Signaling Good Faith

Andrew McClellan and Daniel Rappoport[*]

February 15, 2024

Abstract

A decision maker (DM), who will take a binary decision, cares about his reputation for being "good", i.e., wanting to accord his action choice with public evidence, as opposed to being "bad", i.e., having a fixed partisan agenda regardless of the evidence. While the decision is taken after evidence is realized, the DM has the option to take a "stand" beforehand, i.e., to communicate his intentions via a cheap-talk message. A wide range of equilibria exist and are characterized by how much the good DM reveals about his standards at this initial communication stage. The most informative of these is *ex-ante signaling* which sees the DM effectively commit to a contingent plan as a function of the realized evidence. Our main theorem states that ex-ante signaling minimizes the probability that the DM follows his partisan agenda across all equilibria. We also consider how the design of the investigation—the distribution of evidence—affects outcomes in the presence of communication prior to its realization. The investigation mitigates the DM's partisan behavior more when the distribution of evidence is "unpredictable" as this hinders the DM in targeting his announced contingent plan.

1. Introduction

Across a wide array of institutions, individuals' decisions are scrutinized for whether they align with public objectives. This often comes down to whether the decision accords with some public evidence as opposed to the potentially biased agenda of the decision

[*] McClellan: University of Chicago, Booth School of Business. Rappoport: University of Chicago, Booth School of Business. We would like to thank Steve Callander, Laura Doval, Piotr Dworczak, Wioletta Dziuda, Ben Golub, Alex Frankel, Marina Halac, Thomas Jungbauer, Emir Kamenica, Navin Kartik, Jacob Leshno, Alessandro Pavan, Doron Ravid and Joe Root.

maker. While, the decision is only made after the evidence is made public, in many contexts, it takes time for the evidence to be realized. This gives the opportunity for the decision maker to "take a stand", or state their intentions, before the uncertainty is resolved. Our paper explores how such "ex-ante" signaling efforts affect the choices of these reputationally concerned agents. To make this concrete, consider the following examples.

1. Political scandals often initiate investigations which are then followed by a decision to censure, impeach, or expel the involved politician. Moderate representatives care about their reputation for *integrity*, i.e., wanting to decide based on the objective evidence instead of partisan objectives. Before the investigation concludes, these representatives can make informative statements about how they will decide or defer and only signal with their eventual decision. An example is the impeachment inquiry initiated by Speaker of the House Pelosi in September 2019 concerning a call between President Trump of the United States and President Zelensky of Ukraine.[1] Despite the inquiry being ongoing, various pivotal senators were interviewed and asked to weigh in about their intended impeachment votes.[2]

2. Many government organizations such as the Federal Trade Commission (FTC) or Food and Drug Administration (FDA) are tasked with approval decisions. The officials involved may have their idiosyncratic preferences about each issue, but also have a desire to project integrity rather than appearing to seek a particular outcome regardless of the specifics. These organizations can declare their standards for approval up front or decide on a case-by-case basis after observing the evidence. For example, in 2020, national drug regulatory agencies were eager to approve a safe COVID-19 vaccine but faced credibility worries that they were rushing the process. The FDA laid out a specific efficacy threshold in clinical trials for approval, whereas the European Union's counterpart deliberately provided no such lower bound (Singh

[1] A clear example of these politicians' concern for appearing non-partisan is that the senate voted unanimously to release the transcript of the call and whistleblower complaint despite President Trump and many Republican party operatives urging against it. (See Mcardle, Mairead (2019) "Senate GOP Unanimously Approves Dem Resolution Calling for Release of Whistleblower Complaint" *National Review*, September 24). More broadly, politicians are frequently rewarded for appearing non-partisan, e.g., John Hickenlooper benefited from taking bipartisan positions in the 2020 presidential election (see Bernstein, Jonathan (2013) "Understanding the importance of a reputation for bipartisainship," *Washington Post*, July 24.)

[2] Some made informative statements: Senator Romney reported that the transcript was "troubling" and Senators Graham, Ernst, and Toomey reported their doubts that convincing evidence of a quid pro quo would turn up. Others refused to comment: Senator Sasse criticized his colleagues for jumping to conclusions, and pledged to wait and see until the investigation was concluded. See Costa, Roberts (2019) "Cracks emerge among Senate Republicans over Trump urging Ukrainian leader to investigate Biden" *Washington Post*, September 25.

and Upshur (2021)). Similar issues arise in the decisions of other government agencies, such as the FTC deciding whether to approve a merger.

3. University admissions committees would like to appear as though their decisions are based on academic potential despite being pressured to consider the legacy statuses, donations, or other non-academic features of their applicants. Many American universities practice "holistic" admissions and will not give exact criteria for admission. The lack of transparency in holistic admissions has been criticized for facilitating higher admission rates for unqualified applicants.[3] One alternative is to publicize specific criteria for admission, a practice common in universities throughout Europe and Asia.[4,5]

We study three important questions in such settings. First, how informative can communication be prior to the revelation of evidence, e.g., how much can politicians distinguish their standards during an investigation? Second, how does informative communication about hypothetical plans affect outcomes, e.g., would we expect that Republican senators who indicate conditions for impeachment up front convict more or less than those who wait and see, and would universities admit more donor applicants were they to publicize admissions standards rather than use holistic admissions? Third, how does the type of uncertainty about the evidence affect outcomes in the presence of communication? This is important in settings where the investigation is the choice of some "investigator," e.g., how should Speaker Pelosi conduct the impeachment inquiry to get the most Republican senators to convict, and how should firms provide evidence about potential mergers to the FTC to ensure the highest chance of approval?

Our model features a single decision maker (DM), and an inactive Bayesian observer. The game consists of two stages: a communication stage and a decision stage. At the communication stage, the DM, sends a cheap-talk message about his preferences.[6] At the

[3] Pinker, Steven (2014) "The Trouble With Harvard" *The New Repbulic*, September 4.

[4] Frisancho and Krishna (2016) describes how admission to Delhi University is automatic if an applicant's exam score crosses a social group dependent cut-off.

[5] Other examples abound. Many academic journals have required or offered preregistration (see Warren, Matthew (2018) "First analysis of 'pre-registered' studies shows sharp rise in null findings," *Nature*, October 24.), i.e., specifying the design of the study and conditions for acceptance before the data is observed or analyzed. While preregistration is often discussed in terms of its incentive effects on authors, it will also have effects on which papers are selected by reputationally concerned editors.

[6] While cheap-talk communication fits statements made by pivotal representatives during political investigations, our applications to regulatory agencies are better fit by endowing the DM with the ability to commit to a contingent plan. As we show in Subsection 6.1, the focal equilibrium of our model with cheap talk also prevails in the alternative model where the DM is endowed with commitment, and so our main results apply to both cases.

decision stage, the evidence $e \in \mathbb{R}$ is realized and the DM chooses a binary action, either $a = 1$ or $a = 0$. In addition to the evidence, the DM's preferences over the action also depend on his private type, which is either bad — a "partisan"— or good — a "non-partisan." The non-partisan would like to accord his action with the evidence and his privately known leniency ℓ; more specifically he would like to take $a = 1$ more if e is higher or ℓ is lower. On the other hand, the partisan does not care about taking the right decision and suffers a constant disutility from taking $a = 1$ regardless of the evidence. The leniency can be interpreted in two ways: (i) as idiosyncratic heterogeneity in standards for this particular decision, e.g., different politicians have different views about the appropriate extent of executive power while still maintaining integrity, or (ii) private non-verifiable information about the "right" standard, e.g., FDA officials have specific expertise about the drug being considered. Finally, the DM also cares about his reputation for being a non-partisan in the eyes of the observer who sees the DM's cheap-talk message, the realized evidence, and the DM's chosen action.

In the first part of our paper, we analyze the model for a fixed exogenous distribution of evidence or "investigation". In the second part, we introduce an investigator who specifies the investigation subject to constraints. In our main specification, the investigator seeks to maximize the probability of conviction.[7]

We first show that each equilibrium can be pinned down by how much information the communication stage transmits about the leniency of the non-partisan type. Two salient cases are the extremes: (i) when the communication stage involves babbling, and all signaling is done at the decision stage, and (ii) when the communication stage perfectly communicates his leniency, and there is no additional signaling at the decision stage. We term these equilibria ex-post and ex-ante signaling respectively. It turns out that ex-ante signaling is tantamount to the DM committing to a contingent plan as a function of the evidence revealed, e.g., stating "I will convict if the evidence meets ... standard". Conversely, ex-post signaling could be interpreted as the DM saying "I will not speculate on hypotheticals".

It is not apparent how changing the equilibrium would affect outcomes: if anything, the effective "commitment power" provided by ex-ante signaling would seem to benefit the DM and perhaps allow the partisan choose his preferred action more frequently. However, Theorem 1 shows that ex-ante signaling has the highest probability of $a = 1$ across all equilibria. In addition, ex-ante signaling delivers a higher probability of $a = 1$ than ex-post signaling for *every* evidence realization. This means that politicians who answer interviewers' questions will tend to break with their party more than those who successfully "dodge

[7] Subsection 6.3 shows that many of our conclusions are robust to the case in which the investigator's preferences are evidence dependent.

the cameras"; government agencies that specify approval criteria up-front will go against their appointers' political interests more than those who decide on a case-by-case basis; and setting clear admissions criteria will lead to more meritocratic admissions decisions relative to holistic admissions. The broad intuition is simple: before the realization of evidence, the DM is willing to make stronger claims in order to attain a higher reputation because there are many evidence realizations under which these stronger claims do not require a different action than weaker ones. Conversely, under ex-post signaling, after a "pivotal" evidence realization occurs, obtaining a high reputation requires taking the high action with probability one. While this simple reasoning is sufficient to prove the result with two leniency ℓ types, the full intuition revolves around the "convexity of reputation" which we elaborate on in Subsection 4.2.

We then move to the investigator's design problem. In our main specification we consider the investigator flexibly choosing an information structure about a binary state, e.g., guilt or innocence of a politician. We focus on the ex-ante signaling equilibrium and characterize the investigation that maximizes the probability of $a = 1$. Even in the absence of a designer, our characterization speaks to how the distribution of evidence affects outcomes when the DM takes informative stands, i.e., under ex-ante signaling.

One main takeaway is that the optimal investigation admits no mass points unlike that seen in familiar Bayesian persuasion design problems. This is because the DM responds to changes in the investigation by altering which leniency he claims at the communication stage. This is important from the investigator's perspective: we show that, across all investigations, the investigator's interests (i.e. maximizing probability of conviction) and partisan's interest are exactly misaligned *in equilibrium*. The implication is that the investigator wants to imbue as little predictability as possible to avoid "targeting" from the partisan, i.e., declaring thresholds just above where evidence is likely to be.

The layout of the paper is as follows. Section 2 describes our model. Section 3 describes basic properties of and categorizes all equilibria. Section 4 states our main results comparing equilibria. Section 5 characterizes the investigator's optimal investigation and describes comparative statics. And lastly, Section 6 discusses equilibrium selection, alternative commitment and timing assumptions, and various robustness results.

1.1. Literature Review

We add to the literature studying the impact of reputation concerns (e.g., Holmström (1999), Scharfstein and Stein (1990), Prendergast and Stole (1996)), in particular those papers that include cheap talk (e.g., Sobel (1985), Morris (2001), Ottaviani and Sorensen (2006a), Ottaviani and Sorensen (2006b)). Our decision maker's preferences are closest to those in Morris (2001). He studies an informed sender who seeks a reputation for being responsive to the state—similar to our non-partisan—rather than having a state-independent preference—similar to our partisan. The main difference in our preferences is that we have heterogeneity in the "good" type's preferences, i.e., there is a non-degenerate distribution of leniency types. Importantly, communication has no value in our model when the leniency type distribution is degenerate, but can otherwise change equilibrium outcomes in a significant way.[8]

We are also connected to the costly signaling literature initiated by Spence (1973). As in Bénabou and Tirole (2006), Esteban and Ray (2006), and Frankel and Kartik (2019), the multidimensional type of the DM—namely preference heterogeneity of the non-partisan in our model—precludes separating equilibria. Frankel and Kartik (2022) and Ball (2022), among others bring a design perspective to such settings, studying how to design scoring systems in the presence of strategic manipulation.[9,10]

Previous works in the signaling and communication literatures have studied the impact of exogenous signals about a sender's private type. Daley and Green (2014) study how, in a Spence signaling model, the sender's equilibrium actions are impacted by the revelation of a informative signal on his type after his costly signaling action is chosen. Similar models are studied by Kurlat and Scheuer (2021), who allow receivers to differ in the informativeness of their signal on the sender's type, and Alós-Ferrer and Prat (2012), who allow exogenous signals to be revealed over time via on-the-job learning about the sender's type. Chen (2012) studies how the timing of cheap-talk communication by a privately informed sender relative to an informative signal shapes what is communicated and compares outcomes when communication before the public signal to when it occurs after.

Our results also speak to the literature on the impacts of transparency in the presence of reputational concerns. Papers such as Prat (2005) and Levy (2007) study how a (purely) reputationally motivated agent's action changes when they know their action will be revealed relative to the action being hidden (i.e., transparency increases). Our paper instead

[8] Other papers study different reputation incentives with related interpretations. The advisor in Durbin and Iyer (2009) seeks a reputation for being "incorruptible" (i.e., valuing bribes relatively less as compared with outcomes). Olszewski (2004) and Acemoglu et al. (2013) study a sender who prefers to be seen as honest. In settings with a biased advisor (i.e., one who does not make decisions), a positive reputation for competence (e.g., as in Prendergast (1993), Prat (2005), and Li (2007)) induces a preference for different actions to be taken based on the state or evidence.

[9] Rappoport (2022) considers designing optimal delegation policies for agents engaged in costly signaling.

[10] Ali and Bénabou (2020) considers a costly signaling model where there is a common and, more or less, public variable that affects signaling incentives, but there is no communication prior to its revelation. Kartik and Van Weelden (2018) also features communication before the revelation of uncertainty and subsequent costly signaling, but considers different material and reputation incentives of the DM.

studies how communicating the decision maker's *strategy* impacts actions choices when the decision maker has both material and reputational concerns and actions are always revealed. Increased transparency in our model corresponds to more informative communication about the agent's strategy (i.e., do they specify their strategy before evidence is realized).[11]

Our study of optimal investigations ties the model to the information design literature started by Kamenica and Gentzkow (2011). The impact of uncertainty over the receiver's type on information disclosure, which Kamenica and Gentzkow (2011) show can be handled using their concavication approach, has also been studied in papers such as Alonso and Câmara (2016), Kolotilin et al. (2017) and Kolotilin (2018). We differ from these previous papers by considering how the design of information impacts the DM's choices prior to evidence being realized. Recent papers such as Boleslavsky and Kim (2018) and Zapechelnyuk (2020) study information design in the presence of moral hazard problem while Hörner and Lambert (2020) study feedback design in a dynamic career concerns model. Boleslavsky and Kim (2018) develop concavification techniques analogous to those used in Kamenica and Gentzkow (2011) in the presence of moral hazard. Our model, in contrast, looks at the impact of the investigation on communication strategies (and their subsequent impact on action choices). The impact of information disclosure where agents are concerned with beliefs on their type also arises in mechanism design models with limited commitment (e.g., Doval and Skreta (2022)).

Lastly, there is of course a broad political economy literature concerning partisanship and partisan reputations. In these models (e.g., in Maskin and Tirole (2004), Acemoglu et al. (2013), Kartik and Van Weelden (2018), and Agranov (2016)) electoral incentives push against appearing "partisan," in the sense of having extreme policy preferences relative to the median voter. This reputation incentive could be included in our framework by encoding higher reputation payoffs for some leniency types, namely those close to the median voter, without changing many of our main intuitions (see Section 6). Fox and Van Weelden (2010) models partisans as politicians who want to prop up the reputation of other officials in their own party in addition to their own. Bussing and Pomirchy (2022) consider a similar definition of partisan reputations to our paper in the context of political oversight and checks and balances, but among other differences, do not focus on communication.

[11] Our comparison between ex-ante signaling, which specifies a complete contingent plan, and ex-post signaling, which waits until the evidence is realized echoes themes from the literature on incomplete contracts initiated by Grossman and Hart (1986) and Hart and Moore (1988). There it is assumed to be arbitrarily costly to specify complete contracts/contingent plans. Subsequent papers (e.g., Aghion et al. (1994)) have studied the design of more complex contracts to avoid the inefficiencies caused by contractual incompleteness; our results complement these by highlighting how communication and high reputation incentives can overcome the inability to commit to fully specified contingent plans.

Related incentives also arise in models of the media (e.g., Shapiro (2016)) where journalists want to appear "objective."

2. Model

Overview There are three players: an investigator, a decision maker (DM), and a Bayesian observer. The DM eventually chooses $a = 1$ or $a = 0$. His preferences over this decision depend on his privately known type $\theta \in \Theta$ and the realized evidence $e \in E \equiv \mathbb{R}$. The DM also values his reputation in the eyes of the observer.

The timing is as follows. In the initial communication stage, the evidence is unknown and the DM only knows its CDF F; we assume $\int_E edF(e)$ is well-defined and finite. The DM sends a cheap-talk message $m \in M$ to the observer, where M is some sufficiently large metrizeable space.[12] After the message is sent, the decision stage begins: the evidence e is publicly revealed and then the DM chooses an action a. The observer sees the DM's message and action choice in addition to the realized evidence and forms beliefs, after which payoffs are realized.

Our paper is broken into two main parts. The first part of the paper analyzes the case where the investigation F is exogenous and arbitrary, i.e., the investigator is inactive. The second part of the paper considers an investigator who can design F, with restrictions, to suit his interests.

Preferences The DM can either be a partisan (P) or a non-partisan (N). The prior probability of N types is $q \in (0,1)$. Non-partisan DMs have heterogeneous and privately known leniency $\ell \in \mathbb{R}$. Conditional on being a non-partisan, the distribution of ℓ has CDF G with $L \equiv \mathrm{Supp}(G)$. We assume for expositional convenience that either F or G is atomless. We will refer to non-partisans with leniency ℓ as "ℓ types." Accordingly, the set of types is $\Theta = L \cup \{P\}$ with prior distribution $\nu_0 \in \Delta(\Theta)$.[13] The DM also values his reputation in the eyes of the observer of being an N type. The utility of type $\theta \in \Theta$ from taking action a, given evidence e, and public belief μ that he is type N is given by

$$u(\theta, e, a, \mu) \equiv \begin{cases} -ac + \rho\mu & \text{if } \theta = P, \\ a(e-\ell) + \rho\mu & \text{if } \theta = \ell, \end{cases}$$

[12] We will assume $|\Delta(\Theta)| \leq |M|$ where, for a metrizable space Y, we let $\Delta(Y)$ denote the set of all Borel probability measures over Y, endowed with the weak* topology.

[13] Then $q = \nu_0(L) = 1 - \nu_0(P)$ and $G(\ell) = \nu_0(\{\ell' : \ell' \leq \ell\}|\theta \in L)$.

Partisan DMs always want to choose $a = 0$ and their disutility $c > 0$ from $a = 1$ is independent of the evidence realization. N types prefer $a = 0$ more if (i) the evidence is less convincing (e is lower), or (ii) they are more lenient (ℓ is higher).[14] Subsection 2.1 elaborates, but our leading interpretation of leniency is that it is private non-verifiable information about the "correct" evidence threshold for action $a = 1$. This also justifies the absence of ℓ in the P type's utility for the same reason that e does not appear: the P type does not care about taking the right decision.[15] The weight $\rho > 0$ parameterizes how much the DM values reputation. We refer to the first component of the payoff that depends on the action as the **material payoff** and $\rho\mu$ as the **reputation payoff**. We assume that reputation incentives are strong in the following sense.

Assumption 1. $\rho > 2\max\{\frac{c}{q}, \frac{c}{1-q}\}$.

Broadly, this assumption guarantees that the reputation incentives can be strong enough to convince P to choose $a = 1$. Note that if $\rho < c$, then P will never choose $a = 1$. Assumption 1 is stronger and, as we will show, ensures that, given any public history, P will choose $a = 1$ with positive probability if some ℓ types do as well.

For our main specification, the investigator maximizes the probability of $a = 1$, namely his utility is equal to a. In Subsection 6.3 we extend many of our main takeaways to a model where the investigator's preferences over a depend on e.

Strategies and Equilibrium We study perfect Bayesian equilibria with an additional refinement formalized below—hereafter, simply equilibria. An equilibrium $\mathcal{E}$ consists of a communication-stage strategy $\sigma : \Theta \to \Delta(M)$, a decision-stage strategy $\zeta : M \times E \times \Theta \to \{0, 1\}$, an interim belief after the messaging stage $\nu_1 : M \to \Delta(\Theta)$, and a final belief after the decision stage $\nu_2 : M \times A \times E \to \Delta(\Theta)$, such that for all $\theta \in \Theta$, $m \in M$ and $e \in E$,

1. ν_1 is obtained from σ using Bayes rule.[16]

[14] The utility function over actions of N types is assumed to be $a(e - \ell)$ for convenience. Our results still hold (with notational tweaks) if the utility difference between $a = 1$ and $a = 0$ is increasing in e and decreasing in ℓ.

[15] Nonetheless this introduces an asymmetry between N types and P types in our model in that only N types have privately observed heterogeneity in their preferences. One could envision a model that also endowed P with unobserved heterogeneity in his disutility from taking $a = 1$ denoted c. Such heterogeneity tends to place limits on the amount of informative communication at the communication stage. For example, if there is only one ℓ type, and many c types, one can show that the unique equilibrium involves babbling at the communication stage. Thus, our model omits heterogeneity in c in order to most parsimoniously study pre-play communication.

[16] That is, for all Borel $\hat{\Theta} \subseteq \Theta$ and $\hat{M} \subseteq M$, $\int_{\hat{\Theta}} \sigma(\hat{M}|\theta) d\nu_0(\theta) = \int_{\hat{M}} \nu_1(\hat{\Theta}|m) \int_{\Theta} d\sigma(m|\theta) d\nu_0(\theta)$

2. ν_2 is obtained from ζ using Bayes rule with prior $\nu_1(\cdot|m)$.[17]

3. $\sigma(M_\theta|\theta) = 1$ where $M_\theta \equiv \arg\max_{m\in M} \int_E \left(\max_{a\in\{0,1\}}\ u(\theta, e, a, \nu_2(L|m, e, a))\right) dF(e)$.

4. $\zeta(A_{\theta,m,e}|\theta, m, e) = 1$ where $A_{\theta,m,e} \equiv \arg\max_a\ u(\theta, e, a, \nu_2(L|m, e, a))$.

In addition, we impose a version of the D1 refinement à la Cho and Kreps (1987) and Ramey (1996). Let $\Theta_m \equiv \text{Supp}(\nu_1(\cdot|m)) \subseteq \Theta$ be the support of the interim belief on the DM's type following message m but before an action is chosen. We impose the D1 refinement at the decision stage, after evidence has been realized and message m has been sent, where the type space is Θ_m.[18] In our framework this refinement simplifies to the following: if, after sending message m and observing evidence e, the DM takes an off-path action, the observer believes the DM to be the type(s) in Θ_m who would benefit the most in terms of their material payoff from this deviation relative to their equilibrium payoffs.[19]

We begin by defining some useful notation. Let $U_\theta^{\mathcal{E}}(F)$ be the expected utility of type θ given investigation F and equilibrium $\mathcal{E}$.[20] Let $v^{\mathcal{E}}(e, F)$ be the probability of action $a = 1$ given evidence realization e, investigation F, and equilibrium $\mathcal{E}$, and let $V^{\mathcal{E}}(F) \equiv \int_E v^{\mathcal{E}}(e, F)dF(e)$ be the associated ex-ante probability of $a = 1$ (i.e., the investigator's expected utility).

The **equilibrium outcomes** associated with equilibrium $\mathcal{E}$ are the profile of type-dependent expected utilities and probability of action $a = 1$ as a function of the evidence, i.e., given by $(\{U_\theta^{\mathcal{E}}(F)\}_{\theta\in\Theta}, \{v^{\mathcal{E}}(e, F)\}_{e\in E})$. Two equilibrium outcomes are equivalent if $\{U_\theta^{(\cdot)}(F)\}_{\theta\in\Theta}$ and $\{v^{(\cdot)}(e, F)\}_{e\in E}$ are the same for a probability one set of types and evidence realizations respectively. With some abuse of terminology, we say a set of equilibria admit a **unique** equilibrium outcome if the associated set of equilibrium outcomes are all equivalent to each other.

[17] That is, for all Borel $\hat{\Theta} \subseteq \text{Supp}(\nu_1(\cdot|m))$, $\int_{\hat{\Theta}} \zeta(a|\theta, m, e)d\nu_1(\theta|m) = \nu_2(\hat{\Theta}|m, e, a) \int_\Theta \zeta(a|\theta, m, e)d\nu_1(\theta|m)$ and $\text{Supp}(\nu_2(\cdot|m, e, a)) \subseteq \text{Supp}(\nu_1(\cdot|m))$.

[18] Because our game consists of a communication stage prior to the revelation of an uncertain e, it does not fit in the static signaling games studied in the literature. We are not aware of existing notions that formalize this natural "ex-interim D1" refinement. Another alternative would be to use an "ex-ante D1" refinement, i.e., with the full type space Θ. One can show that in our model this approach yields a less expositionally convenient but essentially identical set of equilibria: every ex-ante D1 equilibrium is also an ex-interim D1 equilibrium, and every ex-interim D1 equilibrium outcome is the limit of some sequence of ex-ante D1 equilibrium outcomes.

[19] We provide the formal definition of D1 in the context of our game in the Appendix.

[20] While our outcome variables depend on all model parameters, the dependence on the investigation F and equilibrium $\mathcal{E}$ is made explicit for expositional clarity.

2.1. Discussion

Partisan Preferences: It is important to note that even though a high ℓ type and P both prefer the $a = 0$ for "essentially" all evidence realizations, this does not mean their preferences are equivalent. This perspective ignores the main tradeoff the DM faces between reputational and material payoffs, a tradeoff that makes the intensity of preferences over actions important. If we instead modeled the "bad" P type as a very high ℓ type, then this would mean P prefers to take action $a = 0$ much more than he values reputation. Indeed, this is the interpretation of bad types in the canonical Spence (1973) education model: bad types have a higher cost of education or, *equivalently and indistinguishably*, a lower value for reputation. Most of our applications do not fit well with this interpretation, e.g., it does not seem appropriate to model partisan politicians as being defined by their lack of office motivation, or a corrupt FTC official as not caring about being fired.

Instead, as mentioned earlier, our preferences mirror those in Morris (2001). The distinction between good and bad types is that good types care more about getting the decision "right" than bad types. For extreme evidence realizations, non-partisan types care more about stakes of the decision, whereas partisans care more about reputation. However, for middling evidence realizations, where the stakes of the decision are low for a non-partisan type, this comparison is flipped. Of course, what counts as high-stakes versus low-stakes evidence depends on ℓ, which is the DM's private information.

Reputation for Leniency: We assume that reputational payoffs are purely determined by the observer's belief that $\theta \in L$ rather than their beliefs about which ℓ type the DM may be. This assumption streamlines our exposition and is natural in applications in which ℓ represents the DM's transitory private information or idiosyncratic preferences that are only relevant for the decision at hand. For example, a politician may possess classified information about the relevant scandal. However, in some settings the DM may have have competing reputation concerns to appear as different leniency types; for example, a politician may value appearing to have positions closer to the median voter *in addition* to appearing non-partisan. In Section 6, we show that, under a modified version of Assumption 1, all of our results extend to a model in which the reputation payoff from the observer holding belief $\mu \in \Delta(\Theta)$ is given by $\mathbb{E}[r(\theta)|\theta \sim \mu]$ for some function $r(\cdot)$ such that $r(\ell) > 0 = r(P)\ \forall \ell$.

Commitment Versus Cheap Talk: We assume that the communication stage involves the DM sending a cheap-talk message. However, in many of our motivating examples the DM may have the option or obligation to commit to a contingent plan before the evidence is realized. For example, the FDA can mandate that its officials specify approval criteria prior to the start of clinical trials and university admissions committees can have a policy of prespecifying admissions criteria prior to receiving applications. In addition, agents may be able to "opt to commit," even when they are not forced to, by verifiably delegating the decision or making publicly enforceable statements. As Section 3 elaborates, the current cheap-talk model admits an equilibrium where the DM effectively commits at the communication stage to a contingent plan as a function of the realized evidence. Subsection 6.1 shows that the unique equilibrium outcomes will be the same as this most informative cheap-talk equilibrium when he has access to commitment power.

3. Equilibrium Characterization

This section characterizes equilibrium behavior. First we establish properties that must hold across all equilibria in Lemma 1. Then we taxonomize the set of equilibria in Lemma 2. It will be useful to make statements in terms of induced mappings from evidence to actions, i.e., $x \in \mathcal{X} \equiv \{x' : E \to \{0,1\}\}$. Define thresholds $\tilde{e}_\ell \equiv \ell - c$ and the threshold contingent plan $x_\ell(e) \equiv \mathbb{1}(e \geq \tilde{e}_\ell)$.

Lemma 1. *For any equilibrium $\mathcal{E}$, the following hold:*

1. *The P type positively mixes over all messages sent by N types, i.e., $\sigma(\cdot|P)$ and $\Sigma_N(\cdot) \equiv \int_L \sigma(\cdot|\ell)dG(\ell)$ are mutually absolutely continuous.*
2. *N types choose actions consistent with x_ℓ with probability one, i.e.,*

$$\int_E \int_L \int_M \zeta(x_\ell(e)|\ell, m, e)d\sigma(m|\ell)dG(\ell)dF(e) = 1.$$

3. *After sending message m, the P type positively mixes over the action choices of ℓ types who also send m, i.e.,*

$$\forall m \in M_P\,,\ \nu_1(L|m) > 0 \text{ and } \ \forall e, a,\ \int_L \zeta(a|\ell, m, e)d\nu_1(\ell|m, \theta \in L) > 0 \iff \zeta(a|P, m, e) > 0.$$

The interpretation of the 1st and 3rd point is that P cannot be distinguished from N following any "on-path" history. A key implication is that P is indifferent across mimicking the behavior of any ℓ type, at both the communication and decision stages. These points follow from the high reputation incentives. If a message is sent only by P then it yields an equilibrium reputation and utility of 0 for P. However, P can obtain an expected utility of at least $\rho q - c$ by mimicking the strategy of some ℓ type, which is strictly preferred by

Assumption 1. Conversely, a message that is sent only by ℓ types yields a reputation of 1, so P's equilibrium utility must be at least $\rho - c$. However, P gets at most an expected reputation payoff, and thereby also utility, of ρq from following the equilibrium strategy,[21] which is strictly less than $\rho - c$ again by Assumption 1. The argument for why, after sending message m, P mixes over the actions chosen by ℓ types who also send m is similar, but has to contend with the subtlety that the relevant utility bounds are now dependent on $\nu_1(L|m)$ instead of the prior q. The proof shows that any equilibrium $\nu_1(L|m)$ is close enough to q such that the above argument goes through.

The second point says the ℓ type's action choice as a function of the evidence (almost surely) follows the fixed rule $x_\ell(e)$.[22] To avoid probability one caveats, going forward we refer to the outcome equivalent equilibria where N's actions correspond with $x_\ell(e)$ *everywhere*, i.e., $\forall e \in E,\ \ell \in L$. The ℓ type's action choice is not only constant across equilibria and messages, but also across parameters of the model such as the investigation and the type distribution of the DM. This independence should not be misunderstood as arising because the ℓ types choose their ideal action unaffected by reputation incentives. Indeed, ℓ types engage in "political correctness" (Morris (2001)): in order to signal non-partisanship they select the partisan's dis-preferred action $a = 1$ for $e \in (\ell - c, \ell)$ where they prefer $a = 0$. Instead, x_ℓ is distinguished by the fact that it provides the highest signaling value to the ℓ type: x_ℓ maximizes the utility difference between ℓ and P types across all contingent plans $x \in \mathcal{X}$.

The intuition behind point 2 is as follows. Suppose first that both actions are on path following some evidence realization e. This implies that P mixes over $a = 1$ and $a = 0$. However, the type $\tilde{\ell} \equiv e + c$ has the same preferences as P given e, i.e., he has the same trade off between the cost of $a = 1$ and reputation. Combined with the fact that N's utility for $a = 1$ is decreasing in ℓ, all $\ell > \tilde{\ell}$ must choose $a = 0$ and $\ell < \tilde{\ell}$ must choose $a = 1$, i.e., ℓ types choose actions consistent with x_ℓ. Alternatively, if $a = 0$ (respectively $a = 1$) is off path, then it must be that $\ell > \tilde{\ell}$ (respectively $\ell < \tilde{\ell}$) for every $\ell \in \text{Supp}(\nu_1(\cdot|m))$; otherwise, by D1, the off-path action would be interpreted as originating from the ℓ type that violates these inequalities, and this off-path action would be a profitable deviation for P.

We next identify and categorize the set of equilibrium outcomes. For any equilibrium, the communication stage conveys information about the leniency of the DM conditional

[21] This follows from corollary 2 in Hart and Rinott (2020) which shows that, for any signal structure, and for any state ω, the expected posterior belief of ω conditional on state ω is higher than the prior probability of ω.

[22] The reason for the almost surely caveat is that action choices are not pinned down for evidence-leniency pairs where $e = \tilde{e}_\ell$. However, this set has zero probability given our assumption that either F or G are atomless. Indeed, this is our only reason for making this assumption.

on them being a non-partisan. We call this induced Bayes plausible information structure $\Lambda \in \Delta(\Delta(L))$ the **Leniency Information Structure** (LIS) associated with the equilibrium $\mathcal{E}$.[23] This is formally defined as, for each Borel $H \subset \Delta(L)$, $\Lambda(H) = \int_{m \in M} \mathbb{1}\big(\nu_1(\cdot|m, \theta \in L) \in H\big) d\Sigma_N(m)$.

Lemma 2. *For each LIS, the set of associated equilibria admit a unique equilibrium outcome.*

There are two main takeaways from the lemma. First, equilibrium outcomes can be uniquely described by the associated information the communication stage conveys about the leniency of the DM. Second, *every* LIS is associated with an (potentially different) equilibrium outcome. That is, unlike familiar cheap-talk models (e.g., Crawford and Sobel (1982)), there is no monotonicity restriction on the equilibrium strategies of ℓ types. More importantly, this permissiveness means that in equilibrium the communication-stage message can convey a wide range of information about ℓ, from the perfectly informative LIS where each ℓ sends a different message to the perfectly uninformative LIS where all DM types send the same message. At the beginning of the next section, we further inspect these salient extreme cases.

Lemma 1 and Lemma 2 provide an effective blueprint for constructing an equilibrium. An equilibrium outcome is pinned down by its LIS which can be directly imputed to the messaging strategies of the ℓ types at the communication stage. Each of these ℓ types follow up with x_ℓ at the decision stage no matter which message they initially chose. P mixes over all messages sent by the ℓ types at the communication stage and all on-path follow up contingent plans at the decision stage in order to ensure their own indifference.

The above heuristic for constructing equilibrium strategies is valid because of the following property: if, for some candidate equilibrium strategies, the P type is indifferent across messages, then each ℓ type's incentive are ensured as well. Figure 1 displays the reasoning. Consider $\underline{\ell} < \overline{\ell}$ who send different messages $\underline{m}$ and $\overline{m}$ respectively. Suppose the P type is indifferent between sending $\overline{m}$ and following up with $x_{\overline{\ell}}$ (i.e., using threshold $\tilde{e}_{\overline{\ell}}$), and sending $\underline{m}$ and following up with $x_{\underline{\ell}}$ (i.e., using threshold $\tilde{e}_{\underline{\ell}}$). This means that the expected reputational difference between the latter and the former strategy must be equal to the material utility difference from switching their action choice for $e \in (\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$, i.e., the absolute value of the area $R_1 + R_2$ measured according to the distribution of evidence F. But notice that if type $\underline{\ell}$ considers deviating from $\underline{m}$ and $x_{\underline{\ell}}$ to $\overline{m}$ and $x_{\overline{\ell}}$, they only gain the absolute value of R_1 in material utility which does not compensate them for the reputational loss of $R_1 + R_2$. Analogously if $\overline{\ell}$ considers deviating from $\overline{m}$ followed by $x_{\overline{\ell}}$ to $\underline{m}$

[23] Formally, Bayes-plausibility is satisfied if for all Borel $\tilde{L} \subset L$, $\nu_0(\tilde{L}) = \int_{\mu \in \Delta(L)} \mu(\tilde{L}) d\Lambda(\mu)$.

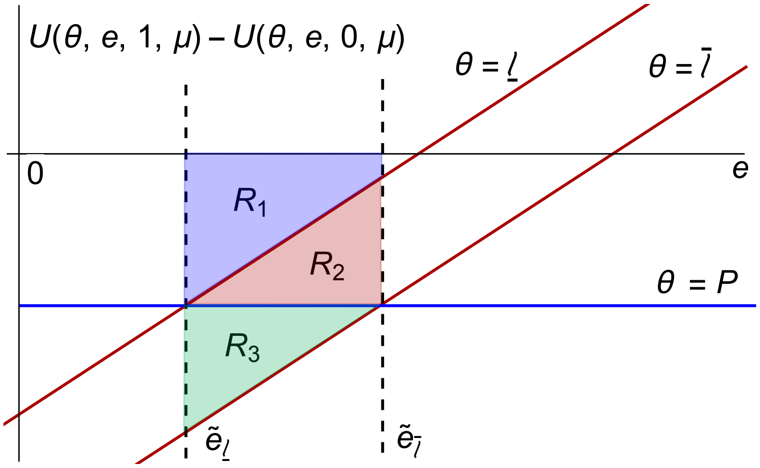


Figure 1: Difference in material utility between $a = 1$ and $a = 0$ for different DM types as a function of evidence.

followed by $x_{\underline{\ell}}$ they lose the absolute value of $R_1 + R_2 + R_3$ in material utility from using the lower threshold, which is greater than the reputational gain $R_1 + R_2$. Thus, P's indifference ensures each ℓ type's incentives.[24]

4. The Effects of Informative Stands

In light of Lemma 2, we refer to equilibria by their associated LIS. We describe the two salient extreme cases below.

Ex-Ante and Ex-Post Signaling: We refer to the equilibrium associated with the perfectly informative LIS as **ex-ante signaling** and denote it as equilibrium α. Under ex-ante signaling, each ℓ type sends a different message m_ℓ. Consistent with Lemma 1, P positively mixes over these messages. After sending m_ℓ, the DM follows x_ℓ at the decision stage. In other words, sending m_ℓ is tantamount to committing to a contingent plan, i.e., saying "I will take action $a = 1$ if and only if $e \geq \tilde{e}_\ell$". While there is still uncertainty about the DM's partisanship following message m_ℓ, the equilibrium has no **residual strategic uncertainty**: there does not exist a positive probability set of m, e for which both actions are on-path after message m and evidence e is realized.

At the other extreme is the equilibrium associated with the uninformative LIS which we term **ex-post signaling** and denote as equilibrium β. Under ex-post signaling the DM

[24] Of course, each ℓ type can consider other follow up contingent plans after deviating at the communication stage. The generalization of the point above is that x_ℓ maximizes the expected utility difference between type ℓ and type P across all contingent plans. The proof of Lemma 2 uses this to show that if P is disincentivized from such deviations then so is ℓ.

"babbles," e.g., regardless of his type, he sends the same message interpreted as "I will wait and see until the investigation concludes." Ex-post signaling admits residual strategic uncertainty under the weak condition that there exist two leniency types ℓ', ℓ'' such that $F(\tilde{e}_{\ell'}) \neq F(\tilde{e}_{\ell''})$. A unique feature of ex-post signaling is that because the communication stage is uninformative, given an evidence realization e, outcomes do not depend on the investigation F, i.e., $v^{\beta}(e, F) \equiv v^{\beta}(e)$ is independent of F (and so we drop the associated dependence).

The above description highlights the extent to which the DM can take "informative stands" before the evidence realizes; under ex-ante signaling, he can effectively publicly commit to his contingent plan. Alternatively, the DM can decide on a case-by-case basis obviating the communication stage. Our main result looks at how different communication protocols impact the probability of taking the action. First, we introduce one more technical condition. We say there is **mild agreement** if for every pair $\ell', \ell'' \in \operatorname{Supp}(G)$, $\exists e \in \operatorname{Supp}(F)$ such that $x_{\ell'}(e) = x_{\ell''}(e)$, i.e. no two ℓ types always choose different actions in equilibrium.

Theorem 1. *Ex-ante signaling delivers the highest probability of $a = 1$ among all equilibria, i.e., $V^{\alpha}(F) \geq V^{\mathcal{E}}(F)$ $\forall \mathcal{E}$. This comparison is strict if $\mathcal{E} \neq \alpha$ has residual strategic uncertainty and there is mild agreement.*

The two actions are only differentiated in the model by the partisan's bias towards $a = 0$. Highlighting the comparison with ex-post signaling, Theorem 1 shows the DM goes against his partisan interests more when he takes the "most informative stands", i.e., pre-specifies his contingent plan, rather than deciding on a case-by-case basis. In terms of the applications, the politician who answers interviewers' questions will tend to break with party more, and universities will admit more donor or legacy applicants when using holistic admissions. Beyond predictive implications, in many contexts it is plausible that whether ex-ante or ex-post signaling outcomes prevail is a design decision which can be informed by Theorem 1. Subsection 6.1 and Subsection 6.2 elaborate, showing how ex-ante signaling outcomes can be achieved.

Depending on the parameters, certain LIS may correspond to the same equilibrium outcomes as ex-ante signaling; for example, all equilibria have the same outcomes if the distribution of evidence is degenerate. However, if equilibrium actions are not completely predictable at the decision stage, then the equilibrium delivers different outcomes than ex-ante signaling; in particular, a strictly lower probability of $a = 1$. *All* imperfectly informative LIS are associated with an equilibrium with residual strategic uncertainty if and only if each ℓ type's threshold results in a different probability of $a = 1$ (i.e., $1 - F(\tilde{e}_{\ell})$). An example is the case where F has full support over $\mathbb{R}$. This also guarantees mild agreement.

Given that $a = 1$ is taken most often under ex-ante signaling, a natural follow up question is whether the same comparison holds for each evidence realization. While it is difficult to make this comparison for arbitrary equilibria, we show such a ranking does indeed hold when comparing ex-ante signaling to ex-post signaling.

Proposition 1. *For a probability one $E' \subseteq E$, $v^{\alpha}(e, F) \geq v^{\beta}(e)$ for all $e \in E'$.*

It is worth noting that there is nothing "mechanical" about ex-ante signaling that leads to a higher probability of $a = 1$. It is also not clear whether ex-ante or ex-post signaling provides higher reputation incentives to take $a = 1$, and why this shouldn't depend on the parameters. Under ex-post signaling, following evidence realization e, P considers whether to choose $a = 1$ and pool with $\ell > e + c$, or to choose $a = 0$ and pool with $\ell < e + c$, while under ex-ante signaling, P can directly target any specific leniency type and effectively commit to that leniency type's threshold. That is, $v^{\beta}(e)$ depends only on $G(e + c)$ whereas $v^{\alpha}(e, F)$ depends on the whole distribution G and the investigation F. The next subsections develop intuition for why the broad comparison in Theorem 1 holds.

4.1. Intuition for Theorem 1 with Binary Leniency Types

Suppose G is supported on two leniency types $\underline{\ell} < \overline{\ell}$, F has full support on $\mathbb{R}$ (which guarantees mild agreement), and, for notational convenience, $c = 1$. Now let us compare the probability of $a = 1$ for each evidence realization between ex-post and ex-ante signaling, i.e., $v^{\alpha}(e, F)$ to $v^{\beta}(e)$. If $e < \tilde{e}_{\underline{\ell}}$ or $e > \tilde{e}_{\overline{\ell}}$, then Lemma 1 implies that all DM types take the same action—$a = 0$ and $a = 1$ respectively—under all equilibria. In addition, by Lemma 1, the N types action choices do not depend on the equilibrium. Thus the comparison turns on P's decision given *pivotal* evidence realizations $e \in [\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$.

Consider such a pivotal evidence realization e. Under ex-ante signaling, P will mix between $m_{\underline{\ell}}$ and $m_{\overline{\ell}}$, and follow through with $x_{\underline{\ell}}$ and $x_{\overline{\ell}}$ respectively. Thus, the probability that P takes $a = 1$ after e is the probability that he mimics the $\underline{\ell}$ at the communication stage, which is pinned down by P's indifference across messages:

$$\rho\bigg(\nu_1^{\alpha}(L|m_{\underline{\ell}}) - \nu_1^{\alpha}(L|m_{\overline{\ell}})\bigg) = F(\tilde{e}_{\overline{\ell}}) - F(\tilde{e}_{\underline{\ell}}).$$

That is, the difference in reputation at $m_{\underline{\ell}}$ relative to $m_{\overline{\ell}}$ is proportional to the difference in probability with which $\underline{\ell}$ takes $a = 1$ relative to $\overline{\ell}$.

Under ex-post signaling, every DM type chooses the same message $m_0 \in M$ at the communication stage. Given evidence realization e at the decision stage, P similarly chooses

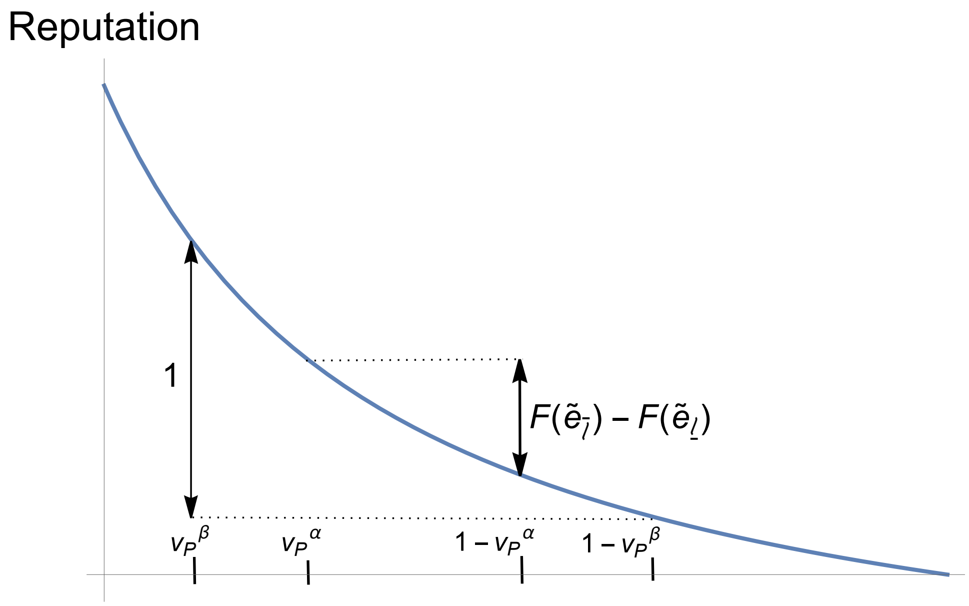


Figure 2: Reputation as a function of the partisan's strategy in the binary example where $v_P^{\mathcal{E}}$ is a shorthand for the probability P takes $a = 1$ given $e \in (\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$ under equilibrium $\mathcal{E}$.

$a = 1$ with the probability that he mimics the $\underline{\ell}$ type, which is determined by

$$\rho\Big(\nu_2^{\beta}(L|m_0, 1, e) - \nu_2^{\beta}(L|m_0, 0, e)\Big) = 1.$$

Figure 2 illustrates how P shifts his strategy so that the reputation incentives compensate him for the difference in material loss between mimicking $\underline{\ell}$ and $\overline{\ell}$. Under ex-post signaling, conditional on evidence e, the difference in $\mathbb{P}(a = 1)$ from pooling with $\underline{\ell}$ or $\overline{\ell}$ is 1, compared with $F(\tilde{e}_{\overline{\ell}}) - F(\tilde{e}_{\underline{\ell}}) < 1$ under ex-ante signaling. In order to create a higher reputation difference in the ex-post case, P must mimic $\underline{\ell}$ less frequently, which in turn means they choose $a = 1$ less frequently. This argument establishes that $V^{\alpha}(F) > V^{\beta}(F)$ in this binary-type example.

The underlying force behind the above argument is that the P type is willing to promise more ex-ante because this promise will only be called for a subset of evidence realizations. Under ex-ante signaling, mimicking $\underline{\ell}$ as compared to $\overline{\ell}$ yields extra reputation regardless of whether these two types take different decisions ex-post. In contrast, under ex-post signaling, the extra reputation from mimicking $\underline{\ell}$ as opposed to $\overline{\ell}$ only realizes when these types take different decisions.[25]

[25] This intuition echoes discussions from the expressive voting literature (e.g., Brennan and Hamlin (1998)) which argue that in elections where the voter is unlikely to be pivotal, the inherent value of expressing certain preferences dominates in their voting decision relative to the instrumental value of implementing a preferred

While this intuition is compelling, it is difficult to extend this argument directly to show a similar ranking holds with more leniency types or across other equilibria. Instead, we now take a different approach, one that proves useful when studying optimal information design and provides additional insights into the forces behind Theorem 1.

4.2. Proof Sketch of Theorem 1

We first establish an inverse relationship between P's equilibrium expected utility and the equilibrium probability of $a = 1$. This result allows us to prove Theorem 1 by showing that P's equilibrium expected utility is lowest under ex-ante signaling. We then show this comparison can be seen from convexity of P's ex-ante signaling utility in the investigation F. As we elaborate below, this convexity is driven by an underlying convexity property of Bayesian updating.

Lemma 3 (Opposing Interests).
For every equilibrium $\mathcal{E}$,

$$V^{\mathcal{E}}(F) = \frac{1}{c}\left(\rho q - U_P^{\mathcal{E}}(F)\right).$$

We label this as opposing interests because the investigator's interests oppose P's interests in *equilibrium*. In particular, it says that P and the investigator cannot be made simultaneously better off through equilibrium selection. This relationship may seem intuitive as P and the investigator have opposing interests concerning the decision. However, the *game* is not one of opposing interests between the investigator and P because (i) there is a third party—the N type—and (ii) even fixing N's equilibrium behavior, P's payoffs also depend on reputation. That is, both the investigator and P could be made better off by P choosing a strategy which provides him with a higher expected reputation and a higher probability of $a = 1$. Lemma 3 shows that this is not possible in equilibrium.

Another notable feature of the relationship between $V^{\mathcal{E}}(F)$ and $U_P^{\mathcal{E}}(F)$ is its simplicity. In particular, conditional on the value of $U_P^{\mathcal{E}}(F)$, $V^{\mathcal{E}}(F)$ does not depend on the investigation F, or the distribution of leniency G. This feature makes the opposing interests lemma useful in thinking about the investigator designing F in Section 5—they will seek to minimize P's utility.

Given Lemma 3, we can focus on analyzing P's equilibrium utility $U_P^{\mathcal{E}}(F)$. We next make two observations. First, note that all equilibria yield equivalent outcomes when F is degenerate, as in this case, there is no difference between the decision stage and the

policy.

communication stage. Second, note that $U_P^\beta(F)$ is linear in F: by definition, nothing happens at the communication stage under ex-post signaling, so F only impacts the outcome separably through the probability of evidence e. Putting these points together gives-

$$U_P^\beta(F) = \mathbb{E}[U_P^\beta(\delta_e)|e \sim F] = \mathbb{E}[U_P^\alpha(\delta_e)|e \sim F],$$

where δ_e denotes the degenerate distribution on e. Thus, the comparison that $U_P^\alpha(F) \leq U_P^\beta(F)$ holds if $U_P^\alpha(F)$ is convex in F, which we establish in the next lemma.

Lemma 4. *$U_P^\alpha(F)$ is convex in the investigation F.*

The intuition for Lemma 4 follows from a fundamental property about Bayesian updating: adding probability that a given type sends some signal changes the corresponding conditional belief on that type less if they already send that signal with high probability. In our setting, this means that the belief that the DM is an N type following any message is convex in the probability that P sends that message. This convexity is illustrated in Figure 2. To see how convexity of reputation relates to convexity of $U_P^\alpha(F)$, consider two investigations F_1 and F_2 with corresponding reputation functions R_1^α and R_2^α. For some $\lambda \in (0,1)$, let $F_\lambda = \lambda F_1 + (1-\lambda)F_2$. P's material utility from sending any message m_ℓ is linear in F: P chooses $a = 1$ under F_λ with probability equal to the average of that under F_1 and F_2. However, P cannot achieve the "average reputation" at every m_ℓ because reputation is convex in the rate at which he declares each message, which yields the convexity of $U_P^\alpha(\cdot)$.[26]

Ex-Ante Signaling vs. Other Equilibria: We have shown that ex-ante signaling has a higher probability of $a = 1$ than ex-post signaling. However, Theorem 1 says that ex-ante signaling delivers a higher conviction probability than *any* other equilibrium. Our proof shows how to use the first comparison to prove the second.

The idea is as follows. Fix an equilibrium $\mathcal{E}$. Note that P's expected utility conditional on sending a message $m \in M_P$ is the ex-post signaling equilibrium utility with prior equal to the interim belief $\nu_1(\cdot|m)$. Using the comparison between ex-post and ex-ante signaling, we obtain that P's expected utility conditional on sending message m is higher than if one were to instead conduct ex-ante signaling with prior $\nu_1(\cdot|m)$.

Now consider an alternative messaging strategy which first selects a message according to the original equilibrium strategy under $\mathcal{E}$ and then sends a follow up message m_ℓ

[26] In order to maintain the reputation $\lambda R_1^\alpha(m_\ell) + (1-\lambda)R_2^\alpha(m_\ell)$, the convexity of the reputation implies P would need to, for all $\ell \in L$, declare m_ℓ at a rate less than the average across the equilibria induced by F_1 and F_2. But this cannot be since the total measure of P's messages must equal one.

according to the ex-ante signaling equilibrium given prior $\nu_1(\cdot|m)$. Conditional on sending each initial message under this new strategy, the above logic implies that P's expected utility is lower than under the original equilibrium $\mathcal{E}$. The only remaining issue, is that P may not be indifferent across messages. However, because this comparison holds for every message, when P adjusts his strategy to reestablish indifference across all messages, the resulting equilibrium is ex-ante signaling, and his new equilibrium expected utility is still lower than in the original equilibrium.

4.3. Comparing the DM's Utility

Combining the investigator's preference for ex-ante signaling with the fact that his interests oppose that of P immediately yields that ex-ante signalling is P's least favorite equilibrium. However, the properties of equilibria in Lemma 1 facilitate extending this comparison to all DM types.

Corollary 1. *For any F and two equilibria $\mathcal{E}, \mathcal{E}'$,*

1. $U_\theta^{\mathcal{E}}(F) - U_\theta^{\mathcal{E}'}(F)$ *is constant across* $\theta \in \Theta$.
2. $U_\theta^{\alpha}(F) \leq U_\theta^{\mathcal{E}}(F)\ \forall \theta \in \Theta$; *this inequality is strict if $\mathcal{E}$ has residual strategic uncertainty and there is mild agreement.*

Given Theorem 1 and Lemma 3, the second point follows directly from the first. The first point says that the difference in utility between any two equilibria is type independent. The idea is that (i) each ℓ type chooses x_ℓ in every equilibrium, so their utility difference is just given by the expected reputation difference, and (ii) P is indifferent between mimicking any ℓ type, and so this expected utility difference must be constant across ℓ. This result provides one rationalization for why politicians may "dodge the cameras" and admissions committees may favor non-transparency—or, in our terminology, favor ex-post signaling. This result also points to interesting questions about equilibrium selection issues, which we address in Subsection 6.1.

5. Optimal Investigations

Having studied the impact of communication for an arbitrary fixed F, we now turn to how the investigation affects the action choice of the DM. For the results in Section 3 and Section 4, we can be relatively agnostic about what the evidence represents: while it is natural to think that it represents a belief about or expected value of an unknown state, nothing in our setup requires such an interpretation. We make this explicit in this section:

we let $e \in [0,1] = E$ represent a posterior belief about a binary state $\omega \in \{0,1\}$ with prior $\overline{e} \in (0,1)$, where each investigation represents an information structure about ω.[27]

The relationship between the investigation and outcomes depends on the equilibrium. It is worth noting that under ex-post signaling, standard "concavification" techniques from the information-design literature can be applied to understand this relationship. In this case, conditional on the evidence realization, the outcome is independent of the investigation, so that $V^{\beta}(F) = \int_E v^{\beta}(e)dF(e)$ is linear in F.

As elaborated further in Subsection 6.1, we view ex-ante signaling outcomes as focal because they arise naturally as either the result of institutional design and commitment, or as the uniquely selected equilibrium under a compelling refinement. In contrast to ex-post signaling, outcomes depend on the investigation even conditional on the evidence realization under ex-ante signaling, i.e., how P chooses which ℓ type to mimic at the communication stage, and hence $v^{\alpha}(e, F)$, depends on F. Thus, the probability of $a = 1$ is not linear in F. This invalidates the use of concavification techniques. In this section we analyze how the investigator chooses an investigation maximize the probability of $a = 1$ under ex-ante signaling. This design framing is sometimes directly relevant to our applications; an impeachment inquiry is often lead by a member of the opposing political party, and the firm seeking a merger is responsible for disclosing information to the FTC. However, beyond the direct design question, our results reveal comparative statics intuitions on how the investigation affects outcomes that are novel and specific to the case in which DM takes informative stands.

5.1. Characterization

For this section we assume the leniency distribution G admits a continuous density g on its support with $[c, 1+c] \subseteq \mathrm{Supp}(G)$. To calculate the investigator's utility, we sum the probability of $a = 1$ given message m_ℓ weighted by the probability that the DM sends message m_ℓ. Letting $\mathcal{F}$ be the set of CDFs with support on $[0,1]$, the investigator's design problem is

$$\max_{F \in \mathcal{F}} \int_L (1 - F(\tilde{e}_\ell))\left(qg(\ell)d\ell + (1-q)d\sigma(m_\ell|P)\right),$$
$$\text{such that} \int_0^1 (1 - F(e))de = \overline{e}.$$

The constraint captures Bayes plausibility: the average posterior is the prior, i.e., F is an information structure. In order to solve this problem, we use Lemma 3, which shows that

[27] In Subsection 6.4, we discuss how our results extend to the case with more than two states.

maximizing the investigator's expected utility is equivalent to minimizing that of P. It is straightforward to derive how F determines $U_P^\alpha(F)$: we show in the proof of Lemma 4 that $U_P^\alpha(F)$ is given by the solution U to $\int_L \frac{\rho q g(\ell)}{U+c(1-F(\tilde{e}_\ell))} d\ell = 1$.[28] These observations allow us to rewrite the investigator's problem as follows:

$$\begin{aligned} &\min_{U \geq 0,\ F \in \mathcal{F}} \quad U\ , \\ &\text{such that } \int_L \frac{\rho q g(\ell)}{U + c - cF(\tilde{e}_\ell)} d\ell = 1, \\ &\int_0^1 (1 - F(e)) de = \overline{e}. \end{aligned} \tag{1}$$

The extra constraint ensures the choice of U in (1) is equal to $U_P^\alpha(F)$. We show that it is without loss to relax both constraints to only hold as inequalities. This relaxed version of the investigator's problem minimizes a linear objective over a convex constraint set. We can construct a Lagrangian which, with some standard ironing techniques, allows us to solve for the optimal investigation.

Define $H : E \to \mathbb{R}_+$ as $H(e) \equiv \int_{-\infty}^{e} g(e' + c) de'$. Denote $\overline{H}$ as the concavification of H,[29] and $\overline{h}$ as its derivative in e, which is continuous because g is continuous.[30]

Theorem 2. *For $k, U \in \mathbb{R}$, define $\widehat{F}(e; k, U) \equiv U/c + 1 - k\sqrt{\overline{h}(e)}$. The uniquely optimal investigation is given, for $e < 1$, by*

$$F^*(e) = \begin{cases} 0 & \text{if } \widehat{F}(e;k,U) < 0, \\ \widehat{F}(e;k,U) & \text{if } \widehat{F}(e;k,U) \in [0,1], \\ 1 & \text{if } \widehat{F}(e;k,U) > 1, \end{cases}$$

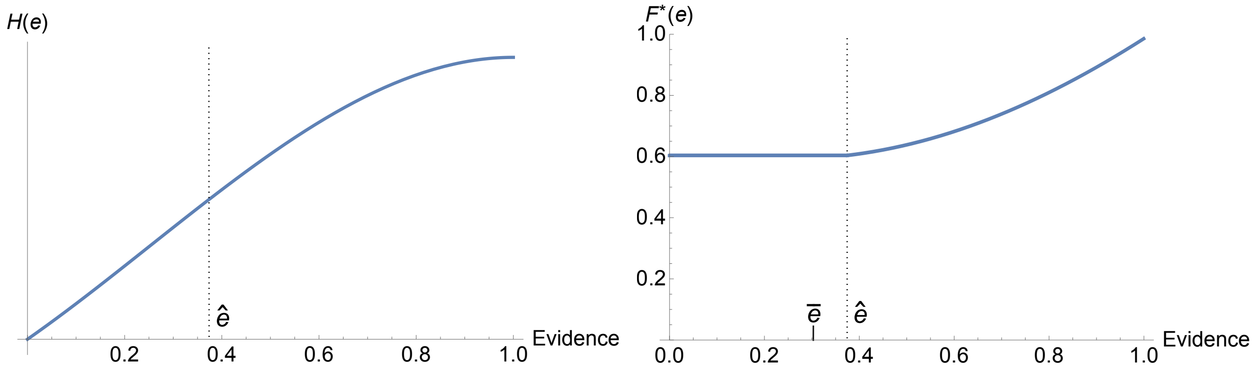
with $U = U_P^\alpha(F^*)$ as the partisan's utility given F^* and some $k > 0$.

Because each ℓ type uses a fixed threshold, H captures the probability that non-partisans choose $a = 1$ given evidence e. It is then well known that the curvature of $\overline{H}$ (or the mono-

[28] The derivation of this equation uses the following logic. P's indifference across messages provides an expression for $\nu_1(L|m_\ell)$ in terms of the probability of $a = 1$ at m_ℓ—namely, $1 - F(\tilde{e}_\ell)$—and $U_P^\alpha(F)$. Because $\frac{g(\ell)q}{\nu_1(L|m_\ell)}$ is equal to the probability or density of m_ℓ, the sum of this fraction over m_ℓ is equal to 1.

[29] The concavification of H is the point-wise lowest function over all concave $\tilde{H} : E \to \mathbb{R}$ such that $\tilde{H}(e) \geq H(e)\ \forall e \in E$.

[30] There are two remaining parameters in the characterization in Theorem 2: $U_P^\alpha(F^*)$ and k. These are jointly pinned down by the two constraints in (1). While an explicit expression is not always feasible, solving these two equations numerically is straightforward.



(a) $H(e)$: CDF of non-partisan thresholds above $e = 0$

(b) F^*: CDF of the Optimal investigation

Figure 3: $L = \mathbb{R}$, $g(\ell)$ is a standard logistic distribution with mean $\frac{1}{2}$, $c = \frac{1}{4}$, $q = \frac{1}{2}$, $\overline{e} = \frac{3}{10}$, and $\rho = 3$.

tonicity of $\overline{h}$) captures the information provision incentives when the investigator faces only non-partisans: providing information over regions where $\overline{h}$ is constant (decreasing) increases (decreases) the probability that N chooses $a = 1$. Our characterization is also in terms of $\overline{h}$ but these incentives are distorted by the fact that the investigator must also persuade the partisan DM.

Figure 3 presents an example of an optimal investigation. In this example, the distribution of non-partisan leniencies is single peaked, and so H is convex for small e, and concave for large e, as illustrated in the left panel. Correspondingly, the concavification of H is linear below $\hat{e}$ and equal to H above $\hat{e}$, i.e., $\overline{h}$ is constant below $\hat{e}$ and strictly decreasing above $\hat{e}$. From the right panel of Figure 3, we see that F^* provides information is consistent with N's information incentives below $\hat{e}$, but in contrast, provides some information, in a smooth way, above $\hat{e}$ at the detriment of N's outcomes. We develop the sense in which these properties are general in the two following immediate corollaries, stated without proof.

Corollary 2. *The optimal investigation admits a continuous density for $e \in (0, 1)$; in particular, F^* has no interior mass points.*

Corollary 2 implies that the uninformative investigation is *never* optimal. This result is counterintuitive, as uninformative experiments can be optimal in the Bayesian persuasion literature (Kamenica and Gentzkow (2011)), in particular, when certain concavity conditions on the distribution of thresholds are met. While given a fixed F, these conditions can be satisfied in our model, the key difference is that the distribution of thresholds is endogenous to the investigation: P will tend to respond to a high probability of a particular evidence level by feigning leniency that is just out of reach of such evidence. Given the opposing interests lemma, this response by P leads the investigator to minimize predictability about the realized evidence. Notice that this tendency hinges on the communication stage being informative. As we show in Appendix F, this "unpredictability" is not a feature of the optimal investigation under ex-post signaling where an uninformative investigation may be optimal.

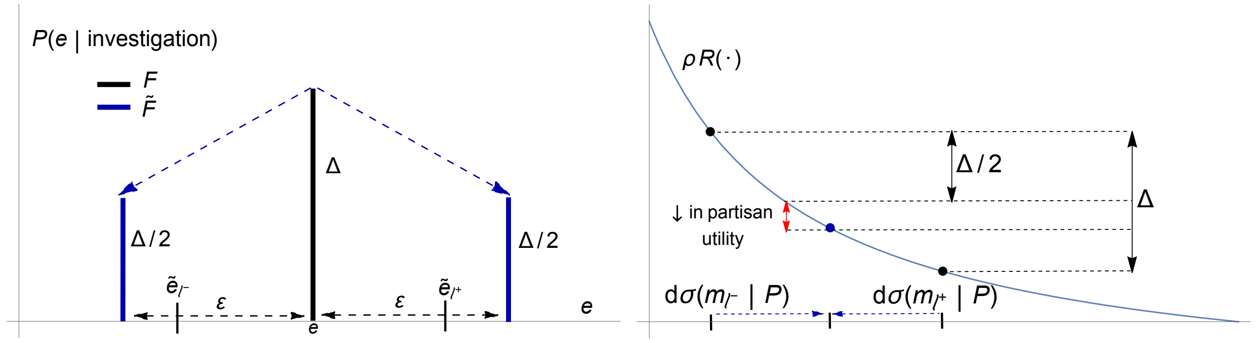
Figure 4 provides intuition for the corollary. It depicts two investigations that differ only around evidence e, with $c = 1$ for convenience. F has an (isolated) mass point of size Δ at evidence e, while $\tilde{F}$ equally splits this mass point on e to $e + \varepsilon$ and $e - \varepsilon$. When $\varepsilon > 0$ is small, because the density of ℓ types g is continuous, the change in $\mathbb{P}(a = 1)$ from ℓ types is second order. However, P increases $\mathbb{P}(a = 1)$ in a first order sense when moving from F to $\tilde{F}$.

To see why, consider two types ℓ^-, ℓ^+ as illustrated in the left panel of the figure, with $e - \varepsilon < \tilde{e}_{\ell^-} < e < \tilde{e}_{\ell^+} < e + \varepsilon$. Under F, ℓ^- chooses $a = 1$ with Δ higher probability than ℓ^+, so, to preserve P's indifference, the equilibrium reputation payoff must be Δ higher from sending m_{ℓ^-} than m_{ℓ^+}. In contrast, under $\tilde{F}$, m_{ℓ^-} and m_{ℓ^+} choose $a = 1$ with the same probability and therefore must command the same reputation. The reputation for these associated messages as a function of $d\sigma(m_\ell|P)$ is illustrated in the right panel of Figure 4. As highlighted in Subsection 4.2, this reputation is convex: as P increases $d\sigma(m_\ell|P)$, the marginal decrease in the reputation for m_ℓ becomes smaller. The right panel illustrates that, because of this convexity, when P equalizes his strategy across m_{ℓ^+} and m_{ℓ^-}, the reputation payoff for m_{ℓ^-} falls by more than $\frac{\Delta}{2}$ and the reputation payoff for m_{ℓ^+} rises by less than $\frac{\Delta}{2}$. That is, P's expected utility at these messages has fallen.[31] Because of the opposing interests lemma, this change benefits the investigator.

Corollary 3. *The optimal investigation is fully informative if and only if $\overline{h}$ is constant.*

This corollary is a direct implication of the fact that $F^*(e) \in (0, 1)$ is constant in e, equivalently F^* is supported on $\{0, 1\}$, if and only if $\overline{h}$ is constant. To understand this result, recall that the monotonicity of $\overline{h}$ captures the investigator's design incentives when only facing non-partisans. Therefore, an alternative statement of Corollary 3 is that the investigator provides full information if and only if full information maximizes the investigator's objective among non-partisans. Because F^* balances design incentives between both types, this means that the investigator's design goals for P align with that for ℓ types when $\overline{h}$ is

[31] There are other messages sent under ex-ante signaling, which now have higher utility for P. To restore equilibrium, P would also have to reallocate some mass from $\{\ell^-, \ell^+\}$ to these other messages. But this would serve to decrease the reputation for these messages preserving the conclusion.



(a) A local spread of investigation F

(b) Change in partisan strategy

Figure 4: Locally spreading any mass point harms the partisan.

constant, but are misaligned when $\overline{h}$ is decreasing.

At a high level, the intuition is as follows. All else equal, P benefits from correlating his strategy with the ℓ types. When the investigator increases the probability of evidence in an interval, i.e., increases $F(\tilde{e}_{\overline{\ell}}) - F(\tilde{e}_{\underline{\ell}})$ for $\overline{\ell} > \underline{\ell}$, P reallocates mass from mimicking types below $\underline{\ell}$ to types above $\overline{\ell}$. If g is increasing, in which case $\overline{h}$ is constant, then this response by P further correlates his strategy with that of the ℓ types, and thereby tends to benefit P. Conversely, if g is decreasing, this change in the investigation tends to miscorrelate P and ℓ types' strategies and thereby harm P. Given the opposing interests lemma, the former change harms the investigator, while the latter change benefits them.

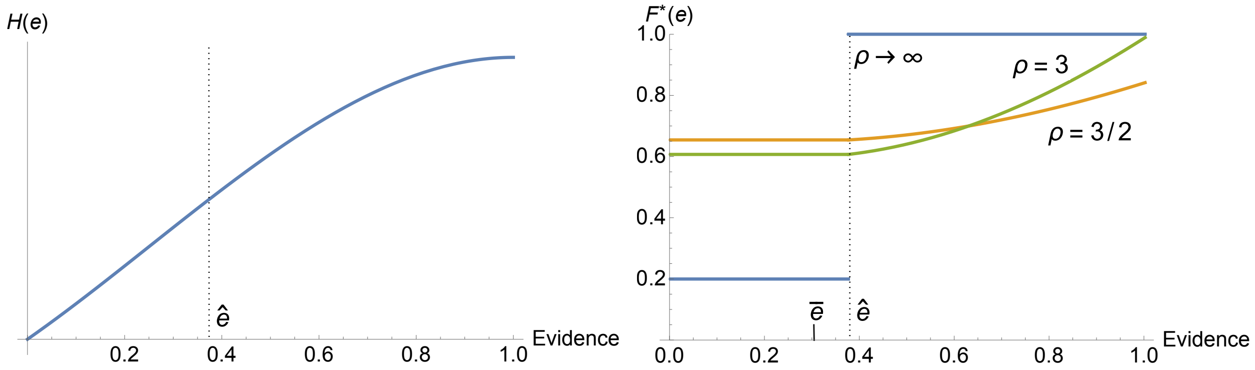
5.2. Comparative Statics

We next explore comparatives statics of the investigation design problem. We begin by documenting some basic changes in the parameters that increase the probability of $a = 1$.

Proposition 2. *Let $\tilde{G}$ be a distribution of ℓ that first-order stochastically dominates G. For any fixed F, $V^{\alpha}(F)$ is higher under G than G' and when ρ or q increases.*[32]

Because these comparisons hold for a fixed investigation F, they also hold for the investigator's value in the design problem. The intuition for these comparative statics is straightforward as each change can be seen as increasing the alignment between the DM and investigator. An increase in q decreases the probability of the P type whose preferences are at odds with the investigator's. Similarly, a first-order stochastic decrease in G

[32] One omitted parameter from this result is c. Although one might naturally conjecture that an increase in c induces less conviction by P and therefore hurts the investigator, the probability of $a = 1$ from ℓ types is increasing in c (as can easily been seen from Lemma 1). Either force can dominate, making comparative statics on c ambiguous.



(a) Distribution of non-partisan commitments

(b) Optimal Investigations

Figure 5: $L = \mathbb{R}$, $g(\ell)$ is a standard logistic distribution with mean $\frac{1}{2}$, $c = \frac{1}{4}$, $q = \frac{1}{2}$, $\overline{e} = \frac{3}{10}$.

means that the non-partisan prefers $a = 1$ more often. By increasing ρ, we are increasing the importance of reputation relative to material payoffs in the DM's utility. This change then reduces the misalignment between the partisan and investigator.[33]

We next look at how the optimal investigation changes with the size of reputation incentives. We can interpret an increase in ρ as a decrease in the relative importance of the decision at hand—i.e., the stakes of the decision are lower. Our next result charchterizes how the informativeness of the investigation changes with the stakes of the decision.

Proposition 3. *The optimal investigation F^* becomes less Blackwell informative as ρ or q increases.*

We illustrate in Figure 5 how the optimal investigation changes with ρ in the example from Figure 3. In the limiting case when $\rho \to \infty$, P will fully mimic the distribution of ℓ types' messages, and so the distribution of thresholds is investigation independent. This means that the optimal investigation converges to the Bayesian persuasion solution for the problem of maximizing conviction from ℓ types: a point mass at 0 and at $\hat{e}$. As ρ decreases, the optimal investigation maintains 0 mass on $[0, \hat{e})$ where $\overline{h}$ is flat, but spreads the point mass on $\hat{e}$ to evidence levels in $[\hat{e}, 1)$.

To see the intuition for Proposition 3, note first that regardless of ρ and q, the optimal investigation puts 0 mass on regions where $\overline{h}$ is increasing. When $\overline{h}$ is decreasing, the investigator balances two opposing incentives for the ℓ types and P: the investigator wants to hedge against P's strategic "targeting" by spreading out the distribution of evidence, but wants to contract the optimal investigation for the ℓ types because their distribution

[33] Despite also affecting their signaling incentives, a change in ρ has no effect on the non-partisan's strategy, and thereby their probability of $a = 1$.

over thresholds is concave. When q is large the contraction incentives for the ℓ types are weighted more, and so the optimal investigation is less informative. When ρ is large, P seeks to mimic the ℓ types more and is therefore less responsive to changes in the investigation. This makes the investigator's hedging incentive with P less significant, again leading to a less informative investigation.

Our final comparative statics looks at the impact of mean-preserving spreads of the distribution ℓ on the investigator's utility. Such spreads can be interpreted as an increase in the polarization of non-partisans. The comparative statics for mean-preserving spreads of ℓ are, in general, ambiguous. However, some of this ambiguity is an artifact of our bounded evidence space: The ℓ types above the support of F^* always take $a = 0$, and so a spread of leniencies in this region can only increase the probability of $a = 1$. Our next result shows that, under a regularity condition on g, and excluding these changes in "non-pivotal" ℓ types, a spread in the distribution of ideologies harms the investigator, i.e., decreases the probability of $a = 1$.

Let F^* be the optimal investigation given $\ell \sim G$. We say that $\tilde{G}$ (with associated density $\tilde{g}$) is a *pivotal mean-preserving contraction* of G if $\tilde{G}$ is a mean-preserving contraction of G and $g(\ell) = \tilde{g}(\ell)$ for all ℓ such that $\tilde{e}_\ell \notin \mathrm{Supp}(F^*)$. One simple type of pivotal mean-preserving contraction is one that contracts probability locally around some ℓ such that $F^*(\tilde{e}_\ell) \in (0, 1)$.

Proposition 4. *Suppose that g is log-concave. If $\tilde{G}$ is a pivotal mean-preserving contraction of G, then the investigator does better under $\tilde{G}$ than G.*

The broad intuition is as follows. Consider spreading $\underline{\ell}$ and $\overline{\ell}$ so that they are further away from each other. This spreads the material payoff difference from mimicking these types for P as there are new evidence realizations between $\tilde{e}_{\underline{\ell}}$ and $\tilde{e}_{\overline{\ell}}$. As a result the equilibrium reputation for $m_{\underline{\ell}}$ and $m_{\overline{\ell}}$ must also spread. However, because reputation is convex, a similar logic to that in Subsection 4.2 shows that this increases the utility of P, and thereby harms the investigator.

6. Discussion and Extensions

6.1. Commitment and Equilibrium Selection

Our framework admits a wide array of equilibrium outcomes—one for each LIS. Under our focal equilibrium—ex-ante signaling—this cheap-talk communication is most informative about the eventual decision. Recall that under ex-ante signaling it is *as if* the DM commits to a contingent plan even though he only has access to cheap talk. However, there are many natural ways in which exogenous commitment power can arise in our setting; for example, the DM could publicly delegate the decision, put the decision plan in a legally binding contract, or simply bear large lying costs (as in Kartik (2009)). In addition, such commitment can be mandated externally; for example, government agencies and publicly funded universities can be required to specify approval and admissions criteria respectively. Motivated by this, we explore how endowing the DM with commitment power at the communication stage affects outcomes in our model. We show that ex-ante signaling outcomes are the unique equilibrium outcome if either (i) commitment is *mandated*, or (ii) commitment is *available* and the DM has any uncertainty about their preferences at the communication stage that is privately revealed at the decision stage.

The Commitment Model In the commitment model, the DM commits to a publicly observed contingent plan $x \in \mathcal{X}$ instead of choosing a messaging and decision strategy. Following the commitment, evidence is realized, the action is taken according to x, and payoffs are realized. The preferences of the DM are the same as that in Section 2. We maintain our focus on equilibria that satisfy the D1 refinement.[34]

Proposition 5. *The commitment model admits a unique equilibrium outcome that is equivalent to that under ex-ante signaling.*

In the proof, we show that there is an equilibrium in which each ℓ type chooses x_ℓ, with P mixing over $\{x_\ell\}_{\ell \in L}$, which then generates the same equilibrium outcome as in ex-ante signaling. We then show that no other equilibria can be sustained; in particular, in equilibrium N will, with probability one, never commit to a contingent plan that yields different outcomes than x_ℓ. Both points follow from the fact that x_ℓ delivers the maximal "signaling value" for type ℓ as formalized in (**??**).

The Optional Commitment Model The optional commitment model has two alterations from our main model. First, at the communication stage, each DM has the option to commit to an arbitrary contingent plan as a function of the evidence, $x \in \mathcal{X}$. However, unlike in the commitment model, the DM can abstain from commitment and send a cheap-talk message instead, in which case the game proceeds as in our main model. We continue to apply the D1 refinement.

Second, the preferences of the DM are perturbed as follows. The utility of the DM of type θ, taking action a, given evidence realization e, and reputation μ is given by $u(\theta, e, a, \mu) +$

[34] In the appendix, we provide a formal definition of equilibrium in the commitment model.

εa where ε is a random variable that is mean 0, independent of other parameters, with support equal to $[-\delta, \delta]$, with $\delta > 0$. The DM does not know ε at the communication stage, but privately observes ε at the decision stage. The variable ε represents changing conditions between the communication and decision stages: a politician may learn that conviction is actually more or less favorable for their party than previously expected, or the admissions officer may learn new revelations about a potential applicant. It can also represent evidence from the investigation that is revealed privately to the DM but not to the public. For example, certain findings of the Trump impeachment inquiry were redacted for the public but revealed to senators making the impeachment decision.

Proposition 6. *For any $\delta > 0$ such that $\rho > 2\max\{\frac{\delta}{q}, \frac{\delta}{1-q}\}$, the optional commitment model admits a unique equilibrium outcome equivalent to that under ex-ante signaling.*

Notice that the proposition holds for arbitrarily small preference shocks, but also for large ones modulated by the weight on reputation ρ.[35] The intuition is as follows. Ex-ante signaling is the unique equilibrium with no residual strategic uncertainty at the decision stage. Because the DM does not know ε at the communication stage, equilibria with residual strategic uncertainty provide the benefit of being able to adjust the action choice to the realization of ε at the decision stage. However, this benefit is greater for P than it is for ℓ types. The reason is that ℓ will only take ε into account for *pivotal* evidence realizations, i.e., when $e - \ell$ is close to the difference in reputation between the two actions, while P, who does not care about evidence, is responsive to ε at any evidence realization. Thus, if there exists some ℓ who faces residual strategic uncertainty in equilibrium and x_ℓ goes unused, then it will be given a reputation of 1, which is not possible given the assumed high value of reputation. This captures the intuition by which "dodging the cameras" is interpreted negatively: being vague about one's standards at the communication-stage signals a desire to be responsive to idiosyncratic partisan preferences (ε) rather than the evidence.[36]

6.2. Timing of Evidence Disclosure

In many settings, the timing of evidence disclosure is a choice of the investigator who can choose to reveal some information before the DM has a chance to announce their contingent plan: an investigation into a political scandal could leak details before the inquiry

[35] When δ is large enough to violate the inequality in Proposition 6, the option value from acting on the realization of ε could exceed the reputational gains from committing at the communication stage. In this case, each x_ℓ commitment would still garner a reputation of 1 according to the D1 refinement, but could go unused.

[36] Committing to a policy ex-ante is also used for signaling value in Callander (2008). There, the policy decision is a scalar rather than a function, however the intuition has similarity in that committing to extreme policies signals a value for material payoff vs. reputation (in that paper, office motivation).

is formally announced, or firms could publicly disclose financial records before submitting their application for a merger to the FTC. When should the investigator release information to the DM and, more broadly, how does the timing of disclosure affect equilibrium outcomes?

To answer this question, we consider a version of our baseline model with two stages of evidence disclosure. Before the DM sends a message, they observe an initial public evidence state $e_0 \sim F_0$. After the message is sent, the final evidence $e_1 \sim F_1(\cdot|e_0)$ is realized, and an action is chosen. The preferences of the DM are the same as in Section 2 with only the final evidence e_1 being payoff relevant. Let $\overline{F}$ be the unconditional distribution of e_1.[37] We maintain the focus on ex-ante signaling equilibria in each subgame following the realization of e_0, and so our timing results also apply to the commitment model.

Consider an investigator who can choose among different (F_0, F_1) with the same $\overline{F}$. By choosing different F_0, he can span various timings of evidence disclosure. When F_0 is degenerate, all information is backloaded until after the DM communicates, in which case equilibrium outcomes correspond to those under ex-ante signaling. When F_1 is degenerate, all information is front-loaded to before communication, in which case equilibrium outcomes correspond to those under ex-post signaling. That is, even though we focus on the ex-ante signaling equilibrium conditional on e_0, front-loading disclosure generates ex-post signaling outcomes due to the fact that when the evidence distribution is degenerate, ex-ante signaling and ex-post signaling are identical. Our next result shows that the investigator prefers to backload information relative to any other timing of disclosure.

Proposition 7. *Among all F_0 and F_1 with the same $\overline{F}$, $F_0 = \overline{F}$ delivers the lowest $\mathbb{P}(a = 1)$, and $F_1(\cdot|\cdot) = \overline{F}$ delivers the highest $\mathbb{P}(a = 1)$.*

This result follows from the convexity of $U_P^\alpha(\cdot)$. Thus, delaying evidence disclosure (while keeping the final distribution of e_1 constant) hurts P and benefits the investigator.

6.3. State-Dependent Investigator Preferences

We have so far assumed that the investigator's preferences are state independent—that is, the investigator always prefers $a = 1$ and has a utility independent of e. While we think this is a reasonable assumption (or approximation) in many settings, it is natural to ask how our results on investigation design depend on this assumption. Indeed, we used this assumption to establish the opposing interests lemma which greatly simplifies our analysis. Nevertheless, many of our main insights continue to hold when the investigator has state-dependent preferences.

[37] More precisely, $\overline{F}(e_1) = \int_{e_0} F_1(e_1|e_0) dF_0(e_0)$.

We maintain that $e \in [0,1]$ represents a posterior belief about a binary state, and G admits a continuous density. The investigator's utility from action a and evidence e is now given by $(e - \ell_I)a$ where $\ell_I < 1$. We make the additional assumption that all ℓ types can be persuaded by some evidence realization in equilibrium, i.e., that $0 < \min_{\ell \in L} \ \tilde{e}_\ell < \max_{\ell \in L} \ \tilde{e}_\ell < 1$. Because of the high reputation incentives, this guarantees that P is also persuadable in equilibrium.

Proposition 8. *The investigator prefers ex-ante signaling to ex-post signaling. For sufficiently high ρ, the optimal investigation under ex-ante signaling has no interior mass points.*

Because the DM is responsive to evidence, there is no "effective" conflict of interest when $\ell_I > 0$, and the state is observed. Therefore, the investigator gets his first best utility from full revelation. The interesting case is when $\ell_I < 0$, i.e. when the investigator prefers conviction in both states, but has stronger preferences in state 1. In this case, the fact that the investigator prefers ex-ante signaling to ex-post signaling follows directly from Proposition 1.

To see why the investigator still wants to ensure unpredictability, i.e. set an investigation with no mass points, recall that the intuition provided for Corollary 2 in Figure 4 used a *local* perturbation. Introducing the investigator's continuous evidence-dependent preferences affect the tradeoff from locally spreading an evidence mass point in a second order way and so it remains beneficial. The one subtlety comes from the fact that P responds by recalibrating the probability with which he mimics ℓ types with non-local thresholds whose conviction probability is unaffected by the perturbation; and, because the opposing interests lemma no longer holds, we cannot simply compare P's utility to determine the investigator's ranking. The proof shows that with high reputation incentives the positive effect illustrated in Figure 4 dominates.

6.4. Optimal Investigations with Multiple States

While, in our main specification we consider an investigation about a binary state, many of our results are robust to the case where the investigator specifies an information structure about a larger state space. As is well known, compactly describing the set of Bayes plausible experiments quickly becomes intractable as the cardinality of the state space increases. We therefore focus on the case where the ℓ types' material preferences over actions depend only the posterior mean about an unknown state. Here, we interpret the evidence $e \in E \equiv [0,1]$ as the posterior mean about some state $\omega \in [0,1]$,[38] where the domain is $[0,1]$

[38] This means that the DM's underlying objective is linear in ω.

for expositional convenience. ω is distributed according to CDF K which has strictly positive density k. Using insights from Gentzkow and Kamenica (2016) and Kolotilin (2018), a CDF over posterior means $F : [0,1] \to [0,1]$ is a feasible choice for the investigator if and only if it satisfies the following Bayes plausibility constraints:

$$\int_0^e F(e')de' \leq \int_0^e K(e')de' \; \forall e \in E, \text{ and}$$
$$\int_0^1 F(e')de' = \int_0^1 K(e')de'. \tag{2}$$

The investigator's problem can then be written in the same manner as in (1) substituting the constraints in (2) for the Bayes plausibility constraint. To avoid ironing complications, we assume that g is strictly decreasing on $[c, 1+c]$. We characterize the optimal investigation in the Appendix (Proposition 9) and show that, despite the more complicated constraint set, the main takeaways from Section 5 hold true.

Corollary 4. *The optimal investigation has no mass points.*

Corollary 5. *If* $\frac{g(e+c)}{(\rho q + c(1-K(e)))^2}$ *is strictly increasing in* e*, then full information is uniquely optimal.*

The first corollary shows that the investigator reduces the predictability of the investigation by avoiding mass points. This is despite the fact that, because g is assumed to be strictly decreasing, providing no information would yield the highest probability of $a = 1$ from ℓ types. The second corollary says that if the cost of providing information to nonpartisans is small, roughly that g decreases slowly (or more specifically, the condition in Corollary 5), then full information is optimal.[39]

6.5. Reputation for Leniency

We now extend the model to allow the DM to differentially value his reputation for appearing as specific leniency types and maintain the same material payoffs. For $r : \Theta \to \mathbb{R}_+$, let $\rho \int_\Theta r(\theta) d\nu(\theta)$ be the reputation payoff when the public holds beliefs $\nu \in \Delta(\Theta)$. We normalize $r(P) = 0$.

We again focus on the case of high reputational concerns. Let $\underline{r} \equiv \inf_{\ell \in L} \; r(\ell)$ and $\overline{r} \equiv \sup_{\ell \in L} \; r(\ell)$. We adapt Assumption 1 as follows.

Assumption 2. $\rho > \max\{\frac{c(\underline{r}+\overline{r})}{\underline{r}^2 - q\overline{r}^2}, \frac{c(\underline{r}+\overline{r})}{q\underline{r}^2}\}$ *and* $q < (\frac{\underline{r}}{\overline{r}})^2$.

[39] While the case in which g is non-monotonic is complicated, the case where g is increasing is tractable, and it can be shown that full information is optimal as in Corollary 3 for the case of two states.

Note that by setting $\underline{r} = \overline{r} = 1$ we recover our baseline model, in which case Assumption 2 is equivalent to Assumption 1. Roughly, Assumption 2 says that the difference between $\underline{r}$ and $\overline{r}$ is not too large relative to the difference between $\underline{r}$ and $r(P) = 0$, i.e., the difference in reputational values for different leniency types does not trump the DMs reputational concern to avoid appearing partisan. This is relatively flexible, e.g., it imposes no monotonicity requirements on $r(\ell)$ with respect to ℓ.

The role of Assumption 2 is identical to that of our original Assumption 1 in our baseline model. It ensures that neither the P type nor the ℓ type will ever fully reveal themselves in equilibrium. In the former case the low reputational payoff is too costly relative to any material gains he could accrue from deviation. In the latter case, it ensures that the *minimum* reputational gain the P type could access from mimicking such an ℓ type would compensate for any material losses he may suffer.

Our appendix proves all our results in this more general environment under Assumption 2. In particular, the statements of results from Section 3, Section 4, Subsection 6.1, and Subsection 6.2 remain unchanged. Other than the comparative statics on G, all results in Section 5, Subsection 6.3 and Subsection 6.4 go through with minor modifications. [40]

7. Conclusion

We study a model of communication by a DM concerned with developing a reputation for taking the right action. Our model sheds light on how communication in the presence of uncertainty over what the right action is shape equilibrium actions. We find that a wide range of communication strategies can be sustained in equilibrium, that the equilibrium in which the DM announces his contingent plan in advance leads to the highest rate of taking the action, and that such communication shapes the design of investigations in ways that are qualitatively distinct from standard information-design problems.

A number of questions remain for future work. Our main results easily extend to the case where ℓ types material payoffs are an arbitrary function strictly increasing in e and decreasing in ℓ; however, whether Theorem 1 continues to hold under alternative specifications of P's utility or when the payoff from reputations is not separable from material payoffs remains to be explored. Another natural extension is to allow for P's cost c to be heterogeneous (and privately observed). In the extreme case where the distribution of ℓ is degenerate, we can show that no informative communication can be sustained. Studying a

[40] More specifically, to account for $r(\theta)$, when using assumptions on $g(\ell)$ (e.g., continuity or monotonicity) in our baseline model, we impose analogous assumptions for $r(\ell)g(\ell)$. We also slightly redefine $\overline{h}$ and the assumptions used in Proposition 6 and, for Corollary 5, require $\frac{r(e+c)g(e+c)}{(\rho q+c(1-K(e)))^2}$ to be increasing.

model with both heterogeneity in both ℓ and c could shed light on what is needed for reputation to sustain informative communication. Finally, our main results have explored how to maximize the probability of $a = 1$. A natural follow-up question is which equilibrium minimizes the probability of $a = 1$. Although one might naturally conjecture that it would be the least-informative equilibrium (i.e., ex-post signaling), we can construct an F such that a partially-informative LIS yields a lower probability of $a = 1$. More generally, exploring whether imposing regularity properties on F can generate clear comparative statics on how the informativeness of the LIS impacts the probability of $a = 1$ is a promising direction for future work.

References

Acemoglu, D., Egorov, G., and Sonin, K. (2013). A political theory of populism. *The Quarterly Journal of Economics*, 128(2):771–805.

Aghion, P., Dewatripont, M., and Rey, P. (1994). Renegotiation design with unverifiable information. *Econometrica: Journal of the Econometric Society*, pages 257–282.

Agranov, M. (2016). Flip-flopping, primary visibility, and the selection of candidates. *American Economic Journal: Microeconomics*, 8(2):61–85.

Ali, S. N. and Bénabou, R. (2020). Image versus information: Changing societal norms and optimal privacy. *American Economic Journal: Microeconomics*, 12(3):116–164.

Alonso, R. and Câmara, O. (2016). Political disagreement and information in elections. *Games and Economic Behavior*, 100:390–412.

Alós-Ferrer, C. and Prat, J. (2012). Job market signaling and employer learning. *Journal of Economic Theory*, 147(5):1787–1817.

Ball, I. (2022). Scoring strategic agents.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Boleslavsky, R. and Kim, K. (2018). Bayesian persuasion and moral hazard. *Available at SSRN 2913669*.

Brennan, G. and Hamlin, A. (1998). Expressive voting and electoral equilibrium. *Public choice*, 95(1-2):149–175.

Bussing, A. and Pomirchy, M. (2022). Congressional oversight and electoral accountability. *Journal of Theoretical Politics*, 34(1):35–58.

Callander, S. (2008). Political motivations. *The Review of Economic Studies*, 75(3):671–697.

Chen, Y. (2012). Value of public information in sender–receiver games. *Economics Letters*, 114(3):343–345.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

Crawford, V. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Daley, B. and Green, B. (2014). Market signaling with grades. *Journal of Economic Theory*, 151:114–145.

Doval, L. and Skreta, V. (2022). Mechanism design with limited commitment. *Econometrica*, 90(4):1463–1500.

Durbin, E. and Iyer, G. (2009). Corruptible advice. *American Economic Journal: Microeconomics*, 1(2):220–42.

Esteban, J. and Ray, D. (2006). Inequality, lobbying, and resource allocation. *American Economic Review*, 96(1):257–279.

Fox, J. and Van Weelden, R. (2010). Partisanship and the effectiveness of oversight. *Journal of Public Economics*, 94(9):674–687.

Frankel, A. and Kartik, N. (2019). Muddled information. *Journal of Political Economy*, 127(4):1739–1776.

Frankel, A. and Kartik, N. (2022). Improving information from manipulable data. *Journal of the European Economic Association*, 20(1):79–115.

Frisancho, V. and Krishna, K. (2016). Affirmative action in higher education in india: targeting, catch up, and mismatch. *Higher Education*, 71:611–649.

Gentzkow, M. and Kamenica, E. (2016). A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review*, 106(5):597–601.

Grossman, S. J. and Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of political economy*, 94(4):691–719.

Hart, O. and Moore, J. (1988). Incomplete contracts and renegotiation. *Econometrica: Journal of the Econometric Society*, pages 755–785.

Hart, S. and Rinott, Y. (2020). Posterior probabilities: Dominance and optimism. *Economics Letters*, 194:109352.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

Hörner, J. and Lambert, N. S. (2020). Motivational Ratings. *The Review of Economic Studies*, 88(4):1892–1935.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395.

Kartik, N. and Van Weelden, R. (2018). Informative Cheap Talk in Elections. *The Review of Economic Studies*, 86(2):755–784.

Kolotilin, A. (2018). Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635.

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., and Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964.

Kurlat, P. and Scheuer, F. (2021). Signalling to experts. *The Review of Economic Studies*, 88(2):800–850.

Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American economic review*, 97(1):150–168.

Li, W. (2007). Changing One's Mind when the Facts Change: Incentives of Experts and the Design of Reporting Protocols. *The Review of Economic Studies*, 74(4):1175–1194.

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Maskin, E. and Tirole, J. (2004). The politician and the judge: Accountability in government. *American Economic Review*, 94(4):1034–1054.

Morris, S. (2001). Political correctness. *Journal of Political Economy*, 109(2):231–265.

Olszewski, W. (2004). Informal communication. *Journal of Economic Theory*, 117(2):180–200.

Ottaviani, M. and Sorensen, P. N. (2006a). Professional advice. *Journal of Economic Theory*, 126(1):120–142.

Ottaviani, M. and Sorensen, P. N. (2006b). Reputational cheap talk. *The RAND Journal of Economics*, 37(1):155–175.

Prat, A. (2005). The wrong kind of transparency. *American economic review*, 95(3):862–877.

Prendergast, C. (1993). A Theory of Yes Men. *American Economic Review*, 83(4):757–70.

Prendergast, C. and Stole, L. (1996). Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of Political Economy*, 104(6):1105–34.

Ramey, G. (1996). D1 signaling equilibria with multiple signals and a continuum of types. *Journal of Economic Theory*, 69(2):508–531.

Rappoport, D. (2022). Reputational delegation. *Working Paper*.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *Amercian Economic Review*, 80(Jun.):465–479.

Shapiro, J. M. (2016). Special interests and the media: Theory and an application to climate change. *Journal of Public Economics*, 144:91–108.

Singh, J. A. and Upshur, R. E. (2021). The granting of emergency use designation to covid-19 candidate vaccines: implications for covid-19 vaccine trials. *The Lancet Infectious Diseases*, 21(4):e103–e109.

Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies*, 52(4):557–573.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Toikka, J. (2011). Ironing without control. *Journal of Economic Theory*, 146(6):2510–2526.

Zapechelnyuk, A. (2020). Optimal quality certification. *American Economic Review: Insights*, 2(2):161–76.

A. Preliminaries

We begin by defining some useful notation. Given an equilibrium, let $q_m \equiv \nu_1(L|m)$ be the interim belief the DM is an ℓ type, $q(m, e, a) \equiv \nu_2(L|m, e, a)$ be the posterior belief that $\theta \in L$ after message m, action a and evidence e. To avoid unnecessary repetition, we prove all of our results under the assumption of heterogeneous reputation for leniency—i.e., the reputational payoff is $R(m, e, a) \equiv \mathbb{E}[r(\theta)|m, e, a] = \int_\Theta r(\theta) d\nu_2(\theta|m, e, a)$, subject to Assumption 2. If $q_m > 0$, take $G_m(\ell) = \frac{\nu_1(\{\ell' : \ell' \leq \ell\}|m)}{\nu_1(L|m)}$ and $L_m = \text{Supp}(G_m)$. For notational simplicity, we will often drop dependence on F in $U^{\mathcal{E}}_\theta(F)$ in the proofs for Section 3 as it is held fixed. Let $U^{\mathcal{E}}_{\theta,m}$ be the equilibrium expected utility to θ from sending message m.

Our first result uses Assumption 2 to place bounds on the reputations that may arise in equilibrium. We say a is *off-path* after m, e if $\int_{\Theta_m} \zeta(a|\theta, m, e) d\nu_1(\theta|m) = 0$.[41] We say that a is on-path after m, e if it is not off-path.

Lemma 5. *Take any $e \in E$. For all $m \in M$ and $a \in \{0, 1\}$, $q_m \leq \frac{\rho q \overline{r} + c}{\rho \underline{r}} < 1$ and $q(m, e, a) < 1$. For all $m \in M_P$ and on-path a after m and e, $q_m > 0$ and $q(m, e, a) > 0$.*

Proof. First, we show that $U^{\mathcal{E}}_{P,m} \in [-c + \rho q_m \underline{r}, \rho q_m \overline{r}]$ for all $m \in M$. By Corollary 2 of Hart and Rinott (2020), conditional on m, e and $\theta = P$, the expected public belief that $\theta \in L$, namely $\sum_{a \in \{0,1\}} q(m, e, a)\zeta(a|P, m, e)$, is at most q_m. Using $R(m, e, a) \leq \overline{r} q(m, e, a)$, we then have

$$
\begin{aligned}
U^{\mathcal{E}}_{P,m} &= \int_E \left(\sum_{a \in \{0,1\}} (-ca + \rho R(m, e, a))\zeta(a|P, m, e) \right) dF(e) \\
&\leq \int_E \left(\sum_{a \in \{0,1\}} \rho \overline{r} q(m, e, a)\zeta(a|P, m, e) \right) dF(e) \\
&\leq \rho q_m \overline{r}.
\end{aligned}
$$

For each message $m \in M$ and e, Bayes plausibility requires there exists an action $\overline{a}_e$ such that $q(m, \overline{a}_e, e) \geq q_m$. $U^{\mathcal{E}}_{P,m}$ must do weakly better than choosing $\overline{a}_e$ after each e, so $U^{\mathcal{E}}_{P,m} \geq \int_E (-ca_e + \rho R(m, \overline{a}_e, e)) dF(e)$. Using $R(m, e, \overline{a}_e) \geq q(m, e, \overline{a}_e)\underline{r} \geq q_m \underline{r}$, we then have $-c + \rho q_m \underline{r} \leq U^{\mathcal{E}}_{P,m}$.

We next derive similar bounds for the expected equilibrium payoff $U^{\mathcal{E}}_P$. Bayes plausibility implies that for some $m \in M_P$, $q_m \leq q$. For $m \in M_P$, $U^{\mathcal{E}}_P = U^{\mathcal{E}}_{P,m}$, which along with

[41] This definition is slightly different than that used in Ramey (1996), who imposes an additional restriction when defining an equilibrium, if $\int_{\Theta_m} \zeta(a|\theta, m, e) d\nu_1(\theta|m) = 0$ but $\zeta(a|\theta, m, e) > 0$ for some $\theta \in \nu_1(\cdot|m)$, then $\text{Supp}(\nu_2(\cdot|m, e, a)) \subseteq \{\theta \in \Theta_m : \zeta(a|\theta, m, e) > 0\}$. Our results would not change if we imposed this additional condition and defined off-path to be such that $\zeta(a|\theta, m, e) = 0$ for all $\theta \in \text{Supp}(\Theta_m)$.

$U^{\mathcal{E}}_{P,m} \leq \rho q_m \overline{r} \leq \rho q \overline{r}$ gives our desired upper bound. Bayes plausibility also implies that for some $m' \in M$, $q_{m'} \geq q$. Because $U^{\mathcal{E}}_P \geq U^{\mathcal{E}}_{P,m'}$, our desired lower bound follows from $U^{\mathcal{E}}_{P,m'} \geq -c + \rho q_{m'} \underline{r} \geq -c + \rho q \underline{r}$.

Next, for any $m \in M$, we show $q_m \leq \frac{\rho q \overline{r} + c}{\rho \underline{r}} < 1$. Using $-c + \rho q_m \underline{r} \leq U^{\mathcal{E}}_{P,m} \leq U^{\mathcal{E}}_P \leq \rho q \overline{r}$, we have $q_m \leq \frac{\rho q \overline{r} + c}{\rho \underline{r}}$. If $\frac{\rho q \overline{r} + c}{\rho \underline{r}} \geq 1$, then $\rho \leq \frac{c}{\underline{r} - q\overline{r}}$, because $\underline{r} - q\overline{r} > 0$ per Assumption 2. Using the same assumption $\rho \geq \frac{c(\underline{r} + \overline{r})}{\underline{r}^2 - q\overline{r}^2}$, so $\frac{c}{\underline{r} - q\overline{r}} \geq \frac{c(\underline{r} + \overline{r})}{\underline{r}^2 - q\overline{r}^2}$, which simplifies to $0 \geq \overline{r}\underline{r}(1-q)$, a contradiction.

Similarly, for any $m \in M_P$, using $-c + \rho q \underline{r} \leq U^{\mathcal{E}}_P = U^{\mathcal{E}}_{P,m} \leq \rho q_m \overline{r}$, we have $\rho q_m \geq \frac{\rho q \underline{r} - c}{\overline{r}}$. By Assumption 2, $\rho \geq \frac{c(\underline{r} + \overline{r})}{q \underline{r}^2} > \frac{c}{q\underline{r}}$, so $\frac{\rho q \underline{r} - c}{\overline{r}} > 0$, which implies $q_m > 0$.

For the sake of contradiction, suppose $q(m, e, a) = 1$ for some $m \in M, e \in E$, which implies $q_m > 0$. Because $q_m < 1$, $q(m, e, a) = 1$ implies $\zeta(a|P, m, e) = 0$. Then $R(m, e, a) \geq \underline{r}$ while, for $a' \neq a$, $R(m, e, a') \leq q_m \overline{r}$. For P not to have a profitable deviation to choose a, it must be that $-ca' + \rho q_m \overline{r} \geq -ca + \rho \underline{r}$, which implies $q_m \geq \frac{c(a'-a) + \rho \underline{r}}{\rho \overline{r}} \geq \frac{\rho \underline{r} - c}{\rho \overline{r}}$. Combining this inequality with $q_m \leq \frac{\rho q \overline{r} + c}{\rho \underline{r}}$ and simplifying, we conclude that $\rho \leq \frac{c(\overline{r} + \underline{r})}{\underline{r}^2 - q\overline{r}^2}$, a contradiction of Assumption 2. We conclude that $q(m, e, a) < 1$.

Next, suppose $q(m, e, a) = 0$ for some on-path a and $m \in M_P$, which implies $\zeta(a|P, m, e) > 0$. P's utility from taking action a is then $-ca \leq 0$. For $a' \neq a$, $R(m, e, a') \geq \int_\Theta r(\theta) d\nu_1(\theta|m) \geq q_m \underline{r}$. For a to be an equilibrium action for P, it must be that $-ca' + \rho q_m \underline{r} \leq -ca$; simplifying, we get $\rho q_m \leq \frac{c(a'-a)}{\underline{r}} \leq \frac{c}{\underline{r}}$. Combining this inequality with $\rho q_m \geq \frac{\rho q \underline{r} - c}{\overline{r}}$ and simplifying, we have $\rho \leq \frac{c(\overline{r} + \underline{r})}{q \underline{r}^2}$, a contradiction of Assumption 2. We conclude that $q(m, e, a) > 0$. *Q.E.D.*

Using Ramey (1996), we now define the D1 refinement formally in the context of our game. Recall that we are imposing the D1 refinement on the signaling game following message $m \in M$ and evidence e with type space Θ_m.

Take any a that is off-path following some m, e (with $a' = 1 - a$). The reputation payoff from a' is then the interim reputation $\mathbb{E}[r(\theta)|m] = \int_\Theta r(\theta) d\nu_1(\theta|m)$, so the equilibrium payoff for each $\theta' \in \Theta_m$ following m, e is $u(\theta', a', e, \mathbb{E}[r(\theta)|m])$.[42] Suppose there exists non-empty $\Theta'_m \subset \Theta_m$ such that, for all $\theta'' \in \Theta_m \backslash \Theta'_m$, there exists $\theta' \in \Theta'_m$ for which

$$\begin{aligned} &\{\mu \in \Delta(\Theta_m) : \ u\big(\theta'', e, a, \int_{\Theta_m} r(\theta) d\mu(\theta)\big) > u\big(\theta'', e, a', \mathbb{E}[r(\theta)|m]\big)\} \\ &\quad \subsetneq \{\mu \in \Delta(\Theta_m) : \ u\big(\theta', e, a, \int_{\Theta_m} r(\theta) d\mu(\theta)\big) > u\big(\theta', e, a', \mathbb{E}[r(\theta)|m]\big)\}. \end{aligned} \tag{3}$$

[42] One might be worried that there exists a measure zero set of $\theta \in \Theta_m$ take the off-path action a (which is allowed by our definition of off-path), in which case we cannot directly infer that their equilibrium payoff is $u(\theta', a', e, \mathbb{E}[r(\theta)|m])$. However, these payoffs are continuous in the type θ and are equal to $u(\theta', a', e, \mathbb{E}[r(\theta)|m])$ on a measure one set of θ, so they must be equal for all θ.

An equilibrium $\mathcal{E}$ violates D1 if the support of $\nu_2(\cdot|m,e,a)$ is not contained in Θ'_m; $\mathcal{E}$ satisfies D1 if it does not violate D1.

We now show some implications of D1 on the equilibrium actions.

Lemma 6. *Take any $m \in M$ such that $q_m > 0$. Let a be an off-path action following m, e and take $a' = 1 - a$. Then $q(m,e,a) = 1$ if $(e - \tilde{e}_\ell)(a' - a) < 0$ for some $\ell \in L_m$ and $q(m,e,a) = 0$ if $(e - \tilde{e}_\ell)(a' - a) > 0$ for all $\ell \in L_m$.*

Proof. Let a be an off-path action following m, e. By $q_m > 0$, $L_m \neq \emptyset$. We note that

$$\begin{aligned} &\{\mu \in \Delta(\Theta_m) : \; u(P, e, a, \int_{\Theta_m} r(\theta) d\mu(\theta)) > u(P, e, a', \mathbb{E}[r(\theta)|m])\} \neq \emptyset \\ &\iff \rho(\max_{\ell \in L_m} r(\ell) - \mathbb{E}[r(\theta)|m]) > c(a - a'). \end{aligned}$$

The last inequality holds if $\rho(\underline{r} - q_m \overline{r}) > c$, or equivalently $q_m < \frac{\rho \underline{r} - c}{\rho \overline{r}}$, which holds because, by Lemma 5, $q_m < \frac{\rho q \overline{r} + c}{\rho \underline{r}}$ and $\frac{\rho q \overline{r} + c}{\rho \underline{r}} < \frac{\rho \underline{r} - c}{\rho \overline{r}}$ by $\rho \geq \frac{c(\underline{r} + \overline{r})}{\underline{r}^2 - q \overline{r}^2}$ (Assumption 2).

D1 requires $\nu_2(P|m,e,a) = 0$ (which implies $q(m,e,a) = 1$) if (3) holds for $\theta'' = P$ and some $\theta' \in L_m$, which simplifies to $(e - \tilde{e}_\ell)(a' - a) < 0$ for some $\ell \in L_m$. Similarly, D1 requires $\nu_2(L_m|m,e,a) = 0$ (which implies $q(m,e,a) = 0$) if (3) holds for $\theta' = P$ and all $\theta'' \in L_m$, which simplifies to $(e - \tilde{e}_\ell)(a' - a) > 0$ for all $\ell \in L_m$. *Q.E.D.*

B. Proofs from Section 3

Proof of Lemma 1

Proof. First, we show point 1. For the sake of contradiction, suppose $\sigma(\cdot|P)$ and Σ_N are not mutually absolutely continuous. Then there exists $M' \subset M$ such that either $\sigma(M'|P) > \Sigma_N(M') = 0$ or $\Sigma_N(M') > \sigma(M'|P) = 0$. In the first case, there exists $m \in M'$ such that $q_m = 0$, contradicting Lemma 5. In the second case, there exists $m \in M'$ such that $q_m = 1$, contradicting Lemma 5. Therefore, $\sigma(\cdot|P)$ and $\Sigma_N(\cdot)$ are mutually absolutely continuous.

Next, we prove point 2. Take any $m \in M$ and $\ell \in \Theta_m$ such that $e \neq \tilde{e}_\ell$. Let $a = x_\ell(e)$ and $a' = 1 - a$. Because $a' > a$ if and only if $e < \tilde{e}_\ell$, $(e - \tilde{e}_\ell)(a' - a) < 0$. For the sake of contradiction, suppose $\zeta(a'|\ell, m, e) > 0$. Then ℓ (weakly) prefers a' over a, so

$$(e - \ell)a' + \rho R(m, e, a') \geq (e - \ell)a + \rho R(m, e, a). \tag{4}$$

Suppose $\zeta(a|P,m,e)>0$. Then P (weakly) prefers a over a', so

$$-ca+\rho R(m,e,a)\geq -ca'+\rho R(m,e,a'). \tag{5}$$

Adding (5) to (4) and simplifying yields $(e-\tilde{e}_\ell)(a'-a)\geq 0$, a contradiction. Therefore, $\zeta(a|P,m,e)=0$. If a is on-path, then $q(m,e,a)=1$. If a is off-path, then, by Lemma 6, $q(m,e,a)=1$ because $(e-\tilde{e}_\ell)(a'-a)<0$. But $q(m,e,a)=1$ contradicts Lemma 5. Therefore, $\zeta(a'|\ell,m,e)=0$, i.e., $\zeta(x_\ell(e)|\ell,m,e)=1$.

By definition of ν_1, there cannot exist a positive probability set of $\ell\in L$ for which $\sigma(\{m\in M:\ell\notin L_m\}|\ell)>0$. Therefore, there exists $L'\subseteq L$ such that $\nu_0(L'|\theta\in L)=1$ and each $\ell\in L'$, with probability one, sends messages for which $\ell\in L_m$ (namely, $\sigma(\{m\in M:\ell\in L_m|\ell)=1$), for which we have shown $\zeta(x_\ell(e)|\ell,m,e)=1$ when $e\neq\tilde{e}_\ell$. Because either F or G is atomless, the probability of (ℓ,e) such that $e=\tilde{e}_\ell$ is zero, so $\int_E\int_L\int_M\zeta(x_\ell(e)|\ell,m,e)d\sigma(m|\ell)dG(\ell)dF(e)=1$.

Finally, we prove point 3. Take any arbitrary $m\in M_P$ and e,a. Then $q_m>0$ by Lemma 5. If $\zeta(a|P,m,e)=0$ and $\int_L\zeta(a|\ell,m,e)dG_m(\ell)>0$, then $q(m,e,a)=1$, a contradiction of Lemma 5. If $\int_L\zeta(a|\ell,m,e)dG_m(\ell)=0$ and $\zeta(a|P,m,e)>0$, then $q(m,e,a)=0$, a contradiction of Lemma 5. *Q.E.D.*

Proof of Lemma 2

Take an arbitrary $\Lambda\in\Delta(\Delta(L))$ that is Bayes plausible with respect to G. Parameterize a subset $M'\subseteq M$ by the induced belief on L, i.e., let $m_\nu\in M'$ be such that $m_\nu\neq m_{\nu'}$ for $\nu,\nu'\in\Delta(L)$ such that $\nu\neq\nu'$ and take $M_\Lambda=\{m_\nu:\ \nu\in\text{Supp}(\Lambda))\}$. Define $\Sigma_N\in\Delta(M)$ as $\forall\tilde{M}\subset M,\ \Sigma_N(\tilde{M})\equiv\Lambda(\{\nu:m_\nu\in\tilde{M}\})$.

For $\tilde{q}\in(0,1)$ and $\tilde{G}$ a CDF over L, define $z(\cdot;\tilde{q},\tilde{G})$ in the following way. For e such that $\tilde{G}(e+c)\in(0,1)$, let $z(e;\tilde{q},\tilde{G})$ be the unique value of $z\in[0,1]$ such that

$$\rho\frac{\tilde{q}\int_L r(\ell)\mathbb{1}(e\geq\tilde{e}_\ell)d\tilde{G}(\ell)}{\tilde{q}\tilde{G}(e+c)+(1-\tilde{q})z}-c=\rho\frac{\tilde{q}\int_L r(\ell)\mathbb{1}(e<\tilde{e}_\ell)d\tilde{G}(\ell)}{\tilde{q}(1-\tilde{G}(e+c))+(1-\tilde{q})(1-z)}, \tag{6}$$

if such a z exists. Otherwise, take $z(e;\tilde{q},\tilde{G})=0$ if the left-hand side of (6) is lower for all z and $z(e;\tilde{q},\tilde{G})=1$ if the opposite holds. For e such that $\tilde{G}(e+c)=0$, we set $z(e;\tilde{q},\tilde{G})=0$ and for e such that $\tilde{G}(e+c)=1$, set $z(e;\tilde{q},\tilde{G})=1$. Adopting the convention that $\frac{0}{0}=0$, define $\tilde{R}_1(e;\tilde{q},\tilde{G})\equiv\frac{\tilde{q}\int_L r(\ell)\mathbb{1}(e\geq\tilde{e}_\ell)d\tilde{G}(\ell)}{\tilde{q}\tilde{G}(e+c)+(1-\tilde{q})z(e;\tilde{q},\tilde{G})}$ and $\tilde{R}_0(e;\tilde{q},\tilde{G})\equiv\frac{\tilde{q}\int_L r(\ell)\mathbb{1}(e<\tilde{e}_\ell)d\tilde{G}(\ell)}{\tilde{q}(1-\tilde{G}(e+c))+(1-\tilde{q})(1-z(e;\tilde{q},\tilde{G}))}$. It is immediate that $z(e;\tilde{q},\tilde{G})$ is continuous in $\tilde{q}$.

For an arbitrary $\tilde{q} \in (0,1)$ and CDF $\tilde{G}$ on L, define

$$w(e;\tilde{q},\tilde{G}) = \begin{cases} \rho\tilde{q}\int_L r(\ell)d\tilde{G}(\ell) - c & \text{if } \tilde{G}(e+c) = 1, \\ \max\{\rho\tilde{R}_1(e;\tilde{q},\tilde{G}) - c, \rho\tilde{R}_0(e;\tilde{q},\tilde{G})\} & \text{if } \tilde{G}(e+c) \in (0,1), \\ \rho\tilde{q}\int_L r(\ell)d\tilde{G}(\ell) & \text{if } \tilde{G}(e+c) = 0. \end{cases}$$

Given our constructed strategy, this will correspond to the P type's utility after evidence realization e and having induced interim beliefs associated with $(\tilde{q},\tilde{G})$ at the messaging stage. We then define the expected payoff from w as

$$W(\tilde{q};\tilde{G}) \equiv \int_E w(e;\tilde{q},\tilde{G})dF(e).$$

Our next result gives some properties of W.

Claim 1. *$W(\tilde{q};\tilde{G})$ is continuous and strictly increasing in $\tilde{q}$ with $W(\tilde{q};\tilde{G}) \in [\rho\tilde{q}\underline{r} - c, \rho\tilde{q}\overline{r}]$.*

Proof. Continuity is easily seen from the fact that $z(e)$ is continuous in $\tilde{q}$ (we will drop dependence on $\tilde{q},\tilde{G}$ in this proof). That W is increasing in $\tilde{q}$ follows from the fact that w is strictly increasing in $\tilde{q}$ for all e.

We now show $w(e) \in [\rho\tilde{q}\underline{r} - c, \rho\tilde{q}\overline{r}]$ (which immediately implies W respects the same bounds). That this holds for w when $\tilde{G}(e+c) \in \{0,1\}$ is obvious. We therefore focus on the e such that $\tilde{G}(e+c) \in (0,1)$.

Suppose $z(e) \in (0,1)$, so (6) holds with equality, which implies $\tilde{R}_1(e) > \tilde{R}_0(e)$ and

$$\tilde{R}_1(e) \geq \tilde{q}\int_L r(\ell)d\tilde{G}(\ell) \geq \tilde{R}_0(e).$$

The above inequalities imply $\tilde{R}_1(e) \geq \tilde{q}\underline{r}$ and $\tilde{R}_0(e) \leq \tilde{q}\overline{r}$. Thus, because $w(e) = \rho\tilde{R}_0(e) = \rho\tilde{R}_1(e) - c$ in this case, $w(e) \in [\rho\tilde{q}\underline{r} - c, \rho\tilde{q}\overline{r}]$ immediately follows.

Now suppose $z(e) = 0$. Then $w(e) = \rho\tilde{R}_0(e) \geq \rho\tilde{R}_1(e) - c$. That $w(e) \leq \rho\tilde{q}\overline{r}$ then follows from

$$\tilde{R}_0(e) \leq \frac{\tilde{q}\int_{e+c}^{\infty}\overline{r}d\tilde{G}(\ell)}{\tilde{q}(1-\tilde{G}(e+c)) + (1-\tilde{q})} = \tilde{q}\frac{\overline{r}(1-\tilde{G}(e+c))}{\tilde{q}(1-\tilde{G}(e+c)) + (1-\tilde{q})} \leq \tilde{q}\overline{r}. \tag{7}$$

That $w(e) \geq \rho\tilde{q}\underline{r} - c$ then follows from the fact that $\tilde{R}_1(e) = \frac{\int_{-\infty}^{e+c} r(\ell)d\tilde{G}(\ell)}{\tilde{G}(e+c)} \geq \underline{r} \geq \tilde{q}\underline{r}$.

Finally, suppose $z(e) = 1$. Then $w(e) = \rho\tilde{R}_1(e) - c \geq \rho\tilde{R}_0(e)$. That $w(e) \leq \rho\tilde{q}\overline{r}$ follows from the fact that, by analogous argument to that in (7), $\tilde{R}_1(e) \le \tilde{q}\bar{r}$. That $w(e) \ge \rho\tilde{q}\underline{r} - c$ follows from $\tilde{R}_0(e) = \frac{\int_{e+c}^{\infty} r(\ell)d\tilde{G}(\ell)}{1-\tilde{G}(e+c)} \ge \underline{r} \ge \tilde{q}\underline{r}$ when $z(e) = 0$. *Q.E.D.*

We construct P's messaging strategy by specifying a Radon-Nikodym derivative $s(\cdot)$ and defining $\sigma(\cdot|P)$ via $\sigma(\hat{M}|P) = \int_{\hat{M}} s(m)d\Sigma_N(m)$ for any Borel $\hat{M} \subseteq M$. When such strategies are used, what will be the interim belief q_m for $m \in M_\Lambda$ is given by $\varphi(s(m)) \equiv \frac{q}{q+(1-q)s(m)}$. These will correspond to "on-path" interim updates following m. For $\nu \in \Delta(L)$, we let G_ν be the cdf over L corresponding to ν.

By Assumption 2, $\rho > \frac{c(\bar{r}+\underline{r})}{\underline{r}^2 - q\bar{r}^2}$, which implies $\rho > \frac{c}{\underline{r}-q\bar{r}}$, or equivalently $\rho\underline{r} - c > \rho q\bar{r}$. Similarly, $\rho > \frac{c(\underline{r}+\bar{r})}{q\underline{r}^2}$ implies $\rho > \frac{c}{q\underline{r}}$, or equivalently $\rho q\underline{r} - c > 0$. Using these bounds and the bounds on W from Claim 1, we have

$$\begin{aligned} &\lim_{s\to 0} W(\varphi(s), G_\nu) \ge \lim_{s\to 0} \rho\varphi(s)\underline{r} - c = \rho\underline{r} - c > \rho q\bar{r}, \\ &\lim_{s\to \infty} W(\varphi(s), G_\nu) \le \lim_{s\to\infty} \rho\varphi(s) = 0 < \rho q\underline{r} - c. \end{aligned} \tag{8}$$

For $U \in [\rho q\underline{r} - c, \rho q\bar{r}]$, define $s^*(U; m_\nu)$ to be the value of s such that $U = W(\varphi(s), G_\nu)$. We note that such an s exists given (8) and the fact that W is continuous in its first argument. Also because W is continuous and strictly increasing in its first argument, and $\varphi(\cdot)$ is continuous and strictly decreasing, $s^*(U; m_\nu)$ is continuous and strictly decreasing in U.

Claim 2. *There exists a unique $U^* \in [\rho q\underline{r} - c, \rho q\bar{r}]$ such that $1 = \int_{M_\Lambda} s^*(U^*; m_\nu)d\Sigma_N(m_\nu)$. Moreover, for $m_\nu \in M_\Lambda$, we have $\rho\varphi(s^*(U^*; m_\nu))\underline{r} - c \ge 0$ and $z(e; \varphi(s^*(U^*; m_\nu)), G_\nu) \in (0,1)$ whenever $G_\nu(e+c) \in (0,1)$.*

Proof. Take any $m_\nu \in M_\Lambda$. We note that $\varphi(s) \lesseqgtr q$ if and only if $1 \lesseqgtr s$. Let $U = \rho q\underline{r} - c$. Because $W(\tilde{q}; G_\nu) \ge \rho\tilde{q}\underline{r} - c$ for all $\tilde{q}$, we have

$$\rho q\underline{r} - c = U = W(\varphi(s^*(U; m_\nu)), G_\nu) \ge \rho\varphi(s^*(U; m_\nu))\underline{r} - c.$$

Thus, $q \ge \varphi(s^*(U; m_\nu))$, which implies $s^*(U; m_\nu) \ge 1$ and $\int_{M_\Lambda} s^*(U; m_\nu)d\Sigma_N(m_\nu) \ge \int_{M_\Lambda} d\Sigma_N(m_\nu) = 1$.

Let $U' = \rho q\bar{r}$. Because, $W(\tilde{q}; G_\nu) \le \rho\tilde{q}\bar{r}$, for all $\tilde{q}$ we have

$$\rho q\bar{r} = U' = W(\varphi(s^*(U'; m_\nu)), G_\nu) \le \rho\varphi(s^*(U'; m_\nu))\bar{r}.$$

Thus, $q \le \varphi(s^*(U'; m_\nu))$, which implies $s^*(U'; m_\nu) \le 1$ and $\int_{M_\Lambda} s^*(U'; m_\nu)d\Sigma_N(m_\nu) \le \int_{M_\Lambda} d\Sigma_N(m_\nu) = 1$. Because $s(\cdot; m_\nu)$ is continuous and strictly decreasing, there exists a unique $U^* \in [\rho q\underline{r} - c, \rho q\bar{r}]$ such that $1 = \int_{M_\Lambda} s^*(U^*; m_\nu)d\Sigma_N(m_\nu)$.

Take any $m_\nu \in M_\Lambda$. Because not all $s(U^*; m_\nu)$ can be greater than one, there exists $m_{\nu'} \in M_\Lambda$ such that $\varphi(s^*(U^*; m_{\nu'})) \geq q$. By Claim 1, $\rho q \underline{r} - c \leq \rho\varphi(s^*(U^*; m_{\nu'}))\underline{r} - c \leq U^* \leq \rho\varphi(s^*(U^*; m_\nu))\overline{r}$, which implies $\rho\varphi(s^*(U^*; m_\nu)) \geq \frac{\rho q \underline{r} - c}{\overline{r}}$. Then $\rho\varphi(s^*(U^*; m_\nu))\underline{r} - c \geq 0$ if $\frac{\rho q \underline{r} - c}{\overline{r}} \underline{r} \geq c$ or $\rho \geq \frac{c(\overline{r} + \underline{r})}{q \underline{r}^2}$, which holds by Assumption 2.

Finally, suppose $G_\nu(e+c) \in (0,1)$. If $z(e; \varphi(s^*(U^*; m_\nu)), G_\nu) = 0$, then $\rho\tilde{R}_0(e; \varphi(s^*(U^*; m_\nu)), G_\nu) \geq \rho\tilde{R}_1(e; \varphi(s^*(U^*; m_\nu)), G_\nu) - c = \rho\frac{\int_{-\infty}^{e+c} r(\ell) d\tilde{G}(\ell)}{\tilde{G}(e+c)} - c \geq \rho\underline{r} - c$. Moreover, $\rho\tilde{R}_0(e; \varphi(s^*(U^*; m_\nu)), G_\nu) \leq \rho\varphi(s^*(U^*; m_\nu))\overline{r}$, so $\rho\varphi(s^*(U^*; m_\nu)) \geq \frac{\rho\underline{r} - c}{\overline{r}}$. There exists $m_{\nu'} \in M_\Lambda$ such that $\varphi(s^*(U^*; m_{\nu'})) \leq q$, so $\rho\varphi(s^*(U^*; m_\nu))\underline{r} - c \leq U^* \leq \rho\varphi(s^*(U^*; m_{\nu'}))\overline{r} \leq \rho q \overline{r}$, which implies $\rho\varphi(s^*(U^*; m_\nu)) \leq \frac{\rho q \overline{r} + c}{\underline{r}}$. Thus, $\frac{\rho q \overline{r} + c}{\underline{r}} \geq \frac{\rho\underline{r} - c}{\overline{r}}$, or equivalently $\frac{c(\overline{r} + \underline{r})}{\underline{r}^2 - q\overline{r}^2} \geq \rho$, a contradiction of Assumption 2. An analogous argument rules out $z(e; \varphi(s^*(U^*; m_\nu)), G_\nu) = 1$, so $z(e; \varphi(s^*(U^*; m_\nu)), G_\nu) \in (0,1)$. *Q.E.D.*

We now construct an equilibrium $\mathcal{E}$ associated with the LIS Λ. As is well-known, for any Bayes-plausible Λ, there exists a signal structure that induces it (Kamenica and Gentzkow (2011)) which corresponds to a set of strategies $\{\sigma(\cdot|\ell)\}_{\ell \in L}$ with support on M_Λ such that the posterior on L (conditional on $\theta \in L$) after $m_\nu \in M_\Lambda$ is ν. In particular $\sigma(m_\nu|\ell) = 0 \; \forall \ell \notin \text{Supp}(\nu)$. Define $\sigma(\cdot|P)$ by $d\sigma(\cdot|P) = s^*(U^*; m) d\Sigma_N(m)$. Let ν_1 be defined as, for $m_\nu \in M_\Lambda$,

$$\nu_1(\tilde{\Theta}|m_\nu) = \varphi(s^*(U^*; m_\nu))\nu(\tilde{\Theta} \backslash \{P\}) + (1 - \varphi(s^*(U^*; m_\nu)))\mathbb{1}(P \in \tilde{\Theta}),$$

and $\nu_1(P|m) = 1$ if $m \notin M_\Lambda$.

The decision-stage strategies are given by

$$\zeta(1|\ell, m, e) = \begin{cases} x_\ell(e) & \text{if } m = m_\nu \in M_\Lambda, \; \ell \in L_m, \\ \mathbb{1}(1 \in \arg\max_a \; u(\ell, e, a, \tilde{R}_a(e; \nu_1(N|m_\nu), G_\nu))) & m = m_\nu \in M_\Lambda, \; \ell \notin L_m. \\ \mathbb{1}(1 \in \arg\max_a \; u(\ell, e, a, 0)) & m \notin M_\Lambda. \end{cases}$$

$$\zeta(1|P, m, e) = \begin{cases} z(e; \varphi(s^*(U^*; m_\nu), G_\nu) & \text{if } m = m_\nu \in M_\Lambda, \\ 0 & \text{else.} \end{cases}$$

For $m_\nu \in M_\Lambda$ and on-path a following m_ν, e, let $\nu_2(\cdot|m, e, a)$ be the Bayes update induced by these strategies, i.e., $\nu_2(\tilde{\Theta}|m_\nu, e, a) = \frac{\int_{\tilde{\Theta}} \zeta(a|\theta, m, e) d\nu_1(\theta|m_\nu)}{\int_{\Theta} \zeta(a|\theta, m, e) d\nu_1(\theta|m_\nu)}$ for all $\tilde{\Theta} \subset \Theta$; otherwise, we set $\nu_2(P|m, e, a) = 1$. This generates a reputation of $R(m, e, a) = \tilde{R}_a(e; \nu_1(N|m_\nu), G_\nu)$ for $m_\nu \in M_\Lambda$ and 0 otherwise.

By construction these strategies generate an expected utility for P of U^*. Our next claim verifies that $\mathcal{E}$ is an equilibrium.

Claim 3. *$\mathcal{E}$ is an equilibrium.*

Proof. We start by verifying that P has no incentive to deviate. First, we consider the decision stage. Take $m \notin M_\Lambda$. Then $\nu_1(P|m) = 1$, then reputation is 0 regardless of the action so $a = 0$ is clearly optimal. Take $m_\nu \in M_\Lambda$. When $G_\nu(e+c) \in (0,1)$, $z(e; \varphi(s^*(U^*; m_\nu), G_\nu) \in (0,1)$ implies P is indifferent over actions by construction. P also clearly has no incentive to deviate to $a = 1$ when $G_\nu(e+c) = 0$ because it is worse from a material and reputational perspective. Finally, P has no incentive to deviate to $a = 0$ when $G_\nu(e+c) = 1$ as his payoff from $a = 1$ is at least $\rho\varphi(s^*(U^*; m_\nu)\underline{r} - c > 0$ and his payoff from $a = 0$ is zero. There is also no incentive to deviate at the communication stage: P is indifferent across all $m_\nu \in M_\Lambda$ by construction and because $U^* \geq \rho q \underline{r} - c > 0$, strictly prefers the expected utility of U^* from any $m_\nu \in M_\Lambda$ to the expected utility of 0 from sending $m \notin M_\Lambda$.

Next, we show that no ℓ type has an incentive to deviate at the decision stage following m such that $\ell \in \text{Supp}(\nu_1(\cdot|m))$ (which implies $m \in M_\Lambda$); that there is no incentive to deviate after any other m follows immediately from the definition of ζ. Take an arbitrary $m_\nu \in M_\Lambda$, $\ell \in L_{m_\nu}$ and e. Set $a = x_\ell(e)$ and $a' = 1 - a$. By the definition of x_ℓ, $(e - \tilde{e}_\ell)(a - a') \geq 0$. By the definition of z and Claim 2 $x_\ell(e) \in \text{Supp}(\zeta(\cdot|P, m_\nu, e))$, so P's incentive constraint implies $-ca + \rho R(m_\nu, e, a) \geq -ca' + \rho R(m_\nu, e, a')$. If ℓ has a strict incentive to deviate to a', then $(e - \ell)a + \rho R(m_\nu, e, a) < (e - \ell)a' + \rho R(m_\nu, e, a')$. Subtracting P's incentive constraint and simplifying, we get $(e - \tilde{e}_\ell)(a - a') < 0$, a contradiction.

Next, we consider ℓ's incentive to deviate at the communication stage. Because $\sigma(\{m_\nu \in M_\Lambda : \ \ell \in L_{m_\nu}\}|\ell) = 1$, it suffices to show that ℓ cannot do better than sending a message $m_\nu \in M_\Lambda$ such that $\ell \in L_{m_\nu}$. Take such an m and suppose ℓ has a profitable deviation to announce message m' and follow contingent plan $x' \in \mathcal{X}$, so that

$$\int_E ((e-\ell)x'(e) + \rho R(m', x'(e), e))dF(e) > \int_E ((e-\ell)x_\ell(e) + \rho R(m_\nu, x_\ell(e), e))dF(e).$$

Because P finds it optimal to send an arbitrary $\tilde{m}_\nu \in M_\Lambda$ and use strategy x_ℓ for $\ell \in L_{\tilde{m}_\nu}$, this means that P prefers to send m_ν and follow with x_ℓ, than send m' and follow with x'.

$$\int_E (-cx_\ell(e) + \rho R(m_\nu, x_\ell(e), e))dF(e) \geq \int_E (-cx'(e) + \rho R(m, x'(e), e))dF(e).$$

Adding these inequalities together and simplifying, we get $\int_E (c+e-\ell)x'(e)dF(e) > \int_E (c+e-\ell)x_\ell(e)dF(e)$, a contradiction of $x_\ell \in \arg\max_{x \in \mathcal{X}} \int_E (c+e-\ell)x(e)dF(e)$. Therefore, ℓ must have no incentive to deviate at the communication stage.

Finally, we show that D1 is satisfied. It is trivially satisfied following $m \notin M_\Lambda$ since

$\nu_1(P|m) = 1$.[43] Take $m_\nu \in M_\Lambda$. The only off-path actions following m_ν occur when $G_\nu(e + c) \in \{0, 1\}$ by construction and Claim 2. If $G_\nu(e + c) = 1$, then $a = 0$ is an off-path action. There are two cases to consider: when $e > \max_{\ell \in L_{m_\nu}} \tilde{e}_\ell$ and when $e = \max_{\ell \in L_{m_\nu}} \tilde{e}_\ell$. In the first case, by Lemma 6, D1 requires $\nu_2(P|m_\nu, e, a) = 1$ because $e - \tilde{e}_\ell > 0$ for all $\ell \in L_m$. For the second case, we now show that $\nu_2(P|m_\nu, e, a) = 1$ is consistent with D1. D1 requires no weight be placed on any $\ell \in \Theta_m$ whenever P has a larger incentive to deviate to a than ℓ, namely

$$\{\mu \in \Delta(\Theta_{m_\nu}) : \ (e - \ell)a' + \rho \int_{\Theta_{m_\nu}} r(\theta) d\mu(\theta) > (e - \ell)a + \rho\mathbb{E}[r(\theta)|m_\nu]\}$$
$$\subsetneq \{\mu \in \Delta(\Theta_{m_\nu}) : \ -ca' + \rho \int_{\Theta_{m_\nu}} r(\theta) d\mu(\theta) > -ca + \rho\mathbb{E}[r(\theta)|m_\nu]\},$$

which rules out all $\ell < \max L_{m_\nu}$ when $e = \max_{\ell \in L_{m_\nu}} \tilde{e}_\ell$. However, the above sets are equal for $\ell' = \max L_{m_\nu}$ at such e, in which case any beliefs that ascribe probability only on ℓ' and P are consistent with D1. Thus, $\nu_2(P|m_\nu, e, a) = 1$ is consistent with D1. An analogous argument holds for when $G_\nu(e + c) = 0$.

We know by Lemma 1 that the N's distribution over actions and evidence is unique (and the same for all LIS) up to zero probability events. That P's equilibrium distribution is unique follows from the fact that that $s^*(U^*; m)$ defines the unique messaging strategy that leaves P indifferent across messages $m_\nu \in M_\Lambda$ and $z(e; \varphi(s^*(U^*; m_\nu)), G_\nu)$ is the unique mixture over equilibrium mixture over actions given interim beliefs $(q_m, G_m) = (\varphi(s^*(U^*; m_\nu)), G_\nu)$. While an equilibrium may feature multiple messages that induces the same G_ν contingent on $\theta \in L$, P must mix over these messages with the same probability inducing the same interim belief $\varphi(s^*(U^*; m_\nu))$ over all such messages; if not, one would have a $\varphi(s^*(U^*; m_\nu))$ higher than the others, which P would then strictly prefer. Thus, in any equilibrium with an LIS of Λ, the equilibrium outcomes are unique. By Lemma 1, P is indifferent between mimicking the strategy of each ℓ type. Therefore, for each $m \in M_\ell$, $U^* = \int_E (-cx_\ell(e) + \rho R(m, e, x_\ell(e))) dF(e)$, so ℓ's equilibrium utility is $\int_0^1 (e - \ell)x_\ell(e) + \rho R(m, e, x_\ell(e))) dF(e) = \int_E (e - \ell + c)x_\ell(e) dF(e) + U^*$. Thus, the expected utility of ℓ is unique by the uniqueness of U^*. *Q.E.D.*

[43] This triviality comes from the fact that our D1 refinement is specified for interim beliefs. Because $\nu_1(P|m) = 1$ after $m \notin M_\Lambda$, there is no uncertainty at the interim stage, and so our D1 refinement has no bite. In general, our D1 refinement cannot restrict the beliefs for actions following "off-path messages," the implication can also be shown for other "ex-ante" D1 refinements.

C. Proofs from Section 4

Lemma 7 (Opposing Interests).
For every equilibrium $\mathcal{E}$, $V^{\mathcal{E}}(F) = \frac{1}{c}\left(\rho\mathbb{E}[r(\theta)] - U_P^{\mathcal{E}}(F)\right)$.

Proof. Take any equilibrium $\mathcal{E}$. Because of Lemma 1, after $m \in M_P$, P is indifferent across mimicking the strategy of a probability one set of leniency type $\ell \in L_m$:

$$\begin{aligned} U_P^{\mathcal{E}}(F) = \int_E \Big(& - c\zeta(1|\ell, m, e) \\ & + \rho\{\zeta(1|\ell, m, e)R(m, e, 1) + \zeta(0|\ell, m, e)R(m, e, 0)\}\Big) dF(e) \end{aligned}$$

The same equality holds if we replace ℓ with P. Taking expectations of both sides with respect to $\nu_1(\cdot|m)$ and using the law of iterated expectations then yields

$$\begin{aligned} U_P^{\mathcal{E}}(F) &= \int_L \Big\{ \int_E \Big(- c\zeta(1|\theta, m, e) \\ &\quad + \rho\{\zeta(1|\theta, m, e)R(m, e, 1) + \zeta(0|\theta, m, e)R(m, e, 0)\}\Big) dF(e)\Big\} d\nu_1(\theta|m) \\ &= -c\mathbb{P}(a = 1|m) + \rho\mathbb{E}[r(\theta)|m]. \end{aligned}$$

Taking the ex-ante expectation of both sides over messages in M_P (which is a probability one set under $\sigma(\cdot|P)$ and $\Sigma_N(\cdot)$ by Lemma 1) and again applying the law of iterated expectations then yields

$$\begin{aligned} U_P^{\mathcal{E}}(F) &= \int_{M_P} (-c\mathbb{P}(a = 1|m) + \rho\mathbb{E}[r(\theta)|m])(qd\sigma(m|P) + (1 - q)d\Sigma_N(m)) \\ &= -c\mathbb{P}(a = 1) + \rho\mathbb{E}[r(\theta)] \end{aligned}$$

Rearranging terms and using $V^{\mathcal{E}}(F) = \mathbb{P}(a = 1)$ then yields our desired result. *Q.E.D.*

Proof of Lemma 4

Proof. We first derive an equation for determining $U_P^{\alpha}(F)$. Because of the uniqueness in Lemma 2, it is without loss to focus on our constructed equilibrium in the proof of that lemma for the perfectly informative LIS. Under this equilibrium each message in M_Λ is associated with a single leniency type ℓ, denote it m_ℓ, in the sense that $L_{m_\ell} = \{\ell\}$. Both ℓ and P follow up each m_ℓ with x_ℓ. This means that $R(m_\ell, e, x_\ell(e))$. Because $\Sigma_N(\cdot)$ and $\sigma(\cdot|P)$ are mutually absolutely continuous, we can describe P's messaging strategy by the

Radon-Nikodym derivative $s(m_\ell) = \frac{d\sigma(m_\ell|P)}{d\Sigma_N(m_\ell)}$ so that $\sigma(\hat{M}|P) = \int_{\hat{M}} s(m_\ell) d\Sigma_N(m_\ell)$ for each $\hat{M} \subseteq M_\Lambda$. Thus, by Bayes rule, $R(m_\ell, e, x_\ell(e)) = \frac{qr(\ell)}{q+(1-q)s(m_\ell)}$ for all e. P's expected material payoff from x_ℓ is $-c(1 - F(\tilde{e}_\ell))$ and so his utility is given by

$$U_P^\alpha(F) = -c(1 - F(\tilde{e}_\ell)) + \rho \frac{qr(\ell)}{q + (1-q)s(m_\ell)} \; \forall \ell \in L.$$

We then have $q + (1-q)s(m_\ell) = \frac{\rho qr(\ell)}{U_P^\alpha(F)+c(1-F(\tilde{e}_\ell))}$. Taking the expectation over both sides with respect to ℓ and using, by $\sigma(m_\ell|\ell') = \mathbb{1}(\ell' = \ell)$, $\int_L s(m_\ell)dG(\ell) = \int_L s(m_\ell)d\Sigma_N(m_\ell) = \int_{M_\Lambda} d\sigma(m|P) = 1$, we have

$$1 = \int_L \frac{\rho qr(\ell) dG(\ell)}{U_P^\alpha(F) + c - cF(\tilde{e}_\ell)}. \tag{9}$$

Take an arbitrary pair of CDFs F_1, F_2 and $\lambda \in (0,1)$ and define $F_\lambda = \lambda F_1 + (1-\lambda)F_2$. Using (9), we then have

$$\begin{aligned}
&\int_L \frac{\rho qr(\ell)dG(\ell)}{U_P^\alpha(F_\lambda) + c - cF_\lambda(\tilde{e}_\ell)} \qquad (10)\\
&= \lambda \int_L \frac{\rho qr(\ell)dG(\ell)}{U_P^\alpha(F_1) + c - cF_1(\tilde{e}_\ell)} + (1-\lambda) \int_L \frac{\rho qr(\ell)dG(\ell)}{U_P^\alpha(F_2) + c - cF_2(\tilde{e}_\ell)}\\
&\geq \int_L \frac{\rho qr(\ell)dG(\ell)}{\lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2) + c - c(\lambda F_1(\tilde{e}_\ell) + (1-\lambda)F_2(\tilde{e}_\ell))}\\
&= \int_L \frac{\rho qr(\ell)dG(\ell)}{\lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2) + c - cF_\lambda(\tilde{e}_\ell)},
\end{aligned}$$

where the inequality follows from the fact that $\frac{1}{y}$ is convex in y. This inequality implies $U_P^\alpha(F_\lambda) \leq \lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2)$. *Q.E.D.*

As discussed in the text, $U_P^\alpha(\delta_e) = U_P^\beta(\delta_e)$ for all e, so Lemma 4 implies

$$U_P^\alpha(F) \leq \int_E U_P^\alpha(\delta_e)dF(e) = \int_E U_P^\beta(\delta_e)dF(e) = U_P^\beta(F). \tag{11}$$

Note that the inequality in (10) is strict if there exists $L' \subseteq L$ such that $\int_{L'} dG(\ell) > 0$ and $F_1(\tilde{e}_\ell) \neq F_2(\tilde{e}_\ell)$ for all $\ell \in L'$, in which case we have $U_P^\alpha(F_\lambda) < \lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2)$. In (11), we are taking a convex combination over δ_e, so the inequality is strict if there exists L', E' such that $\int_{L'} dG(\ell) > 0$, $\int_{E'} dF(e) > 0$ and $\delta_e(\tilde{e}_\ell) = \mathbb{1}(e \geq \tilde{e}_\ell) \neq \mathbb{1}(e' \geq \tilde{e}_\ell) = \delta_{e'}(\tilde{e}_\ell)$ for all $\ell \in L'$ and $e, e' \in E'$. Suppose β has residual strategic uncertainty and mild agreement holds and, for the sake of contradiction, that no such L', E' exist. Then for a probability one set of ℓ types, either $F(\tilde{e}_\ell) = 0$ or $F(\tilde{e}_\ell) = 1$. If $F(\tilde{e}_\ell) = 0$ for a probability one set of ℓ, then there is no residual strategic uncertainty, a contradiction. A similar argument holds if $F(\tilde{e}_\ell) = 1$ for a probability one set of ℓ. Therefore, there must exist a positive probability set of ℓ' such that $F(\tilde{e}_{\ell'}) = 0$ and a positive probability set of ℓ'' such that $F(\tilde{e}_{\ell''}) = 1$. But then there is no $e \in \text{Supp}(F)$ for which $x_{\ell'}(e) = x_{\ell''}(e)$, a contradiction of mild agreement. Thus, under mild agreement and residual strategic uncertainty for β, $U_P^\alpha(F) < U_P^\beta(F)$. We use this observation in the proof of Theorem 1 below.

Proof of Theorem 1

Proof. Take any equilibrium $\mathcal{E}$ with strategies $\{\sigma(\cdot|\theta)\}_{\theta\in\Theta}$ and recall that $\Sigma_N(\cdot) = \int_L \sigma(\cdot|\ell)dG(\ell)$. By Lemma 3, it suffices to show $U_P^{\mathcal{E}}(F) \geq U_P^\alpha(F)$, with a strict inequality if $\mathcal{E}$ has residual strategic uncertainty and there is mild agreement.

Recall that G_m and q_m are the interim beliefs associated after $m \in M_P$ in $\mathcal{E}$ and define $U_P^{\beta,m}(F)$ to be the ex-post signaling utility when $\ell \sim G_m$, $\mathbb{P}(\theta \in L) = q_m$ and $e \sim F$. Note that $U_P^{\mathcal{E}}(F) = U_P^{\beta,m}(F)\ \forall m \in M^P$. P's utility following message m and evidence e is given by $U_P^{\beta,m}(\delta_e)$.

Define $U_P^{\alpha,m}(F)$ to be the (unique) value of U that solves $\int_L \frac{\rho q_m r(\ell)}{U+c-cF(\tilde{e}_\ell)}dG_m(\ell) = 1$.[44] We now show $U_P^{\beta,m}(\delta_e) = U_P^{\alpha,m}(\delta_e)$. It suffices to show $\int_L \frac{\rho q_m r(\ell)}{U_P^{\beta,m}(\delta_e)+c-c\delta_e(\tilde{e}_\ell)}dG_m(\ell) = 1$. Suppose $G_m(e+c) \in (0,1)$. Let z be the probability P selects $a = 1$ when evidence is e; by Lemma 1, $z \in (0,1)$. Then $a = 0$ is an optimal action for P, so $U_P^{\beta,m}(\delta_e) = \rho\frac{q_m \int_L r(\ell)\mathbb{1}(e<\tilde{e}_\ell)dG_m(\ell)}{q_m(1-G_m(e+c))+(1-q_m)z}$, which implies $q_m(1-G_m(e+c))+(1-q_m)z = \frac{\rho q_m \int_L r(\ell)\mathbb{1}(e<\tilde{e}_\ell)dG_m(\ell)}{U_P^{\beta,m}(\delta_e)}$. Similarly, because $a = 1$ is also an optimal action, $U_P^{\beta,m}(\delta_e) = \rho\frac{q_m \int_L r(\ell)\mathbb{1}(e\geq\tilde{e}_\ell)dG_m(\ell)}{q_m G_m(e+c)+(1-q_m)z} - c$, which implies $q_m G_m(e+c)+(1-q_m)z = \frac{\rho q_m \int_L r(\ell)\mathbb{1}(e\geq\tilde{e}_\ell)dG_m(\ell)}{U_P^{\beta,m}(\delta_e)+c}$. Adding these together, we have

$$
\begin{aligned}
1 &= \frac{\rho q_m \int_L r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)dG_m(\ell)}{U_P^{\beta,m}(\delta_e)+c} + \frac{\rho q_m \int_L r(\ell)\mathbb{1}(e < \tilde{e}_\ell)dG_m(\ell)}{U_P^{\beta,m}(\delta_e)} \\
&= \int_L \frac{\rho q_m r(\ell)}{U_P^{\beta,m}(\delta_e)+c-c\mathbb{1}(e<\tilde{e}_\ell)}dG_m(\ell) \\
&= \int_L \frac{\rho q_m r(\ell)}{U_P^{\beta,m}(\delta_e)+c-c\delta_e(\tilde{e}_\ell)}dG_m(\ell).
\end{aligned}
$$

The argument when $G_m(e+c) \in \{0,1\}$ is analogous.

[44] That a unique solution exists follows from the following arguments. As shown in the proof of Lemma 5, $\rho q_m \underline{r} > c$ which implies $\int_L \frac{\rho q_m r(\ell)}{U+c-cF(\tilde{e}_\ell)}dG_m(\ell) > 1$ when $U = 0$. Because $\int_L \frac{\rho q_m r(\ell)}{U+c-cF(\tilde{e}_\ell)}dG_m(\ell)$ is strictly decreasing in U with a limit of 0 as $U \to \infty$, a unique solution to $\int_L \frac{\rho q_m r(\ell)}{U+c-cF(\tilde{e}_\ell)}dG_m(\ell) = 1$ exists.

By the arguments made in Lemma 4, $U_P^{\alpha,m}(\cdot)$ is convex and so,[45] for all $m \in M_P$, we have

$$U_P^{\alpha,m}(F) \leq \int_E U_P^{\alpha,m}(\delta_e)dF(e) = \int_E U_P^{\beta,m}(\delta_e)dF(e) = U_P^{\beta,m}(F) = U_P^{\mathcal{E}}(F). \tag{12}$$

For the sake of contradiction, suppose $U_P^{\alpha}(F) > U_P^{\mathcal{E}}(F)$. Then, by (12), $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ for all $m \in M_P$. By Lemma 1, $\Sigma_N(M_P) = \sigma(M_P|P) = 1$, we can take the expectation over $m \in M_P$ of both sides of $\int_L \frac{\rho q_m r(\ell)}{U_P^{\alpha,m}(F)+c-cF(\tilde{e}_\ell)} dG_m(\ell) = 1$ to get

$$\begin{aligned}
1 &= \int_{M_P} \left[\int_L \frac{\rho q_m r(\ell) dG_m(\ell)}{U_P^{\alpha,m}(F) + c - cF(\tilde{e}_\ell)} \right] (qd\Sigma_N(m) + (1-q)d\sigma(m|P)) \qquad (13)\\
&= \int_L \int_{M_P} \frac{\rho q r(\ell)}{U_P^{\alpha,m}(F) + c - cF(\tilde{e}_\ell)} d\sigma(m|\ell) dG(\ell),\\
&> \int_L \int_{M_P} \frac{\rho q r(\ell)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_\ell)} d\sigma(m|\ell) dG(\ell)\\
&= \int_L \frac{\rho q r(\ell)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_\ell)} dG(\ell)\\
&= 1
\end{aligned}$$

where the second equality follows from Bayes rule, the inequality follows from $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ and the final equality by (9) in the proof of Lemma 4, a contradiction. Therefore, we conclude that $U_P^{\alpha}(F) \leq U_P^{\mathcal{E}}(F)$.

Finally, suppose that $U_P^{\alpha}(F) = U_P^{\mathcal{E}}(F)$ when there is mild agreement and residual strategic uncertainty in $\mathcal{E}$. As we have shown after Lemma 4, mild agreement and residual strategic uncertainty implies $\int_E U_P^{\beta,m}(\delta_e)dF(e) > U_P^{\alpha,m}(F)$ so, by (12), $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ for $m \in M_P$. The same arguments as above in (13) lead to a contradiction. Therefore $U_P^{\alpha}(F) < U_P^{\mathcal{E}}(F)$ when there is mild agreement and residual strategic uncertainty in $\mathcal{E}$. *Q.E.D.*

Proof of Proposition 1

Proof. For notational simplicity, we drop dependence of v^α on F. Let E' be the set of $e \in E$ such that $G(e+c)$ has no mass-point (i.e., G is discontinuous at e); by our assumptions that either F or G is atomless, $\int_{E'} dF(e) = 1$. Take any evidence level $e \in E'$. The proof is immediate if $G(e+c) = 0$ as $v^\alpha(e) = v^\beta(e) = 0$ or if $G(e+c) = 1$ as $v^\alpha(e) = v^\beta(e) = 1$. Suppose $G(e+c) \in (0,1)$. Note that $v^\alpha(e) = \int_L \mathbb{1}(e \geq \tilde{e}_\ell)(qdG(\ell) + (1-q)d\sigma(m_\ell|P))$. Let $s(m_\ell) = \frac{d\sigma(m_\ell|P)}{d\Sigma_N(m_\ell)}$ be the Radon-Nikodym derivative as in the proof of Lemma 4. Using

[45] The arguments in Lemma 4 showing $U_P^{\alpha}(F)$ is convex only relied on the fact that $U_P^{\alpha}(F)$ is the solution to $\int_L \frac{\rho q r(\ell) dG(\ell)}{U+c-cF(\tilde{e}_\ell)} = 1$, and so apply to $U_P^{\alpha,m}$ as well.

the fact that $\sigma(m_\ell|\ell') = \mathbb{1}(\ell = \ell')$ under ex-ante signaling, we have $v^\alpha(e) = \int_{-\infty}^{e+c}(q + (1-q)s(m_\ell))dG(\ell)$. As shown in the proof of Lemma 4, $q + (1-q)s(m_\ell) = \frac{\rho q r(\ell)}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)}$, so $v^\alpha(e) = \int_L \frac{\rho q r(\ell)\mathbb{1}(e\geq\tilde{e}_\ell)}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)}dG(\ell)$.

Let $\underline{G}^r(e) \equiv \int_L r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)dG(\ell)$ and $\overline{G}^r(e) \equiv \int_L r(\ell)\mathbb{1}(e < \tilde{e}_\ell)dG(\ell)$. It is straightforward to show that $v^\beta(e)$ is the unique solution to

$$\frac{\rho q\underline{G}^r(e)}{v^\beta(e)} - c = \frac{\rho q\overline{G}^r(e)}{1-v^\beta(e)}.$$

This means $v^\beta(\cdot)$ does not depend on F and only depends on (G, r) through $\underline{G}^r(\cdot)$ and $\overline{G}^r(\cdot)$.

We show that $v^\alpha(e) - v^\beta(e) \geq 0$ by showing that this inequality holds when we select the distribution of ℓ and reputations $(\widehat{G}, \hat{r})$ to minimize $v^\alpha(e)$ while holding $v^\beta(e)$ fixed. This latter requirement is equivalent to requiring $\int_E \hat{r}(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)d\widehat{G}(\ell) = \underline{G}^r(e)$ and $\int_L \hat{r}(\ell)\mathbb{1}(e < \tilde{e}_\ell)d\widehat{G}(\ell) = \overline{G}^r(e)$ in which case we refer to $(\widehat{G}(\ell), \hat{r})$ as feasible.

It is without loss to focus on F such that $\mathrm{Supp}(F)$ is contained in a compact interval.[46] We then construct a feasible $(\widehat{G}, \hat{r})$ where $\widehat{G}$ has binary support and yields a lower $v^\alpha(e)$ than (G, r). Take some $\ell'' < \min \mathrm{Supp}(F) + c$ and $\ell' > \max \mathrm{Supp}(F) + c$. Define $(\widehat{G}, \hat{r})$ by

$$(\widehat{G}(\ell), \hat{r}(\ell)) = \begin{cases} (0, r(\ell)) & \text{if } \ell < \ell'', \\ (G(e+c), \frac{\underline{G}^r(e)}{G(e+c)}) & \text{if } \ell'' \leq \ell < \ell', \\ (1, \frac{\overline{G}^r(e)}{1-G(e+c)}) & \text{if } \ell \geq \ell'. \end{cases}$$

Let U and $\widehat{U}$ be the corresponding ex-ante signaling equilibrium expected utilities for P under (G, r) and $(\widehat{G}, \hat{r})$ respectively. We will show that the difference between $v^\alpha(e)$ under (G, r) and $(\widehat{G}, \hat{r})$ given by

$$\int_E \frac{\rho q r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)}{U + c - cF(\tilde{e}_\ell)}dG(\ell) - \frac{\rho q\underline{G}^r(e)}{\widehat{U}+c} \geq \max\left\{\frac{\rho q\underline{G}^r(e)}{U+c} - \frac{\rho q\underline{G}^r(e)}{\widehat{U}+c}, \frac{\rho q\overline{G}^r(e)}{\widehat{U}} - \frac{\rho q\overline{G}^r(e)}{U}\right\},$$

which is greater than 0 for any $\hat{U}, U$. To see the the LHS is greater than the first term on the

[46] For any F with unbounded support, we can consider a version of F truncated at $[-z, z]$ for some $z \in \mathbb{R}$; taking $z \to \infty$, it is straightforward to show that the value of $v^\alpha(e)$ under the truncated F will converge to the value of $v^\alpha(e)$ under the original F.

RHS,

$$\begin{aligned}\int_L \frac{\rho q r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)}{U+c-cF(\tilde{e}_\ell)} dG(\ell) - \frac{\rho q \underline{G}^r(e)}{\widehat{U}+c} &\geq \int_L \frac{\rho q r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)}{U+c} dG(\ell) - \frac{\rho q \underline{G}^r(e)}{\widehat{U}+c} \\ &= \frac{\rho q \underline{G}^r(e)}{U+c} - \frac{\rho q \underline{G}^r(e)}{\widehat{U}+c}.\end{aligned}$$

To see that the LHS is greater than the second term on the RHS, note that by the definition of U and $\hat{U}$ $\int_L \frac{\rho q r(\ell)}{U+c-cF(\tilde{e}_\ell)} dG(\ell) = 1 = \int_L \frac{\rho q \hat{r}(\ell)}{\widehat{U}+c-cF(\tilde{e}_\ell)} d\widehat{G}(\ell)$, which implies

$$\int_L \frac{\rho q r(\ell)}{U+c-cF(\tilde{e}_\ell)} dG(\ell) = \frac{\rho q \underline{G}^r(e)}{\widehat{U}+c} + \frac{\rho q \overline{G}^r(e)}{\widehat{U}}.$$

Rearranging terms, we get

$$\begin{aligned}\int_E \frac{\rho q r(\ell)\mathbb{1}(e \geq \tilde{e}_\ell)}{U+c-cF(\tilde{e}_\ell)} dG(\ell) - \frac{\rho q \underline{G}^r(e)}{\widehat{U}+c} &= \frac{\rho q \overline{G}^r(e)}{\widehat{U}} - \int_L \frac{\rho q r(\ell)\mathbb{1}(e < \tilde{e}_\ell)}{U+c-cF(\tilde{e}_\ell)} dG(\ell) \\ &\geq \frac{\rho q \overline{G}^r(e)}{\widehat{U}} - \int_L \frac{\rho q r(\ell)\mathbb{1}(e < \tilde{e}_\ell)}{U} dG(\ell) \\ &= \frac{\rho q \overline{G}^r(e)}{\widehat{U}} - \frac{\rho q \overline{G}^r(e)}{U}.\end{aligned}$$

We conclude that $v^\alpha(e)$ is (weakly) smaller under $(\widehat{G}, \hat{r})$. Thus, $v^\alpha(e)$ is minimized using a binary support $\widehat{G}$. For a binary support $\{\underline{\ell}, \overline{\ell}\}$, $v^\alpha(e) - v^\beta(e)$ is zero for $e \notin [\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$, and constant for $e \in [\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$, so Theorem 1 establishes that $v^\alpha(e) - v^\beta(e) \geq 0$. *Q.E.D.*

D. Proofs from Section 5

We will state the proof of Theorem 2 below for the model allowing for differential type reputations. The results in this more general model are identical to those in our baseline model after, abusing notation slightly, we redefine $H(e) = \int_{-\infty}^{e} r(e+c)g(e+c)$, $h(e) = r(e+c)g(e+c)$ and $\overline{h}(e)$ accordingly. We also assume that $r(\ell)g(\ell)$ is continuous (rather than just g being continuous).

Proof of Theorem 2

Proof. We first do a change of variables, noting that $\int_L \frac{\rho q r(\ell) g(\ell)}{U+c-cF(\tilde{e}_\ell)} d\ell = \int_E \frac{\rho q h(e)}{U+c-cF(e)} de$. We then solve a relaxed version of the investigator's problem where we only require the constraints to hold as inequalities:

$$\begin{aligned} &\min_{U \geq 0, F \in \mathcal{F}} U \\ \text{subject to} \quad &\int_E \frac{\rho q h(e)}{U + c - cF(e)} de \leq 1, \\ &\int_0^1 (1 - F(e)) de \leq \bar{e}. \end{aligned} \tag{14}$$

Both constraints are convex in U and F. By Theorem 1 (Chapter 8) of Luenberger (1997), there exist multipliers $\eta, \lambda \geq 0$ such that any solution U^*, F^* to (14) will solve[47]

$$\min_{U \geq 0, F \in \mathcal{F}} U + \eta \left[\int_E \frac{\rho q h(e)}{U + c - cF(e)} de - 1 \right] + \lambda \left[\int_0^1 (1 - F(e)) de - \bar{e} \right].$$

Complementary slackness conditions imply each multiplier η, λ is strictly positive only if its corresponding constraint binds; if both constraints bind, then the relaxation to inequality constraints is without loss. If $\eta = 0$, then $U^* = 0$ is clearly optimal. However, for any choice of F^*, we have

$$\int_E \frac{\rho q h(e)}{U^* + c - cF^*(e)} de = \int_E \frac{\rho q h(e)}{c - cF^*(e)} de \geq \int_E \frac{\rho q h(e)}{c} de \geq \frac{\rho q \underline{r}}{c} > 1$$

where the final equality follows from, by Assumption 2, $\rho > \frac{c(\underline{r} + \bar{r})}{q \underline{r}^2}$, which implies $\frac{\rho q \underline{r}}{c} > \frac{\underline{r} + \bar{r}}{\underline{r}} > 1$. Thus, $U^* = 0$ is not feasible. Therefore, $\eta > 0$ and $U^* > 0$.

Fixing the optimal value of U^*, the optimal investigation F^* must solve

$$\min_{F \in \mathcal{F}} \int_E \left(\frac{\eta \rho q h(e)}{U^* + c - cF(e)} - \lambda F(e) \right) de - \eta + \lambda - \lambda \bar{e}. \tag{15}$$

It is clear that $\lambda > 0$; otherwise $F^*(e) = 0$ for all e, which violates $\int_0^1 (1 - F^*(e)) de \leq \bar{e}$.

The restriction that F be a CDF requires the use of ironing techniques to solve (15). By Theorem 3.1 of Toikka (2011), $F^*(e) = \arg\min_{x \in [0,1]} \frac{\eta \bar{h}(e) \rho q}{U^* + c - cx} - \lambda x$. Taking the first-order condition, whenever $F^*(e) \in (0, 1)$, we have

$$\frac{\eta \rho q \bar{h}(e)}{(U^* + c - cF^*(e))^2} - \lambda = 0.$$

[47] This theorem requires a Slater condition hold, namely there exist U, F such that both constraints are slack. Such U, F can be found by setting $F(e) = 1$ for all $e > 0$ and $U > \rho q \bar{r}$.

Letting $k = \sqrt{\frac{\eta\rho q}{c\lambda}}$, a bit of algebra gives us $F^*(e) = \frac{U^*}{c}+1-k\sqrt{\bar{h}(e)}$ whenever $F^*(e) \in (0,1)$. $F^*(e) = 0$ whenever $\frac{\eta\bar{h}(e)\rho q}{c(\frac{U^*}{c}+1)^2}-\lambda > 0$; this condition simplifies to $\frac{U^*}{c} < k\sqrt{\bar{h}(e)}-1$. Similarly, $F^*(e) = 1$ whenever $\frac{\eta\bar{h}(e)\rho q}{c(\frac{U^*}{c})^2} - \lambda < 0$, or alternatively, when $\frac{U^*}{c} > k\sqrt{\bar{h}(e)}$. That $U^* = U_P^\alpha(F^*)$ follows from the fact that the first constraint in (14) holds with equality. *Q.E.D.*

We again derive the comparative results when reputational payoffs are $\int_\Theta r(\theta)d\nu_2(\theta|m,a,e)$; the results are identical to those presented in the text, with the exception of comparative statics on G, where we maintain the assumptions of the baseline model (namely, $\overline{r} = \underline{r} = 1$). In the proofs below, we will use the fact, as shown in the proof of Lemma 3, that $U_P^\alpha(F)$ is the unique U that solves $\int_L \frac{\rho q r(\ell)}{U+c-cF(\tilde{e}_\ell)}dG(\ell) = 1$.

Proof of Proposition 2

Proof. Fix an investigation F and distribution G of ℓ. Taking the derivative of the expression in (9) with respect to ρ, we have

$$-\frac{dU_P^\alpha(F)}{d\rho}\int_L \frac{\rho q r(\ell)}{(U_P^\alpha(F)+c-cF(\tilde{e}_\ell))^2}dG(\ell) + \int_L \frac{q r(\ell)}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)}dG(\ell) = 0.$$

After some simplification and using Jensen's inequality, we get[48]

$$\begin{aligned}
(\frac{dU_P^\alpha(F)}{d\rho})^{-1} &= q\int_L r(\ell)dG(\ell)\int_L \left(\frac{\rho^2}{(U_P^\alpha(F)+c-cF(\tilde{e}_\ell))^2}\right)\frac{r(\ell)dG(\ell)}{\int_L r(\ell')dG(\ell')} \\
&\geq q\int_L r(\ell)dG(\ell)\left(\int_L \frac{\rho r(\ell)dG(\ell)}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)}\cdot\frac{1}{\int_L r(\ell')dG(\ell')}\right)^2 \\
&= q\int_L r(\ell)dG(\ell)\left(\frac{1}{q}\cdot\frac{1}{\int_L r(\ell')dG(\ell')}\right)^2 \\
&= \frac{1}{q\int_L r(\ell)dG(\ell)}.
\end{aligned}$$

Thus, $\frac{dU_P^\alpha(F)}{d\rho} \leq q\int_L r(\ell)dG(\ell) = \mathbb{E}[r(\theta)]$. By Lemma 3, we have $\frac{dV^\alpha(F)}{d\rho} = \frac{1}{c}[\mathbb{E}[r(\theta)] - \frac{dU_P^\alpha(F)}{d\rho}] \geq 0$. An analogous argument holds for the comparative static on q.

Next, we look at first-order stochastic dominance shifts of the distribution of ℓ when $\overline{r} = \underline{r} = 1$. By Lemma 3, it suffices to show that P's equilibrium expected utility is lower under G than $\tilde{G}$. Let U and $\tilde{U}$ be P's equilibrium expected utility under G and $\tilde{G}$ respectively. For the sake of contradiction, suppose $U > \tilde{U}$. Because $\tilde{G}$ first-order stochastically dominates

[48] Here we are using $\int_L \frac{qr(\ell)}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)}dG(\ell) = \frac{1}{\rho}$ by (9).

G, we have

$$1 = \int_L \frac{\rho q}{\tilde{U} + c - cF(\tilde{e}_\ell)} d\tilde{G}(\ell) > \int_L \frac{\rho q}{U + c - cF(\tilde{e}_\ell)} d\tilde{G}(\ell) \geq \int_L \frac{\rho q}{U + c - cF(\tilde{e}_\ell)} dG(\ell),$$

which contradicts the fact that $\int_L \frac{\rho q}{U+c-cF(\tilde{e}_\ell)} dG(\ell) = 1$. Therefore, $\tilde{U} \geq U$. *Q.E.D.*

Proof of Proposition 3

Proof. Fix the value of q. Take $\tilde{\rho}, \rho$ with $\tilde{\rho} > \rho$ and corresponding optimal investigations $\tilde{F}$ and F. Let $\tilde{U}$ and U be P's equilibrium expected utility for $\tilde{\rho}, \tilde{F}$ and ρ, F respectively.

We first show that $\tilde{U} \geq U$. For the sake of contradiction, suppose $U > \tilde{U}$. Optimality of F requires $\int_L \frac{\rho q r(\ell) g(\ell)}{U+c-cF(\tilde{e}_\ell)} d\ell \leq \int_L \frac{\rho q r(\ell) g(\ell)}{U+c-c\tilde{F}(\tilde{e}_\ell)} d\ell$: if not, then the investigator could choose F' and some $U'' < U$ such that $\int_L \frac{\rho q r(\ell) g(\ell)}{U''+c-c\tilde{F}(\tilde{e}_\ell)} d\ell < 1$, contradicting the optimality of U and F in (1) when the weight on reputation is ρ. We then have

$$1 = \int_L \frac{\rho q r(\ell) g(\ell)}{U + c - cF(\tilde{e}_\ell)} d\ell \leq \int_L \frac{\rho q r(\ell) g(\ell)}{U + c - c\tilde{F}(\tilde{e}_\ell)} d\ell < \int_L \frac{\tilde{\rho} q r(\ell) g(\ell)}{U' + c - c\tilde{F}(\tilde{e}_\ell)} d\ell = 1,$$

a contradiction. Thus, $\tilde{U} \geq U$.

By Theorem 2, there exists k and $\tilde{k}$ such that $F(e) = \overline{F}(e; k, U)$ and $\tilde{F}(e) = \overline{F}(e; \tilde{k}, \tilde{U})$. F is first-order stochastically increasing in U and first-order stochastically decreasing in k, with a similar comparative statics for $\tilde{F}$ with respect to U' and k'. Because $\tilde{U} \geq U$, Bayes plausibility (namely, $\int_0^1 (1 - F(e)) de = \bar{e} = \int_0^1 (1 - \tilde{F}(e)) de$) then requires $\tilde{k} \geq k$, with strict inequality if and only if $\tilde{U} > U$.

The proposition trivially holds if $\tilde{F} = F$. Suppose $\tilde{F} \neq F$. Then $\tilde{U} > U$ and $\tilde{k} > k$. We now argue that $\tilde{F}$ must cross F once and from below, which implies F second-order stochastically dominates F. For the sake of contradiction, suppose $\tilde{F}$ crosses F from above (which must occur if $\tilde{F}$ crosses F more than once). Then there exists $e_1 < e_2$ such that $F(e_1) < \tilde{F}(e_1)$ and $\tilde{F}(e_2) < F(e_2)$. Because $\tilde{F}(e_1) \leq \tilde{F}(e_2)$, we then must have $\tilde{F}(e_1), \tilde{F}(e_2) \in (0, 1)$, which implies

$$\begin{aligned}
\frac{U}{c} + 1 - k\sqrt{\overline{h}(e_1)} \leq F(e_1) < \tilde{F}(e_1) = \frac{\tilde{U}}{c} + 1 - \tilde{k}\sqrt{\overline{h}(e_1)}, \\
\frac{\tilde{U}}{c} + 1 - \tilde{k}\sqrt{\overline{h}(e_2)} = \tilde{F}(e_2) < F(e_2) \leq \frac{U}{c} + 1 - k\sqrt{\overline{h}(e_2)}.
\end{aligned}$$

Adding these inequalities together and simplifying, we get $\sqrt{\overline{h}(e_1)} < \sqrt{\overline{h}(e_2)}$. But this contradicts the fact that $\overline{h}$ is decreasing. Therefore, $\tilde{F}$ can cross F at most once and only from below. That $\tilde{F}$ must cross F follows from Bayes plausibility: if they did not cross and $\tilde{F} \neq F$, then one distribution would strictly first-order stochastically dominate the other, a contradiction of the fact that they both have the same mean by Bayes plausibility. Because F and $\tilde{F}$ have the same mean and F second-order stochastically dominates $\tilde{F}$, the optimal investigation strategy under $\tilde{\rho}$ is less informative than under ρ. An analogous argument shows that informativeness is decreasing in q holding ρ fixed. *Q.E.D.*

Proof of Proposition 4

Let g and $\tilde{g}$ be the densities corresponding to G and $\tilde{G}$ respectively and let F^* be the optimal investigation under G. Take $h(e) = g(e+c)$ and $\tilde{h}(e) = \tilde{g}(e+c)$ for $e \in (0,1)$ and let $\overline{h}$ be the ironed version of h. We first prove a useful result given the log concavity of g.

Lemma 8. *If g is log concave, then $\frac{1}{\sqrt{\overline{h}(e)}}$ is convex.*

Proof. We note that $\overline{h}$ is decreasing, strictly so on some interval only if $\overline{h} = h$ and h is strictly decreasing on that interval; otherwise $\overline{h}$ is constant. Log concavity of g immediately implies log concavity of h. Because h is log concave, it is single peaked and there exists a cutoff e_c such that $\overline{h}$ is constant on $[0, e_c]$ and decreasing on $[e_c, 1]$. The derivative of $\frac{1}{\sqrt{\overline{h}(e)}}$ is 0 for $e < e_c$ and $\frac{-\overline{h}'(e)}{2\overline{h}(e)^{\frac{3}{2}}} \geq 0$ for $e > e_c$. To establish global convexity, it suffices to show that $\frac{-\overline{h}'(e)}{2\overline{h}(e)^{\frac{3}{2}}}$ is increasing on $(e_c, 1]$.

For $e > e_c$, $\overline{h}(e) = h(e)$. Our desired conclusion follows if $\frac{d}{de}\frac{-h'(e)}{2h(e)^{\frac{3}{2}}} \geq 0$, which holds if and only if $\frac{3}{2}h'(e)^2 \geq h''(e)h(e)$. That this inequality holds follows from $h''(e)h(e) \leq h'(e)^2$ (by log-concavity of h) and $h'(e)^2 \leq \frac{3}{2}h'(e)^2$. *Q.E.D.*

With this result in hand, we turn to the proof of the proposition.

Proof. Let $U_P^{\alpha}(F; g)$ and $U_P^{\alpha}(F; \tilde{g})$ be P equilibrium expected utility with investigation F and distribution g and $\tilde{g}$ respectively. Take a distribution F^* which is optimal given g. By Theorem 2, for some $k \in \mathbb{R}_+$, $F^*(e) = \frac{U_P^{\alpha}(F^*;g)}{c} + 1 - k\sqrt{\overline{h}(e)}$ when in $(0,1)$. We will show $U_P^{\alpha}(F^*; g) \geq U_P^{\alpha}(F^*; \tilde{g})$.

Let $\underline{e}^* = \min\ \mathrm{Supp}(F^*)$ and $\overline{e}^* = \max\ \mathrm{Supp}(F^*)$. Because $\tilde{g}$ is a pivotal mean-preserving contraction of g, $\tilde{h}(e) = h(e)$ for all $e \notin (\underline{e}^*, \overline{e}^*)$, $\tilde{G}(c) = G(c)$ and $\tilde{G}(1+c) = G(1+c)$. Then

$$\begin{aligned}
&\frac{\rho q G(c)}{U_P^\alpha(F^*; g) + c} + \int_0^1 \frac{\rho q h(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de + \frac{\rho q (1 - G(1+c))}{U_P^\alpha(F^*; g)} \\
&\quad - \frac{\rho q \tilde{G}(c)}{U_P^\alpha(F^*; g) + c} + \int_0^1 \frac{\rho q \tilde{h}(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de + \frac{\rho q (1 - \tilde{G}(1+c))}{U_P^\alpha(F^*; g)} \\
&= \int_{\underline{e}^*}^{\overline{e}^*} \frac{h(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de - \int_{\underline{e}^*}^{\overline{e}^*} \frac{\tilde{h}(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de \\
&= \int_{\underline{e}^*}^{\overline{e}^*} \frac{h(e)}{ck\sqrt{\overline{h}(e)}} de - \int_{\underline{e}^*}^{\overline{e}^*} \frac{\tilde{h}(e)}{ck\sqrt{\overline{h}(e)}} de \\
&\geq 0,
\end{aligned}$$

where the inequality follows because $\frac{1}{\sqrt{\overline{h}(e)}}$ is a convex function by Lemma 8 and $\tilde{g}$ is a pivotal mean-preserving contraction of g.[49]

For the sake of contradiction, suppose $U_P^\alpha(F^*; g) < U_P^\alpha(F^*; \tilde{g})$. Then

$$\begin{aligned}
1 &= \int_L \frac{\rho q \tilde{g}(\ell)}{U_P^\alpha(F^*; \tilde{g}) + c - cF^*(\tilde{e}_\ell)} d\ell \\
&= \frac{\rho q \tilde{G}(c)}{U_P^\alpha(F^*; \tilde{g}) + c} + \int_0^1 \frac{\rho q \tilde{h}(e)}{U_P^\alpha(F^*; \tilde{g}) + c - cF^*(e)} de + \frac{\rho q (1 - \tilde{G}(1+c))}{U_P^\alpha(F^*; \tilde{g})} \\
&< \frac{\rho q \tilde{G}(c)}{U_P^\alpha(F^*; g) + c} + \int_0^1 \frac{\rho q \tilde{h}(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de + \frac{\rho q (1 - \tilde{G}(1+c))}{U_P^\alpha(F^*; g)} \\
&\leq \frac{\rho q G(c)}{U_P^\alpha(F^*; g) + c} + \int_0^1 \frac{\rho q h(e)}{U_P^\alpha(F^*; g) + c - cF^*(e)} de + \frac{\rho q (1 - G(1+c))}{U_P^\alpha(F^*; g)} \\
&= \int_L \frac{\rho q g(\ell)}{U_P^\alpha(F^*; g) + c - cF^*(\tilde{e}_\ell)} d\ell
\end{aligned}$$

which is a contradiction of $\int_L \frac{\rho q g(\ell)}{U_P^\alpha(F^*;g)+c-cF^*(\tilde{e}_\ell)} d\ell = 1$. Therefore, $U_P^\alpha(F^*; g) \geq U_P^\alpha(F^*; \tilde{g})$, which, by Lemma 3, implies the investigator is better off under $\tilde{g}$ than g when holding the investigation fixed at F^*. Allowing the investigator to optimize the investigation after moving to $\tilde{g}$ can only make the investigator better off. *Q.E.D.*

E. Proofs from Section 6

Before stating the proof of Proposition 5, we first formally define an equilibrium in the commitment model. We let $R(x)$ denote the reputation from choosing a particular com-

[49] It is easy to see that $\tilde{h}$ is a a mean-preserving contraction of h if $\tilde{g}$ is a pivotal mean-preserving contraction of g.

mitment $x \in \mathcal{X}$ where we endow $\mathcal{X}$ with the metric $d(x, x') = \int_E |x(e) - x'(e)| dF(e)$.[50] An equilibrium is given by a strategy $\xi : \Theta \to \Delta(\mathcal{X})$ and a belief system $\nu : \mathcal{X} \to \Delta(\Theta)$ such that

1. ν is obtained from ξ using Bayes rule whenever possible with $\text{Supp}(\nu(\cdot|x)) \subseteq \{\theta : x \in \text{Supp}(\xi(\cdot|\theta))\}$ if $\{\theta : x \in \text{Supp}(\xi(\cdot|\theta))\} \neq \emptyset$,
2. $\xi(\mathcal{X}^*_\theta|\theta) = 1$ where $\mathcal{X}^*_\theta = \arg\max_{x \in \mathcal{X}} \int_E u(\theta, e, x(e), \int_\Theta r(\theta) d\nu(\theta|x)) dF(e)$.

We continue to impose the D1 refinement on equilibrium (as defined in Ramey (1996)). In the context of our game, this is defined as follows. Let U_θ be type θ's equilibrium payoff. Take any x that is not in the support of $\xi(\cdot|\theta)$ for any $\theta \in \Theta$. Suppose there exists $\Theta' \subseteq \Theta$ such that, for each $\theta \notin \Theta'$, there exists $\theta' \in \Theta'$ such that

$$\{\mu \in \Delta(\Theta) : \int_E u(\theta, e, x(e), \int_\Theta r(\theta) d\mu(\theta)) dF(e) > U_\theta\}$$
$$\subsetneq \{\mu \in \Delta(\Theta) : \int_E u(\theta', e, x(e), \int_\Theta r(\theta) d\mu(\theta)) dF(e) > U_{\theta'}\}.$$

Then an equilibrium with belief system ν violates D1 if the support of $\nu(\cdot|x)$ is not contained in Θ'. An equilibrium satisfies D1 if it does not violate D1.

Proof of Proposition 5

Proof. Throughout, we let $R(x) = \int_\Theta r(\theta) d\nu(\theta|x)$ wherever ν is clear and define $\mathcal{X}_\ell \equiv \arg\max_x \int_E (c + e - \ell) x(e) dF(e)$. We split the proof into several steps.

Step 1 (Equilibrium Construction): Let each $\xi(x_\ell|\ell) = 1$ for all $\ell \in L$ and define P's equilibrium mixing strategy $\xi(\cdot|P) \in \Delta(\{x_\ell\}_{\ell \in L})$ by $d\xi(x_\ell|P) = r(\ell) dG(\ell) \frac{q}{1-q} [\frac{\rho}{U^\alpha_P(F) + c - cF(\tilde{e}_\ell)} - 1]$.[51] Set equilibrium beliefs $\nu(\ell|x) = \frac{U^\alpha_P(F) + c - cF(\tilde{e}_\ell)}{\rho}$, $\nu(P|x) = 1 - \nu(\ell|x)$ for $x \in \{x_\ell\}_{\ell \in L}$, and $\nu(P|x) = 1$ otherwise; this leads to $R(x_\ell) = \frac{qr(\ell) dG(\ell)}{q dG(\ell) + (1-q) d\xi(x_\ell|P)}$ for $\ell \in L$ and $R(x) = 0$ otherwise. We note that $U^\alpha_P(F) = -c \int_E x_\ell(e) dF(e) + \rho R(x_\ell)$ for all $\ell \in L$.

We note that these strategies generates the same outcomes as in ex-ante signaling.[52] P is indifferent across all $\{x_\ell\}_{\ell \in L}$ by construction and has no incentive to deviate to $x \notin \{x_\ell\}_{\ell \in L}$ as all such x generate an expected utility of at most 0, which is lower than his equilibrium utility $U^\alpha_P(F)$.[53] By the arguments in Claim 3, no ℓ type has an incentive to deviate.

[50] Formally, we take the DM's choices to be an equivalence class of functions x that differ only on zero probability events.

[51] It is straightforward to check that $\int_L \xi(x_\ell|P) = 1$ given the definition of $U^\alpha_P(F)$.

[52] It is straightforward to verify that $d\xi(x_\ell|P) = d\sigma^\alpha(m_\ell|P)$.

[53] As shown in Lemma 2, $U^\alpha_P(F) \geq \rho q \underline{r} - c > 0$.

Finally, we show that the off-path reputations are consistent with D1. Take any $x \notin \{x_\ell\}_{\ell \in L}$. D1 rules out $R(x) = 0$ only if there exists an ℓ such that

$$\{\mu \in \Delta(\Theta) : -c\int_E x(e)dF(e) + \rho \int_\Theta r(\theta)d\mu(\theta) \geq U_P^\alpha(F)\}$$
$$\subsetneq \{\mu \in \Delta(\mu) : \int_E (e-\ell)x(e) + \rho \int_\Theta r(\theta)d\mu(\theta) \geq \int_E (e-\ell)x_\ell(e) + \rho R(x_\ell)\}.$$

The left-hand side above is non-empty.[54] Using the fact that $U_P^\alpha(F) = -c\int_E x_\ell(e)dF(e) + \rho R(x_\ell)$, the above statement is equivalent to

$$\int_E (c+e-\ell)x_\ell(e)dF(e) < \int_E (c+e-\ell)x(e)dF(e),$$

which contradicts $x_\ell \in \mathcal{X}_\ell$. Therefore, $R(x) = 0$ is consistent with D1.

Step Two (Outcome Equivalence): We show that all other equilibria are outcome equivalent in two steps. Take any equilibrium with corresponding strategies $\{\xi(\cdot|\theta)\}_{\theta\in\Theta}$ and belief system ν. First, we show that in any equilibrium ℓ types must only choose from $\mathcal{X}_\ell$ (i.e., $\xi(\mathcal{X}_\ell|\ell) = 1$). Second, we show $\xi(\mathcal{X}_\ell|P)$ must take the form specified in Step One.

We first establish that, across all equilibria, a bound on the ex-post belief that $\theta \in L$.

Claim 4. *$\nu(L|x) < 1$ for all $x \in \mathcal{X}$.*

Proof of Claim: For the sake of contradiction, suppose there exists $x \in \mathcal{X}$ such that $\nu(L|x) = 1$. Then $R(x) \geq \underline{r}$. By Bayes' plausibility, there must exist $x' \in \mathcal{X}_P^*$ such that $\nu(L|x') \leq q$, which implies $R(x') \leq q\overline{r}$. For x' to be in $\mathcal{X}_P^*$, we must have

$$-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho R(x). \tag{16}$$

Because $-c\int_E x'(e)dF(e) \leq 0$, $-c\int_E x(e)dF(e) \geq -c$, $R(x') \leq q\overline{r}$ and $R(x) \geq \overline{r}$, (16) implies $-c + \rho\underline{r} \leq \rho q\overline{r}$, or $\rho \leq \frac{c}{\underline{r}-q\overline{r}} \leq \frac{c(\underline{r}+\overline{r})}{\underline{r}^2-q\overline{r}^2}$, a contradiction of Assumption 2.

Next, we show $\mathcal{X}_\ell^* \subseteq \mathcal{X}_\ell$ for all $\ell \in L$. For the sake of contradiction, suppose there exists $x \in \mathcal{X}_\ell^* \backslash \mathcal{X}_\ell$ for some ℓ. Fixing this ℓ and x, there are two cases to consider: $\text{cl}(\mathcal{X}_P^*) \cap \mathcal{X}_\ell \neq \emptyset$ and $\text{cl}(\mathcal{X}_P^*) \cap \mathcal{X}_\ell = \emptyset$ where $\text{cl}(\mathcal{X}_P^*)$ is the closure of $\mathcal{X}_P^*$.

In the first case, where $\text{cl}(\mathcal{X}_P^*) \cap \mathcal{X}_\ell \neq \emptyset$, there exists a sequence of $\{x'_n\}_{n=0}^\infty$ such that

[54] Take $x' \in \{x_\ell\}_{\ell\in L}$ such that $\nu(L|x') \leq q$ (such an x' exists by Bayes plausibility), which implies $U_P^\alpha(F) \leq \rho R(x') \leq \rho q\overline{r}$. Setting μ with mass only on $\arg\max_{\ell\in L}\ r(\ell)$ is associated with a utility of at least $\rho\overline{r} - c$. If the set on the left-hand side was empty, then $\rho\overline{r} - c \leq U_P^\alpha(F) \leq \rho q\overline{r}$ or $\rho \leq \frac{c}{\overline{r}(1-q)}$, which contradicts (using Assumption 2) $\rho \geq \frac{c(\overline{r}+\underline{r})}{\underline{r}^2-q\overline{r}^2} \geq \frac{c}{\underline{r}-q\overline{r}} \geq \frac{c}{\overline{r}(1-q)}$.

$x'_n \in \mathcal{X}^*_P$ for all n and, for $x' = \lim_{n\to\infty} x'_n$, $x' \in \mathcal{X}_\ell$. P then weakly prefers x'_n to x and ℓ weakly prefers x to x'_n:

$$\begin{aligned} -\int_E cx'_n(e)dF(e) + \rho R(x'_n) &\geq -\int_E cx(e)dF(e) + \rho R(x), \\ \int_E (e-\ell)x(e)dF(e) + \rho R(x) &\geq \int_E (e-\ell)x'_n(e)dF(e) + \rho R(x'_n). \end{aligned}$$

Adding these inequalities and simplifying, we get $\int_E (c+e-\ell)(x(e) - x'_n(e))dF(e) \geq 0$ for all n. Taking the limit as $n \to \infty$ yields $\int_E (c+e-\ell)(x(e)-x'(e))dF(e) \geq 0$, a contradiction to $x' \in \mathcal{X}_\ell$ and $x \notin \mathcal{X}_\ell$.

Now consider the second case, when $\text{cl}(\mathcal{X}^*_P) \cap \mathcal{X}_\ell = \emptyset$. Take any $x' \in \mathcal{X}^*_P$. Because $x_\ell \notin \text{cl}(\mathcal{X}^*_P)$, we have $x_\ell \notin \text{Supp}(\xi(\cdot|P))$, so for $\nu(L|x_\ell) < 1$, it must be that $\{\ell' : x_\ell \in \text{Supp}\{\xi(\cdot|\ell')\}\} = \emptyset$. D1 then requires that $\nu(L|x_\ell) = 1$ if

$$\begin{aligned} &\{\mu \in \Delta(\Theta) : -c\int_E x_\ell(e)dF(e) + \rho\int_\Theta r(\theta)d\mu(\theta) > -c\int_E x'(e)dF(e) + \rho R(x')\} \\ &\quad \subsetneq \{\mu \in \Delta(\Theta) : \int_E (e-\ell)x_\ell(e) + \rho\int_\Theta r(\theta)d\mu(\theta) > \int_E (e-\ell)x(e)dF(e) + \rho R(x)\}. \end{aligned} \tag{17}$$

By analogous arguments to those in Step 1, the left-hand side of (17) is non-empty. Because $x' \in \mathcal{X}^*_P$, we have $-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho R(x)$. Therefore, (17) holds if

$$\begin{aligned} &\{\mu \in \Delta(\Theta) : -c\int_E x_\ell(e)dF(e) + \rho\int_\Theta r(\theta)d\mu(\theta) > -c\int_E x(e)dF(e) + \rho R(x)\} \\ &\quad \subsetneq \{\mu \in \Delta(\Theta) : \int_E (e-\ell)x_\ell(e) + \rho\int_\Theta r(\theta)d\mu(\theta) > \int_E (e-\ell)x(e)dF(e) + \rho R(x)\}. \end{aligned} \tag{18}$$

After some simplification, strict inclusion holds if $\int_E (c+e-\ell)(x_\ell(e) - x(e))dF(e) > 0$, which holds because $x \notin \mathcal{X}_\ell$. Thus, $\nu(L|x_\ell) = 1$, which contradicts Claim 4. Therefore, we conclude that $\mathcal{X}^*_\ell \subseteq \mathcal{X}_\ell$ in any equilibrium. Thus, all equilibrium strategies for a probability one set of ℓ types are outcome equivalent to x_ℓ with probability one (the only times they may differ is when $e = \tilde{e}_\ell$, which occurs with only for a probability zero set of (e, ℓ)).

Next, we argue that $\xi(\cdot|P)$ and $\Xi(\cdot) \equiv \int_L \xi(\cdot|\ell)dG(\ell)$ must be mutually absolutely continuous. Claim 4 implies that $\Xi(\cdot)$ is absolutely continuous with respect to $\xi(\cdot|P)$. For the sake of contradiction, suppose $\xi(\cdot|P)$ is not absolutely continuous with respect to $\Xi(\cdot)$. If not, then because $\xi(\mathcal{X}^*_P|P) = 1$, there exists $X' \subset \mathcal{X} \subseteq \mathcal{X}^*_P$ such that $\xi(X'|P) > 0 = \Xi(X')$. Then $R(x) = 0$ for some $x \in X'$ and, because $x \in \mathcal{X}^*_P$, P's equilibrium expected utility is $-c\int_E x(e)dF(e)$. But, by Bayes plausibility, there exists $x' \in \text{Supp}(\Xi)$ such that $\nu(L|x') \geq q$, which implies $R(x') \geq q\underline{r}$, in which case P can achieve a utility of $-c\int_E x'(e)dF(e) + \rho R(x') \geq \rho q\underline{r} - c > 0 \geq -c\int_E x(e)dF(e)$. Thus, choosing x is strictly dominated by x', contradicting $x \in \mathcal{X}_P^*$. This argument also implies that $R(x) > 0$ for all $x \in \mathcal{X}_P^*$. Given that all ℓ must choose only from $\mathcal{X}_\ell$ and, for probability one set of ℓ, all $x \in \mathcal{X}_\ell$ lead to equivalent actions with probability one, the fact that P has a unique mixing strategy over $\mathcal{X}_\ell$ follows from the same arguments as in Lemma 2. *Q.E.D.*

Next, we turn to the optional commitment model. Let $\lambda \in \Delta([-\delta, \delta])$ be the distribution over ε. An equilibrium consists of a strategy at the communication stage $\sigma : \Theta \to \Delta(\mathcal{X} \cup M)$, a follow up strategy at the decision stage $\zeta : M \times E \times [-\delta, \delta] \times \Theta \to \Delta(\{0, 1\})$ and belief systems $\nu_1 : \mathcal{X} \cup M \to \Delta(\Theta)$, $\nu_2 : (M \times E \times A) \to \Delta(\Theta)$ such that

1. ν_1 is obtained from Bayes rule whenever possible, with $\text{Supp}(\nu_1(\cdot|x)) \subseteq \{\theta : x \in \text{Supp}(\sigma(\cdot|\theta))\}$ if $\{\theta : x \in \text{Supp}(\sigma(\cdot|\theta))\} \neq \emptyset$,

2. $\nu_2(\cdot|m, e, a)$ is obtained from Bayes rule whenever possible given prior $\nu_1(\cdot|m)$ for $m \in M$,

3. For each m, θ, e, $\zeta(A^*_{m,e,\varepsilon,\theta}|m, e, \varepsilon, \theta) = 1$ where

$$A^*_{m,e,\varepsilon,\theta} = \arg\max_{a\in\{0,1\}} u(\theta, e, a, \int_\Theta r(\theta)d\nu_2(\theta|m, e, a)) + \varepsilon a,$$

4. For each θ, $\sigma(\mathcal{Y}^*_\theta|\theta) = 1$ where

$$\mathcal{Y}^*_\theta = \arg\max_{y\in M\cup\mathcal{X}} \int_E \Bigg[\mathbb{1}(y \in M)\Big\{\int_{-\delta}^{\delta}\big(\max_{a\in\{0,1\}} u(\theta, e, a, \int_\Theta r(\theta)d\nu_2(\theta|y, a, e)) + \varepsilon a\big)d\lambda(\varepsilon)\Big\} + \mathbb{1}(y \in \mathcal{X})u(\theta, e, y(e), \int_\Theta r(\theta)d\nu_1(\theta|y))\Bigg]dF(e),$$

where ε does not appear in the utility following $y \in \mathcal{X}$ because it is mean zero. Notice that ζ only takes effect if a cheap-talk message is sent. We again impose the D1 refinement on the choice of $x \in \mathcal{X}$ (as defined in the commitment model) and on the choice of a following a cheap talk message (as defined in our baseline model).

Proof of Proposition 6

We prove the result under the differential type reputation model under the assumption in addition to Assumption 2 that $\rho \geq \max\{\frac{2\delta}{q\underline{r}}, \frac{\delta}{\underline{r}-q\bar{r}}\}$.

Proof. Equilibrium existence follows by taking the same strategies (and beliefs following $x \in \mathcal{X}$) as in the commitment model and setting $\nu(P|m) = 1$ following any $m \in M$. Moreover, by Proposition 5, this equilibrium yields the same outcome as ex-ante-signaling. Therefore, we only need to prove that all equilibria are have outcomes that are equivalent to ex-ante signaling.

Take an equilibrium $\mathcal{E}$. If $\Sigma_N(M) = \sigma(\cdot|M) = 0$, then the same arguments as in Proposition 5 show that the equilibrium outcome is equivalent to ex-ante signaling. Suppose that $\Sigma_N(M) > 0$ or $\sigma(M|P) > 0$. We first show $\sigma(\cdot|P)$ and $\Sigma_N(\cdot) = \int_L \sigma(\cdot|\ell)dG(\ell)$ are mutually absolutely continuous over M. Suppose there exists $M' \subseteq M$ such that $\sigma(M'|P) > 0$ or $\Sigma_N(M') > 0$. If $\Sigma_N(M') = 0 < \sigma(M'|P)$, then for some $m \in M'$, $\nu_1(P|m) = 1$ and the reputation for $m \in M'$ following any action at the decision stage is 0. If $\sigma(M'|P) = 0 < \Sigma_N(M')$, then there exists $m \in M'$ such that $\nu_1(L|m) = 1$ and the reputation is at least $\underline{r}$ for each action at the decision stage. By Bayes plausibility, there exists an m' or x with reputation at most $q\bar{r}$ after each action. If the reputation after m is always at least $\underline{r}$, then, m is a profitable deviation from m' or x for any type of DM as they can choose an optimal action for each (e, ε) realization and still have a higher reputation as $q\bar{r} < \underline{r}$ by Assumption 2. If the reputation after m is always 0, P attains a maximum utility of $\max\{-c + \delta, 0\}$ from doing so. However, the P type can attain at least $\rho q\underline{r} - c - \delta$ by mimicking some ℓ type whose expected equilibrium reputation is at least $q\underline{r}$ (such ℓ exist by Bayes plausibility). Because $\rho > \frac{2\delta}{q\underline{r}}$ and $\rho > \frac{c(\underline{r}+\bar{r})}{q\underline{r}^2}$ by Assumption 2, this is a contradiction. Therefore, $\Sigma_N(M') > 0$ if and only if $\sigma(M'|P) > 0$ for all $M' \subseteq M$.

Take ℓ and $m \in M$ such that m is an optimal message for ℓ (i.e., $m \in \mathcal{Y}^*_\ell$) and take $R(m, e, a) = \int_\Theta r(\theta)d\nu_2(\theta|m, e, a)$. Now consider the difference in payoff between sending message m in equilibrium and taking commitment x_ℓ for types ℓ and P as a function of e, ε. For type ℓ, this is given by

$$\max\{e - \ell + \varepsilon + \rho R(m, e, 1), \rho R(m, e, 0)\} - (e - \ell + \varepsilon)\mathbb{1}(e - \ell \geq -c) - \rho R(x_\ell), \tag{19}$$

and for P, it is given by

$$\max\{-c + \varepsilon + \rho R(m, e, 1), \rho R(m, e, 0)\} - (-c + \varepsilon)\mathbb{1}(e - \ell \geq -c) - \rho R(x_\ell). \tag{20}$$

Notice that the expression for ℓ is weakly less than it is for P for every e, ε. The expression is strictly less by $e - \ell + c$ if $e - \ell > -c$ and both ℓ and P choose $a = 0$ following (m, e, ε), or by $-(e - \ell + c)$ if $e - \ell < -c$ and both ℓ and P choose $a = 1$ following (m, e, ε). If the difference is strictly less for ℓ than it is for P, previous arguments imply that $\nu_1(L|x_\ell) = 1$, so $R(x_\ell) \geq \underline{r}$ and P's expected utility in equilibrium is at most $\max\{-c+\delta, 0\}+\rho q\bar{r}$. Because

P's utility from x_ℓ is then at least $-c + \rho\underline{r}$, P would have a strict incentive to deviate to x_ℓ if $\delta > c$ by $\rho > \frac{\delta}{\underline{r} - q\bar{r}}$ and if $\delta \leq c$ by $\rho > \frac{c(\underline{r}+\bar{r})}{\underline{r}^2 - q\bar{r}^2} > \frac{c}{\underline{r} - q\bar{r}}$.

Note that (19) and (20) are equal only if, for every ℓ sending m, both P and ℓ choose $x_\ell(e)$ with probability 1 when $e \neq \tilde{e}_\ell$. Note that this can only occur if a single ℓ types sends m, otherwise the evidence realizations that induce one ℓ type to choose $a = 0$ and the other ℓ type to choose $a = 1$ would necessitate two different actions from P. This means that every on-path message is sent by the P type and a single ℓ type, and this message is followed up by x_ℓ. Thus, the DM either sends cheap-talk messages which lead to actions following x_ℓ or chooses some commitment $x \in \mathcal{X}$. As shown in the proof of Proposition 7, each ℓ type can only choose commitment x_ℓ or something payoff equivalent (i.e., $\sigma(\mathcal{X}_\ell|\ell) = 1$), and the set of optimal commitments for P is contained in the set $\cup_{\ell \in L} \mathcal{X}_\ell$. This means, that at the communication stage, each ℓ type identifies themselves among L and follows up with x_ℓ at the decision stage, and the P type mixes over these options. Outcome equivalence to ex-ante signaling follows from arguments in Lemma 2. *Q.E.D.*

Proof of Proposition 7

Proof. P's expected utility conditional on e_0 is $U_P^\alpha(F_1(\cdot|e_0))$. By an analogous proof to that in Lemma 3, $\mathbb{P}(a = 1|e_0) = \frac{\rho\mathbb{E}[r(\theta)] - U_P^\alpha(F_1(\cdot|e_0))}{c}$. Thus, $\mathbb{P}(a = 1) = \int_E \mathbb{P}(a = 1|e_0) dF_0(e_0) = \frac{1}{c}[\rho\mathbb{E}[r(\theta)] - \int_E U_P^\alpha(F_1(\cdot|e_0)) dF_0(e_0)]$. The proposition then follows immediately from convexity of $U_P^\alpha(\cdot)$. *Q.E.D.*

Proof of Proposition 8

For the differential type reputation model, we now require that $r(\ell)g(\ell)$ be continuous (rather than just g being continuous). For $\ell_I \in [0, 1)$, the investigator can achieve his first-best payoff via full information disclosure: because $0 < \min_{\ell \in L} \tilde{e}_\ell < \max_{\ell \in L} \tilde{e}_\ell < 1$, $e = 0$ leads to $a = 0$ with probability one and $e = 1$ leads to $a = 1$ with probability one. Therefore, let us focus on the case when $\ell_I < 0$, that is, the investigator prefers $a = 1$ at all $e \in [0, 1]$.

That the investigator prefers ex-ante to ex-post signaling follows immediately from Proposition 1. The fact that no mass points are used is shown in the following Lemma.

Lemma 9. *For sufficiently high ρ, the optimal investigation has no mass points in $(0, 1)$.*

Proof. Take any F with a mass point on $\hat{e} \in (0, 1)$ and $U_P^\alpha(F) > 0$. Take some small $\varepsilon > 0$. Suppose $\hat{e} \leq \ell_I$. Then because no DM type will choose $a = 1$ at $e \in [\hat{e} - \varepsilon, \hat{e} + \varepsilon]$ for sufficiently small ε, it is without loss to smooth out the mass point to be a continuous density on $[\hat{e} - \varepsilon, \hat{e} + \varepsilon]$ as doing so will not change the probability of $a = 1$ at such e. Let us therefore suppose $\hat{e} > \ell_I$. Consider F_δ such that $F_\delta(e) = F(e)$ for all $e \notin (\hat{e} - \varepsilon, \hat{e} + \varepsilon)$ and F_δ moves δ mass away from $\hat{e}$ and splits it equally between $\hat{e} - \varepsilon, \hat{e} + \varepsilon$, so $F_\delta(e) = F(e) + \frac{\delta}{2}\mathbb{1}(e \in [\hat{e} - \varepsilon, \hat{e})) - \frac{\delta}{2}\mathbb{1}(e \in [\hat{e}, \hat{e} + \varepsilon])$.

Take a distribution of evidence F and let $\eta(\ell; U, \delta) = \frac{r(\ell)dG(\ell)q\rho}{U + c - cF_\delta(\tilde{e}_\ell)}$. As shown in the proof of Lemma 3, the distribution of m_ℓ is $qdG(\ell) + (1-q)d\sigma(m_\ell|P) = \frac{\rho q r(\ell) dG(\ell)}{U_P^\alpha(F) + c - cF(\tilde{e}_\ell)}$. The investigator's utility is given by

$$\begin{aligned}
&\int_E (e - \ell_I)\left[\int_L \mathbb{1}(\tilde{e}_\ell \le e)(qdG(\ell) + (1-q)d\sigma(m_\ell|P))\right] dF_\delta(e) \\
&= \int_E (e - \ell_I)\left[\int_L \mathbb{1}(\tilde{e}_\ell \le e)\eta(\ell; U_P^\alpha(F), \delta)d\ell\right] dF_\delta(e). \qquad (21)
\end{aligned}$$

For notational ease, we let $U = U_P^\alpha(F_\delta)$. Taking the derivative of (21) at $F = F_\delta$ with respect to δ, we have

$$\begin{aligned}
&\frac{dU}{d\delta}\int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \le e)\frac{\partial \eta(\ell; U, \delta)}{\partial U} d\ell dF_\delta(e) + \int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \le e)\frac{\partial \eta(\ell; U, \delta)}{\partial \delta} d\ell dF_\delta(e) \\
&\quad + \int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \le e)\eta(\ell; U, \delta) d\ell \frac{d}{d\delta} dF_\delta(e). \qquad (22)
\end{aligned}$$

We will show that this expression, for small ε and evaluated at $\delta = 0$, is strictly positive.

We show the first term in (22) is positive. Because $\frac{d\eta(\ell; U, \delta)}{dU} \le 0$, it suffices to show $\frac{dU}{d\delta} \le 0$. Because $U_P^\alpha(F_\delta)$ is characterized by $\int_L \eta(\ell; U_P^\alpha(F_\delta), \delta) d\ell = 1$, we have

$$0 = \frac{dU}{d\delta}\int_L \frac{\partial \eta(\ell; U, \delta)}{\partial U} d\ell + \int_L \frac{\partial \eta(\ell; U, \delta)}{\partial \delta} d\ell.$$

Because $\frac{\partial \eta(\ell; U, \delta)}{\partial U} \le 0$, strictly so when $g(\ell) > 0$, $\frac{dU}{d\delta} \le 0$ if and only if $\int_L \frac{\partial \eta(\ell; U, \delta)}{\partial \delta} d\ell \le 0$. Given the form of F_δ, for sufficiently small ε we have

$$\begin{aligned}
\int_L \frac{\partial \eta(\ell; U, \delta)}{\partial \delta} d\ell &= \frac{1}{2}\Bigg[\int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e} - \varepsilon, \hat{e}))\frac{c\rho q r(\ell) g(\ell)}{(U + c - cF_\delta(\tilde{e}_\ell))^2} d\ell \\
&\qquad - \int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e}, \hat{e} + \varepsilon])\frac{c\rho q r(\ell) g(\ell)}{(U + c - cF_\delta(\tilde{e}_\ell))^2} d\ell\Bigg] \\
&< 0,
\end{aligned}$$

where the inequality follows from the fact that, because F_δ has a mass point on $\hat{e}$, $F(\tilde{e}_\ell)$ is discretely higher for $\tilde{e}_\ell > \hat{e}$ than for $\tilde{e}_\ell < \hat{e}$. Thus, $\frac{dU}{d\delta} \le 0$.

Next, we show that the second term in (22) is positive. Let ΔF be the size of mass point on $\hat{e}$. Next, we note that for small ε

$$
\begin{aligned}
&\int_E (e-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\frac{\partial \eta(\ell;U,\delta)}{\partial \delta} d\ell dF_\delta(e) \\
&= \int_E (e-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\frac{1}{2}\big[\mathbb{1}(\tilde{e}_\ell \in [\hat{e}-\varepsilon,\hat{e}))\frac{cr(\ell)g(\ell)q\rho}{(U+c(1-F_\delta(\tilde{e}_\ell)))^2} \\
&\qquad - \mathbb{1}(\tilde{e}_\ell \in [\hat{e},\hat{e}+\varepsilon])\frac{cr(\ell)g(\ell)q\rho}{(U+c(1-F_\delta(\tilde{e}_\ell)))^2}\big] d\ell dF_\delta(e) \\
&= \frac{1}{2}\int_L \big[\mathbb{1}(\tilde{e}_\ell \in [\hat{e}-\varepsilon,\hat{e}))\frac{cr(\ell)g(\ell)q\rho}{(U+c(1-F_\delta(\tilde{e}_\ell)))^2}\int_{\tilde{e}_\ell}^\infty (e-\ell_I)dF_\delta(e) \\
&\qquad - \mathbb{1}(\tilde{e}_\ell \in [\hat{e},\hat{e}+\varepsilon])\frac{cr(\ell)g(\ell)q\rho}{(U+c(1-F_\delta(\tilde{e}_\ell)))^2}\int_{\tilde{e}_\ell}^\infty (e-\ell_I)dF_\delta(e)\big] d\ell \\
&\approx \frac{\varepsilon cr(\hat{e}+c)g(\hat{e}+c)q\rho}{2}\left[\frac{(\hat{e}-\ell_I)\Delta F + \int_{\hat{e}+\varepsilon}^\infty (e-\ell_I)dF_\delta(e)}{(U+c(1-F_\delta(\hat{e})+\Delta F))^2} - \frac{[\int_{\hat{e}+\varepsilon}^\infty (e-\ell_I)dF_\delta(e)}{(U+c(1-F_\delta(\hat{e})))^2}\right]
\end{aligned}
$$

We claim the last line above is strictly positive for large enough ρ. Pulling out common factors and the denominators and doing a bit of simplification, we get that the above expression is strictly positive if

$$
\begin{aligned}
0 &< \int_{\hat{e}+\varepsilon}^\infty (e-\ell_I)dF_\delta(e)[(U+c(1-F_\delta(\hat{e})))^2 - (U+c(1-F_\delta(\hat{e})+\Delta F))^2] \\
&\qquad + (\hat{e}-\ell_I)\Delta F(U+c(1-F_\delta(\hat{e})))^2 \\
&= \Delta F[(\hat{e}-\ell_I)(U+c-cF(\hat{e}))^2 - \int_{\hat{e}+\varepsilon}^\infty (e-\ell_I)dF_\delta(e)(2(U+c(1-F_\delta(\hat{e}))) + c^2\Delta F)].
\end{aligned}
$$

Because $\hat{e}-\ell_I > 0$ and $U \geq \rho q\underline{r} - c$ in equilibrium, the last line above is strictly positive for sufficiency large ρ.

Finally, we show the final term in (22) is positive. For small enough ε, we have

$$\begin{aligned}
&\int_E (e-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \le e)\eta(\ell;U,\delta)d\ell \frac{d}{d\delta}dF_\delta(e) \\
&= \frac{1}{2}(\hat{e}-\varepsilon-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell < \hat{e}-\varepsilon)\eta(\ell;U,\delta)d\ell + \frac{1}{2}(\hat{e}+\varepsilon-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell < \hat{e}+\varepsilon)\eta(\ell;U,\delta)d\ell \\
&\qquad -(\hat{e}-\ell_I)\int_L \mathbb{1}(\tilde{e}_\ell < \hat{e})\eta(\ell;U,\delta)d\ell \\
&= \frac{1}{2}(\hat{e}-\ell_I)(\int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e},\hat{e}+\varepsilon])\eta(\ell;U,\delta)d\ell - \int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e}-\varepsilon,\hat{e}))\eta(\ell;U,\delta)d\ell) \\
&\qquad + \frac{1}{2}\varepsilon(\int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e}-\varepsilon,\hat{e}+\varepsilon])\eta(\ell;U,\delta)d\ell) \\
&\ge \frac{1}{2}(\hat{e}-\ell_I)(\int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e},\hat{e}+\varepsilon])\eta(\ell;U,\delta)d\ell - \int_L \mathbb{1}(\tilde{e}_\ell \in [\hat{e}-\varepsilon,\hat{e}))\eta(\ell;U,\delta)d\ell) \\
&\ge 0,
\end{aligned}$$

where the final inequality follows from the fact that, for ℓ such that $\tilde{e}_\ell = \hat{e}$, $\eta(\ell - z; U, \delta)$ is discretely lower than $\eta(\ell + z; U, \delta)$ for small z due to the mass point on $\hat{e}$.

Having shown that all terms in (22) are positive, we conclude that F can not have been optimal as moving to F_δ for some $\delta > 0$ strictly increases the investigator's payoff. *Q.E.D.*

Optimal Investigation with Multiple States Proofs

Let $M(F,e) \equiv \frac{r(e+c)g(e+c)\rho q}{\left(U_P^\alpha(F)+c(1-F(e))\right)^2}$, which gives the derivative of the density of the DM's declaration of m_ℓ for $\ell = e + c$. Let $(\underline{e}, \overline{e})$ be the min and max over $\text{Supp}(F^*)$ respectively. Analogously to our baseline model, we assume that $r(\ell)g(\ell)$ is continuous and increasing over $[c, 1+c]$.

Proposition 9. *The optimal investigation F^* exists and must satisfy the following properties.*

1. *F^* is strictly increasing for $e \in (\underline{e}, \overline{e})$*
2. *$M(F^*, e)$ is increasing on $[\underline{e}, \overline{e}]$,*
3. *If $\int_0^e (F(e') - K(e'))de' < 0$ for $e \in (e_1, e_2) \subset [\underline{e}, \overline{e}]$, then $M(F^*, e)$ is constant on $[e_1, e_2]$*

Proof. Note that the optimum exists because the constraint set is compact and the objective is continuous in (F, U). Let $U = U_P^\alpha(F^*)$ for the optimal F^*. Then $\int_L \frac{\rho q r(\ell)g(\ell)}{U+c-cF^*(\tilde{e}_\ell)}d\ell = 1$. The optimal F^* must minimize $\int_L \frac{\rho q r(\ell)g(\ell)}{U+c-cF(\tilde{e}_\ell)}d\ell$ over feasible F; if not, then the investigator could choose an alternative F' such that $\int_L \frac{\rho q r(\ell)g(\ell)}{U+c-cF'(\tilde{e}_\ell)}d\ell < 1$, in which case $U_P^\alpha(F') <$

$U_P^\alpha(F^*)$, a contradiction of the optimality of F^*. The optimal investigation F^* must then solve

$$\begin{aligned} \min_{F\in\mathcal{F}} \quad & \int_L \frac{\rho q r(\ell) g(\ell)}{U + c - cF(\tilde{e}_\ell)} d\ell\ , \qquad (23)\\ \text{such that } BP: & \int_0^e F(e')de' \le \int_0^e K(e')de' \ \forall e \in E, \text{ and} \\ & \int_0^1 F(e')de' = \int_0^1 K(e')de'. \end{aligned}$$

First, suppose for the sake of contradiction that F^* is constant on some interval $[e_1, e_2) = \{e : F^*(e) = F^*(e_1)\}$. For $\varepsilon > 0$, consider a perturbation $\tilde{F}$ of F^* where

$$\tilde{F}(e) = \begin{cases} F^*(e) & e < e_1 - \varepsilon \text{ or } e \ge e_2 + \varepsilon \\ F^*(e_1) - \delta & e \in [e_1 - \varepsilon, (e_1 + e_2)/2], \\ F^*(e_1) + \delta & e \in [(e_1 + e_2)/2, e_2 + \varepsilon], \end{cases} .$$

where δ is taken small so that $\tilde{F}$ is a CDF. $\tilde{F}$ clearly satisfies the BP constraints. For sufficiently small ε, the impact of this perturbation value of this change on the objective in (23) as $\delta \to 0$ is approximately

$$-\int_{e_1}^{(e_1+e_2)/2} M(F^*, e)de + \int_{(e_1+e_2)/2}^{e_2} M(F^*, e)de < 0,$$

where the inequality holds because g is strictly decreasing and F^* is constant on this interval, contradicting the optimality of F^*.

Next, suppose for the sake of contradiction that $M(F^*, e_1) > M(F^*, e_2)$ for $\underline{e} \le e_1 < e_2 \le \overline{e}$. Take $\varepsilon > 0$. If $e_1 = \underline{e}$ and $F^*(\underline{e}) = 0$ then, because M is right continuous, replace e_1 with $e_1 + \varepsilon$ so that the inequality on M still holds. Similarly if $\overline{e} = e_2$ and then replace e_2 with $e_2 - \varepsilon$ so the inequality on M still holds. Now take the perturbation $\tilde{F}$ of F^* given by

$$\tilde{F}(e) = \begin{cases} F^*(e) & e \notin [e_1 - \varepsilon, e_1 + \varepsilon) \cup [e_2 - \varepsilon, e_2 + \varepsilon) \\ F^*(e_1) - \delta & e \in [e_1 - \varepsilon, e_1 + \varepsilon), \\ F^*(e_1) + \delta & e \in [e_2 - \varepsilon, e_2 + \varepsilon), \end{cases}$$

where δ is taken small so that $\tilde{F}$ is a CDF. $\tilde{F}$ clearly satisfies the BP constraints. For small ε, the impact of this perturbation on the objective in (23) as $\delta \to 0$ is approximately $2\varepsilon(-M(F^*, e_1) + M(F^*, e_2)) < 0$, contradicting the optimality of F^*.

Lastly take a region (e', e'') where the BP constraint does not bind, but the constraint binds at e' and e''. Then both the perturbation above and its opposite are available for $e' < e_1 < e_2 < e''$. This means that if $M(F^*, e)$ is not constant on this interval, F^* is not optimal. *Q.E.D.*

Proof of Corollary 4

Proof. If F discontinuously jumps at some e, then the BP constraint must not be binding around e. Because g is continuous, a discontinutiy in F^* implies $M(F^*, e)$ is not constant around e, a contradiction of Proposition 9. *Q.E.D.*

Proof of Corollary 5

Here, for the differential type reputation model we assume $\frac{r(e+c)g(e+c)}{(\rho\mathbb{E}[r(\theta)]+c-cK(e))^2}$ is strictly increasing in e.

Proof. Take e where the BP constraint binds but does not bind for some region above e. Note that the constraint always binds at $e = 0$, so such an e exists. Also at such an e, $K(e) = F^*(e)$. This means that $M(F^*, e')$ must be constant for $e' \in [e, e + \varepsilon)$ with ε sufficiently small, and as long as the BP constraint continues to not bind. Note that the condition that $\frac{r(e'+c)g(e'+c)}{(\rho\mathbb{E}[r(\theta)]+c(1-K(e')))^2}$ is increasing and fact that $U_P^\alpha(F^*) \leq \rho\mathbb{E}[r(\theta)]$[55] implies that $\frac{r(e'+c)g(e'+c)}{\left(U_P^\alpha(F^*)+c(1-K(e'))\right)^2}$ is increasing in e' on $[e, e + \varepsilon)$ which implies in this region that

$$\frac{r(e'+c)g(e'+c)}{\left(U_P^\alpha(F^*)+c(1-K(e'))\right)^2} > M(F^*, e'). \tag{24}$$

From (24), we conclude $K(e') > F^*(e')$. That is F^* grows slower than K, which means the equality BP constraint cannot be satisfied at any higher evidence level violating the equality constraint at $e = 1$. *Q.E.D.*

F. Optimal Design under Ex-Post Signaling

In this appendix we compare the optimal investigation under ex-ante signaling to that under ex-post signaling. This comparison gives us insights into how the structure of optimal investigations is shaped by the presence of communication, or alternatively, the timing of the evidence realization. Note that because of Theorem 1, the investigator will always prefer the investigation in Theorem 2 to the optimal investigation under ex-post signaling.

[55] This inequality is implied by Lemma 3, because $U_P^\alpha(F^*) > \rho\mathbb{E}[r(\theta)]$ implies the probability of $a = 1$ is negative.

However, this does not say anything about the relative informativeness of these investigations, which is especially important in applications where the evidence may be important beyond the DM's choice, e.g., the information a firm submits to the Environmental Protection Agency about its environmental impact. In such settings a planner may want to impose either ex-ante or ex-post signaling depending on which leads to a more informative investigation. We will show that the comparison in informativeness depends on the investigator's design incentives when facing only non-partisans.

Recall $v^{\beta}(e)$ is the probability of conviction as a function of the evidence given ex-post signaling. Due to the simplicity of ex-post signaling, we can explicitly derive this conviction probability in baseline model where $\underline{r} = \overline{r}$ as

$$v^{\beta}(e) = \frac{1}{2c}\left(\rho q + c - \sqrt{(\rho q + c)^2 - 4\rho q c G(e + c)}\right).$$

Because the messaging strategy under ex-post signaling involves babbling, which is independent of F, $v^{\beta}(e)$ does not depend on F. We can write the investigator's design problem as

$$\max_{F \in \mathcal{F}} \int_0^1 v^{\beta}(e) dF(e),$$
$$\text{such that } \int_0^1 (1 - F(e)) de = \overline{e}.$$

This design problem is a standard Bayesian persuasion problem and the following result characterizing the optimal information structure follows immediately from Kamenica and Gentzkow (2011).

Proposition 10. *Let $Cav(v^{\beta})$ be the concavified value of v^{β}. There exists an optimal F with binary support if $v^{\beta}(\overline{e}) < Cav(v^{\beta})(\overline{e})$ and an optimum with degenerate support on $\overline{e}$ if $v^{\beta}(\overline{e}) = Cav(v^{\beta})(\overline{e})$.*

An immediate implication is that if v^{β} is strictly concave in e, then an uninformative investigation is uniquely optimal. Because v^{β} is a convex transformation of G, it is not quite sufficient for the investigator to want to withhold information from the non-partisan. However, if the investigator is significantly harmed by providing information to the non-partisan, i.e., G is "sufficiently concave", then an uninformative investigation will be optimal under ex-post signaling.[56] Note that in these cases (and in general), the optimal investigation under ex-ante signaling provides some information; see Corollary 2. Thus,

[56] An example is when the leniency is distributed according to the standard exponential distribution.

there are cases, namely those in which ℓ types' convicts significantly less when given information, in which the optimal investigation under ex-ante signaling is more informative in a Blackwell sense than that under ex-post signaling.

However, the comparison can also go the other way. Because v^{β} is a convex transformation of G, there will be examples where the ex-post signaling optimal investigation is perfectly informative, but the investigator is harmed by providing information to nonpartisans. In these cases, because concave G implies $\overline{h}$ is decreasing in e, Theorem 2 says that the optimal investigation under ex-ante signaling admits a positive density when F^* is interior, and is thereby imperfectly informative.

A unifying feature between ex-ante and ex-post signaling is that if information increases the ℓ types' conviction probability then full information is optimal under both regimes. This means that, like under ex-ante signaling, P's behavior under ex-post signaling incentivizes the investigator to provide more information.

Corollary 6. *If $\underline{r} = \overline{r} = 1$ and G is convex on $[c, 1+c]$ then the optimal investigation is fully informative under both ex-ante and ex-post signaling.*