \rho c$. The first inequality is due to $c < \bar{c}$. The second is due to $s_2 < \bar{s}$ and (4). Together with (6), the type B agent has a higher payoff in the MPE than in the NCB if and only if $f(s_1) < 0$, where

$$f(s) := e^{-\rho s} - \frac{\beta^B + \rho \exp \left[-\frac{\gamma(\rho + \beta^B)}{\gamma - \beta^B} s \right]}{\rho + \beta^B}. \tag{7}$$

Note that $f(s)$ is first increasing and then decreasing in s. Moreover, $f(0) = 0$ and $\lim_{s \to \infty} f(s) = -\beta^B / (\rho + \beta^B) < 0$. Hence, there exists a $s_1^* = s_1^*(\rho, \gamma, \beta^B) > 0$ such that the type B agent has a higher payoff in the MPE if and only if $s_1 > s_1^*$.

(c) By (5), p^B is continuous and increasing in s_2. By (4), s_2 is continuous and increasing in c, $\lim_{c \to 0} p^B = p^*$ and $\lim_{c \to \bar{c}} p^B = 1$. There exists a unique $\mathfrak{c}_2 \in (0, \bar{c})$ such that $p_0 = p^B$ if and only if $c = \mathfrak{c}_2$. If $c \in [\mathfrak{c}_2, \bar{c})$, $p_0 \leq p^B$. There is no censorship in the MPE, and the type B agent obtains the same payoff in the MPE and the NCB.

Assume $c < \mathfrak{c}_2$ so $p_0 > p^B$. Note that $s_1 < s_1^*$ is equivalent to $\bar{s} < s_2 + s^*$ due to (6), where $s^* := \gamma s_1^* / (\gamma - \beta^B)$. The type B agent has a lower payoff in the MPE if and only if $s_2 > \bar{s} - s^*$. If $\bar{s} \leq s^*$, $s_2 > \bar{s} - s^*$ holds for all $c \in (0, \mathfrak{c}_2)$. Consider $\bar{s} > s^*$. Note that both $\bar{s}$ and s^* do not depend on c. Since s_2 is increasing in c, $\lim_{c \to 0} s_2 = 0$ and $\lim_{c \to \mathfrak{c}_2} s_2 = \bar{s}$, there exists a unique $\mathfrak{c}_1 \in (0, \mathfrak{c}_2)$ such that $s_2 > \bar{s} - s^*$ if $c \in (\mathfrak{c}_1, \mathfrak{c}_2)$, and $s_2 < \bar{s} - s^*$ if $c \in (0, \mathfrak{c}_1)$. We complete the proof by redefining $\mathfrak{c}_1$ as $\mathfrak{c}_1 \mathbb{1}_{\bar{s} > s^*}$. $\square$

C Proofs for Section 4.1

We first prove three lemmas that will be used in the proof of Proposition 3.

Suppose the evaluator uses the strategy T_{p^*}, and the public conjectured strategies are $\tilde{\boldsymbol{x}}^G(p) = \mathbb{1}_{p\in(p^G,1)}$ and $\tilde{\boldsymbol{x}}^B(p) = 0$ for any p. Let $\hat{\Pi}_B(p)$ be the payoff function of the type B when she never censors.

Lemma 5. *Assume $\beta^G > 0$. (a) $\hat{\Pi}_B(p) = 0$ for $p \le p^*$, $\lim_{p\uparrow 1}\hat{\Pi}_B(p) > \rho\ c$; it is continuous in $p \in [0,1)$ and strictly increasing in $p \in [p^*,1)$; (b) there is a unique $\hat{p}^B \in (p^G,1)$ such that $\hat{\Pi}_B(\hat{p}^B) = \rho\ c$.*

Proof of Lemma 5. $\hat{\Pi}_B(p) = 0$ for $p \le p^*$ is obvious. Continuity follows from the same proof in Lemma 2. The second result follows from the first one and the fact that $\hat{\Pi}_B(p) = \Pi_B(p)$ for $p \le p^G$ and $\Pi_B(p^G) < \Pi_G(p^G) = \rho\ c$. It is sufficient to prove that $\hat{\Pi}_B(p)$ is strictly increasing in $p \in (p^G,1)$ and $\lim_{p\uparrow 1}\Pi_B(p) > \rho\ c$.

Let $p \in (p^G,1)$ and $\tau(p)$ be the time that the belief drifts from p to p^G, which is strictly increasing in p. $\hat{\Pi}_B(p)$ can be written as

$$\begin{aligned}&\int_0^{\tau(p)} e^{-(\rho+\beta^B)t}\rho\ w\,\mathrm{d}t + e^{-(\rho+\beta^B)\tau(p)}\hat{\Pi}_B(p^G)\\ =&\frac{\rho\ w}{\rho+\beta^B}\left(1-e^{-(\rho+\beta^B)\tau(p)}\right) + e^{-(\rho+\beta^B)\tau(p)}\hat{\Pi}_B(p^G).\end{aligned}$$

Note that $c < \bar{c}$ implies $\rho\ w/(\rho+\beta^B) > \rho\ c > \hat{\Pi}_B(p^G)$, so $\hat{\Pi}_B(p)$ is strictly increasing in $\tau(p)$ and thus in p. Lastly, $\lim_{p\uparrow 1}\hat{\Pi}_B(p) = \rho\ w/(\rho+\beta^B) > \rho\ c$. □

Lemma 6. *Assume $\beta^G > 0$. In any pure strategy MPE, the evaluator's strategy is T_{p^*}.*

Proof of Lemma 6. Fix a pure strategy MPE. Let $T(\Sigma)$ be the evaluator's strategy and $\tilde{\boldsymbol{x}}$ be the pure conjectured strategy. The admissibility of $\tilde{\boldsymbol{x}}(p)$ implies $\tilde{\boldsymbol{x}}$ is piecewise constant. We have $0 \in \Sigma$. By Lemma 1, $p \notin \Sigma$ for $p > p^*$. We now show that $\Sigma = [0,p^*]$.

Step 1: the value function $U^{\tilde{\boldsymbol{x}}}(p)$ is continuous in $p \in [0,q_1)$ for some $q_1 \in (0,1]$.

$\tilde{\boldsymbol{x}}$ is pure and piecewise constant. There exists a belief interval $(0, q_1)$ where $\tilde{\boldsymbol{x}}$ is a constant pure strategy. Thus, for $p \in (0, q_1)$, we have a Poisson bandit problem where the Poisson processes are homogenous and the belief only goes down or jumps up to 1. As the evaluator's payoff is linear in the belief, his value function is convex and thus continuous in $(0, q_1)$. Since his value function is bounded from above by the value under full information $p\ h + (1-p)m$ and bounded from below by the value under myopic rule $\max\{p\ h, m\}$, and both converge to m when $p \to 0$, the evaluator's value function is also continuous at $p = 0$. Thus, $U^{\tilde{\boldsymbol{x}}}(p)$ is continuous in $[0, q_1)$.

Step 2: there exists a $\hat{p} \in (0, p^*]$ such that $[0, \hat{p}] \subset \Sigma$.

First, I claim that there exists a $\hat{p} \in (0, p^*]$ such that $U^{\tilde{\boldsymbol{x}}}(p) = m$ for $p \in [0, \hat{p}]$. Suppose not. From Step 1 and $0 \in \Sigma$, there exists a $q_2 \in (0, q_1)$ such that $U^{\tilde{\boldsymbol{x}}}(p) > m$ for $p \in (0, q_2)$. Thus, the evaluator never dismisses the agent when $p \in (0, q_2)$. Since $\tilde{\boldsymbol{x}}$ is pure and piecewise constant, we can find a $q_3 \in (0, q_2)$ where both types of agent use a constant pure strategy for $p \in (0, q_3)$. A strategy profile for $p \in (0, q_3)$ must be either (a) $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 1$, or (b) $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 0$, or (c) $\tilde{\boldsymbol{x}}^G(p) = 1$ and $\tilde{\boldsymbol{x}}^B(p) = 0$. Note that $\tilde{\boldsymbol{x}}^G(p) = 0$ and $\tilde{\boldsymbol{x}}^B(p) = 1$ is ruled out as otherwise $J^{\tilde{\boldsymbol{x}}}(p) = 1$. Both (a) and (b) mean that the belief never jumps to 0, so the evaluator never dismisses the agent and her payoff is $p\ h < m$ for small p, which is a contradiction. If (c) is the strategy profile, the evaluator dismisses the agent only when bad news is revealed. His payoff is $p\ h + (1-p)\beta^B m/(\rho + \beta^B) < m$ for small p, which is a contradiction.

Second, it remains to show that it is strictly suboptimal to retain the agent at $p \leq \hat{p}$. It is obvious for $p = 0$. Suppose, by contradiction, that it is optimal for the evaluator to retain the agent at a belief $p \in (0, \hat{p}]$ for a positive time $T > 0$ in the absence of news. Since $\tilde{\boldsymbol{x}}$ is pure and piecewise constant, there exists a $S \in (0, T]$ such that both types of agent use a constant pure strategy (either $\tilde{\boldsymbol{x}}^G = \tilde{\boldsymbol{x}}^B \in \{0, 1\}$, or $\tilde{\boldsymbol{x}}^G = 1$ and $\tilde{\boldsymbol{x}}^B = 0$) in the absence of news before time S. Thus, the arrival rate of bad news for each type, denoted by $\hat{\beta}^\theta = \beta^\theta(1 - \tilde{\boldsymbol{x}}^\theta)$, is constant before time S. Moreover, $J^{\tilde{\boldsymbol{x}}}(p') < p'$ for $p' \in (0, p]$ if defined. Since $U^{\tilde{\boldsymbol{x}}}(p) = m$ for $p \in [0, \hat{p}]$, the evaluators's payoff at belief p can be calculated as a function of S by imposing the continuation

value to be m both after bad news and after time S in the absence of the news:

$$H(S) := p\left[\int_0^S e^{-(\rho+\gamma+\hat{\beta}^G)t}\left(\gamma\ k(\rho+\gamma)+\hat{\beta}^G\ m\right)\mathrm{d}t + e^{-(\rho+\gamma+\hat{\beta}^G)S}\ m\right]$$
$$+\ (1-p)\left[\int_0^S e^{-(\rho+\hat{\beta}^B)t}\ \hat{\beta}^B\ m\ \mathrm{d}t + e^{-(\rho+\hat{\beta}^B)S}\ m\right].$$

It is easy to see that $H'(S) < 0$ for $p \leq p^*$ and $S > 0$. Thus, $H(S) < H(0) = m$ for $S > 0$, which is a contradiction.

Step 3: there exists a $\epsilon > 0$, such that $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 0$ for $p \leq \hat{p} + \epsilon$, where $\hat{p} = \inf\{p : p \notin \Sigma\} \in (0, p^*]$.

Given Step 2 and $p \notin \Sigma$ for any $p > p^*$, $\hat{p}$ is well-defined. Thus, for any $\delta > 0$, there exists a $p_\delta \in (\hat{p}, \hat{p} + \delta)$ such that $p_\delta \notin \Sigma$. The type θ agent's value function $V_\theta^T(p)$ at belief $p \leq p_\delta$ is bounded above by $\bar{V}_\theta(p_\delta) = \int_0^\tau e^{-(\rho+\beta^\theta+\gamma^\theta)t}(\rho\ w + \beta^\theta w + \gamma^\theta w)\,\mathrm{d}t$, where τ is the time it takes for the belief to drift from p_δ to $\hat{p}$. $\bar{V}_\theta(p_\delta)$ is calculated without considering any censoring cost and assuming any news arrival gives the agent her maximal value w. Clearly, $\lim_{\delta\downarrow 0}\tau = 0$ and $\lim_{\delta\downarrow 0}\bar{V}_\theta(p_\delta) = 0$. Thus, there exists a $\hat{\delta} > 0$, such that $\bar{V}_\theta(p_\delta) < \rho\ c$ for any $\delta \leq \hat{\delta}$. Thus, $\Delta_\theta^T(p) \leq V_\theta^T(p) \leq \bar{V}_\theta(p_{\hat{\delta}}) < \rho\ c$ for $p \leq p_{\hat{\delta}}$. Similar to Lemma 3, we must have $\tilde{\boldsymbol{x}}^\theta(p) = 0$ for $p \leq p_{\hat{\delta}}$. Take $\epsilon = p_{\hat{\delta}} - \hat{p}$.

Step 4: $\hat{p} = p^*$ so the result follows.

Given $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 0$ for $p \leq \hat{p} + \epsilon$, Thomas (2016) solves the same problem for $p \leq \hat{p} + \epsilon$ and finds that there exists a unique best response T_{p^*}. Thus, if $\hat{p} < p^*$, retaining the agent at a belief $p \in (\hat{p}, p^*)$ is strictly suboptimal. □

Lemma 7. *Assume $\beta^G > 0$ and $j(p_0) \leq p^*$. In any pure strategy MPE, the type G agent's strategy is $\boldsymbol{x}^G = \mathbb{1}_{p\in(p^G,1)}$, and the type B agent's strategy is $\boldsymbol{x}^B = \mathbb{1}_{p\in(\hat{p}^B,1)}$.*

Proof of Lemma 7. Fix a pure strategy MPE and let $\tilde{\boldsymbol{x}}^\theta$ be the public conjectured strategy. Lemma 6 gives the evaluator's strategy T_{p^*}. For any $p < 1$, using a pure strategy means that $J^{\tilde{\boldsymbol{x}}}(p) = 0$ if $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 1$ or $\tilde{\boldsymbol{x}}^G(p) = 1, \tilde{\boldsymbol{x}}^B(p) = 0$, and $J^{\tilde{\boldsymbol{x}}}(p) = j(p)$ if $\tilde{\boldsymbol{x}}^G(p) = \tilde{\boldsymbol{x}}^B(p) = 0$. $\tilde{\boldsymbol{x}}^G(p) = 0, \tilde{\boldsymbol{x}}^B(p) = 1$ is ruled out. Thus, $j(p_0) \leq p^*$ implies that $V_\theta^{T_{p^*}}(J^{\tilde{\boldsymbol{x}}}(p)) = 0$ for any $p \leq p_0$. According to Lemmas 2 and 3, the type θ agent's best response is a cutoff strategy $\boldsymbol{x}^\theta$, with the cutoff denoted

by $q^\theta \in (p^*, 1)$. Thus, $\tilde{\boldsymbol{x}}^\theta$ must also be a cutoff strategy with cutoff $\tilde{q}^\theta$ and $\tilde{q}^\theta = q^\theta$ in equilibrium. Moreover, we must have $V_\theta^{T_{p^*}}(q^\theta) = \rho\ c$. The rest of the proof shows that $q^G = p^G$ and $q^B = \hat{p}^B$.

(a) We first show that $q^B > q^G$ and $q^G = p^G$. Let $q_m = \min\{q^G, q^B\} \in (p^*, 1)$. For $p \leq q_m$ and $\theta \in \Theta$, we have $\tilde{\boldsymbol{x}}^\theta(p) = \boldsymbol{x}^\theta(p) = 0$ and thus $\Pi_\theta(p) = V_\theta^{T_{p^*}}(p) \leq \rho\ c$. By Lemma 4, $\Pi_G(p) > \Pi_B(p)$ and thus $V_G^{T_{p^*}}(p) > V_B^{T_{p^*}}(p)$ for $p \in (p^*, q_m]$. Suppose $q^B \leq q^G$, then $q_m = q^B$ and $V_G^{T_{p^*}}(q^B) > V_B^{T_{p^*}}(q^B) = \rho\ c$, which contradicts with $V_G^{T_{p^*}}(p_m) \leq \rho\ c$. Thus, we must have $q^B > q^G$. Moreover, $\Pi_G(q^G) = V_G^{T_{p^*}}(q^G) = \rho\ c$, as $q_m = q^G$. By Lemma 4, $\Pi_G(p) = \rho\ c$ if and only if $p = p^G$. Thus, we have $q^G = p^G$.

(b) We now show that $q^B = \hat{p}^B$. Since $q^B > q^G = p^G$ and $\tilde{q}^\theta = q^\theta$ in the MPE, $V_B^{T_{p^*}}(p) = \hat{\Pi}_B(p)$ for $p \leq q^B$, where $\hat{\Pi}_B(p)$ is defined in Lemma 5. Note that $\hat{\Pi}_B(p) = \rho\ c$ if and only if $p = \hat{p}^B$, and $V_B^{T_{p^*}}(q^B) = \rho\ c$. Thus, $q^B = \hat{p}^B$. □

Proof of Proposition 3. The uniqueness result follows from Lemma 6 and Lemma 7. We now verify the equilibrium. Fix the equilibrium public conjectured strategy $\tilde{\boldsymbol{x}}^G = \mathbb{1}_{p \in (p^G, 1)}$ and $\tilde{\boldsymbol{x}}^B = \mathbb{1}_{p \in (\hat{p}^B, 1)}$. Note that $\hat{p}^B > p^G$ and thus $J^{\tilde{\boldsymbol{x}}}(p) = 0$ for $p \in (p^G, 1)$[27] and $J^{\tilde{\boldsymbol{x}}}(p) = j(p)$ for $p \in (0, p^G)$.

(a) Given the evaluator's strategy T_{p^*}, the type B agent's strategy is $\boldsymbol{x}^B = \mathbb{1}_{p \in (\hat{p}^B, 1)}$, we verify the optimality of the type G's strategy.

By Lemma 2, the type G's value function single crosses $\rho\ c$ from below at some belief q^G so that $V_G^{T_{p^*}}(q^G) = \rho\ c$. We first verify that the crossing belief is p^G. Given $\tilde{\boldsymbol{x}}$, we have $V_G^{T_{p^*}}(p) \geq \Pi_G(p)$ for $p \leq p^G$. As $\Pi_G(p)$ is monotone and $\Pi_G(p^G) = \rho\ c$, we have $q^G \leq p^G$. Suppose $q^G < p^G$, then $V_G^{T_{p^*}}(p) = \Pi_G(p)$ for $p \leq q^G$ and $\Pi_G(q^G) < \Pi_G(p^G) = \rho\ c$. This means $V_G^{T_{p^*}}(q^G) < \rho\ c$, which is a contradiction.

We have now $V_G^{T_{p^*}}(p) > \rho\, c$ if and only if $p > p^G$. For $p \in (p^G, 1)$, $\Delta_G^{T_{p^*}}(p) = V_G^{T_{p^*}}(p) > \rho\ c$, as $V_G^{T_{p^*}}(J^{\tilde{\boldsymbol{x}}}(p)) = 0$. By Lemma 3, $\boldsymbol{x}^G(p) = 1$ is optimal for $p \in (p^G, 1)$. For $p < p^G$, $\Delta_G^{T_{p^*}}(p) \leq V_G^{T_{p^*}}(p) < \rho\ c$. Thus, $\boldsymbol{x}^G(p) = 0$ is optimal for $p < p^G$.

(b) Given the evaluator's strategy T_{p^*}, the type G agent's strategy is $\boldsymbol{x}^G = \mathbb{1}_{p \in (p^G, 1)}$, the type B's strategy $\boldsymbol{x}^B = \mathbb{1}_{p \in (\hat{p}^B, 1)}$ is optimal using the same verification above and the fact that $\hat{\Pi}_B(p)$ is monotone and $\hat{\Pi}_B(\hat{p}^B) = \rho\ c$ by Lemma 5.

[27] Note that $J^{\tilde{\boldsymbol{x}}}(p) = 0$ for $p \in (\hat{p}^B, 1)$ is defined off-path. See footnote 11.

(3) Given the agent's strategy, there is no censorship for $p \leq p^G$. Thomas (2016) solves the same problem in the NCB and finds that there exists a unique best response T_{p^*}. Moreover, $p \leq p^G$ never enters the region $(p^G, 1)$. Thus, T_{p^*} is the unique best response when $p \leq p^G$. For $p > p^G$, note that, with probability one, p enters the region $[0, p^G] \cup \{1\}$ by some bounded time. Together with part (a) of Lemma 1, the optimality of T_{p^*} follows. □

Proof of Proposition 4. (a) We show that the evaluator has a higher payoff in the MPE than in the NCB when $p_0 \in (p^G, \hat{p}^B]$.

Given $p \in (p^G, \hat{p}^B]$, it is optimal for the evaluator to retain the agent. His value function $U^{\mathbf{0}}$ in the NCB solves the following Bellman equation (see Thomas (2016) for a derivation):

$$\rho\, U^{\mathbf{0}}(p) = \underbrace{p\,\gamma\,[k(\rho+\gamma) - U^{\mathbf{0}}(p)]}_{\text{Expected benefit from good news}} + \underbrace{\beta(p)\,[U^{\mathbf{0}}(j(p)) - U^{\mathbf{0}}(p)]}_{\text{Expected loss from bad news}} - \underbrace{(\gamma + \beta^G - \beta^B)\,p\,(1-p)\,U^{\mathbf{0}\prime}(p)}_{\text{Deterioration in the absence of news}}, \tag{8}$$

where $\beta(p) := p\beta^G + (1-p)\beta^B$ is the arrival rate of bad news at belief p.

Similarly, his value function $U^{\tilde{\boldsymbol{x}}}$ in the MPE solves the Bellman equation:

$$\rho\, U^{\tilde{\boldsymbol{x}}}(p) = \underbrace{p\,\gamma\,[k(\rho+\gamma) - U^{\tilde{\boldsymbol{x}}}(p)]}_{\text{Expected benefit from good news}} + \underbrace{(1-p)\,\beta^B\,[U^{\tilde{\boldsymbol{x}}}(0) - U^{\tilde{\boldsymbol{x}}}(p)]}_{\text{Expected loss from bad news}} - \underbrace{(\gamma - \beta^B)\,p\,(1-p)\,U^{\tilde{\boldsymbol{x}}\prime}(p)}_{\text{Deterioration in the absence of news}}. \tag{9}$$

The convexity of $U^{\mathbf{0}}(p)$ implies for $p > p^*$, $U^{\mathbf{0}\prime}(p) > [U^{\mathbf{0}}(p) - U^{\mathbf{0}}(j(p))] \,/\, (p - j(p)) > [U^{\mathbf{0}}(p) - U^{\mathbf{0}}(0)] \,/ p > 0$. Together with $p - j(p) = p(1-p)(\beta^B - \beta^G)/\beta(p)$, we have

$$\begin{aligned} &\beta(p)\left[U^{\mathbf{0}}(p) - U^{\mathbf{0}}(j(p))\right] + \beta^G p(1-p)\; U^{\mathbf{0}\prime}(p) - (1-p)\beta^B\left[U^{\mathbf{0}}(p) - U^{\mathbf{0}}(0)\right] \\ &> \left\{\; \beta(p)\frac{p - j(p)}{p} \;+ \frac{\beta^G p(1-p)}{p} \;- (1-p)\beta^B \;\right\}\left[U^{\mathbf{0}}(p) - U^{\mathbf{0}}(0)\right] = 0. \end{aligned} \tag{10}$$

Thus, comparing (8) and (9) at $p = p^G$, using the fact that $U^{\tilde{\boldsymbol{x}}}(p) = U^{\mathbf{0}}(p)$ for $p \leq p^G$ (as no censorship exist in either case) and (10), we have $\lim_{p \downarrow p^G} U^{\mathbf{0}\prime}(p) <$

$\lim_{p \downarrow p^G} U^{\tilde{\boldsymbol{x}}\prime}(p)$. Since the value functions are continuously differentiable in $(p^G, \hat{p}^B)$, the above relation holds for some neighborhood $(p^G, \check{p})$; thus in that neighborhood $U^{\tilde{\boldsymbol{x}}}(p) - U^{\mathbf{0}}(p)$ is increasing in p. We have $U^{\tilde{\boldsymbol{x}}}(p) > U^{\mathbf{0}}(p)$ for $p \in (p^G, \check{p})$.

Suppose, by contradiction, that there exists a smallest $\tilde{p} \in (p^G, \hat{p}^B]$ such that $U^{\tilde{\boldsymbol{x}}}(\tilde{p}) \leq U^{\mathbf{0}}(\tilde{p})$. Thus, $U^{\tilde{\boldsymbol{x}}}(p) > U^{\mathbf{0}}(p)$ for $p \in (p^G, \tilde{p})$ and $U^{\tilde{\boldsymbol{x}}}(\tilde{p}) = U^{\mathbf{0}}(\tilde{p})$ by continuity. Note that we must have $U^{\mathbf{0}\prime}(\tilde{p}) \geq U^{\tilde{\boldsymbol{x}}\prime}(\tilde{p})$. Otherwise, if $U^{\mathbf{0}\prime}(\tilde{p}) < U^{\tilde{\boldsymbol{x}}\prime}(\tilde{p})$, then it holds in a neighborhood of $\tilde{p}$ by continuity so that $U^{\tilde{\boldsymbol{x}}}(p) - U^{\mathbf{0}}(p)$ is increasing in that neighborhood of $\tilde{p}$, which, together with $U^{\tilde{\boldsymbol{x}}}(p) > U^{\mathbf{0}}(p)$ for $p \in (p^G, \tilde{p})$, implies that $U^{\tilde{\boldsymbol{x}}}(\tilde{p}) > U^{\mathbf{0}}(\tilde{p})$, a contradiction. Thus, we have $U^{\mathbf{0}\prime}(\tilde{p}) \geq U^{\tilde{\boldsymbol{x}}\prime}(\tilde{p})$. Comparing (8) and (9) at $p = \tilde{p}$ and using $U^{\mathbf{0}\prime}(\tilde{p}) \geq U^{\tilde{\boldsymbol{x}}\prime}(\tilde{p})$, we have

$$\beta(\tilde{p})\,[U^{\mathbf{0}}(\tilde{p}) - U^{\mathbf{0}}(j(\tilde{p}))] + \beta^G\,\tilde{p}\,(1-\tilde{p})U^{\mathbf{0}\prime}(\tilde{p}) - (1-\tilde{p})\,\beta^B[U^{\mathbf{0}}(\tilde{p}) - U^{\mathbf{0}}(0)] \leq 0,$$

which contradicts (10). We have $U^{\tilde{\boldsymbol{x}}}(p) > U^{\mathbf{0}}(p)$ for any $p \in (p^G, \hat{p}^B]$.

(b) We show the type B agent has a higher payoff in the MPE than in the NCB when $j(p_0) \leq p^*$ and $p_0 \in (p^G, \hat{p}^B]$.

In both the MPE and the NCB, the type B agent does not censor bad news when $p \leq \hat{p}^B$. Thus, in both cases, she is dismissed either when a piece of bad news arrives ($j(p) \leq p^*$) or when the belief drifts down to p^* in the absence of news. In the MPE, the belief drifts down according to the $\dot{p} = d(p, 1, 0)$ for $p \in (p^G, \hat{p}^B]$ and $\dot{p} = d(p, 0, 0)$ for $p \in [p^*, p^G]$. But in the NCB, the belief drifts down according to $\dot{p} = d(p, 0, 0)$ for $p \in [p^*, \hat{p}^B]$. Since the drifting rate is higher in the NCB, i.e., $d(p, 0, 0) < d(p, 1, 0) < 0$, the belief drifts down slower in the MPE than in the NCB, and the type B agent has a higher payoff in the MPE than in the NCB.

(c) Finally, we show the type G agent has a higher payoff in the MPE than in the NCB when $j(p_0) \leq p^*$ and $p_0 \in (p^G, \hat{p}^B]$.

In the MPE, fix $p \in (p^G, \hat{p}^B]$ and let s^1 be the time it takes for the belief to drift from

p to p^G. As $V_G^{T_{p^*}}(p^G) = \Pi_G(p^G) = \rho\ c$ by Lemma 4, the type G agent's payoff is

$$\begin{aligned} V_G^{T_{p^*}}(p) =& \int_0^{s^1} e^{-(\rho+\gamma)t} \left(\rho\ (w - \beta^G\ c) + \gamma\ w\right) \mathrm{d}t + e^{-(\rho+\gamma)s^1} V_G^{T_{p^*}}(p^G) \\ =& \frac{\rho\ (w - \beta^G\ c) + \gamma\ w}{\rho + \gamma} \left(1 - e^{-(\rho+\gamma)s^1}\right) + e^{-(\rho+\gamma)s^1} \rho\ c. \end{aligned}$$

In the NCB, fix $p \in (p^G, \hat{p}^B]$ and let s^0 be the time it takes for the belief to drift from p to p^G. Note that $j(p) \leq p^*$ gives $\Pi_G(j(p)) = 0$. Her payoff in the NCB is

$$\begin{aligned} \Pi_G(p) =& \int_0^{s^0} e^{-(\rho+\beta^G+\gamma)t} \left(\rho\ w + \beta^G\ \Pi_G(j(p)) + \gamma\ w\right) \mathrm{d}t + e^{-(\rho+\beta^G+\gamma)s^0} \Pi_G(p^G) \\ =& \frac{(\rho+\gamma)\ w}{\rho + \beta^G + \gamma} \left(1 - e^{-(\rho+\beta^G+\gamma)s^0}\right) + e^{-(\rho+\beta^G+\gamma)s^0} \rho\ c. \end{aligned}$$

First, $c < \bar{c}$ implies that $\left(\rho\ (w - \beta^G\, c) + \gamma\, w\right) / (\rho+\gamma) > w\,(\rho+\gamma)\,/(\rho+\beta^G+\gamma) > \rho\, c$. Second, according to the definitions of s^1 and s^0, we have

$$\left(\gamma - \beta^B\right)\ s^1 = \ln\left[\frac{p_0}{1-p_0}\frac{1-p^G}{p^G}\right] = \left(\gamma + \beta^G - \beta^B\right)\ s^0,$$

which implies $(\rho+\gamma)\ s^1 > \left(\rho + \beta^G + \gamma\right)\ s^0$. Thus, $V_G^{T_{p^*}}(p) > \Pi_G(p)$ for $p \in (p^G, \hat{p}^B]$. □

References

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

Besley, Timothy and Andrea Prat (2006), "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability." *The American Economic Review*, 96, 720–736.

Bhaskar, V and Caroline Thomas (2019), "Community enforcement of trust with bounded memory." *The Review of Economic Studies*, 86, 1010–1032.

Blackwell, David (1953), "Equivalent comparisons of experiments." *The annals of mathematical statistics*, 265–272.

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

Board, Simon and Moritz Meyer-ter Vehn (2021), "A reputational theory of firm dynamics." *Unpublished paper, UCLA.[430]*, 2381–2462.

Daughety, Andrew F. and Jennifer F. Reinganum (2018), "Evidence Suppression by Prosecutors: Violations of the Brady Rule." *SSRN Electronic Journal*.

Dye, Ronald A. (2017), "Optimal disclosure decisions when there are penalties for nondisclosure." *The RAND Journal of Economics*, 48, 704–732.

Edmond, Chris (2013), "Information manipulation, coordination, and regime change." *Review of Economic Studies*, 80, 1422–1458.

Egorov, Georgy, Sergei Guriev, and Konstantin Sonin (2009), "Why resource-poor dictators allow freer media: A theory and evidence from panel data." *American Political Science Review*, 103, 645–668.

Ekmekci, Mehmet, Leandro Gorno, Lucas Maestri, Jian Sun, and Dong Wei (2020), "Learning from Manipulable Signals." *arXiv:2007.08762 [econ]*.

Eraslan, Hulya and Saltuk Ozerturk (2017), "Information Gatekeeping and Media Bias." Technical report.

Gehlbach, Scott and Konstantin Sonin (2014), "Government control of the media." *Journal of Public Economics*, 118, 163–171.

Guriev, Sergei M. and Daniel Treisman (2018), "Informational Autocracy: Theory and Empirics of Modern Authoritarianism." Technical report.

Hauser, Daniel N (2017), "Promoting a reputation for quality." Technical report.

Hauser, Daniel N. (2021), "Censorship and Reputation." *American Economic Journal: Microeconomics*, forthcoming.

Kartik, Navin, Frances Xu Lee, and Wing Suen (2017), "A Theorem on Bayesian Updating and Applications to Communication Games." Technical report.

Keller, Godfrey and Sven Rady (2015), "Breakdowns." *Theoretical Economics*, 10, 175–202.

Keller, Godfrey, Sven Rady, and Martin Cripps (2005), "Strategic Experimentation with Exponential Bandits." *Econometrica*, 73, 39–68.

Kolotilin, Anton, Tymofiy Mylovanov, and Andriy Zapechelnyuk (2019), "Censorship as Optimal Persuasion." SSRN Scholarly Paper ID 3501474.

Kovbasyuk, Sergey and Giancarlo Spagnolo (2021), "Memory and markets." *Available at SSRN 2756540*.

Kuvalekar, Aditya and Elliot Lipnowski (2020), "Job insecurity." *American Economic Journal: Microeconomics*, 12, 188–229.

Lorentzen, Peter (2014), "China's Strategic Censorship." *American Journal of Political Science*, 58, 402–414.

Marinovic, Iván, Andrzej Skrzypacz, and Felipe Varas (2018), "Dynamic certification and reputation for quality." *American Economic Journal: Microeconomics*, 10, 58–82.

Milgrom, Paul and John Roberts (1986), "Price and advertising signals of product quality." *Journal of Political Economy*, 94, 796–821.

Povel, Paul and Günter Strobl (2019), "Lying to speak the truth: Selective manipulation and improved information transmission." SSRN Scholarly Paper ID 3488734.

Presman, E. L. (1991), "Poisson Version of the Two-Armed Bandit Problem with Discounting." *Theory of Probability & Its Applications*, 35, 307–317.

Redlicki, Jakub (2017), "What Drives Regimes to Manipulate Information: Criticism, Collective Action, and Coordination." Technical report.

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

Smirnov, Aleksei and Egor Starkov (2021), "Bad News Turned Good: Reversal Under Censorship." *American Economic Journal: Microeconomics*, forthcoming.

Sun, Yiman (2021), "A dynamic model of censorship." SSRN Scholarly Paper 4078301, URL https://papers.ssrn.com/abstract=4078301.

Thomas, Caroline (2016), "Career concerns and policy intransigence - a dynamic signalling model." Department of Economics Working Papers 161228, The University of Texas at Austin, Department of Economics.

Varas, Felipe, Iván Marinovic, and Andrzej Skrzypacz (2020), "Random inspections and periodic reviews: Optimal dynamic monitoring." *The Review of Economic Studies*, 87, 2893–2937.