Signaling Good Faith

Andrew McClellan and Daniel Rappoport*

July 13, 2023

Abstract

A decision maker (DM) takes a binary decision, and cares about their reputation for being responsive to evidence—a non-partisan type—as opposed to having a fixed agenda—a partisan type. Non-partisans have heterogenous "leniency," i.e., evidence thresholds above which the high action is preferred, whereas partisans always prefer the low action. Before the evidence is realized the DM can send a cheap talk message to the public about his plans following the evidence. A wide range of communication equilibria exist, including an "ex-ante signaling" in which the DM announces (and follows through on) a threshold above which he will take the high action. We show that the ex-ante signaling equilibrium generates the highest probability of taking the high action across all equilibria. We then endogenize the evidence distribution as the choice of an investigator who seeks to maximize the probability of the high action. Allowing for communication qualitatively changes the structure of the optimal investigation, pushing towards "unpredictability" in the distribution of evidence that hinders the partisan type in targeting his announced action threshold.

1. Introduction

In September 2019, a whistleblower complaint surfaced concerning a call between President Trump of the United States, and President Zelensky of Ukraine. The call prompted allegations that Trump had conditioned aid on Zelensky conducting investigations into

*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, Ben Golub, Alex Frankel, Marina Halac, Thomas Jungbauer, Emir Kamenica, Navin Kartik, Jacob Leshno, Alessandro Pavan, Doron Ravid and Joe Root.

Trump's political rivals. Speaker of the House Pelosi responded by initiating an impeachment inquiry. Like many political scandals, despite the investigation being ongoing, various pivotal senators were asked to weigh in about their intended impeachment votes. 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.[1] These politicians faced cross pressure between appearing to approach the issue objectively and siding with their party.[2]

For the purpose of this paper, the important features of the example above are as follows. A decision maker (the politician) weighs the tradeoff between taking a potentially unfavorable action (voting for conviction), and obtaining a positive reputation in the eyes of the public (projecting integrity to the voters). In addition, there is uncertainty about the value of the decision (the results of the impeachment inquiry). The decision maker can stake a position before this uncertainty is resolved or they can be vague (answer interviewers' questions, or "dodge the cameras"). An additional aspect is that the way uncertainty is resolved is determined by an interested party (Speaker Pelosi). These features describe other political scandals and decisions well, but are not unique to them:

1. 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 set their standards for approval up front or decide on a case-by-case basis after observing the evidence. For example, in late 2020, national drug regulatory agencies were eager to approve a safe COVID-19 vaccine but faced credibility worries that they were rushing approval. In the US, the FDA laid out a specific efficacy threshold in clinical trials for approval, whereas the EU's counterpart deliberately provided no such lower bound (Singh and Upshur (2021)). Another example is a firm defending the benefits of a potential merger which

[1] Costa, Roberts (2019) "Cracks emerge among Senate Republicans over Trump urging Ukrainian leader to investigate Biden" *Washington Post*, September 25.

[2] 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.)

can then be approved or denied by the FTC. Because the firm seeks approval, the procedure by which the FTC communicates its standards will impact the disclosures the firm makes about the impacts of the merger.

2. 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] In addition, there is concern about how the design of standardized tests affects how many applicants are admitted primarily due to non-academic qualifications.[5]

We study two important questions in such settings. First, how does communication about hypothetical plans prior to the revelation of evidence affect outcomes? Would we expect that Republican senators who indicate conditions for impeachment up front convict more or less than those who wait and see? Would the FTC approve more mergers if they were to specify approval conditions in advance, than if they decide on a case-by-case basis? Would universities admit more donor applicants were they to publicize admissions standards rather than use holistic admissions? Second, how does the design of the investigation affect outcomes? How should Nancy Pelosi conduct the impeachment inquiry to get the most Republican senators to convict? How should firms provide evidence about potential mergers to ensure the highest chance of approval? And, how might one design standardized tests to give donor applicants the least advantage?

Our model features a single decision maker (DM), an investigator, 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. At the decision stage the evidence $e \in \mathbb{R}$ is revealed and the DM chooses a binary action, either $a = 1$ — "conviction" — or $a = 0$ —"acquittal". In addition to the evidence, the DM's

[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.

preferences over the action also depend on his private type, which is either partisan or non-partisan. The non-partisan would like to convict if the evidence is sufficiently strong; more specifically if e crosses their privately known idiosyncratic leniency $\ell \in \mathbb{R}$. On the other hand, the partisan does not care about the evidence and suffers a constant disutility from taking $a = 1$. 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 and chosen action along with the realized evidence.

The investigator specifies the distribution over evidence—the investigation—subject to constraints. In our main specification, the investigator seeks to maximize the probability of conviction.[6] The first part of our paper develops results for a fixed investigation, i.e., where the investigator is inactive, and the second part allows the investigator to flexibly design an information structure.

One novelty of our model is that the DM seeks a reputation for being *responsive to evidence* rather than for being inherently biased toward one action or another. While a non-partisan type can prefer $a = 1$ more or less than a partisan type depending on the parameters, the distinctive feature of non-partisan types is that their preferences depend on evidence.[7] Our reputation incentives capture the desire to avoid the common accusation of opposition as arguing in "bad faith", i.e., that they have a fixed agenda and simply find arguments to suit it—like the partisan in our model. "Good faith" opposition may have different principles or a different set of standards, but is still interested in the objective evidence on a particular issue—like the non-partisan in our model. Because, as in reality, the non-partisans have heterogeneous principles, it is often difficult to determine whether someone is arguing in bad faith: the partisan will pool with the non-partisan type. Our model explores how uncertainty about the evidence, and the precision of communication at this uncertain stage, affect outcomes.

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 when this communication stage involves babbling, i.e., nothing is communicated; and when it is fully informative, i.e., in the case the DM has a threshold, the observer knows it before the evidence is revealed. We term these equilibria ex-post and ex-ante signaling respectively. We show that ex-ante signaling is tantamount to the DM committing to a

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

[7] While this is our preferred interpretation, an alternative perspective of our model is as a standard costly signaling framework, where there is (i) heterogeneity in the "good" types preferences, and (ii) communication at some stage where the costs of signaling are uncertain.

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, let's wait and see". There are also other imperfectly informative communication equilibria that admit interpretations between ex-post and ex-ante signaling.

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 to convict less frequently. However, Theorem 1 shows that ex-ante signaling not only has a higher probability of conviction than ex-post signaling, but than every other equilibrium as well. This means that politicians who successfully "dodge the cameras" as opposed to answering reporters' questions, will tend to break with party less frequently; government agencies that specify approval criteria up-front will go against their officials' interests more frequently than those who decide on a case by case basis; and setting clear admissions criteria will lead to more socially accepted admission decisions relative to holistic policies. The broad intuition is simple: before the realization of evidence, the DM is more willing to make stronger claims in order to attain a higher reputation because there are many evidence realizations under which these stronger claims are indistinguishable from weaker ones. Conversely, after a particular evidence realization occurs, obtaining a high reputation requires convicting with probability one, so the DM convicts less frequently under ex-post signaling. 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.[8] We focus on the ex-ante signaling equilibrium for a couple reasons, (i) Theorem 1 shows that this is the investigator's preferred equilibrium and in many contexts whether ex-ante or ex-post signaling outcomes prevail may be a design decision, and (ii) we argue in Subsection 6.1 that the ex-ante signaling outcome is likely to be selected under a natural refinement. Theorem 2 characterizes the investigation that maximizes conviction. Even in the absence of a designer, our characterization speaks to how the distribution of evidence affects the probability of conviction.

One main takeaway is that the optimal investigation admits no mass points unlike that seen in standard Bayesian persuasion design problems. This is because the partisan DM responds to changes in the investigation by altering which principles he claims at the com-

[8] In Subsection 6.4, we show that our main takeaways carry over to information design over multiple states.

munication 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 on the realized evidence as possible to avoid "targeting" from the partisan who can declare a threshold for conviction just above where evidence is likely to be.

Theorem 2 also yields a simple characterization of when more information increases the probability of conviction. Roughly, this occurs when the distribution of non-partisan thresholds is convex, or alternatively if and only if more information increases the probability of conviction among non-partisans. The intuition again uses the fact that the partisan's equilibrium interests are exactly misaligned with that of the investigator: we show that when the distribution of non-partisan thresholds is convex, interior evidence realizations induce the partisan to further correlate his strategy with the non-partisan thereby increasing his reputation and harming the investigator.

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. Subsection 5.2 describes comparative statics. And lastly, Section 6 discusses equilibrium selection and extensions of the information design problem.

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 reputation incentives are closest to Morris (2001). He studies an informed sender who seeks a reputation for being a type similar to our non-partisan. The main difference is that our model includes 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. [9]

Since the DM in our model eventually takes an action, we are also connected to the costly

[9] 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.

signaling literature initiated by Spence (1973). As in Bénabou and Tirole (2006), Esteban and Ray (2006), 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.[10] The key departures in our model from these papers are (i) there is uncertainty at some initial stage about a variable that impacts the DM's preferences, (ii) we allow the DM to communicate via cheap talk at this stage, and (iii) this uncertainty is publicly realized before the DM takes an action.[11]

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).[12] Our paper instead 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).[13]

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)

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

[11] 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.

[12] Daley and Green (2014) look at a similar question, studying how signaling incentives are changed when an (exogenous) informative signal about the agent's type is revealed after the agent chooses a costly signaling action.

[13] 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.

and Zapechelnyuk (2020) study information design in the presence of moral hazard problem while Hörner and Lambert (2020) study feedback design in 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 agent's 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 being "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. 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. Related incentives also arise in 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 decision maker will eventually choose $a = 1$ ("conviction") or $a = 0$ ("acquittal"). The preferences of the DM over the action depend on his privately known type, and the evidence $e \in E \equiv \mathbb{R}$. The DM also values his reputation in the eyes of the observer.

In the initial communication stage, the evidence is unknown, and the DM only knows its CDF F; we assume $\int_E e dF(e) < \infty$. The DM sends a cheap talk message $m \in M$ to the observer, where M is some sufficiently large metrizeable space. After the message is sent, the decision stage begins, where the evidence e is publicly revealed and the DM then 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 $q \in (0,1)$. Partisan DMs always want to choose $a = 0$, and their preference over the action does not depend on the evidence realization. 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 $L \equiv \text{Supp}(G)$. We assume for expositional convenience that either F or G is atomless. N types prefer to acquit more if (i) the evidence is less convincing (e is lower), or (ii) they are more lenient (ℓ is higher).[14] For convenience we will often refer to non-partisans with leniency ℓ as "ℓ types". Accordingly, we denote the set of all types by $\Theta = L \cup \{P\}$. The DM also values his reputation in the eyes of the observer of being an N type. More specifically, 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

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

where c and ρ are strictly positive.[15] 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**.

Remark 1. Notably we assume that there is no reputation incentive to appear as specific ℓ types on top of being perceived as non-partisan. We impose this for tractability and to focus on the novel reputation incentive for being responsive to evidence. In many of our applications ℓ could be thought of as issue specific and transitory, while partisanship is interpreted as more persistent. In Subsection 4.1 we discuss how some of our driving intuitions would extend to leniency-dependent reputation values.

We assume that reputation incentives are strong in the following sense.

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

[14] There is 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 the decision 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.

[15] 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.

Broadly, this assumption guarantees that the reputation incentives are strong enough to convince P to convict. Note that if $\rho < c$, then P will never convict. Assumption 1 is stronger and, as we will show, ensures that given any public history, P will convict 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 It will be useful to break up the DM's strategy between the communication stage and decision stage: the DM chooses a messaging strategy $\sigma : \Theta \to \Delta(M)$, and an action strategy $\zeta : M \times E \times \Theta \to \Delta(\{0,1\})$.[16] Given the DM's strategy, the publicly realized evidence e, the message m, and the chosen action a, denote $R(m,a,e)$ as the observer's belief that the DM is an N type.

We study perfect Bayesian equilibria with an additional refinement formalized below—hereafter, simply equilibria. An equilibrium $\mathcal{E}$ consists of σ, ζ, R such that for all $\theta \in \Theta$, $m \in M$ and $e \in E$

1. R is obtained from σ, ζ using Bayes rule whenever possible.
2. $\zeta(\cdot|m,e,\theta)$ is supported on $\arg\max_a \; u(\theta,e,a,R(m,a,e))$.
3. $\sigma(\cdot|\theta)$ is supported on $\arg\max_{m\in M} \int_E \left(\max_{a\in\{0,1\}} \; u(\theta,e,a,R(m,a,e))\right) dF(e)$.

In addition, we impose a version of the D1 refinement à la Cho and Kreps (1987) and Ramey (1996). For a fixed messaging strategy σ, let $\Theta_m \subset \Theta$ 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.[17] In our simple framework this refinement simplifies

[16] Let us describe some notation we adopt throughout the paper. For a metrizable space Y, we let $\Delta(Y)$ denote the set of all Borel probability measures over Y, endowed with the weak* topology. For any two measurable spaces Y and Z, transition probability $\chi : Y \to \Delta(Z)$, point $y \in Y$, and Borel measurable $\hat{Z} \subset Z$, we let $\chi(\hat{Z}|y) \equiv \chi(y)(\hat{Z})$. For example, $\zeta(0|m,e,\ell)$ is the probability that type ℓ takes action 0 after sending message m and evidence e is realized.

[17] 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 (as defined below) is the limit of some sequence of ex-ante D1 equilibrium outcomes.

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.

Given an equilibrium $\mathcal{E}$, let $U_\theta^{\mathcal{E}}(F)$ be the expected utility of the DM of type θ, and let $V^{\mathcal{E}}(F)$ be the probability of $a = 1$ (i.e., the investigator's expected utility).[18]

3. Equilibrium Details

This section details features that hold across all equilibria. Our results concern **equilibrium outcomes**, i.e., the distribution over equilibrium actions and reputation as a function of the DM's type and evidence realization. Our results below apply to a particular member of the equivalence class of equilibria that have the same outcomes with probability one. Given a strategy of the DM, we let G_m be the interim CDF over L conditional on message m and $\theta \in L$, with $L_m \equiv \text{Supp}(G_m)$, and q_m be the interim probability $\theta \in L$ conditional on m. Define $\tilde{e}_\ell \equiv \ell - c$.

Lemma 1. *For any equilibrium, the following hold:*

1. *Following any on-path message m, every $\ell \in L_m$ takes $a = 1$ $(a = 0)$ if $e > \tilde{e}_\ell$ $(e < \tilde{e}_\ell)$.*
2. *P positively mixes over all messages sent by N, i.e., $\text{Supp}(\sigma(\cdot|P)) = \cup_{\ell \in L} \text{Supp}(\sigma(\cdot|\ell))$. Following any on-path message m, P takes $a = 1$ $(a = 0)$ with positive probability $\iff$ $\exists \ell \in L_m : e > \tilde{e}_\ell$ $(e < \tilde{e}_\ell)$.*

It will be useful to describe the DM's decision stage strategy as choosing a distribution over contingent plans as a function of the evidence, i.e., over $x \in \mathcal{X} \equiv \{x' : E \to \{0, 1\}\}$. The first point says the ℓ type's action choice as a function of the evidence (almost surely) follows a fixed rule $x_\ell(e) \equiv \mathbb{1}(e \geq \tilde{e}_\ell)$.[19] This is not only constant across equilibria and on-path 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": in order to signal non-partisanship they select $a = 1$ for $e \in (\ell - c, \ell)$ where they would prefer $a = 0$. Instead, x_ℓ is distinguished by the fact

[18] As a notational convention, we superscript strategies, payoffs, outcomes, etc. with $\mathcal{E}$ to denote when they correspond to the equilibrium $\mathcal{E}$.

[19] The one caveat is that point 1 in the lemma does not pin down behavior at $\tilde{e}_\ell$. However, the set of joint type evidence realizations where $(\theta, e) = (\ell, \tilde{e}_\ell)$ arises with probability zero per our assumptions on F and G. Thus, the choice of ℓ's strategy at $e = \tilde{e}_\ell$ does not affect equilibrium outcomes.

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 second point says that the P cannot be distinguished from the ℓ types at 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.

The intuition behind point 2 of Lemma 1 follows from the high reputation incentives. If an on-path message is sent only by P then it yields an equilibrium reputation, and thereby also utility, of 0 for P. However, P can obtain at least utility $\rho q - c$ by mimicking the strategy some ℓ type, which is strictly preferred by Assumption 1. Conversely, an on-path 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,[20] which is strictly less than $\rho - c$ again by Assumption 1. The argument for why, after message m, P mixes over the actions chosen by $\ell \in L_m$ at the decision phase is similar, but has to contend with the subtlety that the relevant utility bounds are now dependent on q_m instead of the prior q. The proof shows that any equilibrium q_m is close enough to q such that the above argument goes through.

The intuition behind point 1 of Lemma 1 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 $\ell = e + c$ has the same preferences as P, i.e., he has the same trade off between the cost of conviction and reputation at this e. Combined with the fact that N's utility for conviction is increasing in ℓ, this means that ℓ types must make their action choice consistent with x_ℓ. Alternatively, if $a = 0$ (respectively $a = 1$) is off path, then it must be that $\ell > e + c$ (respectively $\ell < e + c$) for every $\ell \in L_m$; 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, note that $\{\sigma(\cdot|\ell)\}_{\ell \in L}$ induces an information structure about ℓ, henceforth the associated **Leniency Information Structure** (LIS). Formally, an LIS refers to a measure $\Sigma \in \Delta(M)$ and set of CDFs $\{G_m\}_{m \in \text{Supp}(\Sigma)}$ such that $\int_M G_m d\Sigma = G$. An LIS is associated with $\{\sigma(\cdot|\ell)\}_{\ell \in L}$ if $\Sigma(\cdot) = \int_L \sigma(\cdot|\ell) dG(\ell)$ and, $\forall m \in \text{Supp}(\Sigma)$, G_m is the interim CDF conditional on m derived from $\{\sigma(\cdot|\ell)\}_{\ell \in L}$.

Lemma 2. *For any LIS, there is a unique equilibrium outcome associated with it.*

The lemma says that equilibria can be uniquely described by the amount of information the communication stage conveys about the ℓ types. The intuition for why *any* LIS

[20] This follows from corollary 2 in Hart and Rinott (2020).

can be sustained at the communication stage comes from the fact that x_ℓ maximizes the expected utility difference between type ℓ and type P across all contingent plans. The proof uses this to show that ensuring P's indifference across mimicking any ℓ type also ensures incentive compatibility for *every* ℓ type regardless of N's strategy. Because the LIS "essentially" pins down the ℓ types' messaging strategies, and Lemma 1 pins down their decision stage strategies, the uniqueness in Lemma 2 refers to the uniqueness of the P's equilibrium strategy when paired with an LIS. Each LIS will require different strategies for P to sustain indifference, which then naturally leads to different equilibrium outcomes. The next section compares these outcomes.

4. Comparing Equilibria

Our main comparison concerns the equilibrium probability of $a = 1$, or $V^{\mathcal{E}}(F)$. Recall that this is the utility of the investigator in our main specification. However, even with a fixed exogenous investigation, this probability is an important outcome: it represents the probability of politicians taking votes against their party's interest, the probability that skeptical bureaucrats approve requests, or the amount of donor applicants admitted to university.

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 announcing 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. In addition, we say there is **mild agreement** if for every pair $\ell', \ell'' \in \text{Supp}(G)$, $\exists e \in \text{Supp}(F) : \; x_{\ell'}(e) = x_{\ell''}(e)$, i.e. no two ℓ types always choose different actions in equilibrium.

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

Depending on the parameters, many 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. On the other hand, the theorem says that whenever the

action is not completely predictable at the decision stage, the equilibrium delivers different outcomes than ex-ante signaling; in particular, a strictly lower conviction probability.[21] *All* imperfectly informative LIS are associated with an equilbrium with residual strategic uncertainty if and only if each ℓ type's conviction probability (i.e., $1 - F(\tilde{e}_\ell)$) is different.

Note that Theorem 1 implies that if the investigator could select the equilibrium, then he would select ex-ante signaling. This is important because in many settings, whether ex-ante signaling outcomes prevail is a design decision. In Subsection 6.1 we discuss how endowing the DM with commitment to a contingent action plan (before e is realized) can select ex-ante signaling outcomes. In Subsection 6.2, we discuss how whether ex-post or ex-ante signaling outcomes arise depends on the investigator's timing of information disclosure. For example, government agencies like the FDA and FTC could require that officials commit to an acceptance threshold, at the inception of a proposal, rather than deciding on a case-by-case basis after the results of the trial or investigation are realized. If these agencies are worried that their officials are unduly biased against approval then such a policy would be beneficial.

Given that $a = 1$ is taken most often under ex-ante signaling, a natural follow up question is whether the same holds for each level of evidence. While the answer to this question is difficult to answer in general, we show such a ranking does indeed hold when comparing ex-ante signaling to the equilibrium associated with the uninformative LIS (i.e., a babbling equilibrium). We refer to this equilibrium as **ex-post signaling** and denote it equilibrium β. Under ex-post signaling, it is as if every type of DM sends the same message interpreted as "I will wait and see until the investigation concludes." Ex-post signaling "usually" admits residual strategic uncertainty; if and only if there exist two leniency types that convict with different probability. Denote $\psi^{\mathcal{E}}(e)$ as the probability of $a = 1$ given evidence e under equilibrium $\mathcal{E}$.

Proposition 1. $\psi^{\alpha}(e) \geq \psi^{\beta}(e)$ *for all* $e \in \text{Supp}(F)$.

Thus, even if the investigator has a higher preference for $a = 1$ after higher evidence realizations (but still always prefers $a = 1$ to $a = 0$), ex-ante signaling will be investigator preferred relative to ex-post signaling. When viewing the distinction between these two equilibria as a difference in the timing of communication (i.e., whether communication happens before or after evidence is disclosed), Proposition 1 implies that such an evidence-sensitive investigator will prefer to delay evidence disclosure until after communication

[21] Mild agreement rules out cases in which P's decision over which ℓ type to mimic is unchanged between the communication stage and the decision stage. An example of such a case is where ℓ is supported on some interval $[\underline{\ell}, \overline{\ell}]$ and F is completely supported outside of $[\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}}]$.

takes place.

It is worth noting that there is nothing "mechanical" about ex-ante signaling that leads to lower conviction probability. Under ex-ante signaling, P could announce a very high threshold. 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 convict and pool with $\ell > e + c$, or to acquit and pool with $\ell < e + c$, while under ex-ante signaling, P can directly target any specific leniency type. That is, $\psi^\beta(e)$ depends only on $G(e + c)$ whereas $\psi^\alpha(e)$ 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 is supported on the entire real line which guarantees minimal agreement, and $c = 1$ for notational convenience. Under ex-ante signaling, P will mix between $m_{\underline{\ell}}$ and $m_{\overline{\ell}}$ so that

$$\rho\left(R^\alpha(m_{\underline{\ell}}) - R^\alpha(m_{\overline{\ell}})\right) = F(\tilde{e}_{\overline{\ell}}) - F(\tilde{e}_{\underline{\ell}}),$$

where $R^\alpha(m_\ell) \equiv \mathbb{P}(\theta \in L | m_\ell) = q_{m_\ell}$. That is, the difference in reputation at $m_{\underline{\ell}}$ relative to $m_{\overline{\ell}}$ is proportional to the difference in probability with which type $\underline{\ell}$ takes $a = 1$ relative to $\overline{\ell}$.

Now let us compare the probability of conviction for each evidence realization between ex-post and ex-ante signaling. If $e < \tilde{e}_{\underline{\ell}}$ or $e > \tilde{e}_{\overline{\ell}}$, then Lemma 1 implies that all DM types take the same action—acquittal and conviction respectively—under all equilibria. Thus the comparison turns on that for $e \in [\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$. Under ex-ante signaling, P convicts with probability $\sigma^\alpha(m_{\underline{\ell}}|P)$. Under ex-post signaling, P similarly convicts with the probability that he mimics the $\underline{\ell}$ type which is determined by

$$\rho\left(R^\beta(1, e) - R^\beta(0, e)\right) = 1,$$

where $R^\beta(a, e)$ is the reputation from choosing action a given evidence e under ex-post signaling. Like under ex-ante signaling, the difference in reputation between mimicking $\underline{\ell}$, i.e., choosing $a = 1$, and mimicking $\overline{\ell}$, i.e., choosing $a = 0$, is proportional to the difference in probabilities that these ℓ types convict. Figure 1 illustrates how P shifts his strategy so that the reputation incentives compensate him for the difference in conviction rates between $\underline{\ell}$ and $\overline{\ell}$. Under ex-post signaling, conditional on evidence $e \in [\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$, the difference

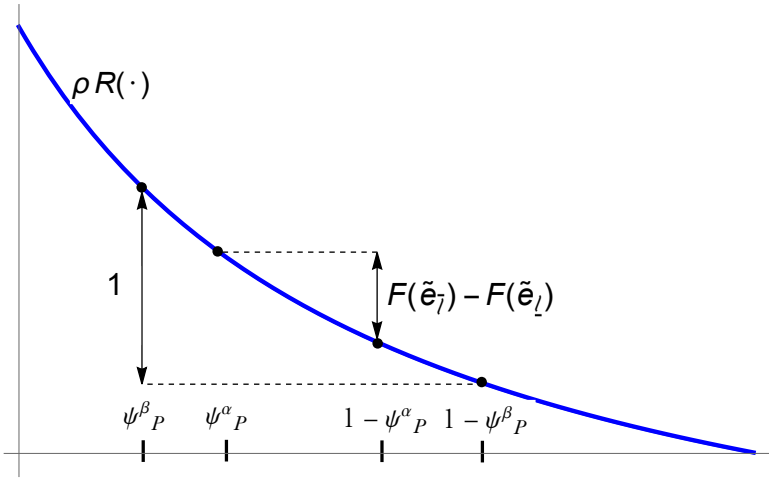


Figure 1: Reputation as a function of the partisan's strategy in the binary example where $\psi_P^{\mathcal{E}}$ is the probability P takes $a = 1$ at $e \in (\tilde{e}_{\underline{\ell}}, \tilde{e}_{\overline{\ell}})$ in $\mathcal{E}$.

in conviction rates 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 convict less frequently. Since the ℓ types' conviction rates do not depend on the equilibrium, this argument establishes that $V^{\alpha}(F) - V^{\beta}(F) > 0$ 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 some 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.[22,23]

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.

[22] The intuition in the binary type case generalizes beyond our baseline model. For instance, one can show that it extends to the case in which different leniency types merited different reputational payoffs in the eyes of the public relative to the partisan type, and to the case in which there is some heterogeneity in the disutility c partisans face from taking $a = 1$.

[23] 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 policy.

4.2. Proof Sketch of Theorem 1

We first establish an inverse relationship between the P's equilibrium utility and the equilibrium probability of conviction. We think this relationship has independent interest, and it turns out to be useful in the subsequent information design problem as well. Here, it means that proving Theorem 1 comes down to proving that P's equilibrium utility is lowest under ex-ante signaling. We then show this comparison can be seen from convexity of P's equilibrium expected utility under ex-ante signaling utility in the investigation F. This convexity is driven by an underlying convexity 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 present the proof here because of its simplicity.

Proof. Take any equilibrium $\mathcal{E}$. Because of Lemma 1, P is indifferent across mimicking the strategy of any leniency type ℓ, i.e.,

$$U_P^{\mathcal{E}}(F) = \int_E \big(- cx_\ell(e) + \rho R(m, x_\ell(e), e)\big) dF(e) \;\; \forall m, \ell : \;\ell \in L_m.$$

Taking the expectation over $(m, x) \in M \times \mathcal{X}$ with respect to the equilibrium distribution, and applying the law of iterated expectations, gives $U_P^{\mathcal{E}}(F) = \rho q - cV^{\mathcal{E}}(F)$. *Q.E.D.*

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 investigator have opposing interests concerning the probability of conviction. 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 conviction. 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 $U_P^{\mathcal{E}}(F)$, it does not depend on the investigation F, or the distribution of leniency G. This 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 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.*

Combining Lemma 3 and Lemma 4 completes the argument for the fact that ex-ante signaling has a higher conviction probability than ex-post signaling.

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 update 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 a P type following any message is concave in the probability that P sends that message, or equivalently, the DM's reputation for being an N type is convex in P's strategy. This convexity is illustrated in Figure 1. 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 equilibrium expected utility from sending message m_ℓ for any F is $\rho R^{\alpha}(m_\ell) - c(1 - F(\tilde{e}_\ell))$. This means that, when sending m_ℓ, 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, and therefore does worse under F_λ.[24]

Ex-Ante Signaling vs. Other Equilibria We have shown that ex-ante signaling has a higher conviction probability than ex-post signaling. However, Theorem 1 says that ex-

[24] 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.

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 an on path message m, i.e., $U_P^{\mathcal{E}}(F)$, is the ex-post signaling equilibrium utility with prior equal to the interim belief (q_m, G_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 (q_m, G_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 according to the ex-ante signaling equilibrium given prior (q_m, G_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. *$U_\theta^{\alpha}(F) \leq U_\theta^{\mathcal{E}}(F)\ \forall \theta \in \Theta$; under mild agreement, the inequality is strict if and only if $\mathcal{E}$ has residual strategic uncertainty.*

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 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 points to interesting questions about equilibrium selection issues, which we address in Subsection 6.1.

5. Optimal Investigations

Having studied the impact of communication on signaling incentives for arbitrary fixed F, we now turn to the investigator's problem of designing an optimal investigation. 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. However, when considering the investigator's problem, what e represents will impose different restrictions on the set of F which are feasible. For simplicity, we focus on the case where e represents a belief about an unknown binary state, and adopt the standard Bayesian persuasion approach of allowing the investigator to choose any F that satisfies a corresponding Bayes plausibility constraint. In Subsection 6.4, we discuss how our results extend to an investigator designing an experiment over multiple states.

Let $\omega \in \{0,1\}$ with prior $\overline{e} \in (0,1)$ that $\omega = 1$. In this case, the Bayes plausibility constraint takes the form $\int_0^1 e dF(e) = \overline{e}$, which we can rewrite as $\int_0^1 (1-F(e))de = \overline{e}$. To ease exposition, for the rest of the section we assume henceforth that the leniency distribution G admits a continuous density g with $[c, 1+c] \subseteq \text{Supp}(G)$; this means that increasing e always increases the equilibrium rate of conviction among non-partisans.

The design of an optimal investigation depends on the equilibrium that is being played (which could in principle change depending on F). We focus on optimal design with the assumption that the communication equilibrium will be the investigator's preferred one, namely ex-ante signaling.[25] As mentioned, ex-ante signaling is also salient, because it is uniquely selected in a natural perturbation of our game (see Subsection 6.1).

5.1. Characterization

To calculate the investigator's utility, we sum the conviction probability given message m_ℓ weighted by the probability that the DM sends message m_ℓ. Letting $\mathcal{F}$ be the set of

[25] In Appendix E, we compare the optimal investigation between ex-post and ex-ante signaling.

CDFs with support on $[0,1]$, the investigator's design problem is

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

While the equilibrium strategies of ℓ types are fixed across all choices of F, this is not true for P, i.e., $\sigma(m_\ell|P)$ depends on the choice of F. Consider $\ell' > \ell''$ who use the corresponding decision rules $x_{\ell'}$ and $x_{\ell''}$. Because $x_{\ell'}$ chooses $a=1$ less often than $x_{\ell''}$, in terms of material payoffs P prefers sending $m_{\ell'}$ to sending $m_{\ell''}$. The reputation difference between $m_{\ell''}$ and $m_{\ell'}$ must compensate P in equilibrium to ensure indifference, i.e., $R^\alpha(m_{\ell''}) \geq R^\alpha(m_{\ell'})$. Exactly how much the reputation must compensate P depends on the choice of F: in switching from mimicking ℓ'' to ℓ', the DM changes his action from $a=1$ to $a=0$ with probability $F(\tilde{e}_{\ell'}) - F(\tilde{e}_{\ell''})$. If the investigator increases this probability then the equilibrium reputation difference $R^\alpha(m_{\ell''}) - R^\alpha(m_{\ell'})$ must also increase. This change in reputations is achieved through P decreasing the probability with which he mimics ℓ'' relative to ℓ'. Because P must be indifferent across all messages, this change also has nonlocal effects on the equilibrium. Intuitively, if the investigation makes more convincing evidence realizations (higher e) more likely, P will pretend to possess more forgiving values (higher ℓ). This response mitigates the gains the investigator realizes from increasing the probability of high e.

The dependence of $\sigma(\cdot|P)$ on F means we cannot solve the investigator's problem using standard information design techniques. The probability of $a=1$ as a function of the evidence depends on the investigation and so, in the language of the Bayesian persuasion literature, the investigator's value is not linear in F.

In order to make the problem tractable, we use Lemma 3, which shows that maximizing the investigator's value is equivalent to minimizing that of P. We can therefore take the investigator's problem to be one of choosing F to minimize $U_P^\alpha(F)$. This is convex in F by Lemma 4. We then need to pin down how the choice of 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 $\int_L \frac{g(\ell)q\rho}{U+c(1-F(\tilde{e}_\ell))}d\ell = 1$.[26] These

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

observations allow us to rewrite the investigator's problem as follows:

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

The first 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.

Let $h(e) \equiv g(e+c)\ \forall e \in (0,1)$ with $H(e) \equiv \int_0^e h(e)de$. Since each $\ell \in (c, 1+c)$ chooses $a=1$ if and only if $e \geq \tilde{e}_\ell$, $H(e)$ represents the mass of extra ℓ types that choose $a=1$ given evidence e on top of the $G(c)$ who always choose $a=1$. Define $\overline{h}$ as the decreasing ironed value of h: the smallest decreasing function $\tilde{h} : [0,1] \to \mathbb{R}_+$ such that $\int_0^e h(x)dx \leq \int_0^e \tilde{h}(x)dx\ \forall e \in (0,1)$. Thus, $\int_c^e \overline{h}(e)de$ is the concavification of H. Because the non-partisans use a fixed threshold at the decision stage, H and its concavification capture the investigator's design incentives conditional on the DM being an ℓ type.

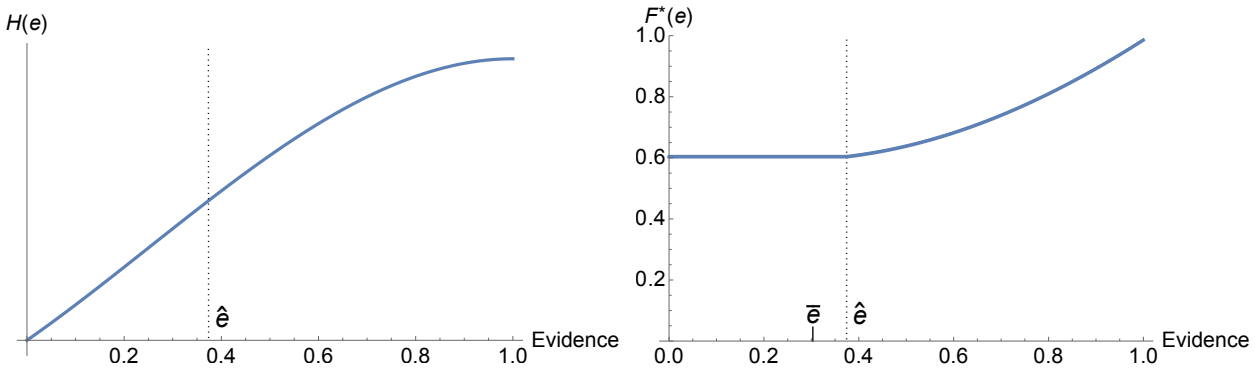
Theorem 2. *For $k, U \in \mathbb{R}$, define $\overline{F}(e;k,U) \equiv U/c + 1 - k\sqrt{\overline{h}(e)}$. The uniquely optimal investigation is given, for $e<1$, by*

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

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

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.

Figure 2 presents an example of an optimal investigation. In this example, the distribution of non-partisan ideologies 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 $\hat{e}$. This means that the investigator obtains a higher conviction probability from



(a) Distribution of non-partisan commitments

(b) Optimal investigation

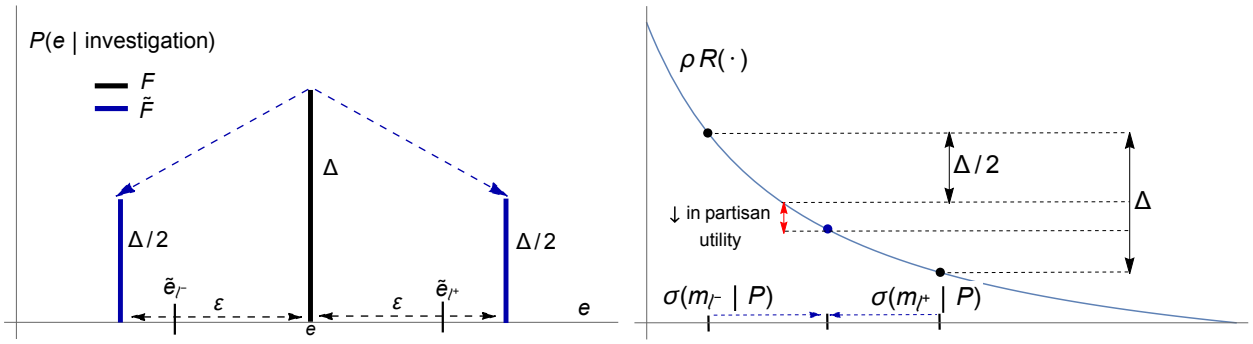
Figure 2: $L = \mathbb{R}$, $g(\ell)$ is a standard logistic distribution with mean $\frac{1}{2}$, $c = \frac{1}{4}$, $q = \frac{1}{2}$, $\overline{e} = \frac{3}{10}$, and $\rho = 3$. We numerically solve for $U_P^{\alpha}(F^*) = 1.381$ and for $k = 11.853$. By Lemma 3 the investigator obtains utility $V^{\alpha}(F^*) = .478$.

N types by providing information below $\hat{e}$, and withholding information above $\hat{e}$. From the right panel of Figure 2, we see that F^* is consistent with this incentive below $\hat{e}$, but in contrast, provides some information, in a smooth way, above $\hat{e}$ to the detriment of N's conviction probability. 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 on* $(0, 1)$*; 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 E, this "unpredictability" is not a feature of the optimal investigation under ex-post signaling, i.e., when the communication stage is uninformative.

Of course, because of the reputational consequences, it is not clear that P would benefit from mass points in the investigation. For example, when feigning principles just above a



(a) A local spread of investigation F

(b) Change in partisan strategy

Figure 3: Locally spreading any mass point harms the partisan.

mass point of evidence, the equilibrium reputation will adjust downwards to reflect the use of such declarations by P. The ability to target such announcements helps P in equilibrium because of the same "convexity of reputation" discussed in Subsection 4.2.

Figure 3 depicts two investigations that differ only around evidence e, with $c = 1$ for convenience. F has an isolated (for ease of exposition) 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 ε is small, because the density of ℓ types g is continuous, the change in conviction from ℓ types is second order. However, P's conviction rate increases 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^- convicts 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^+} convict with the same probability and therefore must command the same reputation. The reputation for these associated messages as a function of $\sigma(m_\ell|P)$ is illustrated in the right panel of Figure 3. This reputation is convex: as P increases $\sigma(m_\ell|P)$, 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 utility at these messages has fallen.[27] Because of the opposing interests lemma, this change benefits the investigator.

Corollary 3. *The optimal investigation produces full information if and only if $\overline{h}$ is constant.*

[27] 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.

This corollary is a direct implication of the fact that $F^*(e) \in (0,1)$ is constant in e 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 ℓ types. Therefore, an alternative statement of Corollary 3 is that the investigator provides full information if and only if full information maximizes conviction among ℓ types. Since 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 constant, but are misaligned when $\overline{h}$ is decreasing.

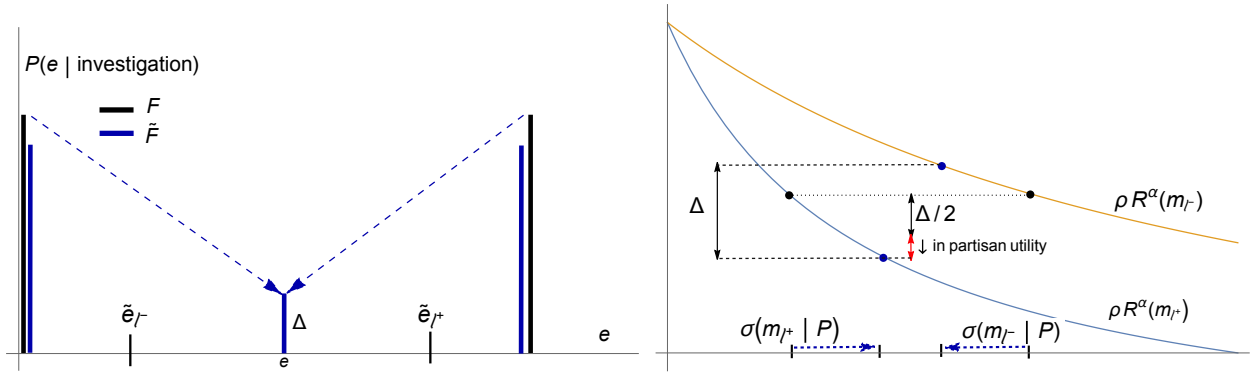
At a high level, the intuition is as follows. All else equal, P benefits from correlating his strategy with the ℓ types. When the investigator increases the probability of evidence in an interval, i.e., increases $F(\tilde{e}_{\ell''}) - F(\tilde{e}_{\ell'})$ for $\ell'' > \ell'$, P reallocates mass from mimicking $\ell < \ell'$ to $\ell > \ell''$. 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 N strategies and thereby harm P. Given the opposing interests lemma, the former change harms the investigator, while the latter change benefits them. Again, the more precise intuition revolves around the convexity of reputation mentioned above.

Figure 4 illustrates why the investigator benefits from partial information when g is decreasing. The left panel shows the fully informative investigation F and a deviation $\tilde{F}$ which contracts Δ mass to some interior evidence level e. Consider two types ℓ^- and ℓ^+ such that $\tilde{e}_{\ell^-} < e < \tilde{e}_{\ell^+}$. Because g is decreasing, ℓ types are more prevalent at m_{ℓ^-} than at m_{ℓ^+}. The right panel illustrates the reputation for ℓ^+ (in blue) and ℓ^- (in orange) as a function of σ^P. Under F, both m_{ℓ^-} and m_{ℓ^+} convict with the same probability and therefore must have the same equilibrium reputation. This means that P must send m_{ℓ^-} with higher probability than m_{ℓ^+}. Under $\tilde{F}$, m_{ℓ^-} convicts with Δ higher probability than m_{ℓ^+}. In order to preserve incentives, P must shift probability away from m_{ℓ^-} towards m_{ℓ^+}. The right panel shows that because of the convexity of reputation, this change harms P: the proportional decrease in the rate P sends m_{ℓ^-}, where the rate is already high, is small, while the proportional increase in the rate P sends m_{ℓ^+}, where the rate is low, is large. The net effect is that P's utility decreases and $\tilde{F}$ is worse for P.

5.2. Comparative Statics

We next explore comparatives statics of the investigation design problem and begin by documenting some basic changes in the parameters that increase the conviction rate.

Proposition 2. *Let $\tilde{G}$ be a distribution of ℓ that first-order stochastic dominates G. For any fixed F,*



(a) Contraction of the fully informative investigation F

(b) Change in Partisan's strategy

Figure 4: Contracting from full information harms P when g is decreasing.

the investigator's equilibrium expected utility is higher under G than G' and when ρ or q increases.[28]

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 means that non-partisan types prefer $a = 1$ more often. By increasing ρ, we are increasing the importance of reputation relative to material payoffs in the DM's utility, which can alternatively be interpreted as a *decrease* in the stakes of the action a. This change then reduces the misalignment between the partisan and investigator.[29]

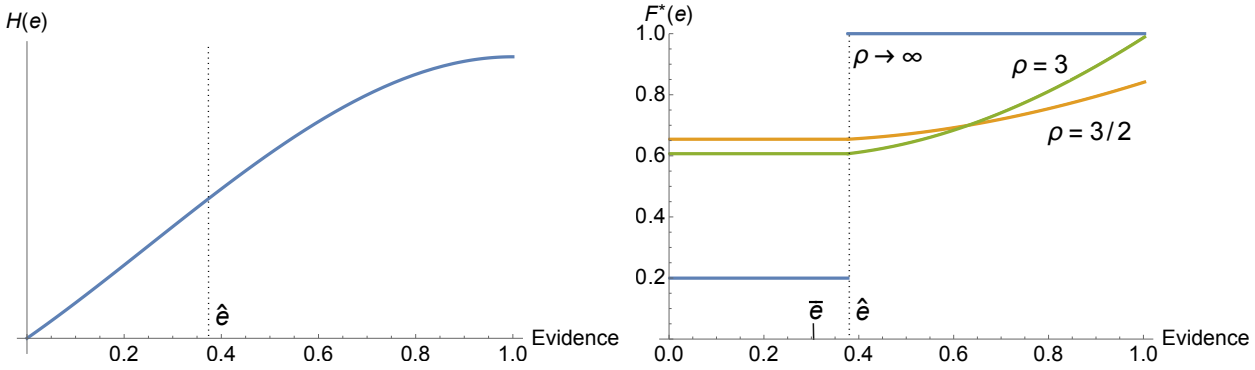
Our next result looks at how the optimal investigation changes with the size of reputation incentives. We can interpret changes in ρ as changes the "stakes" of the decision: lower ρ corresponds to a higher relative importance of the action. Our next result shows that higher stake decisions will have more informative investigations.

Proposition 3. *The optimal investigation F^* is decreasing in informativeness as ρ or q increases.*

We illustrate in Figure 5 how the optimal investigation changes with ρ in the example from Figure 2. In the limiting case when $\rho \to \infty$, P will fully mimic the distribution of ℓ

[28] 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 conviction 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.

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



(a) Distribution of non-partisan commitments

(b) Optimal Investigations

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

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 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 and thereby also leads 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 below the support of F^* always take $a = 1$, and so a spread of principles in this region can only increase the probability of $a = 0$. Similarly, a spread of the distribution of ℓ types in the region above the support of F^* 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 conviction.

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 $\tilde{e}_\ell \notin \text{Supp}(F^*)$. 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 ℓ' and ℓ'' so that they are further away from each other. This spreads the conviction probability from mimicking these types for P as there are new evidence realizations between $\tilde{e}_{\ell'}$ and $\tilde{e}_{\ell''}$. As a result the equilibrium reputation for $m_{\ell'}$ and $m_{\ell''}$ 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. This suggests that P will mimic ℓ'' more under a spread, however this spread also changes the conviction rate from ℓ types in a generally ambiguous way: ℓ'' convicts less frequently and ℓ' convicts more frequently. Our regularity assumptions guarantee that under the optimal investigation, the former effect dominates.

6. Discussion and Extensions

6.1. Commitment and Equilibrium Selection

As with many cheap talk models, 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. There is a strong intuition that providing less information at the communication stage or "dodging the cameras" would be interpreted negatively, pushing against equilibrium multiplicity. This section shows that allowing for commitment to a contingent plan uniquely selects ex-ante signaling outcomes. Recall that ex-ante signaling admits no residual strategic uncertainty, i.e., it is *as if* the DM commits to a contingent plan when communicating. 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 respec-

tively. Motivated by this, we show that ex-ante signaling outcomes are the unique equilibrium outcomes 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.

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.[30]

Proposition 5. *The commitment model has a unique equilibrium outcome in which each ℓ type chooses x_ℓ and P positively mixes over $\{x_\ell\}_{\ell \in L}$. The equilibrium outcomes are the same as under ex-ante signaling in our baseline model.*

The interpretation of the proposition can be broken down into two points. First, ex-ante signaling remains an equilibrium when the DM actually commits to x_ℓ, instead of sending a message that is interpreted as such a commitment (as in ex-ante signaling). Second, no other equilibria can be sustained; in particular, in equilibrium each ℓ type cannot 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; i.e., it maximizes the material utility difference between ℓ and P across all contingent plans. This means that x_ℓ is always more tempting for ℓ than for P, so (roughly) whether x_ℓ is on or off path, it will always be the best option for type ℓ.

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.[31] For ease of exposition, we assume F has full support on $\mathbb{R}$, which guarantees mild agreement.

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 $u(\theta, e, a, \mu) + \varepsilon a$

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

[31] We continue to apply the D1 refinement, which now has implications at the communication stage. Since different types have different value from choosing different commitments, the D1 refinement can be used to rule out certain beliefs following the option to commit.

where ε 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. 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. The variable ε 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.[32] 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. Not committing to a contingent plan signals a desire to be responsive to ε instead of a desire to be responsive to the evidence.[33]

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 is formally announced, or firms could publicly disclose financial records before submitting

[32] With arbitrarily large δ, the option value from acting on the realization of ε could exceed the reputational gains from committing at the communication stage. When the parameters violate the assumption in Proposition 6, each x_ℓ commitment would still garner a reputation of 1 according to the D1 refinement, but could go unused.

[33] 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).

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 $e_i \sim F_i$. After the message is sent, the final evidence $e_f \sim F_f(\cdot|e_i)$ 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_f being payoff relevant. Let $\overline{F}$ be the unconditional distribution of e_f.[34] We maintain the focus on ex-ante signaling equilibria in each subgame following the realization of e_i, and so our timing results also apply to the commitment model.

Consider an investigator who can choose among different (F_i, F_f) with the same $\overline{F}$. By choosing different F_i, he can span various timings of evidence disclosure. When F_i 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_f 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_i, 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_i and F_f with the same $\overline{F}$, $F_i = \overline{F}$ delivers the lowest $\mathbb{P}(a = 1)$, and $F_f(\cdot|\cdot) = \overline{F}$ 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_f 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.

[34] More precisely, $\overline{F}(e_f) = \int_{e_i} F_f(e_f|e_i)dF_i(e_i)$.

We maintain that $e \in [0,1]$ represents a posterior belief about a binary state, and G admits a 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 3 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 3 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 $\omega \in [0,1]$,[35] where the domain is $[0,1]$

[35] 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.[36]

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.

[36] 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.

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

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.

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

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.

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. Exogenous Investigation Proofs

We let $M^{\mathcal{E}}$ denote the set of on-path messages in equilibrium $\mathcal{E}$. For $m \notin M^{\mathcal{E}}$, we set $R(m, a, e) = 0$ for all a, e (with $\Theta_m = \{P\}$). For notational simplicity, we will often drop dependence on F in $U_P^{\mathcal{E}}(F)$ in the proofs for Section 3 as it is held fixed.

We first show that a weaker version of the condition in Assumption 1 holds with respect to each interim belief q_m.

Lemma 5. *For all $m \in M^{\mathcal{E}}$, $\rho > \max\{\frac{c}{q_m}, \frac{c}{1-q_m}\}$.*

Proof. The Lemma holds trivially by Assumption 1 if there is only one on-path message. Suppose $|M^{\mathcal{E}}| \geq 2$. Take any message $m \in \text{Supp}(\sigma(\cdot|P))$. P's expected equilibrium utility from sending message m must be $U_P^{\mathcal{E}}$. By Corollary 2 of Hart and Rinott (2020), P's expected reputation conditional on message m is at most q_m. Because P's expected material payoff conditional on sending any message is weakly negative, $U_P^{\mathcal{E}} \leq \rho q_m$. This bound must hold for all $m' \in M^{\mathcal{E}}$, as any $m' \in M^{\mathcal{E}} \setminus \text{Supp}(\sigma(\cdot|P))$ has $q_{m'} = 1$.

We now argue that $U_P^{\mathcal{E}} \geq \rho q_m - c$ for all $m \in M^{\mathcal{E}}$. Take any such m. Bayes plausibility of the ex-post reputations implies that for each realization of e, there exists an action a_e for which $R(m, a_e, e) \geq q_m$. For following $x(e) = a_e$ to not be profitable deviation for P, we must have

$$U^{\mathcal{E}_{\mathcal{P}}} \geq \int_E (-ca_e + \rho R(m, a_e, e))dF(e) \geq \int_E (-c + \rho q_m)dF(e) = \rho q_m - c.$$

Similar bounds on $U_P^{\mathcal{E}}$ hold with respect to q rather than the interim beliefs q_m. Bayes plausibility of the interim reputation implies there exists $m, m' \in M^{\mathcal{E}}$ such that $q_{m'} \leq q \leq q_m$. That $U_P^{\mathcal{E}} \leq \rho q_{m'}$ then implies $U_P^{\mathcal{E}} \leq \rho q$. Similarly, that $U_P^{\mathcal{E}} \geq \rho q_m - c$ implies $U_P^{\mathcal{E}} \geq \rho q - c$.

We now show that $\rho > \frac{c}{q_m}$ for all $m \in M^{\mathcal{E}}$. Suppose, for the sake of contradiction, $\rho q_m \leq c$. Because $U_P^{\mathcal{E}} \leq \rho q_m$, we then have $U_P^{\mathcal{E}} \leq c$. Combining this inequality with $\rho q - c \leq U_P^{\mathcal{E}}$ implies $\rho q - c \leq c$ or $\rho \leq \frac{2c}{q}$, a contradiction of Assumption 1.

Finally, we show that $\rho > \frac{c}{1-q_m}$ for all $m \in M^{\mathcal{E}}$. Suppose, for the sake of contradiction, $\rho(1 - q_m) \leq c$, which implies $\rho - c \leq \rho q_m$. We know $U_P^{\mathcal{E}} \geq \rho q_m - c$. Because $\rho q_m \geq \rho - c$, we have $U_P^{\mathcal{E}} \geq \rho - 2c$. This inequality, when combined with $U_P^{\mathcal{E}} \leq \rho q$, implies $\rho - 2c \leq \rho q$ or $\rho \leq \frac{2c}{1-q}$, a contradiction of Assumption 1. *Q.E.D.*

The next lemma uses the belief restrictions in Lemma 5 to derive an upper-bound on reputations (on- and off-path) and a lower bound for on-path reputations.

Lemma 6. *Take any $m \in M^{\mathcal{E}}$ and $e \in \text{Supp}(F)$. $R(m,a,e) < 1$ for all $a \in \{0,1\}$. For each on path a, $R(m,a,e) > 0$.*

Proof. Fix any $m \in M^{\mathcal{E}}$ and e. For the sake of contradiction, suppose $R(m,a,e) = 1$ for some $a \in \{0,1\}$ (on- or off-path). Bayes plausibility implies $R(m,a',e) \leq q_m$ and P must choose $a' \neq a$ with probability one after e, m. P's utility from choosing a when e is realized is $\rho - ca$, while his utility from a' is at most $\rho q_m - ca'$. For a' to be an equilibrium strategy for P, it must be that $\rho - ca \leq \rho q_m - ca'$, which simplifies to $\rho(1-q_m) \leq c(a-a') \leq c$, a contradiction of Lemma 5.

For the sake of contradiction, suppose a is on-path and $R(m,a,e) = 0$. Then P must choose a with positive probability after m, e, yielding a payoff of $-ca + \rho R(m,a,e) = -ca \leq 0$. By Bayes plausibility, $R(m,a',e) \geq q_m$ for $a' \neq a$. Choosing a' after m, e yields a payoff of $-ca' + \rho R(m,a',e) \leq -c + \rho q_m$. For P to weakly prefer a to a', we must have $\rho q_m - c \leq 0$, a contradiction of Lemma 5. *Q.E.D.*

We now show some implications of D1 on the equilibrium actions. Remember that we are imposing the D1 refinement on the signaling game induced by type space Θ_m in the subgame following message $m \in M^{\mathcal{E}}$ and evidence e.

Lemma 7. *Let a' be an off-path action after evidence $e \in \text{Supp}(F)$ and message $m \in M^{\mathcal{E}}$ and $a = 1 - a'$. Then $R(m,a',e) = 1$ if $(e - \tilde{e}_\ell)(a - a') < 0$ for some $\ell \in L_m$ and $R(m,a',e) = 0$ if $(e - \tilde{e}_\ell)(a - a') > 0$ for all $\ell \in L_m$.*

Proof. Consider the subgame following $m \in M^{\mathcal{E}}$ and e. Let a' be an off-path action. Then $R(m,a,e) = q_m$ for $a = 1 - a'$. Each $\theta \in \Theta_m$ receives an equilibrium utility of $u(\theta, e, a, q_m)$. D1 requires that, if after observing an off-path action a', the observer place no weight in their belief on any type θ for which there exists $\theta' \in \Theta_m$ such that

$$\{\mu \in [0,1] : \; u(\theta, e, a', \mu) > u(\theta, e, a, q_m)\} \subsetneq \{\mu \in [0,1] : \; u(\theta', e, a', \mu) > u(\theta', e, a, q_m)\}. \quad (3)$$

We note that $\{\mu \in [0,1] : \; u(P, e, a', \mu) > u(P, e, a, q_m)\}$ being non-empty is, after some simplification, equivalent to $\rho(1-q_m) \geq c(a'-a)$, which holds because $\rho > \frac{c}{1-q_m}$ by Lemma 5.

D1 requires $R(m,a',e) = 1$ if (3) holds for $\theta = P$ and some $\theta' \in L_m$, which simplifies to $(e - \tilde{e}_\ell)(a - a') < 0$. Similarly, D1 requires $R(m,a',e) = 0$ 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.*

Proof of Lemma 1

Proof. An immediate implication of Lemma 5 is that $\cup_{\ell \in L} \text{Supp}(\sigma(\cdot|\ell)) = \text{Supp}(\sigma(\cdot|P)) = M^{\mathcal{E}}$: otherwise, there would exist m such that $q_m \in \{0,1\}$, which would violate $\rho > \max\{\frac{c}{q_m}, \frac{c}{1-q_m}\}$. Thus, $P \in \Theta_m$ for all $m \in M^{\mathcal{E}}$.

We now prove point 1 of the lemma. Take any $m \in M^{\mathcal{E}}$ and $\ell \in L_m$. Consider the subgame after message m is sent and evidence e is realized. We note that $\cup_{\ell \in L_m} \text{Supp}(\zeta(\cdot|m,e,\ell)) = \text{Supp}(\zeta(\cdot|m,e,P))$; otherwise, $R(m,a,e) \in \{0,1\}$ for some on-path a, a contradiction of Lemma 6.

Suppose for the sake of contradiction, that $e \neq \tilde{e}_\ell$ and ℓ chooses $a = \mathbb{1}(e < \tilde{e}_\ell)$ with positive probability. Then, for $a' = 1 - a$, we must have

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

If P is indifferent between a and a', then

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

Subtracting (5) from (4) and simplifying yields $(e - \tilde{e}_\ell)(a - a') \geq 0$, a contradiction since $a - a' > 0$ if and only if $e - \tilde{e}_\ell < 0$. Therefore, P must choose a with probability one and a' must be an off-path action. But, by Lemma 7, $R(m,a',e) = 1$ because $(e - \tilde{e}_\ell)(a - a') < 0$. Therefore, ℓ cannot choose $a = \mathbb{1}(e < \tilde{e}_\ell)$ with positive probability in equilibrium.

Finally, we note that the rest of point 2 of the lemma follows from the characterization of ℓ's strategy and that $\cup_{\ell \in L_m} \text{Supp}(\zeta(\cdot|m,e,\ell)) = \text{Supp}(\zeta(\cdot|m,e,P))$. *Q.E.D.*

Proof of Lemma 2

Let $(\Sigma, \{G_m\}_{m \in \text{Supp}(\Sigma)})$ be an LIS and let $M' = \text{Supp}(\Sigma)$. Take $\{\sigma(\cdot|\ell)\}_{\ell \in L}$ to be a set of messaging strategies such that $(\Sigma, \{G_m\}_{m \in M'})$ is the LIS associated with $\{\sigma(\cdot|\ell)\}_{\ell \in L}$ and $\sigma(\cdot|\ell)$ is supported on $\{m : \ell \in L_m\}$—that is, no type ℓ sends a message m for which $\ell \notin L_m$. We will construct an equilibrium using these messaging strategies for ℓ types. The key step in the construction is to pin down P's strategies and the reputation function. To do so, we will set the ℓ types' action strategies to follow x_ℓ and look for P's strategies (and corresponding reputations) that leave P indifferent across all messages in M' and across all action strategies corresponding to $x \in \{x_\ell\}_{\ell \in L_m}$ after sending message $m \in M'$.

Given Lemma 1, we know that in any equilibrium inducing interim beliefs $\{q_m\}_{m \in M'}$, P will be indifferent across all actions used by the ℓ types. We now define a function z that

will determine the strategy used by P.

Take an arbitrary $\tilde{q} \in [0,1]$ and CDF $\tilde{G}$ over L. If $\ell \sim \tilde{G}$ and $\mathbb{P}(\theta = L) = \tilde{q}$, the reputation for $a = 1$ when all $\ell \leq e + c$ choose $a = 1$ and P chooses $a = 1$ with probability z is $\frac{\tilde{q}\tilde{G}(e+c)}{\tilde{q}\tilde{G}(e+c)+(1-\tilde{q})z}$ while the reputation for $a = 0$ is $\frac{\tilde{q}(1-\tilde{G}(e+c))}{\tilde{q}(1-\tilde{G}(e+c))+(1-\tilde{q})(1-z)}$. 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

$$\rho\frac{\tilde{q}\tilde{G}(e+c)}{\tilde{q}\tilde{G}(e+c)+(1-\tilde{q})z} - c = \rho\frac{\tilde{q}(1-\tilde{G}(e+c))}{\tilde{q}(1-\tilde{G}(e+c))+(1-\tilde{q})(1-z)}, \tag{6}$$

if such a z exists. Otherwise, the left-hand side above must be lower than the right-hand side for all z, in which case we take $z(e;\tilde{q},\tilde{G}) = 0$. For e such that $\tilde{G}(e+c) = 0$, we set $z(e;\tilde{q},\tilde{G}) = 0$ and for e such that $\tilde{G}(e+c) = 1$, we set $z(e;\tilde{q},\tilde{G}) = 1$. Adopting the convention that $\frac{0}{0} = 0$, we define $\tilde{R}_1(e;\tilde{q},\tilde{G}) = \frac{\tilde{q}\tilde{G}(e+c)}{\tilde{q}\tilde{G}(e+c)+(1-\tilde{q})z(e;\tilde{q},\tilde{G})}$ and $\tilde{R}_0(e;\tilde{q},\tilde{G}) = \frac{\tilde{q}(1-\tilde{G}(e+c))}{\tilde{q}(1-\tilde{G}(e+c))+(1-\tilde{q})(1-z(e;\tilde{q},\tilde{G}))}$.

We note some properties of z. First, it is easy to see that $z(e;\tilde{q},\tilde{G})$ is continuous in q and increasing in e. Second, Bayes plausibility and (6) imply

$$\tilde{R}_1(e;\tilde{q},\tilde{G}) \geq \tilde{q} \geq \tilde{R}_0(e;\tilde{q},\tilde{G}).$$

Finally, for any $\tilde{q}$ such that $\rho(1-\tilde{q}) > c$, $\tilde{G}(e+c) \in (0,1)$ implies $z(e;\tilde{q},\tilde{G}) \in (0,1)$.[37]

In order to find a messaging strategy that leaves P indifferent across $m \in M'$, we first need to calculate P's expected equilibrium utility from announcing a particular message given the interim beliefs and P's action strategy as described above. For an arbitrary $\tilde{q} \in [0,1]$ and CDF $\tilde{G}$ on L, we define

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

This corresponds to the payoffs given ℓ's strategies and P choosing $a = 1$ after e with probability $z(e;\tilde{q},\tilde{G})$. We then define the expected payoff from w as

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

[37] This fact follows from the observation that at $z = 1$, the right-hand side of (6) is ρ, which is strictly higher than the left-hand side, while for $z = 0$ the left-hand side is $\rho - c$, which is strictly higher $\rho\tilde{q}$, which is an upper-bound for the left-hand side at $z = 0$.

Our next lemma gives some properties of W.

Lemma 8. *$W(\tilde{q};\tilde{G})$ is continuous and strictly increasing in $\tilde{q}$ with $W(\tilde{q};\tilde{G}) \in [\rho\tilde{q}-c, \rho\tilde{q}]$.*

Proof. Continuity is easily seen from the fact that $z(e;\tilde{q},\tilde{G})$ is continuous in $\tilde{q}$. That W is increasing follows from the fact that w is strictly increasing in $\tilde{q}$ for all e.[38] The fact that $W(\tilde{q},\tilde{G}) \in [\rho\tilde{q}-c, \rho\tilde{q}]$ easily follows if we can show $\rho\tilde{R}_0(e;\tilde{q},\tilde{G}) \in [\rho\tilde{q}-c, \rho\tilde{q}]$. The upper bound follows from $\tilde{R}_0(e;\tilde{q},\tilde{G}) \leq \tilde{q}$. To establish the lower bound, using $\tilde{R}_1(e;\tilde{q},\tilde{G}) \geq \tilde{q}$ we have[39]

$$\rho\tilde{R}_0(e;\tilde{q},\tilde{G}) \geq \rho\tilde{R}_1(e;\tilde{q},\tilde{G}) - c \geq \rho\tilde{q} - c.$$

Q.E.D.

In any equilibrium, in order to have $R(m,a,e) \in (0,1)$ for all e, $m \in M^{\mathcal{E}}$ and on-path a, $\sigma(\cdot|P)$ and Σ must be absolutely continuous. We therefore construct P's messaging strategy by specifying a Radon-Nikodym derivative $r(\cdot)$ and defining $\sigma(\cdot|P)$ via $\sigma(\hat{M}|P) = \int_{\hat{M}} r(m)d\Sigma(m)$ for any Borel measurable $\hat{M} \subset M$. When such strategies are used, the interim belief q_m for m with $r(m) = r$ is given by $\varphi(r) \equiv \frac{q}{q+(1-q)r}$.

For $U \in [\rho q - c, \rho q]$, define $s(U;m)$ to be the value of r such that $U = W(\varphi(r), G_m)$. The function $s(U;m)$ gives the rate r at which P must declare m (inducing an interim belief $q_m = \varphi(r)$) such that P's expected utility from announcing m in $W(q_m;G_m)$ is equal to U. We note that such an r exists as W is continuous in q. Using Assumption 1 and the bounds on W from Lemma 8, we have

$$\begin{aligned} &\lim_{r\to 0} W(\varphi(r), G_m) \geq \lim_{r\to 0} \rho\varphi(r) - c = \rho - c > \rho q, \\ &\lim_{r\to \infty} W(\varphi(r), G_m) \leq \lim_{r\to \infty} \rho\varphi(r) = 0 < \rho q - c. \end{aligned} \tag{7}$$

Because $W(q_m, G_m)$ is continuous and strictly increasing in q_m, $s(U;m)$ is continuous and strictly decreasing in U. By (7), $s(U;m)$ exists for all $U \in [\rho q - c, \rho q]$.

We next identify a unique U^* such that $1 = \int_M s(U^*;m)d\Sigma(m)$. U^* correspond to P's equilibrium expected utility and we will use $s(U^*;m)$ as the Radon-Nikodym derivative defining P's messaging strategy. Uniqueness of the solution implies that $s(U^*;m)$ defines the unique messaging strategy leaving P indifferent across $m \in M'$ when payoffs from m are $W(q_m, G_m)$.

[38] This is obvious when $\tilde{G}(e+c) \in \{0,1\}$. When $\tilde{G}(e+c) \in (0,1)$, it is straightforward from (6) that increasing $\tilde{q}$ must strictly increase $\tilde{R}_0(e;\tilde{q},\tilde{G})$.

[39] The first inequality below holds with equality if $z(e;\tilde{q},\tilde{G}) \in (0,1)$.

Lemma 9. *There exists a unique $U^* \in [\rho q - c, \rho q]$ such that $1 = \int_M s(U^*; m) d\Sigma(m)$. Moreover, for $q_m = \varphi(s(U^*; m))$, we have $\rho > \max\{\frac{c}{q_m}, \frac{c}{1-q_m}\}$.*

Proof. Take any $m \in M'$. We note that $\varphi(r) \lesseqgtr q$ if and only if $1 \lesseqgtr r$. Let $U = \rho q - c$. Because $W(\tilde{q}; G_m) \geq \rho \tilde{q} - c$ for all $\tilde{q}$, we have

$$\rho q - c = U = W(\varphi(s(U; m)), G_m) \geq \rho \varphi(s(U; m)) - c.$$

Thus, $q \geq \varphi(s(U; m))$, which implies $s(U; m) \geq 1$ and $\int_M s(U; m) d\Sigma(m) \geq \int_M d\Sigma(m) = 1$.

Let $U' = \rho q$. Because, $W(\tilde{q}; G_m) \leq \rho \tilde{q}$, for all $\tilde{q}$ we have

$$\rho q = U' = W(\varphi(s(U'; m)), G_m) \leq \rho \varphi(s(U'; m)) - c.$$

Thus, $q \leq \varphi(s(U'; m))$, which implies $s(U'; m) \leq 1$ and $\int_M s(U'; m) d\Sigma(m) \leq \int_M d\Sigma(m) = 1$. Because $s(\cdot; m)$ is continuous and strictly decreasing, there exists a unique $U^* \in [\rho q - c, \rho q]$ such that $1 = \int_M s(U^*; m) d\Sigma(m)$.

We conclude by showing that, for all $m \in M'$, $\rho > \max\{\frac{c}{q_m}, \frac{c}{1-q_m}\}$ where $q_m = \varphi(s^*(U; m))$. By definition of $s(U^*; m)$, $W(q_m, G_m) = W(q_{m'}, G_{m'})$ for all $m, m' \in M'$. Because $W(q_m, G_m) \in [\rho q_m - c, \rho q_m]$, we can apply the same arguments as in Lemma 5 to conclude $\rho > \max\{\frac{c}{q_m}, \frac{c}{1-q_m}\}$ for all $m \in M'$. *Q.E.D.*

We now construct an equilibrium $\mathcal{E}$ corresponding to the LIS. Let $\sigma^{\mathcal{E}}(\cdot|\ell) = \sigma(\cdot|\ell)$ and define $\sigma^{\mathcal{E}}(\cdot|P)$ by $d\sigma^{\mathcal{E}}(\cdot|P) = s(U^*; m) d\Sigma(m)$. Let $q_m = \varphi(s(U^*; m))$. We define reputations as $R^{\mathcal{E}}(m, a, e) = \tilde{R}_a(e; q_m, G_m)$ for $m \in M^{\mathcal{E}}$ and 0 for $m \notin M^{\mathcal{E}}$. For action strategies, we set[40]

$$\zeta^{\mathcal{E}}(1|m, e, \ell) = \begin{cases} x_\ell(e) & \text{if } m \in \text{Supp}(\sigma(\cdot|\ell)), \\ \mathbb{1}(1 \in \arg\max_a \ u(\ell, e, a, R^{\mathcal{E}}(m, a, e))) & \text{else.} \end{cases}$$

$$\zeta^{\mathcal{E}}(1|m, e, P) = \begin{cases} z(e; q_m, G_m) & \text{if } m \in M', \\ 0 & \text{else.} \end{cases}$$

It is clear that these strategies generate an expected utility for P of $W(q_m, G_m) = U^*$ when declaring message $m \in M'$ and that, after sending message m, P is indifferent across following any x_ℓ for $\ell \in L_m$. Our next lemma completes the proof of Lemma 2 by showing that $\mathcal{E}$ is an equilibrium; uniqueness of the equilibrium follows immediately from the fact

[40] The strategies for ℓ after sending an out of equilibrium message $m \notin \text{Supp}(\sigma(\cdot|\ell))$ are a selection from the set of optimal strategies for ℓ in such subgames. Any selection here will suffice.

that by Lemma 1, action strategies are pinned down on a probability one set of e, ℓ, m and that $s(U^*; m)$ defines the unique messaging strategy that leaves P indifferent across messages.

Lemma 10. *$\mathcal{E}$ is an equilibrium.*

Proof. We start by verifying that P has no incentive to deviate. First, we consider the action stage. It is clear that P is indifferent over actions when $G_m(e+c) \in (0,1)$. P also clearly has no incentive to deviate to $a = 1$ when $G_m(e+c) = 0$. Finally, P has no incentive to deviate to $a = 0$ when $G_m(e+c) = 1$ as his payoff from $a = 1$ is $\rho q_m - c > 0$ and his payoff from $a = 0$ is zero. There is also no incentive to deviate at the communication stage: P is indifferent across all $m \in M'$ and strictly prefers the expected utility of U^* from any $m \in M'$ to the expected utility of 0 from sending $m \notin M'$ because $U^* \geq \rho q - c > 0$.

Next, we show that no ℓ type has an incentive to deviate at the action stage following $m \in M'$; that there is no incentive deviate after $m \notin M'$ follows immediately from the definition of $\zeta^{\mathcal{E}}$. Take an arbitrary ℓ, e. Set $a = x_\ell(e)$ and $a' \neq a$. By definition of $z(e; q_m, G_m)$ and $\zeta^{\mathcal{E}}(\cdot|m, e, P)$, $x_\ell(e) \in \text{Supp}(\zeta^{\mathcal{E}}(\cdot|m, e, P))$ so P's incentive constraint implies $-ca + \rho R^{\mathcal{E}}(m, a, e) \geq -ca' + \rho R^{\mathcal{E}}(m, a', e)$. If ℓ has a strict incentive to deviate to a', then $(e-\ell)a + \rho R^{\mathcal{E}}(m, a, e) < (e-\ell)a' + \rho R^{\mathcal{E}}(m, a', e)$. Subtracting P's incentive constraint and simplifying, we get $(e - \tilde{e}_\ell)(a - a') < 0$, a contradiction of the fact that $e - \tilde{e}_\ell < 0$ if and only if $a - a' < 0$ by $a = x_\ell(e)$.

Next, we consider ℓ's incentive to deviate at the communication stage. If ℓ has a profitable deviation to announce message m' and follow contingent plan x' then, for $m \in \text{Supp}(\sigma(\cdot|\ell))$, we have

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

Because P is indifferent across all $m'' \in M'$ and x_ℓ for $\ell \in L_{m''}$, the fact that P does not have a profitable deviation to sending message m' and following x' implies

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

Adding these inequalities together and simplifying, we get $\int_E (c+e-\ell)x'(e)dF(e) > \int_E (c+e-\ell)x_\ell(e)dF(e)$, contradicting the fact that $x_\ell \in \arg\max_x \int_E (c+e-\ell)x(e)dF(e)$. Therefore, ℓ must have no incentive to deviate at the communication stage.

Finally, we show that D1 is satisfied.[41] The only off-path actions (after $m \in M'$) occur when $G_m(e+c) \in \{0,1\}$. If $G_m(e+c) = 1$, then $a = 0$ is an off-path action. There are two cases to consider: when $e > \max_{\ell \in L_m} \tilde{e}_\ell$ and when $e = \max_{\ell \in L_m} \tilde{e}_\ell$. In the first case, by Lemma 7, D1 requires $R^{\mathcal{E}}(m,1,e) = 0$ because $e - \tilde{e}_\ell < 0$ for all $\ell \in L_m$. For the second case, any reputation 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

$$\{\mu \in [0,1] : \ \mu \geq \frac{e-\ell}{\rho} + q_m\} \subsetneq \{\mu \in [0,1] : \ \mu \geq -\frac{c}{\rho} + q_m\},$$

which rules out all $\ell < \max L_m$. However, the above sets are equal for $\ell' = \max L_m$, in which case any beliefs that ascribe probability only on ℓ' and P are consistent with D1. Thus, $R^{\mathcal{E}'}(m,1,e) = 0$ is consistent with D1. An analogous argument holds for when $G_m(e+c) = 0$. *Q.E.D.*

Proof of Lemma 4

Proof. We first derive an equation for determining $U_P^\alpha(F)$. P's expected material payoff from following strategy x_ℓ is $-c(1 - F(\tilde{e}_\ell))$. Conditional on following strategy x_ℓ after message m_ℓ, both P and ℓ pool in their action strategy after each evidence realization, so $R^\alpha(m_\ell, x_\ell(e), e) = \frac{qdG(\ell)}{qdG(\ell)+(1-q)d\sigma^\alpha(m_\ell|P)}$ for all e. Because P is indifferent across all $m \in \{m_\ell\}_{\ell \in L}$, we have

$$U_P^\alpha(F) = -c(1 - F(\tilde{e}_\ell)) + \rho \frac{qdG(\ell)}{qdG(\ell) + (1-q)d\sigma^\alpha(m_\ell|P)} \ \forall \ell \in L.$$

Solving for the probability of message m_ℓ, namely $qdG(\ell)+(1-q)d\sigma^\alpha(m_\ell|P)$ and integrating over L, we have

$$1 = \int_L \frac{\rho q dG(\ell)}{U_P^\alpha(F) + c - cF(\tilde{e}_\ell)}. \tag{8}$$

[41] We restrict attention to verifying this for $m \in M'$. For $m \notin M'$, because $\Theta_m = \{P\}$, $R(m,a,e) = 0$ satisfies D1.

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 (8), we then have

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

where the inequality follows from the fact that $\frac{1}{y}$ is convex in y. Moreover, the inequality is strict if and only 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 $\ell \in L'$. This inequality implies $U_P^\alpha(F_\lambda) \leq \lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2)$, with strict inequality if and only if such an L' exists. *Q.E.D.*

As discussed in the text, Lemma 4 implies $U_P^\alpha(F) \leq U_P^\beta(F)$. When $|L| \geq 2$, ex-post signaling under G has some residual strategic uncertainty if and only if there exists $L' \subseteq L$ such that $\int_{L'} dG(\ell) > 0$ and $F(\tilde{e}_\ell) \neq F(\tilde{e}_{\ell'})$ for all $\ell, \ell' \in L'$. Under mild agreement, for each $\ell', \ell'' \in L'$, there exists $e \in \text{Supp}(F)$ such that $\delta_e(\tilde{e}_{\ell'}) = \delta_e(\tilde{e}_{\ell''})$; residual strategic uncertainty then implies that there exists $e' \in \text{Supp}(F)$ such that $\delta_{e'}(\tilde{e}_{\ell'}) \neq \delta_{e'}(\tilde{e}_{\ell''})$. This implies that $\delta_e(\tilde{e}_\ell) \neq \delta_{e'}(\tilde{e}_\ell)$ for some $\ell \in L'$. The proof of Lemma 4 shows that $U_P^\alpha(F) \geq \int_E U_P^\alpha(\delta_e)dF(e) = \int_E U_P^\beta(\delta_e)dF(e)$; under mild agreement, the inequality is strict if and only if there is residual strategic uncertainty. We use this observation in the proof of Theorem 1 below.

Proof of Theorem 1

Proof. Take any equilibrium $\mathcal{E}$ and define $\Sigma^\mathcal{E}(\cdot) = \int_L \sigma^\mathcal{E}(\cdot|\ell)dG(\ell)$. By Lemma 3, it suffices to show $U_P^\mathcal{E}(F) \geq U_P^\alpha(F)$, with, under mild agreement, a strict inequality if and only if there is residual strategic uncertainty.

For $m \in M^\mathcal{E}$, let G_m and q_m be the interim beliefs associated with m in $\mathcal{E}$ and define $U_{P,m}^\beta(\delta_e)$ to be the ex-post signaling utility when $\ell \sim G_m$, $\mathbb{P}(\theta \in L) = q_m$ and $F = \delta_e$. Note that P's utility in $\mathcal{E}$ after sending message $m \in M^\mathcal{E}$ when e is realized is equal to $U_{P,m}^\beta(\delta_e)$. P's expected utility from sending message $m \in M^\mathcal{E}$ is then $\int_E U_{P,m}^\beta(\delta_e)dF(e)$, which, by basic equilibrium considerations, must be equal to $U_P^\mathcal{E}(F)$.

Define $U_{P,m}^\alpha(F)$ to be the (unique) value of U that solves $\int_L \frac{\rho q_m}{U+c-cF(\tilde{e}_\ell)}dG_m(\ell) = 1$; this corresponds to the solution to the equation for $U_P^\alpha(F)$ when $\ell \sim G_m$ and $\mathbb{P}(\theta \in L) =$

q_m. That a unique solution exists follows from the fact that $\rho q_m > c$ by Lemma 5, so $\int_L \frac{\rho q_m}{U+c-cF(\tilde{e}_\ell)} dG_m(\ell) > 1$ when $U = 0$ and that $\int_L \frac{\rho q_m}{U+c-cF(\tilde{e}_\ell)} dG_m(\ell)$ is strictly decreasing in U with a limit of 0 as $U \to \infty$.

Note that $U^\beta_{P,m}(\delta_e) = U^\alpha_{P,m}(\delta_e)$. By the arguments made in Lemma 4, $U^\alpha_{P,m}(\cdot)$ is convex[42] and so, for all $m \in M^\mathcal{E}$, we have

$$U^\alpha_{P,m}(F) \leq \int_E U^\alpha_{P,m}(\delta_e) dF(e) = \int_E U^\beta_{P,m}(\delta_e) dF(e) = U^\mathcal{E}_P(F),$$

with, under mild agreement, strict inequality on a $\Sigma^\mathcal{E}$-probability one set of m if and only if there is residual strategic uncertainty in equilibrium $\mathcal{E}$.

We now relate $U^\alpha_{P,m}(F)$ to $U^\alpha_P(F)$. Because $\int_L \frac{\rho q_m}{U^\alpha_{P,m}(\delta_e)+c-c\mathbb{1}(e\geq\tilde{e}_\ell)} dG_m(\ell) = 1$, we can take the expectation over both sides with respect to e and m to get

$$\begin{aligned} 1 &= \int_M \left(\int_E \left[\int_L \frac{\rho q_m dG_m(\ell)}{U^\alpha_{P,m}(\delta_e) + c - c\mathbb{1}(e \geq \tilde{e}_\ell)} \right] dF(e) \right) (qd\Sigma^\mathcal{E}(m) + (1-q) d\sigma^\mathcal{E}(m|P)) \qquad (9) \\ &\geq \int_M \left[\int_L \frac{\rho q_m dG_m(\ell)}{U^\alpha_{P,m}(F) + c - cF(\tilde{e}_\ell)} \right] (qd\Sigma^\mathcal{E}(m) + (1-q) d\sigma^\mathcal{E}(m|P)) \\ &= \int_M \left[\int_L \frac{\rho q dG(\ell)}{U^\alpha_{P,m}(F) + c - cF(\tilde{e}_\ell)} \right] d\sigma^\mathcal{E}(m|\ell), \end{aligned}$$

where the inequality follows from the convexity of $U^\alpha_{P,m}(F)$ and is strict (under mild agreement) if and only if there is residual strategic uncertainty.

Because $\int_L \frac{\rho q dG(\ell)}{U^\alpha_P(F)+c-cF(\tilde{e}_\ell)} = 1$ as shown in Lemma 4, (9) implies $U^\alpha_P(F) \leq U^\alpha_{P,m}(F)$ on a $\Sigma^\mathcal{E}$-probability one set of m, strictly so (under mild agreement) if and only if there is some residual strategic uncertainty (i.e., the inequality in (9) is strict). Our desired result then follows from the fact that $U^\alpha_{P,m}(F) \leq U^\mathcal{E}_P(F)$ on any probability one set of m, with, under mild agreement, strict inequality if and only if there is residual strategic uncertainty.

Q.E.D.

Proof of Proposition 1

Proof. Take any evidence level e. The proof is immediate if $G(e+c) = 0$ as $\psi^\alpha(e) = \psi^\beta(e) = 0$ by Lemma 1 or if $G(e+c) = 1$ as $\psi^\alpha(e) = \psi^\beta(e) = 1$. Suppose $G(e+c) \in (0,1)$. Note that $\psi^\alpha(e) = \int_{-\infty}^{e+c} (qdG(\ell) + (1-q)d\sigma^\alpha(m_\ell|P))$. As can easily be seen from the proof of Lemma 4, $qdG(\ell) + (1-q)d\sigma^\alpha(m_\ell|P) = \frac{\rho q dG(\ell)}{U^\alpha_P(F)+c-cF(\tilde{e}_\ell)}$, so $\psi^\alpha(e) = \int_{-\infty}^{e+c} \frac{q\rho}{U^\alpha_P(F)+c-cF(\tilde{e}_\ell)} dG(\ell)$. We also

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

note that $\psi^\beta(e)$ is pinned down by

$$\rho\frac{qG(e+c)}{\psi^\beta(e)}-c=\rho R^\beta(m,1,e)-c=\rho R^\beta(m,0,e)=\rho\frac{q(1-G(e+c))}{1-\psi^\beta(e)}. \tag{10}$$

This probability does not depend on F and only depends on G through the value of $G(e+c)$.

We show that $\psi^\alpha(e)-\psi^\beta(e)\geq 0$ by showing that this inequality holds when we select the distribution of ℓ types $G\in\mathcal{G}_e\equiv\{\widetilde{G}:\ \widetilde{G}(e+c)=G(e+c)\}$ in order to minimize $\psi^\alpha(e)$.

It is without loss to focus on F such that $\mathrm{Supp}(F)$ lies in a compact interval. 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 easy to see that the value of $\psi^\alpha(e)$ under the truncated F will converge to the value of $\psi^\alpha(e)$ under the original F.

We construct a minimizing G in two steps. First, we show it is without loss to focus on G with no mass on $(e+c,\max\mathrm{Supp}(F)+c]$ and at most one point in its support above $\max\mathrm{Supp}(F)+c$. Take any $\widetilde{G}\in\mathcal{G}_e$. For some $e'>\max\mathrm{Supp}(F)+c$, let $\widehat{G}$ be such that $\widehat{G}(\ell)=\widetilde{G}(\ell)$ for $\ell\leq e+c$ and $\widehat{G}(\ell)=\widetilde{G}(e+c)+\mathbb{1}(e\geq e')(1-\widetilde{G}(e+c))$ for $\ell\geq e+c$; thus, $\widehat{G}\in\mathcal{G}_e$. Let $\widetilde{U}$ and $\widehat{U}$ be the corresponding equilibrium expected utility under ex-ante signaling for P when ℓ is distributed according to $\widetilde{G}$ and $\widehat{G}$ respectively. For the sake of contradiction, suppose $\widehat{U}<\widetilde{U}$. Then

$$\begin{aligned}
1&=\int_L\frac{\rho q}{\widetilde{U}+c-cF(\tilde{e}_\ell)}d\widetilde{G}(\ell)\\
&\leq\int_{-\infty}^{e+c}\frac{\rho q}{\widetilde{U}+c-cF(\tilde{e}_\ell)}d\widetilde{G}(\ell)+\int_{e+c}^{\infty}\frac{\rho q}{\widetilde{U}}d\widetilde{G}(\ell)\\
&=\int_{-\infty}^{e+c}\frac{\rho q}{\widetilde{U}+c-cF(\tilde{e}_\ell)}d\widehat{G}(\ell)+\int_{e+c}^{\infty}\frac{\rho q}{\widetilde{U}}d\widehat{G}(\ell)\\
&<\int_{-\infty}^{e+c}\frac{\rho q}{\widehat{U}+c-cF(\tilde{e}_\ell)}d\widehat{G}(\ell)+\int_{e+c}^{\infty}\frac{\rho q}{\widehat{U}}d\widehat{G}(\ell)\\
&=\int_L\frac{\rho q}{\widehat{U}+c-cF(\tilde{e}_\ell)}d\widehat{G}(\ell),
\end{aligned}$$

a contradiction of $\int_L\frac{\rho q}{\widehat{U}+c-cF(\tilde{e}_\ell)}d\widehat{G}(\ell)=1$. Therefore, $\widehat{U}\geq\widetilde{U}$. Because $\psi^\alpha(e)$ is decreasing in $U_P^\alpha(F)$, we conclude that $\widehat{G}$ must yield weakly lower probability $a=1$ at e than $\widetilde{G}$.

Next, we argue that that is without loss in our minimization problem to focus on G with binary supports. Take any $\widetilde{G}\in\mathcal{G}_e$ with no mass on $(e+c,\max\mathrm{Supp}(F)+c]$ and at most one point in its support above $\max\mathrm{Supp}(F)+c$. Let $\widehat{G}$ be a distribution in $\mathcal{G}_e$ with a mass point of size $G(e+c)$ strictly below $\min\mathrm{Supp}(F)+c$ and a mass point of size $1-G(e+c)$ at some

point strictly above $\max \text{Supp}(F) + c$. Let $\widetilde{U}$ and $\widehat{U}$ be the equilibrium expected utility for P under ex-ante signaling when the distribution of ℓ types is $\widetilde{G}$ and $\widehat{G}$ respectively.

For the sake of contradiction, suppose $\int_{-\infty}^{e+c} \frac{q\rho}{\widetilde{U}+c-cF(\ell-c)} d\widetilde{G}(\ell) < \int_{-\infty}^{e+c} \frac{q\rho}{\widehat{U}+c-cF(\ell-c)} d\widehat{G}(\ell)$. $\widetilde{U}$ is given by the solution to $\int_L \frac{q\rho}{U+c(1-F(\tilde{e}_\ell))} d\widetilde{G}(\ell) = 1$. Because $\int_{e+c}^{\infty} \frac{q\rho}{\widetilde{U}+c(1-F(\tilde{e}_\ell))} d\widetilde{G}(\ell) = \frac{\rho q(1-G(e+c))}{\widetilde{U}}$, we have

$$\int_{-\infty}^{e+c} \frac{q\rho}{\widetilde{U} + c(1 - F(\tilde{e}_\ell))} d\widetilde{G}(\ell) = 1 - \frac{\rho q(1 - G(e + c))}{\widetilde{U}}.$$

A similar expression holds for $\widehat{U}$. Therefore, $\int_{-\infty}^{e+c} \frac{q\rho}{\widetilde{U}+c-cF(\ell-c)} d\widetilde{G}(\ell) > \int_{-\infty}^{e+c} \frac{q\rho}{\widehat{U}+c-cF(\ell-c)} d\widehat{G}(\ell)$ implies $\frac{\rho q(1-G(e+c))}{\widetilde{U}} > \frac{\rho q(1-G(e+c))}{\widehat{U}}$ or equivalently, $\widetilde{U} < \widehat{U}$. We then have

$$\begin{aligned} 1 = \int_L \frac{q\rho}{\widetilde{U} + c - cF(\tilde{e}_\ell)} d\widetilde{G}(\ell) &\geq \frac{\rho q}{\widetilde{U} + c} G(e + c) + \frac{\rho q(1 - G(e + c))}{\widetilde{U}} \\ &> \frac{\rho q}{\widehat{U} + c} G(e + c) + \frac{\rho q(1 - G(e + c))}{\widehat{U}} = \int_L \frac{q\rho}{\widehat{U} + c - cF(\tilde{e}_\ell)} d\widehat{G}(\ell), \end{aligned}$$

a contradiction of $\int_L \frac{q\rho}{\widehat{U}+c-cF(\tilde{e}_\ell)} d\widehat{G}(\ell) = 1$. Therefore, we must have $\int_{-\infty}^{e+c} \frac{q\rho}{\widetilde{U}+c-cF(\ell-c)} d\widetilde{G}(\ell) \geq \int_{-\infty}^{e+c} \frac{q\rho}{\widehat{U}+c-cF(\ell-c)} d\widehat{G}(\ell)$, that is, $\psi^\alpha(e)$ is weakly lower under $\widehat{G}$ than $\widetilde{G}$.

The above arguments imply that it is without loss to consider $G \in \mathcal{G}_e$ with binary support for our minimization problem. For binary support $\{\underline{\ell}, \overline{\ell}\}$, $\psi^\alpha(e) - \psi^\beta(e)$ 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 $\psi^\alpha(e) - \psi^\beta(e) \geq 0 \; \forall e \in E$, which completes the proof. *Q.E.D.*

B. Optimal Investigation Proofs

Proof of Theorem 2

Proof. We first note that

$$\int_L \frac{g(\ell)q\rho}{U + c - cF(\tilde{e}_\ell)} d\ell = \frac{G(c)q\rho}{U + c} + \int_0^1 \frac{h(e)q\rho}{U + c - cF(e)} de + \frac{(1 - G(1 + c))q\rho}{U}.$$

We solve a relaxed version of the investigator's problem where we only require the

constraints to hold as inequalities:

$$\begin{aligned} &\min_{U\geq 0, F\in\mathcal{F}} U \\ &\text{subject to } \frac{G(c)q\rho}{U+c}+\int_0^1 \frac{h(e)q\rho}{U+c-cF(e)}de+\frac{(1-G(1+c))q\rho}{U}\leq 1, \\ &\int_0^1 (1-F(e))de\leq \bar{e}. \end{aligned} \tag{11}$$

Both constraints are convex in U and F. By Theorem 1 (Chapter 8) of Luenberger (1997), there exist multipliers $\eta \geq 0$ and $\lambda \geq 0$ (on the first and second constraints respectively) such that any solution U^*, F^* to (11) will solve[43]

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

Moreover, complementary slackness conditions imply $\eta, \lambda > 0$ only if both constraints bind, in which case 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

$$\int_L \frac{g(\ell)q\rho}{U^*+c-cF^*(\tilde{e}_\ell)}d\ell = \int_L \frac{g(\ell)q\rho}{c-cF^*(\tilde{e}_\ell)}d\ell \geq \int_L \frac{g(\ell)q\rho}{c}d\ell = \frac{q\rho}{c} > 0,$$

where the final inequality holds by Assumption 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

$$\min_{F\in\mathcal{F}} \int_0^1 \left(\frac{\eta h(e)q\rho}{U^*+c-cF(e)}-\lambda F(e)\right)de-\eta+\lambda-\lambda\bar{e}. \tag{12}$$

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 (12). By Theorem 3.1 of Toikka (2011), $F^*(e) = \arg\min_{x\in[0,1]} \frac{\eta\bar{h}(e)q\rho}{U^*+c-cx} - \lambda x$. Taking the first-order condition, whenever $F^*(e) \in (0,1)$, we have

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

[43] 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 $U > q\rho$.

Letting $k = \sqrt{\frac{\eta q \rho}{c\lambda}}$, a bit of algebra gives us $F^*(e) = \frac{U^*}{c} + 1 - k\sqrt{\overline{h}(e)}$ whenever $F^*(e) \in (0,1)$. $F^*(e) = 0$ whenever $\frac{\eta \overline{h}(e) q\rho}{c(\frac{U^*}{c}+1)^2} - \lambda > 0$; this condition simplifies to $\frac{U^*}{c} < k\sqrt{\overline{h}(e)} - 1$. Similarly, $F^*(e) = 1$ whenever $\frac{\eta \overline{h}(e) q\rho}{c(\frac{U^*}{c})^2} - \lambda < 0$, or alternatively, when $\frac{U^*}{c} > k\sqrt{\overline{h}(e)}$. That $U^* = U_P^\alpha(F^*)$ follows from the fact that the first constraint in (11) holds with equality. *Q.E.D.*

C. Comparative Statics Proofs

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 $\int_L \frac{\rho q}{U + c - cF(\tilde{e}_\ell)} dG(\ell) = 1$.

Proof of Proposition 2

Proof. Fix an investigation F and distribution G of ℓ. Taking the derivative of the expression in (8) with respect to ρ, we have

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

After some simplification and using Jensen's inequality, we get[44]

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

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

Next, we look at first-order stochastic dominance shifts of the distribution of ℓ. 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 G, we have

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

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

[44] Here we are using $\int_L \frac{q}{U_P^\alpha(F) + c - cF(\tilde{e}_\ell)} dG(\ell) = \frac{1}{\rho}$ by (8).

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

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

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, $\int_0^1 (1 - F(e))de = \overline{e} = \int_0^1 (1 - \tilde{F}(e))de$) 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 11. *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 $\underline{e}^* = \min \text{ Supp}(F^*)$ and $\overline{e}^* = \max \text{ Supp}(F^*)$. 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

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

where the inequality follows because $\frac{1}{\sqrt{\tilde{h}(e)}}$ is a convex function by Lemma 11 and $\tilde{g}$ is a pivotal mean-preserving contraction of g.[45]

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

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

which is a contradiction of $\int_L \frac{g(\ell)q\rho}{U_P^\alpha(F^*;g)+c-cF^*(\tilde{e}_\ell)} d\ell = 1$. 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.*

D. Extension Proofs

Before stating the proof of Proposition 7, we first formally define an equilibrium in the commitment model. We let $R(x)$ denote the reputation from choosing a particular commitment $x \in \mathcal{X}$ where we endow $\mathcal{X}$ with the metric $d(x,x') = \int_E |x(e) - x'(e)| dF(e)$.[46] Abusing notation somewhat, an equilibrium is given by a strategy $\xi : \Theta \to \Delta(\mathcal{X})$ and $R : \mathcal{X} \to [0,1]$ such that

1. R is obtained from ξ using Bayes rule whenever possible,
2. $\xi(\cdot|\theta)$ is supported on $\arg\max_{x\in\mathcal{X}} \int_E u(\theta, e, x(e), R(x)) dF(e)$.

Proof of Proposition 5

Proof. We split the proof into several steps.

[45] 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.

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

Step 1 (Equilibrium Verification): We first establish that the strategies described constitute an equilibrium and deliver an equilibrium expected utility of $U_P^\alpha(F)$ to P. Let each ℓ type choose x_ℓ with probability one and define P's equilibrium mixing strategy $\xi(\cdot|P) \in \Delta(\{x_\ell\})$ by $d\xi(x_\ell|P) = dG(\ell)\frac{q}{1-q}[\frac{\rho}{U_P^\alpha(F)+c-cF(\tilde{e}_\ell)} - 1]$. Define the equilibrium reputations by $R(x) = \frac{U_P^\alpha(F)+c-cF(\tilde{e}_\ell))}{\rho}$ for $x \in \{x_\ell\}_{\ell\in L}$, and $R(x) = 0$ otherwise. Note that $R(x_\ell) = \frac{qdG(\ell)}{qdG(\ell)+(1-q)d\xi(x_\ell|P)}$ accords with Bayes rule given the definition of $\xi(\cdot|P)$.

We note that this equilibrium generates the same outcomes as in ex-ante signaling.[47] 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 on-path equilibrium utility $U_P^\alpha(F)$.[48] By the arguments in Lemma 10, no ℓ type has an incentive to deviate.

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

$$\{\mu \in [0,1] : \mu > \frac{-c\int_E (x_\ell(e) - x(e))dF(e)}{\rho} + R(x_\ell)\}$$
$$\subsetneq \{\mu \in [0,1] : \mu > \frac{\int_E (e-\ell)(x_\ell(e) - x(e))dF(e)}{\rho} + R(x_\ell)\}.$$

After some simplification, this is equivalent to

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

which contradicts $x_\ell \in \arg\max_{x\in\mathcal{X}} \int_E (c+e-\ell)x(e)dF(e)$. Therefore, $R(x) = 0$ is consistent with D1.

Step Two (Outcome Equivalence): We show that all other equilibria are outcome equivalent in two steps. First, we show that in any equilibrium ℓ types must only choose from $\mathcal{X}_\ell \equiv \arg\max_x \int_E (c+e-\ell)x(e)dF(e)$. Second, we show that, given the first conclusion, $\xi(\mathcal{X}_\ell|P)$ must take the form specified in Step One.

In order to show uniqueness we first establish that, across all equilibria, the P type's strategy puts positive probability on every commitment made by some ℓ type.

Claim 1. *$R(x) < 1$ for all $x \in \mathcal{X}$.*

Proof of Claim: For the sake of contradiction, suppose there exists $x \in \mathcal{X}$ such that $R(x) =$

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

[48] As shown in Lemma 2, $U_P^\alpha(F) \geq \rho q - c > 0$.

1. By Bayes' plausibility, there must exist $x' \in \text{Supp}(\xi(\cdot|P))$ such that $R(x') \leq q$. For x' to be an equilibrium strategy for P, we must have

$$-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho. \tag{13}$$

Because $-c\int_E x'(e)dF(e) \leq 0$, $-c\int_E x(e)dF(e) \geq -c$ and $R(x') \leq q$, (13) implies $\rho q \geq -c+\rho$, a contradiction of Assumption 1.

Claim 1 implies that $\text{Supp}(\Xi) \subseteq \text{Supp}(\xi(\cdot|P))$ where $\Xi(\cdot) = \int_L \xi(\cdot|\ell)dG(\ell)$. Next, we use Claim 1 to show $\text{Supp}(\xi(\cdot|\ell)) \subseteq \mathcal{X}_\ell$. Suppose, for the sake of contradiction, that in some equilibrium, $\exists \ell \in L, x \notin \mathcal{X}_\ell$ such that $x \in \text{Supp}(\xi(\cdot|\ell))$. There are two cases to consider: $\text{Supp}(\Xi) \cap \mathcal{X}_\ell \neq \emptyset$ and $\text{Supp}(\Xi) \cap \mathcal{X}_\ell = \emptyset$. In the first case, take $x' \in \text{Supp}(\Xi) \cap \mathcal{X}_\ell$. Because $\text{Supp}(\Xi) \subseteq \text{Supp}(\xi(\cdot|P))$, we have $x, x' \in \text{Supp}(\xi(\cdot|P))$. This implies P is indifferent between x and x' and ℓ prefers x to x':

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

Subtracting these inequalities and simplifying, we get $\int_E (c+e-\ell)(x(e)-x'(e))dF(e) \geq 0$, a contradiction to $x' \in \mathcal{X}_\ell$ and $x \notin \mathcal{X}_\ell$.

Now consider the second case, when $\text{Supp}(\Xi) \cap \mathcal{X}_\ell = \emptyset$. Take any $x \in \text{Supp}(\xi(\cdot|\ell))$. D1 requires that $R(x_\ell) = 1$ if

$$\begin{aligned} \{\mu \in [0,1] :\ \mu > \frac{-c\int_E [x(e)-x_\ell(e)]dF(e)}{\rho} + R(x)\} \\ \subsetneq \{\mu \in [0,1] :\ \mu > \frac{\int_E (e-\ell)[x(e)-x_\ell(e)]dF(e)}{\rho} + R(x)\}. \end{aligned} \tag{14}$$

The RHS set of (14) is non-empty,[49] so, 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, $R(x_\ell) = 1$ in any D1 equilibrium. But this contradicts Claim 1. Therefore, we conclude that $\text{Supp}(\xi(\cdot|\ell)) \subseteq \mathcal{X}_\ell$ 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 probabiliy zero set of (e, ℓ)).

[49] Take $x' \in \text{Supp}(\sigma(\cdot|P))$ such that $R(x') \leq q$ (such an x' exists by Bayes plausibility). This means P's equilibrium utility is less than ρq. But setting $\mu = 1$ is associated with a utility of at least $\rho - c$ which is strictly larger by Assumption 1. This means P would be willing to deviate for a reputation of 1—namely, $\mu = 1$ is an element of the RHS set.

Next, we argue that any equilibrium must have $\text{Supp}(\xi(\cdot|P)) \subseteq \text{Supp}(\Xi)$. This inclusion follows from the fact that any $x \in \text{Supp}(\xi(\cdot|P)) \setminus \text{Supp}(\Xi)$ must have $R(x) = 0$ and so yields P an expected utility of 0 for P. But, by Bayes plausibility, there exists $x' \in \text{Supp}(\Xi)$ such that $R(x') \geq q$, in which case P can achieve a utility of $-c \int_E x'(e)dF(e) + \rho R(x') \geq \rho q - c > 0$. Thus, choosing x is strictly dominated by x' and so cannot be an equilibrium strategy for P. This argument also implies that $R(x) > 0$ for all $x \in \text{Supp}(\xi(\cdot|P))$. 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 ν 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 reputation $R : \mathcal{X} \cup (M \times E \times A) \to [0, 1]$ such that

1. R is obtained from Bayes rule whenever possible.
2. For each m, θ, e, $\zeta(\cdot|m, e, \varepsilon, \theta)$ is supported on $\arg\max_{a \in \{0,1\}} \; u(\theta, e, a, R(m, a, e)) + \varepsilon a$.
3. For each θ, $\sigma(\cdot|\theta)$ is supported on

$$\arg\max_{m \in M} \int_E \int_{-\delta}^{\delta} \left(\max_{a \in \{0,1\}} u(\theta, e, a, R(m, a, e)) + \varepsilon a \right) d\nu(\varepsilon) dF(e) \cup \; \arg\max_{x \in \mathcal{X}} \; \int_E u(\theta, e, x(e), R(x)) dF(e),$$

where ε does not appear in the second maximization because it is mean zero. Notice that ζ only takes effect if a cheap talk message is sent.

Proof of Proposition 6

Proof. Suppose there exists an equilibrium where some cheap talk message $m \in \cup_{\theta \in \Theta} \text{Supp}(\sigma(\cdot|\theta))$. Note that $m \in \text{Supp}(\sigma(\cdot|\ell)) \cap \text{Supp}(\sigma(\cdot|P))$ for some ℓ. Otherwise, the reputation for m following any action at the decision stage would be always 1 or always 0. By Bayes plausibility, there exists an on-path m' or x with reputation less than q. If the reputation after m is always 1, 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. If the reputation after m is 0, it must be sent by only the P type; P attains a maximum utility of $\max\{-c + \delta, 0\}$ from doing so. However, the P type can attain at least $\rho q - c - \delta$ by mimicking some ℓ type whose expected equilibrium reputation is at least q (such ℓ exist by Bayes plausibility). Because $\rho q > 2\delta$ and $\rho q > 2c$ by Assumption 1, this is a contradiction.

Let type ℓ send message m in the candidate equilibrium. 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

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

and for P, it is given by

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

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 $R(x_\ell) = 1$ and P would deviate to x_ℓ. The above expressions 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, and the set of on path commitments chosen by P is contained in the set $\{x_\ell\}_{\ell \in L}$. 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_i is $U_P^\alpha(F_f(\cdot|e_i))$. By an analogous proof to that in Lemma 3, $\mathbb{P}(a = 1|e_i) = \frac{\rho q - U_P^\alpha(F_f(\cdot|e_i))}{c}$. Thus, $\mathbb{P}(a = 1) = \int_E \mathbb{P}(a = 1|e_i) dF_i(e_i) = \frac{1}{c}[\rho q - \int_E U_P^\alpha(F_f(\cdot|e_i)) dF_i(e_i)]$. The proposition then follows immediately from convexity of $U_P^\alpha(\cdot)$. *Q.E.D.*

Proof of Proposition 8

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

$$\begin{aligned}
&\int_E (e - \ell_I)\left[\int_L \mathbb{1}(\tilde{e}_\ell \leq e)(qdG(\ell) + (1 - q)d\sigma(m_\ell|P))\right] dF_\delta(e) \\
&= \int_E (e - \ell_I)\left[\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\eta(\ell; U_P^\alpha(F), \delta)d\ell\right] dF_\delta(e). \qquad (15)
\end{aligned}$$

For notational ease, we let $U = U_P^\alpha(F_\delta)$. Taking the derivative of (15) at $F = F_\delta$ with respect to δ, we have

$$\begin{aligned}
&\frac{dU}{d\delta}\int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\frac{\partial \eta(\ell; U, \delta)}{\partial U}d\ell dF_\delta(e) + \int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\frac{\partial \eta(\ell; U, \delta)}{\partial \delta}d\ell dF_\delta(e) \\
&\quad + \int_E (e - \ell_I)\int_L \mathbb{1}(\tilde{e}_\ell \leq e)\eta(\ell; U, \delta)d\ell \frac{d}{d\delta}dF_\delta(e). \qquad (16)
\end{aligned}$$

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

We show the first term in (16) is positive. Because $\frac{d\eta(\ell;U,\delta)}{dU} \leq 0$, it suffices to show $\frac{dU}{d\delta} \leq 0$. 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 $\frac{\partial \eta(\ell;U,\delta)}{\partial U} \leq 0$, $\frac{dU}{d\delta} \leq 0$ if and only if $\int_L \frac{\partial \eta(\ell;U,\delta)}{\partial \delta}d\ell \leq 0$. Given the form of F_δ, for

sufficiently small ε we have

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

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, $\frac{dU}{d\delta} \leq 0$.

Next, we show that the second term in (16) is positive. Let ΔF be the size of mass point on $\hat{e}$. Next, we note that for small ε

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

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

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

Because $\hat{e}-\ell_I > 0$, the final line is strictly positive for U large enough. Because $U \geq \rho q - c$ (as shown in the proof of Lemma 5), the last line above is strictly positive for sufficiency large ρ.

Finally, we show the final term in (16) is positive. For small enough ε, we have

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

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 (16) 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 $M(F,e)\equiv\frac{g(e+c)\rho q}{\left(U_P^\alpha(F)+c(1-F(e))\right)^2}$, 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.

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 $\int_L\frac{g(\ell)q\rho}{U+c-cF^*(\tilde{e}_\ell)}d\ell=1$. The optimal F^* must minimize $\int_L\frac{g(\ell)q\rho}{U+c-cF(\tilde{e}_\ell)}d\ell$ over feasible F; if not, then the investigator could choose an alternative F' such that $\int_L\frac{g(\ell)q\rho}{U+c-cF'(\tilde{e}_\ell)}d\ell<1$, 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

$$\min_{F\in\mathcal{F}} \quad \int_L \frac{g(\ell)q\rho}{U+c-cF(\tilde{e}_\ell)}d\ell\ , \tag{17}$$
$$\text{such that } BP: \int_0^e F(e')de' \leq \int_0^e K(e')de' \ \forall e\in E, \text{ and}$$
$$\int_0^1 F(e')de' = \int_0^1 K(e')de'.$$

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

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

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 (17) 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 $\underline{e} \leq e_1 < e_2 \leq \overline{e}$. Take $\varepsilon > 0$. If $e_1 = \underline{e}$ and $F^*(\underline{e}) = 0$ then, because M is right continuitous, 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 (17) 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 discontinutity in F^* implies $M(F^*, e)$ is not constant around e, a contradiction of Proposition 9. *Q.E.D.*

Proof of Corollary 5

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 $\frac{g(e'+c)}{(\rho q + c(1-K(e')))^2}$ is increasing and fact that $U_P^\alpha(F^*) \leq \rho q$[50] implies that $\frac{g(e'+c)}{\left(U_P^\alpha(F^*)+c(1-K(e'))\right)^2}$ is increasing in e' on $[e, e + \varepsilon)$ which implies in this region that

$$\frac{g(e' + c)}{\left(U_P^\alpha(F^*) + c(1 - K(e'))\right)^2} > M(F^*, e'). \tag{18}$$

From (18), 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.*

E. 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. 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

[50] This inequality is implied by Lemma 3, since $U_P^\alpha(F^*) > \rho q$ implies the probability of $a = 1$ is negative.

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 $\psi^\beta(e)$ 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 as

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

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

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

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

An immediate implication is that if ψ^β is strictly concave in e, then an uninformative investigation is uniquely optimal. Because ψ^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.[51] Note that in these cases (and in general), the optimal investigation under ex-ante signaling provides some information; see Corollary 2. Thus, 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.

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

However, the comparison can also go the other way. Because ψ^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 G is convex on $[c, 1+c]$ then the optimal investigation is fully informative under both ex-ante and ex-post signaling.*