Signaling Good Faith by Taking Stands

Andrew McClellan and Daniel Rappoport*

October 23, 2024

Abstract

A decision maker (DM), who will take a binary decision, cares about his reputation for being "good," i.e., wanting to accord his action choice with public evidence, as opposed to being "bad," i.e., having a fixed partisan agenda regardless of the evidence. While the decision is taken after evidence is realized, the DM has the option to take a "stand" beforehand, i.e., to communicate his intentions via a cheap-talk message. A wide range of equilibria exist and are characterized by how much the good DM reveals about his standards at this initial communication stage. The most informative of these is *ex-ante signaling*, which sees the DM effectively commit to a contingent plan as a function of the realized evidence. Our main theorem shows that, across all equilibria, ex-ante signaling minimizes the probability that the DM follows his partisan agenda.

1. Introduction

Across a wide array of institutions, individuals' decisions are scrutinized for whether they align with public objectives. This often comes down to whether the decision accords with the evidence as opposed to the potentially biased agenda of the decision maker. While, the decision is only made after the evidence is made public, in many contexts the decision maker has an opportunity to "take a stand," or state their intentions, before the uncertainty is resolved. Our paper explores how such "ex-ante" signaling efforts affect outcomes. To make this concrete, consider the following examples.

*McClellan and Rappoport are both at the University of Chicago, Booth School of Business. We thank Steve Callander, Laura Doval, Piotr Dworczak, Wioletta Dziuda, Ben Golub, Alex Frankel, Marina Halac, Thomas Jungbauer, Emir Kamenica, Navin Kartik, Jacob Leshno, Elliot Lipnowski, Alessandro Pavan, Doron Ravid and Joe Root.

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

2. 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 private information about the issue at hand, but also have a desire to project integrity rather than appearing to seek a particular outcome regardless of the specifics. These organizations can declare their standards for approval up front or decide on a case-by-case basis after observing the evidence. For example, in 2020, national drug regulatory agencies were eager to approve a safe COVID-19 vaccine but faced credibility worries that they were rushing the process. The FDA laid out a specific efficacy threshold in clinical trials for approval, whereas the European Union's counterpart deliberately provided no such lower bound (Singh and Upshur (2021)). Similar issues arise in the decisions of other government agencies, such as the FTC designing the specificity of merger guidelines.[3]

3. University admissions committees would like to appear as though their decisions are based on academic potential despite being pressured to consider the legacy statuses, donations, or other non-academic features. Many American universities

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

[2] Some made informative statements: Senator Romney reported that the transcript was "troubling." Others refused to comment: Senator Sasse criticized his colleagues for jumping to conclusions. See Costa, Roberts (2019) "Cracks emerge among Senate Republicans over Trump urging Ukrainian leader to investigate Biden" *Washington Post*, September 25.

[3] See U.S. Department of Justice and Federal Trade Commission (2023).

practice "holistic" admissions and will not give exact criteria for admission. This lack of transparency has been criticized for facilitating higher admission rates for unqualified applicants.[4] One alternative is to publicize specific criteria for admission, a practice common in universities throughout Europe and Asia.[5,6]

We focus on two important questions in such settings. First, how informative can communication be prior to the revelation of evidence, e.g., how much can politicians distinguish their standards during an investigation? Second, how does informative communication about hypothetical plans affect outcomes, e.g., would we expect that Republican senators who indicate conditions for impeachment up front convict more or less than those who wait and see, would more drugs be approved with prespecified efficacy and safety standards, and would universities admit more donor applicants were they to publicize admissions criteria rather than use holistic admissions?

Our model features a single decision maker (DM), and an inactive Bayesian observer. The game consists of two stages: a communication stage and a decision stage. At the communication stage, the DM, sends a cheap-talk message about his preferences.[7] At the decision stage, the evidence $e \in \mathbb{R}$ is realized and the DM chooses a binary action $a \in \{0, 1\}$. In addition to the evidence, the DM's preferences over the action also depend on his two dimensional private type consisting of whether he is a bad type—a "partisan"—or a good type—a "non-partisan"—as well as privately known evidence standards $s \in \mathbb{R}$. The non-partisan would like to accord his action with the evidence and the standard (i.e., acts in "good faith"). On the other hand, the partisan does not care about taking the right decision and suffers a constant disutility from taking $a = 1$ regardless of the evidence or standard. Finally, the DM also cares about his reputation for being a non-partisan in the eyes of the observer who sees the DM's cheap-talk message, the realized evidence, and the DM's chosen action.

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

[5] 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.

[6] 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.

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

The DM's standards can be interpreted in two ways: (i) as private non-verifiable information about the "right" evidence threshold, e.g., FDA officials have specific expertise about the drug being considered, or (ii) as idiosyncratic heterogeneity in preferences with respect to this particular decision, e.g., different politicians can have different views about the appropriate extent of executive power while still maintaining integrity. Depending on the parameters, the non-partisans can prefer $a = 1$ more or less on average relative to the partisan. That is, the partisan is not distinguished from the non-partisan by the direction of his bias, but instead by a lack of responsiveness to evidence and standards.[8] In this sense 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 a different ideology or information, but is still interested in the objective evidence on a particular issue—like the non-partisan in our model.

We first show that each equilibrium can be pinned down by how much information the communication stage transmits about the non-partisan's standards. Two important extreme cases are (i) when the communication stage involves babbling, and all signaling occurs at the decision stage, and (ii) when the communication stage perfectly communicates his standards, and there is no additional signaling at the decision stage. 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 contingent plan with respect to the realized evidence, e.g., stating "I will convict if the evidence meets ... standard." Conversely, ex-post signaling is equivalent to the DM saying "I will not speculate on hypotheticals." While we view these extremes as most salient, Lemma 2 establishes that any intermediately informative communication about standards can be sustained in some equilibrium and we note that these alternative equilibria will tend to induce different outcomes than ex-ante or ex-post 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 choose his preferred action more frequently. However, Theorem 1 shows that ex-ante signaling has the highest probability of $a = 1$, i.e., it features the DM most strongly rejecting his partisan bias. In addition, ex-ante signaling delivers a higher probability of $a = 1$ than ex-post signaling for *every* evidence realization. This means that politicians who answer interviewers' questions will tend to break with their party more than those who successfully "dodge the cameras"; govern-

[8] We elaborate on this distinction at the end of Section 2.

ment agencies that specify approval criteria up-front will go against their appointer's political interests more than those who decide on a case-by-case basis; and setting clear admissions criteria will lead to more meritocratic admissions decisions relative to holistic admissions.

The broad intuition for Theorem 1 is simple: before the realization of evidence, i.e., under ex-ante signaling, the DM is willing to make stronger claims in order to attain a higher reputation because there are many evidence realizations under which these stronger claims do not require a different action than weaker ones. Conversely, under ex-post signaling, after a "pivotal" evidence realization occurs, obtaining a high reputation requires taking the high action with probability one. While this simple reasoning is sufficient with binary standards, the full intuition revolves around a "convexity of reputation" detailed in Subsection 4.2.

Ex-ante signaling and ex-post signaling are focal because their outcomes have simple implementations. Ex-post signaling, which delivers a lower probability of $a = 1$ by Theorem 1, can be induced by banning the interim communication introduced by our model, e.g., the FTC can prohibit agents from making public comments on ongoing merger cases. As we show in Proposition 7, ex-ante signaling outcomes, which conversely delivers a higher probability of $a = 1$ by Theorem 1, arise uniquely when the DM commits to a contingent plan, e.g., the FTC can publicly specify binding merger standards, or verifiably delegate the decision to a known third party. In Subsection 6.2, we show that "intermediate outcomes" can be implemented by regulating how much information from the investigation is revealed before or after the DM commits to a contingent plan. Lastly, in Proposition 8 we argue that, without institutional intervention that restricts communication, ex-ante signaling outcomes are likely to prevail. The intuition is that taking a clear stand is the best way to signal a responsiveness to evidence rather than to one's esoteric agenda.

With this focus on ex-ante signaling, we explore how outcomes change with parameters of our model. We highlight two such comparative statics. Proposition 4 provides conditions under which a mean-preserving spread in the distribution of standards decreases the probability that the DM goes against his partisan interests, i.e., decreases the probability of $a = 1$. Conversely, Proposition 5 provides conditions under which a spread in the evidence distribution increases the probability of $a = 1$. The intuitions for these results reflect two sides of the same coin: a spread in the distribution of standards *relative* to evidence makes the DM's private information in decision making more important relative to public information. This, in turn makes committing to higher ev-

idence thresholds less costly in terms of reputation and induces a lower probability of $a = 1$. Conversely, a spread in the distribution of evidence relative to standards makes the DM's private information less important and has the opposite effect.[9]

With this focus on ex-ante signaling, we explore how outcomes change with parameters of our model. We highlight two such comparative statics. Proposition 4 shows that under a regularity condition, a mean-preserving spread in the distribution of standards decreases the probability that the DM goes against his partisan interests, i.e., decreases the probability of $a = 1$. This change can be interpreted as an increase in polarization in political contexts, or an increase in the importance of expertise relative to public evidence in agency contexts. The rough intuition is that after such a spread the probability that evidence falls between two different standards increases which increases the cost for the bad type (in terms of taking $a = 1$) of mimicking the lower standard. Conversely, Proposition 5 provides conditions under which a spread in the evidence distribution increases the probability of $a = 1$. This change can be interpreted as evidence becoming more informative, e.g., an impeachment inquiry calling more witnesses. The intuition is that a concentrated evidence distribution allows "targeting" by the bad type: if there is a high probability of a given evidence realization, the bad type will tend to feign standards just out of reach of this realization.

The layout of the paper is as follows. Section 2 describes our model. Section 3 characterizes equilibria and taxonomizes them by how much information is revealed about the evidence standard at the communication stage. Section 4 states our main result comparing ex-ante signaling to other equilibria and provides intuition and a proof sketch. Section 5 then explores comparative statics in ex-ante signaling outcomes. And lastly, Section 6 discusses equilibrium selection, alternative commitment and timing assumptions, and robustness results.

1.1. Literature Review

We add to the literature studying career 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), Ottaviani and Sorensen (2006a), Ottaviani and Sorensen

[9] Our conditions for Proposition 4 and for Proposition 5 are satisfied under a uniform distribution of evidence and standards respectively. When the distributions are not uniform, these conditions separately deal with the subtlety that a mean-preserving spread in standards (resp. evidence) is no longer a mean-preserving spread in the probability that the standard is above the realized evidence (resp. evidence is above the realized standard).

(2006b)). Our DM's preferences are closest to those in Morris (2001).[10] He studies an informed sender who seeks a reputation for being responsive to the state—similar to our non-partisan—rather than having a state-independent preference—similar to our partisan.[11] The main difference in our preferences is that we have heterogeneity in the "good" type's preferences, i.e., there is a non-degenerate distribution of standards. Importantly, communication has no value in our model when the standards distribution is degenerate, but can otherwise change equilibrium outcomes in a significant way.

We are also connected to the costly signaling literature initiated by Spence (1973). As in Bénabou and Tirole (2006), Esteban and Ray (2006), and Frankel and Kartik (2019), the multidimensional type of the DM—namely preference heterogeneity of the non-partisan in our model—precludes separating equilibria. Frankel and Kartik (2022) and Ball (2022), among others bring a design perspective to such settings, studying how to design scoring systems in the presence of strategic manipulation.[12,13]

Our model differs from standard costly signaling frameworks by introducing the opportunity to communicate intentions prior to the public revelation of a payoff relevant state.[14] We show that equilibria with more informative communication induce more transparent decision criteria.[15] In this sense, our main result that the DM acts against his partisan biases more when decision criteria are endogenously transparent is related to conclusions in Levy (2007) and Prat (2005) about exogenous changes in transparency.

[10] The preferences in Bussing and Pomirchy (2022) also take this form and, as we do, refer to the "bad" type as a partisan. Many political economy models (e.g., Maskin and Tirole (2004), Kartik and Van Weelden (2018), Agranov (2016)) have an alternative definition of a partisan as one with preferences far away from the median voter. We describe how to incorporate such preferences at the end of Section 2.

[11] Other papers study different reputation incentives with related interpretations. The advisors in Durbin and Iyer (2009) and Acemoglu et al. (2013) seek a reputation for being "incorruptible" (i.e., valuing bribes relatively less as compared with outcomes). A reputation for competence (e.g., Prendergast (1993), Li (2007)) can induce a preference to match the action with the state.

[12] Rappoport (2022) studies optimal delegation policies for agents engaged in costly signaling.

[13] 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.

[14] The impact of exogenous signals have been studied in costly signaling in Daley and Green (2014) and in cheap talk by Chen (2012), who also looks at how the timing of cheap talk impacts outcomes.

[15] Our comparison between ex-ante signaling, which specifies a complete contingent plan, and ex-post signaling, which defers the decision until the evidence is realized, echoes themes from the literature on incomplete contracts (e.g., Grossman and Hart (1986), Hart and Moore (1988)), where complete contracts are assumed to be prohibitively costly. Our results complement these by highlighting how communication and high reputation incentives can overcome the inability to commit to fully specified contingent plans.

2. Model

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

The game takes place in two stages. In the initial communication stage, the evidence is unknown and the DM only knows its CDF F; we assume $\int_E e dF(e)$ is well-defined and finite. The DM sends a cheap-talk message $m \in M$ to the observer, where M is some sufficiently large metrizeable space.[16] After the message is sent, the decision stage begins: the evidence e is publicly revealed and then the DM chooses an action a. The observer sees the DM's message and action choice in addition to the realized evidence and forms beliefs, after which payoffs are realized.

Preferences The DM's type θ depends on two dimensions. The DM can either be a partisan (P) or a non-partisan (N). In addition, the DM privately observes the relevant "standard" $s \in S \subseteq \mathbb{R}$ for the decision. His type is thus given by $\theta \in \{N, P\} \times \mathbb{R}$. The utility from taking action a, given evidence e, standard s, and public belief $\nu \in [0, 1]$ held by the observer that the DM is an N type is given by

$$u(\theta, e, a, \nu) = \begin{cases} -ac + \rho\nu & \text{if } \theta = (P, s), \\ a(e - s) + \rho\nu & \text{if } \theta = (N, s). \end{cases} \tag{1}$$

N types prefer $a = 0$ more if (i) the evidence is less convincing (e is lower), or (ii) their standards are higher (s is higher).[17] We provide additional discussion at the end of this section, but our leading interpretation of a non-partisan's standards is that they are private non-verifiable information about the "correct" evidence threshold for action $a = 1$. In stark contrast, the P type always wants to choose $a = 0$ and his disutility $c > 0$ from $a = 1$ is independent of the evidence realization e and the relevant standard s, i.e., the P type does not care about taking the right decision. The weight $\rho > 0$ parameterizes how much the DM values reputation. We refer to the first component of the payoff that depends on the action as the **material payoff** and $\rho\nu$ as the **reputation payoff**.

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

[17] The payoff from a for N types is assumed to be $a(e - s)$ for convenience. Our results still hold (with minor notational changes) if the utility difference between $a = 1$ and $a = 0$ is increasing in e and decreasing in s.

Because the P type does not care about the standard s, we redefine the type space more compactly as $\theta \in \Theta \equiv S \cup \{P\}$ where s is a shorthand for an N type who observes standard s. The prior distribution over types is denoted $\nu_0 \in \Delta(\Theta)$. We let $q \equiv \nu_0(S)$ be the prior probability of the good N type, and $G(s) \equiv \nu_0(\{s' : s' \leq s\}|\theta \in S)$ be the CDF of standards; without loss, we take $S \equiv \text{Supp}(G)$. We assume for expositional convenience that either F or G is atomless.

We make the following assumption on the strength of reputational incentives.

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

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

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

1. ν_1 is obtained from σ using Bayes rule.[18]
2. ν_2 is obtained from ζ using Bayes rule with prior $\nu_1(\cdot|m)$.[19]
3. $\sigma(M^*_\theta|\theta) = 1$ where $M^*_\theta \equiv \arg\max_{m\in M} \int_E \left(\max_{a\in\{0,1\}} u(\theta, e, a, \nu_2(S|m, e, a))\right) dF(e)$.
4. $\zeta(A_{\theta,m,e}|\theta, m, e) = 1$ where $A_{\theta,m,e} \equiv \arg\max_a u(\theta, e, a, \nu_2(S|m, e, a))$.

In addition, we impose a version of the D1 refinement à la Cho and Kreps (1987) and Ramey (1996). Let $\Theta_m \equiv \text{Supp}(\nu_1(\cdot|m)) \subseteq \Theta$ be the support of the interim belief on the DM's type following message m but before an action is chosen. We impose the D1 refinement at the decision stage, after evidence has been realized and message m has

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

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

been sent, where the type space is Θ_m.[20] In our framework, this refinement simplifies to the following: if, after sending message m and observing evidence e, the DM takes an off-path action, the observer believes the DM to be the type(s) in Θ_m who would benefit the most in terms of their material payoff from this deviation relative to their equilibrium payoffs.[21]

We next define some useful notation. For an equilibrium $\mathcal{E}$ and investigation F, let $U_\theta^{\mathcal{E}}(F)$ be the expected utility of type θ,[22] $v^{\mathcal{E}}(e, F)$ be the probability of action $a = 1$ given evidence realization e, and $V^{\mathcal{E}}(F) \equiv \int_E v^{\mathcal{E}}(e, F)dF(e)$ be the associated ex-ante probability of $a = 1$.

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

Discussion

Partisan Preferences: It is important to note that even though a high s type and P both prefer $a = 0$ for "essentially" all evidence realizations, this does not mean their preferences are equivalent. This perspective ignores the main tradeoff the DM faces between reputational and material payoffs, a tradeoff that makes the intensity of preferences over actions important. If we instead modeled the "bad" P type as the highest s type, then P would prefer to take action $a = 0$ much more than he values reputation as compared with non-partisans. Indeed, this is the interpretation of bad types in the canonical Spence (1973) education model: bad types have a higher cost of education or, *equivalently and indistinguishably*, a lower value for reputation. Our applications do

[20] 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., after a deviation, the observer considers that the DM's type could lie in the full type space Θ. One can show that in our model this approach yields a less expositionally convenient but essentially identical set of equilibria: every ex-ante D1 equilibrium is also an ex-interim D1 equilibrium, and every ex-interim D1 equilibrium outcome is the limit of some sequence of ex-ante D1 equilibrium outcomes.

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

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

not fit well with this interpretation, e.g., it does not seem appropriate to model partisan politicians as being defined by their lack of office motivation, or a corrupt regulator as not caring about being fired.

Instead, as mentioned in our literature review, our preferences mirror those in Morris (2001). The distinction between good and bad types is that good types care more about getting the decision "right" than bad types, that is, their preferences are responsive to the evidence and the relevant standards. For extreme evidence realizations, non-partisan types care more about stakes of the decision, whereas partisans care more about reputation. However, for middling evidence realizations, where the stakes of the decision are low for a non-partisan type, this comparison is flipped.

Reputation for Non-Partisanship: We assume that reputational payoffs are determined by the observer's belief that $\theta \in S$ and does not depend on their beliefs conditional on $s \in S$. This assumption streamlines our exposition and is natural in applications in which s represents the DM's transitory private information or idiosyncratic preferences that are only relevant for the decision at hand. For example, a politician may possess classified information about the relevant scandal. However, in some settings the DM may have have competing reputation concerns to appear as different s types; for example, a politician may also value appearing to have positions closer to the median voter. In Section 6, we introduce a generalized version of Assumption 1 for the case in which reputational payoffs depend on the observer's belief about the DM's standards. Our appendix proves all of our main results in this more general framework.

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

3. Equilibrium Characterization

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

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

1. *The P type positively mixes over all messages sent by N types, i.e., $\sigma(\cdot|P)$ and $\Sigma_N(\cdot) \equiv \int_S \sigma(\cdot|s)dG(s)$ are mutually absolutely continuous.*

2. *N types choose actions consistent with x_s with probability one, i.e.,*

$$\int_S \int_E \int_M \zeta(x_s(e)|s,m,e)d\sigma(m|s)dF(e)dG(s) = 1.$$

3. *For all $m \in M_P^*$, $\nu_1(S|m) > 0$ and, after sending m, the P type positively mixes over the action choices of s types who also send m, i.e.,*

$$\int_S \zeta(a|s,m,e)d\nu_1(s|m) > 0 \iff \zeta(a|P,m,e) > 0 \; \forall e, a.$$

The interpretation of the 1st and 3rd point is that P cannot be distinguished from N following any "on-path" history. A key implication is that P is indifferent across mimicking the behavior of any s type, at both the communication and decision stages. These points follow from the high reputation incentives. If a message is sent only by P, then it yields an equilibrium reputation and utility of 0 for P. However, P can obtain an expected utility of at least $\rho q - c$ by mimicking the strategy of some s type, which is strictly preferred by Assumption 1. Conversely, a message that is sent only by s types yields a reputation of one, 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,[23] which is strictly less than $\rho - c$ again by Assumption 1. The argument for why, after sending message m, P mixes over the actions chosen by s types who also send m is similar, but has to contend with the subtlety that the relevant utility

[23] This follows from Francetich and Kreps (2014): for any signal structure, and for any state ω, the expected posterior belief of ω conditional on state ω is higher than the prior probability of ω.

bounds are now dependent on the interim probability of N types instead of the prior, i.e., on $\nu_1(S|m)$ instead of on q.

The second point states that the s type's realized contingent plan (almost surely) follows the fixed rule $x_s(e)$.[24] To avoid probability one caveats, going forward we focus on equilibria where N's actions correspond with $x_s(e)$ *everywhere*, i.e., $\forall e \in E,\ s \in S$. The s type's action choice is not only constant across equilibria and messages, but also across parameters of the model such as the investigation and the type distribution of the DM. This independence should not be misunderstood as arising because the s types choose their ideal action unaffected by reputation incentives. Indeed, s types engage in "political correctness" (Morris (2001)): in order to signal non-partisanship, they select the partisan's less preferred action $a = 1$ for $e \in (s - c, s)$ even though they prefer $a = 0$. Instead, the reason that x_s is selected by the s types is because it provides the highest "signaling value" regardless of the equilibrium and investigation—that is, x_s maximizes the utility difference between s and P types over all contingent plans $x \in \mathcal{X}$.

To provide intuition for point 2, consider the case in which both actions are on path following some evidence realization e.[25] Point 3 of Lemma 1 implies that P mixes over $a = 1$ and $a = 0$. However, the type $\tilde{s} \equiv e + c$ has the same preferences as P given e, i.e., he has the same trade off between the cost of $a = 1$ and reputation. Combined with the fact that N's utility for $a = 1$ is decreasing in s, all $s > \tilde{s}$ must choose $a = 0$ and $s < \tilde{s}$ must choose $a = 1$, i.e., s types choose actions consistent with x_s.

We next categorize the set of equilibrium outcomes. For any equilibrium, the communication stage conveys information about the standards of the DM conditional on them being a non-partisan. We call this induced Bayes-plausible information structure $\Lambda \in \Delta(\Delta(S))$ the **information structure on standards** (ISS) associated with the equilibrium $\mathcal{E}$.[26] Formally, for each Borel $H \subset \Delta(S)$, $\Lambda(H) = \int_{m \in M} \mathbb{1}\big(\nu_1(\cdot|m, \theta \in S) \in H\big) d\Sigma_N(m)$.

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

There are two main takeaways from the lemma. First, equilibrium outcomes can be uniquely described by the associated information the communication stage conveys

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

[25] If action $a = 0$ (respectively $a = 1$) is off-path, the proof uses the analogous logic in conjunction with the D1 refinement to establish that $s \geq e + c$ (respectively $s \leq e + c$) $\forall s \in \Theta_m$, i.e., actions are consistent with x_s.

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

about the standards of the DM. Second, *every* ISS is associated with an (potentially different) equilibrium outcome. Unlike familiar cheap-talk models (e.g., Crawford and Sobel (1982)), there is no monotonicity restriction on the equilibrium strategies of s types. More importantly, this permissiveness means that in equilibrium the communication-stage message can convey a wide range of information about s, from the perfectly informative ISS where each s sends a different message to the perfectly uninformative ISS where all DM types send the same message. At the beginning of the next section, we provide further details about these salient extreme equilibria.

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

The above heuristic for constructing equilibrium strategies is valid because of the following property: if, for some candidate equilibrium strategies, P is indifferent across messages, then each s type's incentive constraint (to follow their prescribed strategy) is ensured as well. Figure 1 displays the reasoning. Consider $\underline{s} < \overline{s}$ who send different messages $\underline{m}$ and $\overline{m}$ respectively. Suppose P is indifferent between sending $\overline{m}$ and following up with $x_{\overline{s}}$ (i.e., using threshold $\tilde{e}_{\overline{s}}$), and sending $\underline{m}$ and following up with $x_{\underline{s}}$ (i.e., using threshold $\tilde{e}_{\underline{s}}$). This indifference implies that expected reputational difference between the latter and the former strategy must be equal to the material utility difference from switching their action choice for $e \in (\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$, i.e., the absolute value of the area $K_1 + K_2$ measured according to the distribution of evidence F. But notice that if type $\underline{s}$ considers deviating from $\underline{m}$ and $x_{\underline{s}}$ to $\overline{m}$ and $x_{\overline{s}}$, they only gain the absolute value of K_1 in material utility which does not compensate them for the reputational loss of $K_1 + K_2$. Analogously if $\overline{s}$ considers deviating from $\overline{m}$ followed by $x_{\overline{s}}$ to $\underline{m}$ followed by $x_{\underline{s}}$ they lose the absolute value of $K_1 + K_2 + K_3$ in material utility which is greater than the reputational gain $K_1 + K_2$. Thus, P's indifference ensures each s type's incentives.[27]

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

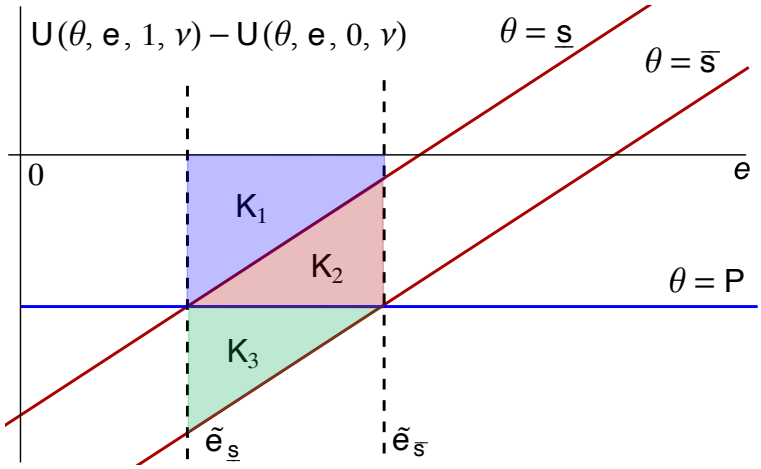


Figure 1: Material utility difference between $a = 1$ and $a = 0$ as a function of evidence.

4. The Effects of Informative Stands

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

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

At the other extreme is the equilibrium associated with the uninformative ISS, which we term **ex-post signaling** and denote as equilibrium β. Under ex-post signaling the DM "babbles," e.g., regardless of his type, he sends the same message interpreted as "I will wait and see until the investigation concludes." Ex-post signaling admits residual strategic uncertainty under the weak condition that there exist two types s', s'' such that $F(\tilde{e}_{s'}) \neq F(\tilde{e}_{s''})$. A distinctive feature of ex-post signaling is that because the communication stage is uninformative, the distribution of actions conditional on an evidence realization e does not depend on the investigation F, i.e., $v^{\beta}(e, F) \equiv v^{\beta}(e)$ is independent of F (and so we drop the associated dependence in this case).

These salient extremes highlight the extent to which the DM can take "informative

stands"; under ex-ante signaling, he can effectively publicly commit to his contingent plan. Alternatively, under ex-post signaling, the DM can decide on a case-by-case basis, obviating the communication stage. Our main result looks at how different communication protocols impact the probability of $a = 1$. First, we introduce a technical condition. We say there is **mild agreement** if for every pair $s', s'' \in S$, $\exists e \in \text{Supp}(F)$ such that $x_{s'}(e) = x_{s''}(e)$, i.e. no two s types always choose different actions in equilibrium.

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

The two actions are only differentiated by P's bias towards $a = 0$; indeed, if the partisan preferred $a = 1$ instead ($c < 0$), then the comparison in Theorem 1 would flip. Highlighting the comparison with ex-post signaling, Theorem 1 then says that the DM goes against his partisan interests more when he takes the "most informative stands," i.e., pre-specifies his contingent plan, rather than deciding on a case-by-case basis. In terms of the applications, the politician who answers interviewers' questions will tend to break with their party more, and universities will admit more donor or legacy applicants when using holistic admissions. Beyond predictive implications, in many contexts it is plausible that whether ex-ante or ex-post signaling outcomes prevail is a design decision. Theorem 1 gives the implications for such decisions. Subsection 6.1 and Subsection 6.2 elaborate, showing how ex-ante signaling outcomes arise uniquely for minor variations in the current model.

Depending on the parameters, certain ISS may correspond to the same equilibrium outcomes as ex-ante signaling; e.g., all equilibria have the same outcomes if the distribution of evidence is degenerate. However, under mild agreement, if equilibrium actions are not completely predictable at the decision stage, then the equilibrium delivers different outcomes than ex-ante signaling; in particular, a strictly lower probability of $a = 1$. *All* imperfectly informative ISS are associated with equilibria with residual strategic uncertainty if and only if each s type's threshold results in a different probability of $a = 1$ (i.e., $1 - F(\tilde{e}_s)$). This holds when F has full support over $\mathbb{R}$, which also guarantees mild agreement.[28]

Given that $a = 1$ is taken most often under ex-ante signaling, a natural follow up question is whether the same comparison holds for each evidence realization. While

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

it is difficult to make this comparison for arbitrary equilibria, we show such a ranking does indeed hold when comparing ex-ante signaling to ex-post signaling.

Proposition 1. *$v^\alpha(e,F) \geq v^\beta(e)$ for all $e \in E$.*

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

4.1. Intuition for Theorem 1 with Binary Standards

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

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

$$\rho\big(\nu_1^\alpha(S|m_{\underline{s}}) - \nu_1^\alpha(S|m_{\overline{s}})\big) = F(\tilde{e}_{\overline{s}}) - F(\tilde{e}_{\underline{s}}).$$

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

Under ex-post signaling, every DM type chooses the same message $m^0 \in M$ at the communication stage. Given evidence realization e at the decision stage, P similarly chooses $a = 1$ with the probability that he mimics the $\underline{s}$ type, which is determined by

$$\rho\big(\nu_2^\beta(S|m^0, 1, e) - \nu_2^\beta(S|m^0, 0, e)\big) = 1.$$

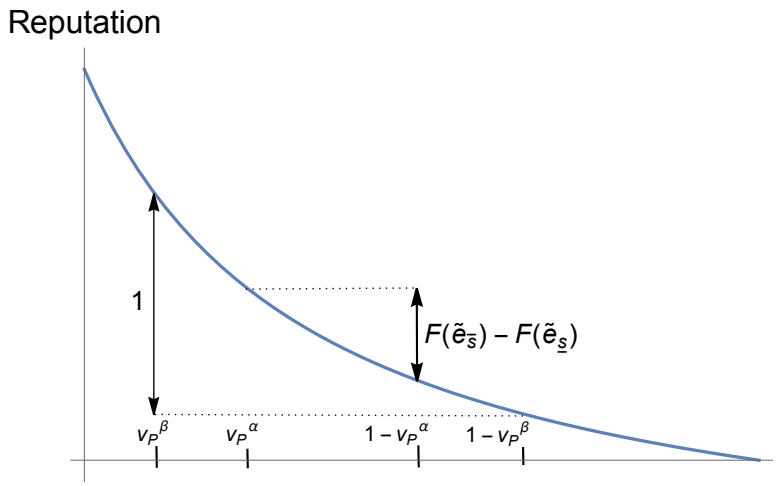


Figure 2: Reputation as a function of P's strategy in the binary example when $\overline{s}, \underline{s}$ are equally likely and where $v_P^{\mathcal{E}}$ is a shorthand for the probability P takes $a = 1$ given $e \in (\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$ under equilibrium $\mathcal{E}$.

Figure 2 illustrates how P shifts his strategy so that the reputation incentives compensate him for the difference in material loss between mimicking $\underline{s}$ and $\overline{s}$. Under ex-post signaling, conditional on evidence e, the difference in $\mathbb{P}(a = 1)$ from pooling with $\underline{s}$ or $\overline{s}$ is 1, compared with $F(\tilde{e}_{\overline{s}}) - F(\tilde{e}_{\underline{s}}) < 1$ under ex-ante signaling. To create a higher reputation difference in the ex-post case, P must mimic $\underline{s}$ less frequently, which in turn means he chooses $a = 1$ less frequently. Thus, $v^{\alpha}(e, F) > v^{\beta}(e)$ for $e \in [\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}}$, which implies $V^{\alpha}(F) > V^{\beta}(F)$ in the binary-standards environment.

The underlying force behind the above argument is that the P type is willing to promise more ex-ante because he will be called to act on this promise for a subset of evidence realizations. The key difference between ex-ante and ex-post signaling is when P chooses what type of s to mimic: in ex-ante signaling, P makes this decision prior to the revelation of e, while, in ex-post signaling, P can adjust his decision to the evidence realization. Under ex-ante signaling, mimicking $\underline{s}$ as compared to $\overline{s}$ yields extra reputation regardless of whether these two types take different decisions ex-post. In contrast, under ex-post signaling, the extra reputation from mimicking $\underline{s}$ as opposed to $\overline{s}$ only realizes when these types take different decisions.[29]

While this intuition is compelling, it is difficult to extend this argument directly to

[29] 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.

show a similar ranking holds with more standards types or across other equilibria. Instead, we now take a different approach, one that proves useful when studying optimal information design and provides additional insights into the forces behind Theorem 1.

4.2. Proof Sketch of Theorem 1

We first establish an inverse relationship between P's equilibrium expected utility and the equilibrium probability of $a = 1$.

Lemma 3. *The probability of $a = 1$ is inversely related to P's equilibrium utility:*

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

The negative relationship between the probability of $a = 1$ and P's utility implies how the other component of P's utility changes—namely his reputation. In particular, Lemma 3 says that in equilibrium there is no way to offer the P type a favorable trade (either by changing the equilibrium focus, the investigation, or the distribution of standards) where he increases his reputation in return for increasing the probability he takes $a = 1$. It is worth noting that such a trade would exist for certain pairs of strategies, but in equilibrium, the P type will tend to push any increase in his utility to decreasing the probability of $a = 1$ at the detriment of his reputation.

Given Lemma 3, proving Theorem 1 reduces to establishing that ex-ante signaling is P's least favorite equilibrium. 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 owing to the fact that $v^{\beta}(e)$ is independent of F. These points imply

$$U_P^{\beta}(F) = \int_E U_P^{\beta}(\delta_e)dF(e) = \int_E U_P^{\alpha}(\delta_e)dF(e),$$

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

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

The intuition for Lemma 4 follows from a fundamental property about Bayesian updating: adding probability that a given type sends some signal changes the corresponding conditional belief on that type less if they already send that signal with high probability. In our setting, this means that the belief that the DM is an N type following

any message is convex in the probability that P sends that message. This convexity is illustrated in Figure 2. To see how convexity of reputation relates to convexity of the P type's ex-ante signaling utility $U_P^\alpha(F)$, consider two investigations $\overline{F}$ and $\underline{F}$ and, for some $\lambda \in (0,1)$, let $F_\lambda = \lambda\overline{F} + (1-\lambda)\underline{F}$. P's material utility from sending message m_s is linear in F: P chooses $a = 1$ under F_λ with probability equal to the average of that under $\overline{F}$ and $\underline{F}$. However, P cannot achieve the "average reputation" at every m_s because reputation is convex in the rate at which he declares each message, which yields the convexity of $U_P^\alpha(\cdot)$.[30]

Ex-Ante Signaling vs. Other Equilibria: We have shown that ex-ante signaling has a higher probability of $a = 1$ than ex-post signaling. We next discuss how to use this comparison to deliver the full strength of Theorem 1: that ex-ante signaling has the highest probability of $a = 1$ among *all* equilibria.

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

Now consider an alternative messaging strategy which first selects a message according to the original equilibrium strategy under $\mathcal{E}$ and then sends a follow up message m_s according to the ex-ante signaling equilibrium given prior $\nu_1(\cdot|m)$. Conditional on sending each initial message under this new strategy, the above logic implies that P's expected utility is lower than under the original equilibrium $\mathcal{E}$. 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

As an intermediate step, the argument above delivers that ex-ante signaling is P's least favorite equilibrium. Using Lemma 1, we extend this comparison to all DM types.

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

[30] Let $\overline{\nu}_1(S|m_s)$ and $\underline{\nu}_1(S|m_s)$ be the corresponding reputations under $\overline{F}$ and $\underline{F}$. In order to maintain the reputation $\lambda\overline{\nu}_1(S|m_s) + (1-\lambda)\underline{\nu}_1(S|m_s)$, the convexity of the reputation implies P would need to, for all $s \in S$, declare m_s at a rate less than the average across the equilibria induced by $\overline{F}$ and $\underline{F}$. But this is impossible since the total measure of P's messages must be preserved.

1. $U_\theta^{\mathcal{E}}(F) - U_\theta^{\mathcal{E}'}(F)$ *is constant across* $\theta \in \Theta$.

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

Given P's preference over equilibria, 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 s type chooses x_s in every equilibrium, so their utility difference is just given by the expected reputation difference from following x_s, and (ii) P is indifferent between mimicking any s type in any equilibrium, and so, similarly, P's expected utility difference across equilibria is given by their expected reputation difference from following x_s. In cases where the DM can design communication protocols, this result provides a rationalization for why politicians may "dodge the cameras" and admissions committees may favor non-transparency—or, in our terminology, favor ex-post signaling. This result also points to interesting questions about equilibrium selection issues, which we address in Subsection 6.1.

5. Comparative Statics

We next explore comparative statics in our focal equilibrium of ex-ante signaling with a focus on the probability of action $a = 1$. We first document how the value of reputation and the prior probability of N types affect this outcome. We then turn to explore changes in the distribution of evidence F and the distribution of standards G; first considering first order stochastic dominance (FOSD) changes and then considering spreads in these distributions.[31]

Proposition 2. *The probability of* $a = 1$ *is higher when* ρ *or* q *increases.*

Proposition 3.

1. *Let* G_1 *and* G_2 *be two distributions of standards such that* G_2 *FOSD* G_1*. Then the probability of* $a = 1$ *is higher under* G_1 *than* G_2*.*

2. *Let* F_1 *and* F_2 *be two distributions of evidence such that* F_2 *FOSD* F_1*. Then the probability of* $a = 1$ *is higher under* F_2 *than* F_1*.*

[31] One omitted parameter from these results is c. Although one might naturally conjecture that an increase in c induces a lower probability of $a = 1$ from P, the probability of $a = 1$ from s types is increasing in c (as can easily been seen from Lemma 1). Either force can dominate, making comparative statics on c ambiguous.

The intuition for these comparative statics is straightforward as each change can be seen as increasing the DM's endogenous preference for $a = 1$. By increasing ρ, we are increasing the importance of reputation relative to material payoffs in the DM's utility. Since reputational incentives push towards avoiding the appearance of a P type who prefers $a = 0$, an increase in ρ also serves to increase the probability of $a = 1$. An increase in q decreases the probability of the P type who is biased against $a = 1$. Similarly, an FOSD decrease in the distribution of standards or an FOSD increase in the distribution of evidence means that the non-partisan prefers $a = 1$ more often.

Our next comparative statics consider spreads in the distribution of standards and evidence. A CDF $\tilde{\mu} : \mathbb{R} \to [0, 1]$ is a mean-preserving spread (MPS) of another CDF $\mu : \mathbb{R} \to [0, 1]$ if $\int_{-\infty}^{y} (\tilde{\mu}(y) - \mu(y))dy \geq 0 \ \forall y \in \mathbb{R}$ with equality at $y = \infty$. We say that $\tilde{\mu}$ is an MPS of μ on an interval $I \subset \mathbb{R}$ if $\tilde{\mu}$ is an MPS of μ and $\mu(y) = \tilde{\mu}(y) \ \forall y \notin I$.

We first consider the distribution of standards. In our agency examples, spreads in the distribution of standards can be interpreted as an increase in the importance of expertise, e.g., the DM obtains better information about the true correct threshold. In political contexts, they can be interpreted as an increase in polarization, e.g., the ideological positions of moderates in a party are further away from those of more extreme members. To ease exposition, we assume for the next proposition that the distribution of evidence admits a density f supported on an interval $[\underline{e}, \overline{e}]$.[32]

Proposition 4. *Suppose $\tilde{G}$ is an MPS of G on $[\underline{e} + c, \overline{e} + c]$. If $\frac{f(e)}{\rho q + c(1-F(e))}$ is increasing on $[\underline{e}, \overline{e}]$, then the probability of $a = 1$ is lower under $\tilde{G}$ than under G.*

To see the intuition consider the case in which F is uniform on $[\underline{e}, \overline{e}]$. Consider a particular type of MPS on $[\underline{e} + c, \overline{e} + c]$ of G given by $\tilde{G}$ where there exists a function $\ell(s)$ with $\ell'(s) > 1$ such that $G(s) = \tilde{G}(\ell(s)) \ \forall s$. That is, the quantiles of $\tilde{G}$ are literally spread out to be further apart from one another than under G. Let Σ be the equilibrium CDF of P's mixing strategy over S when s has CDF G and consider an alternative strategy under $\tilde{G}$ given by the CDF $\tilde{\Sigma}$ defined by $\Sigma(s) = \tilde{\Sigma}(\ell(s))$, i.e., the P type spreads their strategy in the same way that the N types spread their standards. Notice that, because $\tilde{G}$ is a mean-preserving spread of G and evidence is uniformly distributed, the expected probability of choosing $a = 1$ under $\tilde{G}$ when P uses $\tilde{\Sigma}$ is the same as under G when P uses Σ under G. Similarly, P's reputation is the same under both because $\tilde{\Sigma}$ has followed the spread of N types. However, since $\ell(s) - \ell(s') > s - s'$ for $s > s'$, mimicking a higher standard obtains the same decrease in reputation, but a larger decrease in the probability of taking

[32] With abuse of notation, the bounds of this interval are allowed to be infinite.

the action, and so the P types will deviate towards higher standards and take $a = 1$ with lower probability. The regularity condition that $\frac{f}{\rho q + c(1-F)}$ is increasing extends this result beyond uniform evidence distributions. This condition is similar to the commonly used increasing hazard rate condition i.e., $\frac{f}{1-F}$ being increasing.

We next turn to how mean-preserving spreads in the distribution of evidence affect outcomes under ex-ante signaling. A leading example is when the evidence is a posterior about a binary state (or a posterior mean), and a mean-preserving spread in the evidence distribution constitutes the investigation being more informative. Outside of informational contexts, such spreads indicate evidence being more influential in the decision relative to standards, e.g., when the expertise of the DM is relatively less important. Let $H(e) \equiv G(e + c)$ be the probability that the N types use evidence thresholds below e, and thereby the probability that the N types take action $a = 1$ given evidence e. To simplify the exposition, we assume that H is twice continuously differentiable, with $h(e) \equiv H'(e)$.

Proposition 5. *If $\tilde{F}$ is an MPS of F on $[e_1, e_2]$ and $\frac{h(e)}{\rho q + c(1-F(e))}$ is increasing on $[e_1, e_2]$, then $\tilde{F}$ has a higher probability of $a = 1$ than F.*

There are two main takeaways from the above results. The first is that if F is not sufficiently diffuse then a mean-preserving spread increases the probability of $a = 1$. Indeed if there is a large amount of mass within a small interval then $1 - F(e)$ will decrease "too fast" over this interval satisfying the condition in the proposition. That is, if evidence is sufficiently concentrated around a point e, then spread the evidence will increase the rate of taking the action. An extreme case of such concentration is when F has a mass point. Our next result highlights that such evidence can always be spread to increase the probability of $a = 1$. We say that a local MPS at e increases the probability of $a = 1$ if for all $\varepsilon > 0$ sufficiently small, there exists an MPS of F on $[e - \varepsilon, e + \varepsilon]$ that increases the probability of $a = 1$.

Proposition 6. *If F has a mass point at e, then a local MPS at e increases the probability of $a = 1$.*

In order to contextualize this result, consider applications in which the investigation is the choice of some third party who attempts to maximize the probability of $a = 1$. For example, the Speaker of the House can design an impeachment inquiry, and firms can control which information they submit in their application to the FTC for a merger.[33] The

[33] In the appendix we fully characterize optimal investigations in this problem when the evidence is a

above result stands in stark contrast with insights from similar exercises in the Bayesian persuasion literature. There, it is often optimal to have simple experiments which admit mass points at few realizations of evidence. In particular, no information, or a degenerate distribution of evidence, is optimal when certain concavity conditions on the distribution of thresholds are met. While associated conditions can be satisfied given a fixed F 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 standards that are just out of reach of such evidence. Given Lemma 3, this response by P leads to a lower probability of $a = 1$. Thus minimizing predictability in the investigation avoids such targeting.

The second takeaway of Proposition 5 concerns how the distribution of standards impacts the comparative statics of spreading evidence.

Corollary 2. *If H is convex on some $[e_1, e_2]$, and $\tilde{F}$ is a mean-preserving spread of F such that $\tilde{F}(e) = F(e)\ \forall e \notin [e_1, e_2]$, then the probability of $a = 1$ is higher under $\tilde{F}$ than under F.*

If H is convex then h is weakly increasing which means the condition in Proposition 5 is satisfied.[34] To interpret this corollary, recall that H is the probability that N types choose $a = 1$ given evidence e. Thus, if H is convex on some interval then a spread of the evidence distribution on that interval increases the probability that N types take $a = 1$. The corollary is essentially saying that under these conditions such a spread increases the probability of $a = 1$ from P types as well.

At a high level, the intuition is as follows. All else equal, P benefits from correlating his strategy with the N type. Consider a mean-preserving spread that uniformly decreases the probability of evidence between some $[e', e'']$ and uniformly increases the probability of evidence above e'' and below e', i.e., it decreases $F(e'') - F(e')$. This spread means that P's cost of mimicking N types with standards $s' \equiv e' - c$ is now lower relative to $s'' \equiv e'' - c$. As a consequence, P reallocates mass from mimicking standards above s'' to standards below s'. If H is convex, then N types are more prevalent at these standards above s'' than at those below s', and so this response by P serves to miscorrelate his strategy with that of the N type, and thereby tends to harm P, which, by Lemma 3, implies an increase in the probability of $a = 1$.

posterior about a binary state. The assumptions on F stated before Proposition 5 hold—indeed, the optimal investigation admits a density over non-degenerate posteriors. We note that this tendency towards unpredictability hinges on the communication stage being informative; as we show in Subsection F.2, this is not a feature of the optimal investigation under ex-post signaling, for which an uninformative investigation may be optimal.

[34] As this completes the proof of this corollary, we omit its proof from the Appendix.

6. Discussion and Extensions

6.1. Commitment and Equilibrium Selection

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

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

Proposition 7. *The commitment model admits a unique equilibrium outcome which is equivalent to that under ex-ante signaling.*

In the proof, we show that there is an equilibrium in which each s type chooses x_s, with P mixing over $\{x_s\}_{s \in S}$. The interpretation of the proposition can be broken down into two points. First, ex-ante signaling outcomes remain when the DM actually commits to some x_s, instead of sending a message that is interpreted as such a commitment (as in ex-ante signaling). Second, no other equilibrium outcomes can be sustained despite the introduction of commitment.

The Optional Commitment Model The optional commitment model has two alterations from our main model. First, at the communication stage, each DM has the option to commit to an arbitrary contingent plan as a function of the evidence, $x \in \mathcal{X}$, which is

publicly observed. If the DM chooses this option, then the game proceeds as in the commitment model. However, unlike in the commitment model, the DM can abstain from commitment and send a cheap-talk message instead, in which case the game proceeds as in our main model. We continue to apply the D1 refinement. In the appendix, we provide a formal definition of equilibrium in the optional commitment model.

Second, the preferences of the DM are perturbed as follows. The utility of the DM of type θ, taking action a, given evidence e, and belief $\nu \in \Delta(\Theta)$ is given by $u(\theta, e, a, \nu) + \varepsilon a$ where ε is a random variable that is mean 0, independent of other parameters, with support equal to $[-\delta, \delta]$ for some $\delta > 0$ and an atomless distribution. The DM does not know ε at the communication stage, but privately observes ε at the decision stage. The variable ε represents changing conditions between the communication and decision stages that are not made public; e.g., a politician may privately learn that convicting a fellow party member under investigation is actually more or less favorable for their party than previously expected. It can also represent evidence from the investigation that is revealed privately to the DM but not to the public. For example, certain findings of the Trump impeachment inquiry were redacted for the public but revealed to senators making the impeachment decision.

Proposition 8. *If, in addition to Assumption 1, $\rho > 2\max\{\frac{\delta}{q}, \frac{\delta}{1-q}\}$, then the optional commitment model admits a unique equilibrium outcome equivalent to that under ex-ante signaling.*

The intuition for the result 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. The key observation is that this "option value" is greater for the "bad" P types than it is for the "good" N types. The reason is that the good type will only take ε into account for *pivotal* evidence realizations, i.e., when $e - s$ 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 s who faces residual strategic uncertainty in equilibrium and x_s goes unused, then it will be given a reputation of one, which is not possible in equilibrium given the assumed high value of reputation. This captures the intuition by which "dodging the cameras" is interpreted negatively: being vague about one's standards at the communication-stage signals a desire to be responsive to idiosyncratic partisan preferences (ε) rather than the evidence.[35]

[35] Committing to a policy ex-ante is also used for signaling value in Callander (2008). There, the policy

Notice that the proposition holds for arbitrarily small preference shocks, but also for large ones modulated by the weight on reputation ρ. When δ is large enough to violate the inequality in Proposition 8, the option value from acting on the realization of ε could exceed the reputational gains from committing at the communication stage. In this case, each x_s commitment would still garner a full reputation given the argument above, but could go unused. That is, "dodging the cameras" is always interpreted negatively as compared with stating your principals up front, but depending on the reputation incentives, this negative perception may not provide sufficient deterrence for the DM.

6.2. Timing of Evidence Disclosure

Our leading interpretation of our model is that all of the evidence is revealed after the DM communicates. However, in practice, some evidence is often revealed before the DM has a chance to take a stand, e.g., an investigation into a political scandal could leak details before the inquiry is formally announced, or firms could publicly disclose financial records before announcing their intention to apply for a merger. This section takes this interpretation seriously and asks how the timing of evidence disclosure affects outcomes.

To answer this question, we consider a version of our baseline model with two stages of evidence disclosure. Before the DM sends a message, they observe an initial public evidence state $e_0 \sim F_0$. After the message is sent, the final evidence $e_1 \sim F_1(\cdot|e_0)$ is realized, and an action is chosen. The preferences of the DM are the same as in Section 2 with only the final evidence e_1 being payoff relevant. Let $\overline{F}$ be the unconditional distribution of e_1.[36] We maintain the focus on ex-ante signaling equilibria in each subgame following the realization of e_0, and so our results also apply to the commitment model.

Consider different (F_0, F_1) with the same $\overline{F}$. By varying F_0, we can span various timings of evidence disclosure. When F_0 is degenerate, all information is "back-loaded" until after the DM communicates, in which case equilibrium outcomes correspond to those under ex-ante signaling in our baseline model. When F_1 is degenerate, all information is "front-loaded" to before communication, in which case equilibrium outcomes correspond to those under ex-post signaling in our original model. That is, even though we focus on the ex-ante signaling equilibrium conditional on e_0, front-loading disclosure generates ex-post signaling outcomes due to the fact that when the evidence distribu-

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).

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

tion is degenerate, ex-ante signaling and ex-post signaling are identical. Our next result examines how the timing of information disclosure impacts the actions taken.

Proposition 9. *Among all F_0 and F_1 with the same $\overline{F}$, $F_0 = \overline{F}$ delivers the lowest probability of $a = 1$, and $F_1(\cdot|\cdot) = \overline{F}$ delivers the highest probability of $a = 1$.*

This result follows from the convexity of $U_P^{\alpha}(\cdot)$. Thus, delaying evidence disclosure (while keeping the final distribution of e_1 constant) hurts P and leads to a higher probability of taking the action.

6.3. Reputation for Standards

We now extend the model to allow the DM to differentially value his reputation for appearing as specific s types. For some $r : \Theta \to \mathbb{R}_+$, let $\rho \int_\Theta r(\theta) d\nu(\theta)$ be the reputation payoff when the public holds beliefs $\nu \in \Delta(\Theta)$. We normalize $r(P) = 0$. Let $\underline{r} \equiv \inf_{s \in S} r(s)$ and $\overline{r} \equiv \sup_{s \in S} r(s)$. We adapt Assumption 1 as follows.

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

Note that by setting $\underline{r} = \overline{r} = 1$ we recover our baseline model, in which case Assumption 2 is equivalent to Assumption 1. Roughly, Assumption 2 says that the difference between $\underline{r}$ and $\overline{r}$ is not too large relative to the difference between $\underline{r}$ and $r(P) = 0$, i.e., the difference in reputational values for different s types does not trump the DM's reputational concern to avoid appearing partisan. This specification is relatively flexible, e.g., it imposes no monotonicity requirements on $r(s)$ with respect to s. The role of Assumption 2 is identical to that of our original Assumption 1 in our baseline model: it ensures that neither the P type nor the s type will ever fully reveal themselves in equilibrium.

Our appendix proves our main results in this more general environment under Assumption 2. In particular, the statements of results from Section 3, Section 4, Subsection 6.1, and Subsection 6.2 remain unchanged. One exception is our comparative statics results which we keep in the context of our main text model.

7. Conclusion

We introduce the possibility for a decision maker to communicate his intentions before payoff relevant evidence realizes. We show that this communication is permissive: any information structure about standards is feasible in some equilibrium and such information can be highly informative about one's intentions, namely, it can completely

reveal the DM's contingent plan. Our main result compares outcomes across equilibria, establishing that the most informative stands lead the DM to break with his partisan interests most frequently.

A number of questions remain for future work. We have studied the effects of communication prior to the evidence realization and decision for one type of reputational incentives: our DM cares about his reputation for taking the right decision similar to the agent in Morris (2001). But this question is relevant for other signaling interactions and preferences such as that in Spence and Zeckhauser (1971). For example, a college student seeking to signal his ability (i.e., tolerance of difficult classes) can be forced to select their major before or after experimenting with a few courses.

References

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

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.

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

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

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

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

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

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

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

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

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

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.

Francetich, A. and Kreps, D. (2014). Bayesian inference does not lead you astray… on average. *Economics Letters*, 125(3):444–446.

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.

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.

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

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.

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.

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.

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.

Spence, M. and Zeckhauser, R. (1971). Insurance, Information, and Individual Action. *American Economic Review*, 61(2):380–87.

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

U.S. Department of Justice and Federal Trade Commission (2023). Merger guidelines. https://www.ftc.gov/system/files/ftc_gov/pdf/2023_merger_guidelines_final_12.18.2023.pdf. Accessed: 2024-10-23.

Appendix

A. Preliminaries

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

Our first result uses Assumption 2 to place bounds on the reputations that may arise in equilibrium. We say a is *off-path* after m,e if $\int_{\Theta_m} \zeta(a|\theta,m,e) d\nu_1(\theta|m) = 0$.[37] We say that a is on-path after m,e if it is not off-path.

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

Proof. First, we show that $U_{P,m}^{\mathcal{E}} \in [-c + \rho q_m \underline{r}, \rho q_m \overline{r}]$ for all $m \in M$. As shown in Francetich and Kreps (2014), conditional on m,e and $\theta = P$, the expected public belief that $\theta \in S$, namely $\sum_{a\in\{0,1\}} q(m,e,a)\zeta(a|P,m,e)$, is at most q_m. Using $R(m,e,a) \leq$

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

$\overline{r}q(m,e,a)$, we then have

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

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

We next derive similar bounds for the expected equilibrium payoff $U^{\mathcal{E}}_P$. Bayes plausibility implies $q_m \leq q$ for some $m \in M^*_P$. For $m \in M^*_P$, $U^{\mathcal{E}}_{P,m} = U^{\mathcal{E}}_P$, which along with $U^{\mathcal{E}}_{P,m} \leq \rho q_m \overline{r} \leq \rho q \overline{r}$ gives our desired upper bound. Bayes plausibility also implies $q_{m'} \geq q$ for some $m' \in M$. Because $U^{\mathcal{E}}_{P,m'} \leq U^{\mathcal{E}}_P$, our desired lower bound follows from $U^{\mathcal{E}}_{P,m'} \geq -c + \rho q_{m'}\underline{r} \geq -c + \rho q \underline{r}$.

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

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

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

Next, take $m \in M^*_P, e \in E$ and suppose $q(m,e,a) = 0$ for some on-path a following m,e, which implies $\zeta(a|P,m,e) > 0$. P's utility from taking action a is then $-ca \leq 0$. Let $a' \neq a$. Because $q_m > 0$, Bayes plausibility requires $q(m,e,a') \geq q_m$, so $R(m,e,a') \geq q(m,e,a')\underline{r} \geq q_m\underline{r}$. For a to be an equilibrium action for P, it must be that $-ca' + \rho q_m \underline{r} \leq -ca$; simplifying, we get $\rho q_m \leq \frac{c(a'-a)}{\underline{r}} \leq \frac{c}{\underline{r}}$. Combining this inequality with $\rho q_m \geq \frac{\rho q\underline{r}-c}{\overline{r}}$

and simplifying, we have $\rho \leq \frac{c(\overline{r}+\underline{r})}{q\underline{r}^2}$, a contradiction of Assumption 2. We conclude that $q(m,e,a) > 0$. *Q.E.D.*

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

Take any a that is off-path following some m, e (with $a' = 1-a$). The reputation payoff from a' is then the interim reputation $\mathbb{E}[r(\theta)|m] = \int_\Theta r(\theta) d\nu_1(\theta|m)$, so the equilibrium payoff for each $\theta' \in \Theta_m$ following m, e is $u(\theta', a', e, \mathbb{E}[r(\theta)|m])$.[38] Suppose there exists non-empty $\Theta'_m \subset \Theta_m$ such that, for all $\theta'' \in \Theta_m \backslash \Theta'_m$, there exists $\theta' \in \Theta'_m$ for which

$$\begin{aligned} &\{\nu \in \Delta(\Theta_m) : \ u\big(\theta'', e, a, \int_{\Theta_m} r(\theta) d\nu(\theta)\big) > u\big(\theta'', e, a', \mathbb{E}[r(\theta)|m]\big)\} \\ &\subsetneq \{\nu \in \Delta(\Theta_m) : \ u\big(\theta', e, a, \int_{\Theta_m} r(\theta) d\nu(\theta)\big) > u\big(\theta', e, a', \mathbb{E}[r(\theta)|m]\big)\}. \end{aligned} \tag{2}$$

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

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

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

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

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

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

D1 requires $\nu_2(P|m,e,a) = 0$ (which implies $q(m,e,a) = 1$) if (2) holds for $\theta'' = P$ and some $\theta' \in S_m$, which simplifies to $(e - \tilde{e}_s)(a' - a) < 0$ for some $s \in S_m$. Similarly, D1

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

requires $\nu_2(S_m|m,e,a) = 0$ (which implies $q(m,e,a) = 0$) if (2) holds for $\theta' = P$ and all $\theta'' \in S_m$, which simplifies to $(e - \tilde{e}_s)(a' - a) > 0$ for all $s \in S_m$. *Q.E.D.*

B. Proofs from Section 3

Proof of Lemma 1

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

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

$$(e - s)a' + \rho R(m,e,a') \geq (e - s)a + \rho R(m,e,a). \tag{3}$$

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

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

Adding (4) to (3) and simplifying yields $(e - \tilde{e}_s)(a' - a) \geq 0$, a contradiction. Therefore, $\zeta(a|P,m,e) = 0$. If a is on-path, then $q(m,e,a) = 1$. If a is off-path, then, by Lemma 6, $q(m,e,a) = 1$ because $(e - \tilde{e}_s)(a' - a) < 0$. But $q(m,e,a) = 1$ contradicts Lemma 5. Therefore, $\zeta(a'|s,m,e) = 0$, i.e., $\zeta(x_s(e)|s,m,e) = 1$.

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

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

a contradiction of Lemma 5. *Q.E.D.*

Proof of Lemma 2

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

Let $\overline{q} = \frac{\rho \underline{r} - c}{\rho \overline{r}}$ and, adopting the convention that $\frac{0}{0} = 0$, define

$$\tilde{R}_a(e, z; \tilde{q}, \tilde{G}) \equiv \begin{cases} \frac{\tilde{q} \int_S r(s) \mathbb{1}(e \geq \tilde{e}_s) d\tilde{G}(s)}{\tilde{q}\tilde{G}(e+c) + (1-\tilde{q})z} & \text{if } a = 1, \\ \frac{\tilde{q} \int_S r(s) \mathbb{1}(e < \tilde{e}_s) d\tilde{G}(s)}{\tilde{q}(1-\tilde{G}(e+c)) + (1-\tilde{q})(1-z)} & \text{if } a = 0. \end{cases}$$

For $\tilde{q} \in (0, \overline{q})$ and $\tilde{G}$ a CDF over S, define $z(\cdot; \tilde{q}, \tilde{G})$ in the following way. For e such that $\tilde{G}(e+c) = 0$, set $z(e; \tilde{q}, \tilde{G}) = 0$ and for e such that $\tilde{G}(e+c) = 1$, set $z(e; \tilde{q}, \tilde{G}) = 1$. For e such that $\tilde{G}(e+c) \in (0, 1)$, let $z(e; \tilde{q}, \tilde{G})$ be the solution to

$$\rho \tilde{R}_1(e, z, \tilde{q}, \tilde{G}) - c = \rho \tilde{R}_0(e, z, \tilde{q}, \tilde{G}). \tag{5}$$

We now show such a solution z exists, is unique and in $(0, 1)$. Note that $\rho \tilde{R}_1(e, 0, \tilde{q}, \tilde{G}) - c \geq \rho \underline{r} - c$ and $\rho \tilde{R}_0(e, 0, \tilde{q}, \tilde{G}) \leq \rho \tilde{q} \overline{r}$. By $\tilde{q} < \overline{q}$, at $z = 0$, the LHS of (5) is strictly greater than the RHS. Moreover, $\rho \tilde{R}_1(e, 1, \tilde{q}, \tilde{G}) - c \leq \rho \tilde{q} \overline{r} - c$ and $\rho \tilde{R}_0(e, 1, \tilde{q}, \tilde{G}) \geq \rho \underline{r}$. By $\tilde{q} < \overline{q}$, at $z = 1$, the LHS of (5) is strictly less than the RHS. Because $\tilde{R}_1$ is continuous and strictly decreasing in z ande $\tilde{R}_0$ is continuous and strictly increasing in z, there exists a unique $z \in (0, 1)$ such that (5) holds. It is immediate that $z(e; \tilde{q}, \tilde{G})$ is continuous in $\tilde{q}$.

With some abuse of notation, let $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ be equal to $\tilde{R}_a(e, z(e; \tilde{q}, \tilde{G}); \tilde{q}, \tilde{G})$. For an arbitrary $\tilde{q} \in (0, \overline{q})$ and CDF $\tilde{G}$ on S, define

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

Given our constructed strategy, this will correspond to the P type's utility after evidence realization e and having induced interim beliefs associated with $(\tilde{q}, \tilde{G})$ at the messaging

stage. We then define the expected payoff from w as

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

Our next result gives some properties of W.

Claim 1. *$W(\tilde{q};\tilde{G})$ is continuous and strictly increasing in $\tilde{q}$ with $W(\tilde{q};\tilde{G}) \in [\rho\tilde{q}\underline{r} - c, \rho\tilde{q}\overline{r}]$ for $\tilde{q} \in (0,\overline{q})$.*

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

We now show $w(e;\tilde{q},\tilde{G}) \in [\rho\tilde{q}\underline{r} - c, \rho\tilde{q}\overline{r}]$ (which immediately implies W respects the same bounds). That these bounds hold for w when $\tilde{G}(e+c) \in \{0,1\}$ is obvious. Take e such that $\tilde{G}(e+c) \in (0,1)$. From (5), we have

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

These inequalities imply $\tilde{R}_1(e;\tilde{q},\tilde{G}) \geq \tilde{q}\underline{r}$ and $\tilde{R}_0(e;\tilde{q},\tilde{G}) \leq \tilde{q}\overline{r}$. Our desired bounds then follow from $w(e;\tilde{q},\tilde{G}) = \rho\tilde{R}_0(e;\tilde{q},\tilde{G}) = \rho\tilde{R}_1(e;\tilde{q},\tilde{G}) - c$. *Q.E.D.*

We construct P's messaging strategy by specifying a Radon-Nikodym derivative $\psi(\cdot)$ and defining $\sigma(\cdot|P)$ via $\sigma(\hat{M}|P) = \int_{\hat{M}} \psi(m)d\Sigma_N(m)$ for any Borel $\hat{M} \subseteq M$. When such strategies are used, what will be the interim belief q_m for $m \in M_\Lambda$ is given by $\varphi(\psi(m)) \equiv \frac{q}{q+(1-q)\psi(m)}$. These will correspond to "on-path" interim updates following m. For $\nu \in \Delta(S)$, we let G_ν be the cdf over S corresponding to ν. We note that for all $t > \underline{t} \equiv \frac{q(1-\overline{q})}{\overline{q}(1-q)}$, we have $\varphi(t) < \overline{q}$.

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

$$\begin{aligned} &\lim_{t\downarrow\underline{t}} W(\varphi(t),G_\nu) \geq \lim_{t\downarrow\underline{t}} \rho\varphi(t)\underline{r} - c = \rho\overline{q}\underline{r} - c = \frac{\rho\underline{r}^2 - c\underline{r}}{\overline{r}} - c > \rho q\overline{r}, \\ &\lim_{t\to\infty} W(\varphi(t),G_\nu) \leq \lim_{t\to\infty} \rho\varphi(t)\overline{r} = 0 < \rho q\underline{r} - c. \end{aligned} \tag{6}$$

For $U \in [\rho q\underline{r} - c, \rho q\overline{r}]$, define $\psi^*(U;m_\nu)$ to be the value of t such that $U = W(\varphi(t),G_\nu)$. We note that such an t exists and is unique follows from (6) and the fact that $W(\cdot,G_\nu)$ is continuous. Because, in addition, $\varphi(\cdot)$ is continuous and strictly decreasing, $\psi^*(U;m_\nu)$

is continuous and strictly decreasing in U. By (6), this implies $\psi^*(U;m_\nu) > \underline{t}$ for all $U \in [\rho q\underline{r} - c, \rho q\overline{r}]$ (and hence $\varphi(\psi^*(U;m_\nu)) \in (0,\overline{q})$).

Claim 2. *There exists a unique $U^* \in [\rho q\underline{r} - c, \rho q\overline{r}]$ such that $1 = \int_{M_\Lambda} \psi^*(U^*;m_\nu)d\Sigma_N(m_\nu)$. Moreover, $\rho\varphi(\psi^*(U^*;m_\nu))\underline{r} - c > 0$ for all $m_\nu \in M_\Lambda$.*

Proof. Take any $m_\nu \in M_\Lambda$. We note that $\varphi(t) \lesseqgtr q$ if and only if $1 \lesseqgtr t$. Let $U = \rho q\underline{r} - c$. Because $W(\tilde{q};G_\nu) \geq \rho\tilde{q}\underline{r} - c$ for all $\tilde{q} \in (0,\overline{q})$, we have

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

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

Let $U' = \rho q\overline{r}$. Because, $W(\tilde{q};G_\nu) \leq \rho\tilde{q}\overline{r}$ for all $\tilde{q} \in (0,\overline{q})$, we have

$$\rho q\overline{r} = U' = W(\varphi(\psi^*(U';m_\nu)),G_\nu) \leq \rho\varphi(\psi^*(U';m_\nu))\overline{r}.$$

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

Because $\psi^*(U^*;m)$ can not be strictly greater than one for all $m \in M_\Lambda$, there exists $m_{\nu'} \in M_\Lambda$ such that $\varphi(\psi^*(U^*;m_{\nu'})) \geq q$. By Claim 1, we have

$$\rho q\underline{r} - c \leq \rho\varphi(\psi^*(U^*;m_{\nu'}))\underline{r} - c \leq U^* \leq \rho\varphi(\psi^*(U^*;m_\nu))\overline{r}$$

which implies $\rho\varphi(\psi^*(U^*;m_\nu)) \geq \frac{\rho q\underline{r}-c}{\overline{r}}$. Then $\rho\varphi(\psi^*(U^*;m_\nu))\underline{r} - c > 0$ if $\frac{\rho q\underline{r}-c}{\overline{r}}\underline{r} > c$ or $\rho > \frac{c(\overline{r}+\underline{r})}{q\underline{r}^2}$, which holds by Assumption 2. *Q.E.D.*

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

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

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

The decision-stage strategies are given by

$$\zeta(1|s,m,e) = \begin{cases} x_s(e) & \text{if } m = m_\nu \in M_\Lambda,\ s \in S_m, \\ \mathbb{1}\big(1 \in \arg\max_a\ u(s,e,a,\tilde{R}_a(e;\nu_1(S|m_\nu),G_\nu))\big) & \text{if } m = m_\nu \in M_\Lambda,\ s \notin S_m, \\ \mathbb{1}\big(1 \in \arg\max_a\ u(s,e,a,0)\big) & \text{if } m \notin M_\Lambda, \end{cases}$$

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

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

Our next claim verifies that $\mathcal{E}$ is an equilibrium.

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

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

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

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

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

Because P is indifferent across all $m' \in M_\Lambda$ and, following m', using strategy x_s for all $s \in S_{m'}$, P (weakly) prefers to send m_ν and follow with x_s than send m' and follow with x':

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

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

Finally, we show that D1 is satisfied. It is trivially satisfied following $m \notin M_\Lambda$ since $\nu_1(P|m) = 1$.[39] Take $m_\nu \in M_\Lambda$. The only off-path actions following m_ν occur when $G_\nu(e+c) \in \{0,1\}$ by construction. If $G_\nu(e+c) = 1$, then $a = 0$ is an off-path action. There are two cases to consider: when $e > \max_{s\in S_{m_\nu}} \tilde{e}_s$ and when $e = \max_{s\in S_{m_\nu}} \tilde{e}_s$. In the first case, by Lemma 6, D1 requires $\nu_2(P|m_\nu, e, a) = 1$ because $e - \tilde{e}_s > 0$ for all $s \in S_{m_\nu}$. For the second case, we now show that $\nu_2(P|m_\nu, e, a) = 1$ is consistent with D1. D1 requires no weight be placed on any $s \in \Theta_m$ whenever P has a larger incentive to deviate to a than s, namely

$$\begin{aligned}\{\nu' \in \Delta(\Theta_{m_\nu}) : \ (e-s)a' + \rho\int_{\Theta_{m_\nu}} r(\theta)d\nu'(\theta) > (e-s)a + \rho\int_{\Theta_{m_\nu}} r(\theta)d\nu_1(\theta|m_\nu)\} \\ \subsetneq \{\nu' \in \Delta(\Theta_{m_\nu}) : \ -ca' + \rho\int_{\Theta_{m_\nu}} r(\theta)d\nu'(\theta) > -ca + \rho\int_{\Theta_{m_\nu}} r(\theta)d\nu_1(\theta|m_\nu)\},\end{aligned}$$

which rules out all $s < \max S_{m_\nu}$ when $e = \max_{s\in S_{m_\nu}} \tilde{e}_s$. However, the above sets are equal for $s' = \max S_{m_\nu}$ at such e, in which case any beliefs that ascribe probability only on s' and P are consistent with D1. Thus, $\nu_2(P|m_\nu, e, a) = 1$ is consistent with D1. An

[39] This triviality comes from the fact that our D1 refinement is specified for interim beliefs. Because $\nu_1(P|m) = 1$ after $m \notin M_\Lambda$, there is no uncertainty at the interim stage, and so our D1 refinement has no bite. In general, our D1 refinement cannot restrict the beliefs for actions following "off-path messages" when they place full weight on a single type, so a similar conclusion can be shown to hold for other "ex-ante" D1 refinements.

analogous argument holds for when $G_\nu(e+c)=0$. *Q.E.D.*

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

C. Proofs from Section 4

Proof of Lemma 3

The proof of Lemma 3 follows immediately from its generalization below.

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

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

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

The same equality holds if we replace s with P. Taking expectations of both sides with

respect to $\nu_1(\cdot|m)$ and using the law of iterated expectations then yields

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

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

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

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

Proof of Lemma 4

Proof. We first derive an equation for determining $U_P^{\alpha}(F)$. Because of the uniqueness in Lemma 2, it is without loss to focus on our constructed equilibrium in the proof of that lemma for the perfectly informative ISS. Under this equilibrium each message in M_Λ is associated with a single non-partisan type s, denote it m_s, in the sense that $S_{m_s} = \{s\}$. Both s and P follow up each m_s with x_s. This means that their reputation after m, e is $R(m_s, e, x_s(e)) = \nu_1(s|m_s)$. Because $\Sigma_N(\cdot)$ and $\sigma(\cdot|P)$ are mutually absolutely continuous, we can describe P's messaging strategy by the Radon-Nikodym derivative $\psi(m_s) = \frac{d\sigma(m_s|P)}{d\Sigma_N(m_s)}$ so that $\sigma(\hat{M}|P) = \int_{\hat{M}} \psi(m_s) d\Sigma_N(m_s)$ for each $\hat{M} \subseteq M_\Lambda$. Thus, by Bayes rule, $R(m_s, e, x_s(e)) = \frac{qr(s)}{q+(1-q)\psi(m_s)}$ for all e. P's expected material payoff from x_s is $-c(1-F(\tilde{e}_s))$ and so his utility is given by

$$U_P^{\alpha}(F) = -c(1-F(\tilde{e}_s)) + \rho \frac{qr(s)}{q+(1-q)\psi(m_s)} \ \forall s \in S.$$

We then have $q + (1-q)\psi(m_s) = \frac{\rho q r(s)}{U_P^{\alpha}(F) + c(1-F(\tilde{e}_s))}$. Taking the expectation over both sides with respect to s and using, by $\sigma(m_s|s') = \mathbb{1}(s'=s)$, $\int_S \psi(m_s) dG(s) = \int_S \psi(m_s) d\Sigma_N(m_s) =$

$\int_M d\sigma(m|P) = 1$, we have

$$1 = \int_S \frac{\rho qr(s)dG(s)}{U_P^\alpha(F) + c - cF(\tilde{e}_s)}. \tag{7}$$

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 (7), we then have

$$\begin{aligned}
&\int_S \frac{\rho qr(s)dG(s)}{U_P^\alpha(F_\lambda) + c - cF_\lambda(\tilde{e}_s)} \\
&= \lambda \int_S \frac{\rho qr(s)dG(s)}{U_P^\alpha(F_1) + c - cF_1(\tilde{e}_s)} + (1-\lambda)\int_S \frac{\rho qr(s)dG(s)}{U_P^\alpha(F_2) + c - cF_2(\tilde{e}_s)} \\
&\geq \int_S \frac{\rho qr(s)dG(s)}{\lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2) + c - c(\lambda F_1(\tilde{e}_s) + (1-\lambda)F_2(\tilde{e}_s))} \\
&= \int_S \frac{\rho qr(s)dG(s)}{\lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2) + c - cF_\lambda(\tilde{e}_s)},
\end{aligned} \tag{8}$$

where the inequality follows from the fact that $\frac{1}{y}$ is convex in y. Thus, (8) implies $U_P^\alpha(F_\lambda) \leq \lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2)$. *Q.E.D.*

Proof of Theorem 1

Claim 4. *$U_P^\alpha(F) \leq U_P^\beta(F)$, with strict inequality if there is residual strategic uncertainty and mild agreement.*

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

$$U_P^\alpha(F) \leq \int_E U_P^\alpha(\delta_e)dF(e) = \int_E U_P^\beta(\delta_e)dF(e) = U_P^\beta(F). \tag{9}$$

Note that the inequality in (8) is strict if there exists $S' \subseteq S$ such that $\int_{S'} dG(s) > 0$ and $F_1(\tilde{e}_s) \neq F_2(\tilde{e}_s)$ for all $s \in S'$, in which case we have $U_P^\alpha(F_\lambda) < \lambda U_P^\alpha(F_1) + (1-\lambda)U_P^\alpha(F_2)$. In (9), we are taking a convex combination over δ_e, so the inequality is strict if there exists S', E' such that $\int_{S'} dG(s) > 0$, $\int_{E'} dF(e) > 0$ and $\delta_e(\tilde{e}_s) = \mathbb{1}(e \geq \tilde{e}_s) \neq \mathbb{1}(e' \geq \tilde{e}_s) = \delta_{e'}(\tilde{e}_s)$ for all $s \in S'$ and $e, e' \in E'$. Suppose β has residual strategic uncertainty and mild agreement holds and that no such S', E' exist. Then for a probability one set of s types, either $F(\tilde{e}_s) = 0$ or $F(\tilde{e}_s) = 1$. If $F(\tilde{e}_s) = 0$ for a probability one set of s, then there is no residual strategic uncertainty, a contradiction. A symmetric argument holds if $F(\tilde{e}_s) = 1$ for a probability one set of s. Therefore, there must exist a positive probability set of s' such that $F(\tilde{e}_{s'}) = 0$ and a positive probability set of s'' such that

$F(\tilde{e}_{s''}) = 1$. But then there is no $e \in \text{Supp}(F)$ for which $x_{s'}(e) = x_{s''}(e)$, a contradiction of mild agreement. Thus, under mild agreement and residual strategic uncertainty for β, $U_P^{\alpha}(F) < U_P^{\beta}(F)$. *Q.E.D.*

We now turn to the proof of Theorem 1.

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

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

Define $U_P^{\alpha,m}(F)$ to be the (unique) value of U that solves $\int_S \frac{\rho q_m r(s)}{U+c-cF(\tilde{e}_s)}dG_m(s) = 1$.[40] We now show $U_P^{\beta,m}(\delta_e) = U_P^{\alpha,m}(\delta_e)$. It suffices to show $\int_S \frac{\rho q_m r(s)}{U_P^{\beta,m}(\delta_e)+c-c\delta_e(\tilde{e}_s)}dG_m(s) = 1$. Suppose $G_m(e+c) \in (0,1)$. Let z be the probability (in $\mathcal{E}$) that P selects $a = 1$ after m and e; by Lemma 1, $z \in (0,1)$. Then $a = 0$ is an optimal action for P, so $U_P^{\beta,m}(\delta_e) = \frac{\rho q_m \int_S r(s)\mathbb{1}(e<\tilde{e}_s)dG_m(s)}{q_m(1-G_m(e+c))+(1-q_m)z}$, which implies $q_m(1-G_m(e+c)) + (1-q_m)z = \frac{\rho q_m \int_S r(s)\mathbb{1}(e<\tilde{e}_s)dG_m(s)}{U_P^{\beta,m}(\delta_e)}$. Similarly, because $a = 1$ is also an optimal action, $U_P^{\beta,m}(\delta_e) = \frac{\rho q_m \int_S r(s)\mathbb{1}(e\geq\tilde{e}_s)dG_m(s)}{q_m G_m(e+c)+(1-q_m)z} - c$, which implies $q_m G_m(e+c) + (1-q_m)z = \frac{\rho q_m \int_S r(s)\mathbb{1}(e\geq\tilde{e}_s)dG_m(s)}{U_P^{\beta,m}(\delta_e)+c}$. Adding these together, we have

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

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

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

By the arguments made in Lemma 4, $U_P^{\alpha,m}(\cdot)$ is convex and so,[41] for all $m \in M_P^*$,

$$U_P^{\alpha,m}(F) \leq \int_E U_P^{\alpha,m}(\delta_e)dF(e) = \int_E U_P^{\beta,m}(\delta_e)dF(e) = U_P^{\beta,m}(F) = U_P^{\mathcal{E}}(F). \tag{10}$$

For the sake of contradiction, suppose $U_P^{\alpha}(F) > U_P^{\mathcal{E}}(F)$. Then, by (10), $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ for all $m \in M_P^*$. By Lemma 1, $\Sigma_N(M_P^*) = \sigma(M_P^*|P) = 1$, we can take the expectation over $m \in M_P^*$ of both sides of $\int_S \frac{\rho q_m r(s)}{U_P^{\alpha,m}(F)+c-cF(\tilde{e}_s)} dG_m(s) = 1$ to get

$$\begin{aligned}
1 &= \int_{M_P^*} \left[\int_S \frac{\rho q_m r(s) dG_m(s)}{U_P^{\alpha,m}(F) + c - cF(\tilde{e}_s)} \right] (q d\Sigma_N(m) + (1-q) d\sigma(m|P)) \qquad (11)\\
&= \int_S \int_{M_P^*} \frac{\rho q r(s)}{U_P^{\alpha,m}(F) + c - cF(\tilde{e}_s)} d\sigma(m|s) dG(s),\\
&> \int_S \int_{M_P^*} \frac{\rho q r(s)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} d\sigma(m|s) dG(s)\\
&= \int_S \frac{\rho q r(s)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} dG(s)\\
&= 1,
\end{aligned}$$

where the second equality follows from Bayes rule, the inequality from $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ and the final equality from (7), a contradiction. Therefore $U_P^{\alpha}(F) \leq U_P^{\mathcal{E}}(F)$.

Finally, suppose there is mild agreement and residual strategic uncertainty in $\mathcal{E}$ and $U_P^{\alpha}(F) = U_P^{\mathcal{E}}(F)$. As we have shown after Lemma 4, mild agreement and residual strategic uncertainty implies $\int_E U_P^{\beta,m}(\delta_e)dF(e) > U_P^{\alpha,m}(F)$ for a positive probability set of messages so, by (10), $U_P^{\alpha}(F) \geq U_P^{\alpha,m}(F)$ with strict inequality for a positive probability set of messages. The same arguments as above in (11) lead to a contradiction. *Q.E.D.*

Proof of Proposition 1

Proof. For notational simplicity, we drop dependence of v^{α} on F. Take any $e \in E$. The proof is immediate if $G(e+c) = 0$ as $v^{\alpha}(e) = v^{\beta}(e) = 0$ or if $G(e+c) = 1$ as $v^{\alpha}(e) = v^{\beta}(e) = 1$. Suppose $G(e+c) \in (0,1)$. Note that $v^{\alpha}(e) = \int_S \mathbb{1}(e \geq \tilde{e}_s)(q dG(s) + (1-q)d\sigma(m_s|P))$. Let $\psi(m_s) = \frac{d\sigma(m_s|P)}{d\Sigma_N(m_s)}$ be the Radon-Nikodym derivative as in the proof of Lemma 4. Using the fact that $\sigma(m_s|s') = \mathbb{1}(s = s')$ under ex-ante signaling, we have $v^{\alpha}(e) = \int_{-\infty}^{e+c}(q + (1-q)\psi(m_s))dG(s)$. As shown in the proof of Lemma 4,

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

$q + (1-q)\psi(m_s) = \frac{\rho q r(s)}{U_P^\alpha(F)+c-cF(\tilde{e}_s)}$, so $v^\alpha(e) = \int_S \frac{\rho q r(s)\mathbb{1}(e \geq \tilde{e}_s)}{U_P^\alpha(F)+c-cF(\tilde{e}_s)} dG(s)$.

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

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

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

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

It is without loss to focus on F such that $\mathrm{Supp}(F)$ is contained in a compact interval.[42] Take some $s'' < \min \mathrm{Supp}(F) + c$ and $s' > \max \mathrm{Supp}(F) + c$. Define $(\widehat{G}, \hat{r})$ by

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

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

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

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

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

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

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

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

Rearranging terms, we get

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

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

Proof of Corollary 1

Proof. Let $M^*_\theta, M^{*'}_\theta$ and ν_1, ν_1' be the optimal messages for θ and interim-beliefs in $\mathcal{E}$ and $\mathcal{E}'$ respectively. By Lemma 1, there exists a probability one set $\tilde{S} \subseteq S$ such that, for each $s \in \tilde{S}$, there exists message $m \in M^*_P \cap M^*_s, m' \in M^{*'}_P \cap M^{*'}_s$ such that $s \in \text{Supp}(\nu_1(\cdot|m))$ and $s \in \text{Supp}(\nu_1'(\cdot|m'))$ and s follows the contingent plan x_s after sending m and m' in $\mathcal{E}$ and $\mathcal{E}'$ respectively. By $m \in M^*_s, m' \in M^{*'}_s$, we have

$$\begin{aligned}
U^{\mathcal{E}}_s(F) &= \int_E \big((e-s)x_s(e) + \rho R(m, e, x_s(e))\big) dF(e), \\
U^{\mathcal{E}'}_s(F) &= \int_E \big((e-s)x_s(e) + \rho R'(m', e, x_s(e))\big) dF(e).
\end{aligned}$$

where R and R' are the equilibrium reputation functions in $\mathcal{E}$ and $\mathcal{E}'$ respectively. Thus, $U^{\mathcal{E}}_s(F) - U^{\mathcal{E}'}_s(F) = \rho \int_E (R(m, e, x_s(e)) - R'(m', e, x_s(e))) dF(e)$ for all $s \in \tilde{S}$. By Lemma 1, P is indifferent over actions that are taken with positive probability following any optimal message. Under $\mathcal{E}$, by $s \in \text{Supp}(\nu_1(\cdot|m))$ and $m \in M^*_P$, following x_s is an optimal

contingent plan for P after sending m. A similar argument holds for $\mathcal{E}'$, which implies

$$U_P^{\mathcal{E}}(F) = \int_E \big(- cx_s(e) + \rho R(m, e, x_s(e))\big) dF(e),$$
$$U_P^{\mathcal{E}'}(F) = \int_E \big(- cx_s(e) + \rho R'(m', e, x_s(e))\big) dF(e).$$

These inequalities imply $U_P^{\mathcal{E}}(F) - U_P^{\mathcal{E}'}(F) = \rho \int_E (R(m, e, x_s(e)) - R'(m', e, x_s(e)))dF(e)$, proving the first part of the corollary for $s \in \tilde{S}$; the result extends to all S by noting that $U_s^{\mathcal{E}}(F), U_s^{\mathcal{E}'}(F)$ are continuous in s. The second part follows immediately from the first part of the corollary and the the fact that ex-ante signaling is P's least preferred equilibrium by Theorem 1 and Lemma 7. *Q.E.D.*

D. Proofs from Section 5

For this section we revert to our main text model where $r(s) = 1 \; \forall s \in S$. 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 $\int_S \frac{\rho q}{U + c - cF(\tilde{e}_s)} dG(s) = 1$.

Proof of Proposition 2

Proof. Fix an investigation F and distribution G of s. Taking the derivative of the expression in (7) with respect to ρ, we have

$$-\frac{dU_P^{\alpha}(F)}{d\rho} \int_S \frac{\rho q}{(U_P^{\alpha}(F) + c - cF(\tilde{e}_s))^2} dG(s) + \int_S \frac{q}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} dG(s) = 0. \tag{12}$$

By (7), $\int_S \frac{q}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} dG(s) = \frac{1}{\rho}$. Substituting this into (12) and simplifying, we get

$$\begin{aligned}
\left(\frac{dU_P^{\alpha}(F)}{d\rho}\right)^{-1} &= q \int_S \left(\frac{\rho}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} \right)^2 dG(s) \\
&\geq q \left(\int_S \frac{\rho dG(s)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} \right)^2 \\
&= \frac{1}{q},
\end{aligned}$$

where the inequality follows by Hölder's inequality and the final equality follows from $\int_S \frac{\rho dG(s)}{U_P^{\alpha}(F) + c - cF(\tilde{e}_s)} = \frac{1}{q}$ by (7). 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 shows the result for q. *Q.E.D.*

Proof of Proposition 3

Proof. Fix F and let G_2 FOSD G_1. By Lemma 3, it suffices to show that P's equilibrium expected utility is lower under G_1 than G_2. Let U_i be P's equilibrium expected utility under G_i. For the sake of contradiction, suppose $U_1 > U_2$. We have that

$$\begin{aligned} 1 = \int_S \frac{\rho q}{U_2 + c - cF(\tilde{e}_s)} dG_2(s) &> \int_S \frac{\rho q}{U_1 + c - cF(\tilde{e}_s)} dG_2(s) \\ &\geq \int_S \frac{\rho q}{U_1 + c - cF(\tilde{e}_s)} dG_1(s). \end{aligned} \tag{13}$$

These inequalities hold because the integrand is increasing in s and G_2 FOSD G_1. But (13) contradicts $\int_S \frac{\rho q}{U_1 + c - cF(\tilde{e}_s)} dG_1(s) = 1$. Therefore, $U_2 \geq U_1$. The comparative static for F follows once we note that P, when deciding which s type to mimic, his material payoff is purely determined by the probability of $a = 1$. A FOSD shift upward of G is equivalent to a FOSD shift downward of F for this probability, so the result for F follows from that for G. *Q.E.D.*

Proof of Proposition 4

Proof. Assume $\frac{f(e)}{\rho q + c(1-F(e))}$ is increasing in e on $[\underline{e}, \overline{e}]$. After changing variables, this implies $\frac{f(\tilde{e}_s)}{\rho q + c(1-F(\tilde{e}_s))}$ is increasing in s on $[\underline{e} + c, \overline{e} + c]$. First, we show that $\frac{\rho q}{U_P^\alpha(F) + c(1-F(\tilde{e}_s))}$ is convex in s on $[\underline{e}+c, \overline{e}+c]$. Taking the derivative with respect to s yields $\frac{\rho q c f(\tilde{e}_s)}{(U_P^\alpha(F) + c(1-F(\tilde{e}_s)))^2}$. Because $\frac{\rho q c}{U_P^\alpha(F) + c(1-F(\tilde{e}_s))}$ is increasing in s, convexity follows if $\frac{f(\tilde{e}_s)}{U_P^\alpha(F) + c(1-F(\tilde{e}_s))}$ is increasing in s, which follows from the assumption that $\frac{f(\tilde{e}_s)}{\rho q + c(1-F(\tilde{e}_s))}$ is increasing in s and the fact that $\rho q \geq U_P^\alpha(F)$, which is in turn implied by Lemma 3.

Let $\tilde{G}$ be a mean-preserving spread of G with corresponding utilities $\tilde{U}_P^\alpha(F)$ and $U_P^\alpha(F)$ for P. By convexity of $\frac{\rho q}{U_P^\alpha(F) + c(1-F(\tilde{e}_s))}$, we then have

$$\begin{aligned} \int_S \frac{\rho q}{U_P^\alpha(F) + c(1 - F(\tilde{e}_s))} d\tilde{G}(s) &\geq \int_S \frac{\rho q}{U_P^\alpha(F) + c(1 - F(\tilde{e}_s))} dG(s) \\ &= 1 \\ &= \int_S \frac{\rho q}{\tilde{U}_P^\alpha(F) + c(1 - F(\tilde{e}_s))} d\tilde{G}(s). \end{aligned}$$

This inequality implies $\tilde{U}_P^\alpha(F) \geq U_P^\alpha(F)$, which by Lemma 7 implies the probability of $a = 1$ is lower under $\tilde{G}$ than G. *Q.E.D.*

Proof of Proposition 5

Proof. Let $F_\lambda \equiv \lambda\tilde{F} + (1-\lambda)F$. Given Lemma 3, the proposition follows immediately if $\frac{dU_P^\alpha(F_\lambda)}{d\lambda} < 0$ for all $\lambda \in [0,1]$. First, we note that $\frac{h(e)}{(U_P^\alpha(F_\lambda)+(1-F_\lambda(e)))^2}$ is increasing in $[e_1, e_2]$: because $\frac{1}{U_P^\alpha(F)+c(1-F(e))}$ is increasing in e, this follows from the assumption that $\frac{h(e)}{\rho q + c(1-F(e))}$ and, as is implied by Lemma 3, $\rho q \geq U_P^\alpha(F)$. We have shown that $1 = \int_E \frac{\rho q h(e)}{U_P^\alpha(F_\lambda)+(1-F_\lambda(e))} de$. Taking the derivative of both sides with respect to λ and rearranging yields

$$\begin{aligned}
&\frac{dU_P^\alpha(F_\lambda)}{d\lambda}\int_E \frac{h(e)}{(U_P^\alpha(F_\lambda)+(1-F_\lambda(e)))^2}de \\
=&\int_{e_1}^{e_2}(\tilde{F}(e)-F(e))\frac{h(e)}{(U_P^\alpha(F_\lambda)+(1-F_\lambda(e)))^2}de \\
=&-\int_{e_1}^{e_2}\frac{d}{de}\left(\frac{h(e)}{(U_P^\alpha(F_\lambda)+(1-F_\lambda(e)))^2}\right)\left(\int_{e_1}^{e}(\tilde{F}(\tilde{e})-F(\tilde{e}))d\tilde{e}\right)de \\
<&0.
\end{aligned}$$

The second equality follows from integration by parts, and the inequality follows from the fact that $\frac{h(e)}{(U_P^\alpha(F_\lambda)+(1-F_\lambda(e)))^2}$ is increasing on $[e_1, e_2]$ and, because $\tilde{F}$ is an mean-preserving spread of F and equal to F outside of $[e_1, e_2]$, we have $\int_{e_1}^{e}(\tilde{F}(e)-F(e))de \geq 0 \ \ \forall e \in [e_1, e_2]$ with equality at e_2. This inequality implies $\frac{dU_P^\alpha(F_\lambda)}{d\lambda} < 0$. *Q.E.D.*

Proof of Proposition 6

Proof. Suppose that F has a mass point of size Δ at e'. Define $F_{\varepsilon,\delta}(e) = F(e) + \frac{\delta}{2}\mathbb{1}(e \in [e'-\varepsilon, e')) - \frac{\delta}{2}\mathbb{1}(e \in [e', e'+\varepsilon))$, which is a MPS of F for all $\varepsilon, \delta > 0$. Taking the derivative of $\int_E \frac{\rho q h(e)}{U_P^\alpha(F)+c(1-F_{\varepsilon,\delta}(e))}de$ with respect to δ at $\delta = 0$, yields

$$\begin{aligned}
&\int_{e'-\varepsilon}^{e'}\frac{\rho q h(e)}{2(U_P^\alpha(F)+c(1-F(e)))^2}de - \int_{e'}^{e'+\varepsilon}\frac{\rho q h(e)}{2(U_P^\alpha(F)+c(1-F(e)))^2}de \\
=&\int_{e'-\varepsilon}^{e'}\frac{\rho q h(e')}{2(U_P^\alpha(F)+c(1-F(e)))^2}de - \int_{e'}^{e'+\varepsilon}\frac{\rho q h(e')}{2(U_P^\alpha(F)+c(1-F(e)))^2}de + O(\varepsilon^2) \\
\leq&\frac{\rho q h(e')\varepsilon}{2}\left[\frac{1}{(U_P^\alpha(F)+c(1-F(e')+\Delta))^2} - \frac{1}{(U_P^\alpha(F)+c(1-F(e')))^2}\right] + O(\varepsilon^2).
\end{aligned}$$

The last line is strictly negative for all sufficiently small $\varepsilon > 0$. Then for all sufficiently small $\varepsilon, \delta > 0$ we have

$$\begin{aligned}\int_E \frac{\rho q h(e)}{U_P^\alpha(F) + c(1 - F_{\varepsilon,\delta}(e))} de &< \int_E \frac{\rho q h(e)}{U_P^\alpha(F) + c(1 - F(e))} de \\ &= 1 \\ &= \int_E \frac{\rho q h(e)}{U_P^\alpha(F_{\varepsilon,\delta}) + c(1 - F_{\varepsilon,\delta}(e))} de.\end{aligned}$$

Therefore, $U_P^\alpha(F_{\varepsilon,\delta}) < U_P^\alpha(F)$. By Lemma 7, this implies the probability of $a = 1$ is lower under $F_{\varepsilon,\delta}$ than F. *Q.E.D.*

E. Proofs from Section 6

We first formally define an equilibrium in the commitment model. We endow $\mathcal{X}$ with the metric $d(x, x') = \int_E |x(e) - x'(e)| dF(e)$.[43] An equilibrium is given by a strategy $\xi : \Theta \to \Delta(\mathcal{X})$ and a belief system $\nu : \mathcal{X} \to \Delta(\Theta)$ such that

1. ν is obtained from ξ using Bayes rule whenever possible[44] with $\text{Supp}(\nu(\cdot|x)) \subseteq \{\theta : x \in \text{Supp}(\xi(\cdot|\theta))\}$ if $\{\theta : \ x \in \text{Supp}(\xi(\cdot|\theta))\} \neq \emptyset$,
2. $\xi(\mathcal{X}_\theta^*|\theta) = 1$ where $\mathcal{X}_\theta^* \equiv \arg\max_{x \in \mathcal{X}} \int_E u(\theta, e, x(e), \int_\Theta r(\theta) d\nu(\theta|x)) dF(e)$.

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

$$\begin{aligned}\{\nu \in \Delta(\Theta) : \ &\int_E u\big(\theta'', e, x(e), \int_\Theta r(\theta) d\nu(\theta)\big) dF(e) > U_{\theta''}\} \\ &\subsetneq \{\nu \in \Delta(\Theta) : \ \int_E u\big(\theta', e, x(e), \int_\Theta r(\theta) d\nu(\theta)\big) dF(e) > U_{\theta'}\}.\end{aligned}$$

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

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

[44] That is, for all Borel $\hat{\Theta} \subseteq \Theta$ and $\hat{X} \subseteq \mathcal{X}$, $\int_{\hat{\Theta}} \sigma(\hat{X}|\theta) d\nu_0(\theta) = \int_{\hat{X}} \nu_1(\hat{\Theta}|x) \int_\Theta d\xi(x|\theta) d\nu_0(\theta)$.

Proof of Proposition 7

Proof. Throughout, given an equilibrium (ξ, ν), we denote $R(x) = \int_\Theta r(\theta) d\nu(\theta|x)$ as the reputation payoff for x. Also define $\mathcal{X}_s \equiv \arg\max_x \int_E (c + e - s)x(e)dF(e)$. We split the proof into several steps.

Step 1 (Equilibrium Construction): Let each $\xi(x_s|s) = 1$ for all $s \in S$ and define P's equilibrium mixing strategy $\xi(\cdot|P) \in \Delta(\{x_s\}_{s\in S})$ by the Radon-Nikodym derivative $\frac{d\xi(x_s|P)}{dG(s)} = \frac{q}{1-q}[\frac{\rho r(s)}{U_P^\alpha(F)+c-cF(\tilde{e}_s)} - 1]$.[45] Set equilibrium beliefs $\nu(s|x) = \frac{U_P^\alpha(F)+c-cF(\tilde{e}_s)}{\rho r(s)}$, $\nu(P|x) = 1 - \nu(s|x)$ for $x \in \mathcal{X}_s$, and $\nu(P|x) = 1$ otherwise; this leads to $R(x) = \frac{U_P^\alpha(F)+c-cF(\tilde{e}_s)}{\rho}$ for $x \in \mathcal{X}_s$ and $R(x) = 0$ otherwise. We note that $U_P^\alpha(F) = -c\int_E x_s(e)dF(e) + \rho R(x_s)$ for all $s \in S$.

It is clear that these strategies generate the same outcomes as in ex-ante signaling. By Lemma 2 there is no incentive to deviate to any $x \in \cup_{s\in S}\mathcal{X}_s$, and no incentive to deviate to other x because $R(x) = 0$.[46]

Finally, we show that the off-path reputations are consistent with D1. Take any $x \notin \{x_s\}_{s\in S}$. D1 rules out $\nu(P|x) = 1$ only if there exists an s such that

$$\begin{aligned}&\{\nu \in \Delta(\Theta) : -c\int_E x(e)dF(e) + \rho\int_\Theta r(\theta)d\nu(\theta) \geq U_P^\alpha(F)\}\\ &\subsetneq \{\nu \in \Delta(\Theta) : \int_E (e-s)x(e) + \rho\int_\Theta r(\theta)d\nu(\theta) \geq \int_E (e-s)x_s(e) + \rho R(x_s)\}.\end{aligned}$$

The left-hand side above is non-empty.[47] Using the fact that $U_P^\alpha(F) = -c\int_E x_s(e)dF(e) + \rho R(x_s)$, the above set inclusion is equivalent to

$$\int_E (c + e - s)x_s(e)dF(e) < \int_E (c + e - s)x(e)dF(e),$$

which contradicts $x_s \in \mathcal{X}_s$. Therefore, $\nu(P|x) = 1$ is consistent with D1.

Step Two (Outcome Equivalence): We show that all equilibria are outcome equivalent in two steps. Take any equilibrium with corresponding strategies $\{\xi(\cdot|\theta)\}_{\theta\in\Theta}$ and belief system ν. First, we show that in any equilibrium s types must only choose from

[45] It is straightforward to check that $\int_S d\xi(x_s|P) = 1$ given the definition of $U_P^\alpha(F)$.

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

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

$\mathcal{X}_s$ (i.e., $\xi(\mathcal{X}_s|s) = 1$). Second, we show $\xi(\mathcal{X}_s|P)$ must take the form specified in Step One.

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

Claim 5. *$\nu(S|x) < 1$ for all $x \in \mathcal{X}$.*

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

$$-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho R(x). \tag{14}$$

Using our bounds on $R(x)$ and $R(x')$, this implies $-c + \rho\underline{r} \leq \rho q\overline{r}$, or $\rho \leq \frac{c}{\underline{r}-q\overline{r}} \leq \frac{c(\underline{r}+\overline{r})}{\underline{r}^2-q\overline{r}^2}$, a contradiction of Assumption 2. *Q.E.D.*

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

In the first case, where $\text{cl}(\mathcal{X}_P^*) \cap \mathcal{X}_s \neq \emptyset$, there exists a sequence of $\{x_n'\}_{n=0}^{\infty}$ such that $x_n' \in \mathcal{X}_P^*$ for all n and, for $x' = \lim_{n\to\infty} x_n'$, $x' \in \mathcal{X}_s$. P then weakly prefers x_n' to x and s weakly prefers x to x_n':

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

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

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

$$\begin{aligned} \{\nu \in \Delta(\Theta) : -c\int_E x_s(e)dF(e) + \rho\int_\Theta r(\theta)d\nu(\theta) > -c\int_E x'(e)dF(e) + \rho R(x')\} \\ \subsetneq \{\nu \in \Delta(\Theta) : \int_E (e-s)x_s(e) + \rho\int_\Theta r(\theta)d\nu(\theta) > \int_E (e-s)x(e)dF(e) + \rho R(x)\}. \end{aligned} \tag{15}$$

By analogous arguments to those in Step 1, the left-hand side of (15) is non-empty. Because $x' \in \mathcal{X}_P^*$, we have $-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho R(x)$. Therefore, (15) holds if

$$\{\nu \in \Delta(\Theta) : \ -c\int_E x_s(e)dF(e) + \rho\int_\Theta r(\theta)d\nu(\theta) > -c\int_E x(e)dF(e) + \rho R(x)\}$$
$$\subsetneq \{\nu \in \Delta(\Theta) : \ \int_E (e-s)x_s(e) + \rho\int_\Theta r(\theta)d\nu(\theta) > \int_E (e-s)x(e)dF(e) + \rho R(x)\}.$$

After some simplification, strict inclusion holds if $\int_E (c+e-s)(x_s(e) - x(e))dF(e) > 0$, which holds because $x \notin \mathcal{X}_s$. Thus, $\nu(S|x_s) = 1$, which contradicts Claim 5. Therefore, $\mathcal{X}_s^* \subseteq \mathcal{X}_s$. All $x \in \mathcal{X}_s$ lead to the same actions with probability one if $\tilde{e}_s$ has no mass-point under F. Because F or G is atomless, the set of s for which this is true is a probability one set. Thus, all equilibrium strategies for a probability one set of s types are outcome equivalent to x_s with probability one.

Next, we argue that $\xi(\cdot|P)$ and $\Xi(\cdot) \equiv \int_S \xi(\cdot|s)dG(s)$ must be mutually absolutely continuous. Claim 5 implies that $\Xi(\cdot)$ is absolutely continuous with respect to $\xi(\cdot|P)$. For the sake of contradiction, suppose $\xi(\cdot|P)$ is not absolutely continuous with respect to $\Xi(\cdot)$. Then, because $\xi(\mathcal{X}_P^*|P) = 1$, there exists $\mathcal{X}' \subseteq \mathcal{X}_P^*$ such that $\xi(\mathcal{X}'|P) > 0 = \Xi(\mathcal{X}')$. Then $R(x) = 0$ for some $x \in \mathcal{X}'$ and, because $x \in \mathcal{X}_P^*$, P's equilibrium expected utility is $-c\int_E x(e)dF(e)$. But, by Bayes plausibility, there exists $x' \in \text{Supp}(\Xi)$ such that $\nu(S|x') \geq q$, which implies $R(x') \geq q\underline{r}$, in which case P can achieve a utility of $-c\int_E x'(e)dF(e) + \rho R(x') \geq \rho q\underline{r} - c > 0 \geq -c\int_E x(e)dF(e)$. Thus, choosing x is strictly dominated by x', contradicting $x \in \mathcal{X}_P^*$. Given that all s must choose only from $\mathcal{X}_s$ and, for probability one set of s, all $x \in \mathcal{X}_s$ lead to equivalent actions with probability one, analogous arguments to those in Lemma 2 imply P has a unique mixing strategy over $\mathcal{X}_s$. *Q.E.D.*

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

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

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

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

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

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

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

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

Proof of Proposition 8

In addition to Assumption 2, we impose the following assumption (reduces to our assumption in Proposition 8 when $\overline{r} = \underline{r} = 1$):

Assumption 3. $\rho > 2\max\{\frac{\delta}{q\underline{r}}, \frac{\delta}{\underline{r}-q\overline{r}}\}$.

Proof. We first show equilibrium existence. Take the same strategies (and beliefs following $x \in \mathcal{X}$) as in the commitment model with $\sigma(M|\theta) = 0$ for all θ and set $\nu_1(P|m) = 1$ following any $m \in M$, $\zeta(1|\theta, m, e, a) = \mathbb{1}(1 \in \arg\max_a u(\theta, e, a, 0))$ and $\nu_2(P\times[-\delta, \delta]|m, e, a) = 1$. By Proposition 7, no type has an incentive to deviate to any other commitment $x \in \mathcal{X}$ and the equilibrium satisfies D1. We only need to check that no type has an incentive to deviate to send $m \in M$. For P, his expected equilibrium utility is $U^\alpha_P(F)$ and, as shown in the proof of Lemma 2, $U^\alpha_P(F) \geq \rho q\underline{r} - c$. P's value of m is then at most

$$\int_E \int_{-\delta}^{\delta} \max\{-c + \varepsilon, 0\} d\lambda(\varepsilon) dF(e) \leq \max\{-c + \delta, 0\}.$$

That sending such an m is sub-optimal follows from $\rho q\underline{r} - c > 0$ by $\rho > \frac{c(\underline{r}+\overline{r})}{q\underline{r}^2} > \frac{c}{q\underline{r}}$ (Assumption 2) and $\rho q\underline{r} - c > -c + \delta$ by $\rho > \frac{\delta}{q\underline{r}}$ (Assumption 3). Now consider $s \in S$. By $U^\alpha_P(F) = -c(1 - F(\tilde{e}_s)) + \rho R(x_s)$, we have $\rho R(x_s) \geq U^\alpha_P(F) + c(1 - F(\tilde{e}_s)) \geq \rho q\underline{r} - cF(\tilde{e}_s) \geq \rho q\underline{r} - c$. Type s's utility from sending m is $\int_E \int_{-\delta}^{\delta} \max\{e - s + \varepsilon, 0\} d\lambda(\varepsilon) dF(e) \leq$

$\int_E \max\{e - s + \delta, 0\} dF(e)$. He has no incentive to deviate if

$$\int_E (e - s) x_s(e) dF(e) + \rho R(x_s) \geq \int_E \max\{e - s + \delta, 0\} dF(e).$$

After simplifying and substituting in our bound for $\rho R(x_s)$, it suffices to show $\rho q \underline{r} - c \geq \delta$, or $\rho \geq \frac{\delta + c}{q\underline{r}}$. This follows by Assumption 2 if $c > \delta$ and Assumption 3 if $\delta \geq c$. Therefore, s has no incentive to deviate.

To complete the proof, we only need to prove that all equilibria are outcome equivalent to ex-ante signaling. Take an equilibrium $\mathcal{E}$ and recall that $\Sigma_N(\cdot) = \int_S \sigma(\cdot|s) dG(s)$. If $\Sigma_N(M) = \sigma(M|P) = 0$, then the same arguments as in Proposition 7 show that the equilibrium outcome is equivalent to ex-ante signaling. Suppose that $\Sigma_N(M) > 0$ or $\sigma(M|P) > 0$. We first show $\sigma(\cdot|P)$ and $\Sigma_N(\cdot)$ are mutually absolutely continuous over M. Suppose there exists $M' \subseteq M$ such that $\sigma(M'|P) > 0$ or $\Sigma_N(M') > 0$. If $\Sigma_N(M') = 0 < \sigma(M'|P)$, then for some $m \in M'$, $\nu_1(P|m) = 1$ and the reputation following m is always 0. Thus, P attains a maximum utility of $\max\{-c + \delta, 0\}$ from sending m. However, the P type can attain at least $\rho q \underline{r} - c - \delta$ by mimicking some s type whose expected equilibrium reputation is at least $q\underline{r}$ (such s exist by Bayes plausibility). For mimicking m to be optimal for P, we must have $\rho q \underline{r} - c - \delta \geq \max\{-c + \delta, 0\}$. But this contradicts either $\rho > \frac{2\delta}{q\underline{r}}$ by Assumption 3 or $\rho > \frac{c(\underline{r} + \overline{r})}{q\underline{r}^2}$ by Assumption 2.

If $\sigma(M'|P) = 0 < \Sigma_N(M')$, then there exists $m \in M'$ such that $\nu_1(S|m) = 1$ and the reputation is at least $\underline{r}$ for each action at the decision stage. By Bayes plausibility, there exists $y \in M \cup \mathcal{X}$ such that $\nu_1(S|y) \leq q$. If $y \in \mathcal{X}$, then the reputation following y is at most $q\overline{r}$. If $y \in M$, then the expected equilibrium reputation for P following y is at most $q\overline{r}$ (Francetich and Kreps (2014)). Then m is a profitable deviation from y for any type of DM as they can choose an optimal action for each (e, ε) realization and still have a strictly higher reputation as $q\overline{r} < \underline{r}$ by Assumption 2. Therefore, $\Sigma_N(M') > 0$ if and only if $\sigma(M'|P) > 0$ for all $M' \subseteq M$.

Suppose there exists s and $m \in M$ such that m is an optimal message for s and P (i.e., $m \in \mathcal{Y}_s^* \cap \mathcal{Y}_P^*$) and s or P do not choose actions consistent with $x_s(e)$ following m (for a probability one set of e, ε); namely, $\int_E \int_{-\delta}^{\delta} \zeta(x_s(e)|\theta, m, e, \varepsilon) d\lambda(\varepsilon) dF(e) < 1$ for $\theta \in \{s, P\}$. Take $R(m, e, a) = \int_\Theta \int_{-\delta}^{\delta} r(\theta) d\nu_2((\theta, \varepsilon)|m, e, a)$ and $R(x) = \int_\Theta r(\theta) d\nu_1(\theta|x)$. Now consider the difference in payoff between sending message m in equilibrium and

taking commitment x_s for types s and P as a function of e, ε. For type s, this is given by

$$\int_E \int_{-\delta}^{\delta} \big(\max\{e - s + \varepsilon + \rho R(m, e, 1), \rho R(m, e, 0)\} - (e - s + \varepsilon) x_s(e) - \rho R(x_s) \big) d\lambda(\varepsilon) dF(e), \tag{16}$$

and for P, it is given by

$$\int_E \int_{-\delta}^{\delta} \big(\max\{-c + \varepsilon + \rho R(m, e, 1), \rho R(m, e, 0)\} - (-c + \varepsilon) x_s(e) - \rho R(x_s) \big) d\lambda(\varepsilon) dF(e). \tag{17}$$

Notice that the integrand for s in (16) is weakly less than the integrand for P in (17) for every e, ε and so this comparison holds for the expressions themselves.

Suppose, for the sake of contradiction, that (16) is strictly less than (17). First, consider that the commitment $x_s \in \mathcal{Y}_P^*$. Since $m \in \mathcal{Y}_P^*$, (17) is 0 and so by hypothesis (16) is strictly negative which contradicts that $m \in \mathcal{Y}_s^*$. Therefore, $x_s \notin \mathcal{Y}_P^*$, which we will show implies $\nu_1(P|x_s) = 0$. The only way $\nu_1(P|x_s) > 0$ given $x_s \notin \mathcal{Y}_P^*$ is if $x_s \notin \text{Supp}(\sigma(\cdot|s))$ for all $s \in S$. The strict inequality holds between (16) and (17) for every $R(x_s)$. If $R(x_s) = \bar{r}$, then because P's equilibrium payoff is less than $\rho q \bar{r} + \max\{\delta - c, 0\}$, (17) is less than $\rho q \bar{r} + \max\{\delta - c, 0\} - (\rho \bar{r} - c) < 0$, by Assumption 2 and Assumption 3. If $R(x_s) = 0$ then because P's equilibrium payoff is greater than $\rho q \underline{r} - c - \delta$, (17) is greater than $\rho q \underline{r} - c - \delta > 0$ again by Assumption 2 and Assumption 3. So the s type prefers x_s for a strictly larger set of beliefs than the P type, so by the D1 refinement, $\nu_1(P|x_s) = 0$.

By $\nu_1(P|x_s) = 0$, $R(x_s) \geq \underline{r}$ and P's utility from x_s is at least $\rho \underline{r} - c$. His equilibrium utility is again bounded above by $\rho q \bar{r} + \max\{\delta - c, 0\}$ which we have already shown is less than $\rho \underline{r} - c$ by Assumption 2 and Assumption 3. Thus, P prefer x_s, a contradiction.

Therefore, (16) is equal to (17). Because the integrand in (16) is weakly point wise lower than (17) for each (e, ε), these integrands are equal to a probability one set of (e, ε). Thus, $e > s - c \implies -c + \varepsilon + \rho R(m, e, 1) > \rho R(m, e, 0)$ (namely, $e > \tilde{e}_s$ implies $a = 1$ is optimal for P after m, e, which itself implies $a = 1$ is optimal for s) and $e < s - c \implies -c + \varepsilon + \rho R(m, e, 1) < \rho R(m, e, 0)$ (namely, $e < \tilde{e}_s$ implies $a = 0$ is optimal for P after m, e, which itself implies $a = 0$ is optimal for s) for a probability one set of (e, ε). Thus, if (16) is equal to (17), then $\int_E \int_{-\delta}^{\delta} \zeta(x_s(e)|\theta, m, e, \varepsilon) d\lambda(\varepsilon) dF(e) = 1$ for $\theta \in \{s, P\}$.

Let $t(s) \equiv \max\{t : \ x_t \in \mathcal{X}_s\}$ and $\tilde{M}_s \equiv \{m : \int_E \int_{-\delta}^{\delta} \zeta(x_s(e)|P, m, e, \varepsilon) d\lambda(\varepsilon) dF(e) = 1\}$. Our previous conclusion implies $\sigma(\tilde{M}_{t(s)}|s) = \sigma(M|s)$ for all s: otherwise, if $\sigma(M'|s) > 0$

for some $M' \subseteq M \backslash \tilde{M}_{t(s)}$, there exists $m \in M'$ such that $\int_E \int_{-\delta}^{\delta} \zeta(x_s(e)|P, m, e, \varepsilon) d\lambda(\varepsilon) dF(e) < 1$, which contradicts our previous conclusion.

Next, we show $\nu(S|x) < 1$ for all $x \in \mathcal{X}$. Suppose not, then $R(x) \geq \underline{r}$, so P's expected utility from x is at least $-c + \rho\underline{r}$. By Bayes' plausibility, there exists $y \in \mathcal{Y}_P^*$ such that $\nu_1(S|y) \leq q$. P's expected material payoff following any y is at most $\max\{-c + \delta, 0\}$ and his expected reputational payoff is at most $\rho q \bar{r}$, so his expected utility from y is at most $\max\{-c + \delta, 0\} + \rho q \bar{r}$, which is less than $-c + \rho\underline{r}$ as argued above. Thus, $\nu(S|x) < 1$.

If $x \in \text{Supp}(\sigma(\cdot|s))$ for some $s \in S$, then $x \in \text{Supp}(\sigma(\cdot|P))$; otherwise, $\nu(S|x) = 1$. By analogous arguments as in Proposition 7, $\text{Supp}(\sigma(\cdot|s)) \cap \mathcal{X} \subseteq \mathcal{X}_s$. Thus, for a probability one set of s and e, the distribution over actions is the same as under ex-ante signaling.

Therefore, $\sigma(\tilde{M}_{t(s)} \cup \mathcal{X}_{t(s)}|s) = 1$ for all $s \in S$. Since we established that P also follows x_s with probability one after $y \in \tilde{M}_{t(s)} \cup \mathcal{X}_{t(s)}$, P's expected utility from $y \in \tilde{M}_{t(s)} \cup \mathcal{X}_{t(s)}$ is $-c(1 - F(\tilde{e}_s)) + \rho \int_\Theta r(\theta) d\nu_1(S|y)$. We can specify P's mixing probability over $M_{t(s)} \cup \mathcal{X}_{t(s)}$ via a Radon-Nikodym derivative. By the arguments in Lemma 2, there is a unique such Radon-Nikodym derivative that leaves P indifferent; this mixing only depends on the expected material utility and the interim beliefs over S at each y, and since these are the same as that under ex-ante signaling, we obtain the same outcomes for P. Therefore, the equilibrium outcome must be the same as in ex-ante signaling. *Q.E.D.*

Proof of Proposition 9

Proof. P's expected utility conditional on e_0 is $U_P^\alpha(F_1(\cdot|e_0))$. By an analogous proof to that in Lemma 3, $\mathbb{P}(a = 1|e_0) = \frac{1}{c}(\rho\mathbb{E}[r(\theta)] - U_P^\alpha(F_1(\cdot|e_0)))$. Thus,

$$\mathbb{P}(a = 1) = \int_E \mathbb{P}(a = 1|e_0) dF_0(e_0) = \frac{1}{c}[\rho\mathbb{E}[r(\theta)] - \int_E U_P^\alpha(F_1(\cdot|e_0)) dF_0(e_0)].$$

The proposition then follows immediately from convexity of $U_P^\alpha(\cdot)$. *Q.E.D.*

F. Investigation Design

As mentioned, in many salient applications the investigation is determined by an interested third party who has opposing interests to the Partisan, i.e., they want to maximize the probability of $a = 1$.[48] This is the case for the Speaker of the House designing

[48] Our earlier working paper also analyzes the case in which the investigator has evidence dependent preferences.

an impeachment inquiry, or firms submitting information in a merger approval request. The relationship between the investigation and outcomes depends on the equilibrium. In this appendix, we provide characterization results for our two focal equilibria — ex-ante signaling and ex-post signaling — and compare these two solutions. For this section, we revert to our main text model where $r(s) = 1 \; \forall s in S$ and assume the standards distribution G admits a continuous density g with $g(s) > 0 \; \forall s \in [c, 1+c]$

Before analyzing this problem, we must specify the set of available investigations. While our methodology is not specific any particular type of constraint, to fix ideas we study the case in which evidence $e \in [0,1]$ represents a posterior about a binary state $\omega \in \{0,1\}$, with prior $\overline{e} \in (0,1)$. The set of available investigations is the set of Bayes plausible distributions of evidence (posteriors).[49] Thus a CDF F on $[0,1]$ is a valid investigation if $\int_0^1 e dF(e) = \overline{e}$.

F.1. Optimal Investigations Under Ex-Ante Signaling

To calculate the investigator's utility, we sum the probability of $a = 1$ given message m_s weighted by the rate at which the DM sends message m_s. Letting $\mathcal{F}$ be the set of CDFs with support on $[0,1]$, and σ be the DM's communication strategy under ex-ante signaling, the investigator's design problem is

$$\max_{F \in \mathcal{F}} \int_S \left(1 - F(\tilde{e}_s)\right) \left(qg(s)ds + (1-q)d\sigma(m_s|P)\right),$$
$$\text{such that } \int_0^1 e dF(e) = \overline{e}.$$

In order to solve this problem, we use Lemma 3, which shows that maximizing the probability of $a = 1$ is equivalent to minimizing that of P. It is straightforward to derive how F determines $U_P^\alpha(F)$: we show in the proof of Lemma 4 that $U_P^\alpha(F)$ is given by the solution U to $\int_S \frac{\rho q g(s)}{U + c - cF(\tilde{e}_s)} ds = 1$.[50] Using $\int_0^1 e dF(e) = \int_0^1 (1 - F(e)) de$, these observations

[49] In our earlier working paper we show how our main takeaways extend to the case in which the state lies in a compact set and the evidence represents the posterior mean of a given belief.

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

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

$$\min_{U\geq 0,\ F\in\mathcal{F}} \quad U\ , \tag{18}$$
$$\text{such that } \int_S \frac{\rho q g(s)}{U + c - cF(\tilde{e}_s)} ds = 1,$$
$$\int_0^1 (1 - F(e))de = \bar{e}.$$

The extra constraint ensures the choice of U in (18) 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.

Recall that $H(e) = G(e+c)$ is the probability that N types choose $a = 1$ given evidence e. Denote $\overline{H}$ as the concavification of H,[51] and $\overline{h}$ as its derivative in e, which is weakly decreasing and continuous in e.

Proposition 10. *For $k, U \in \mathbb{R}$, define $\widehat{F}(e; k, U) \equiv U/c + 1 - k\sqrt{\overline{h}(e)}$. The uniquely optimal investigation is given, for $e < 1$, by*

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

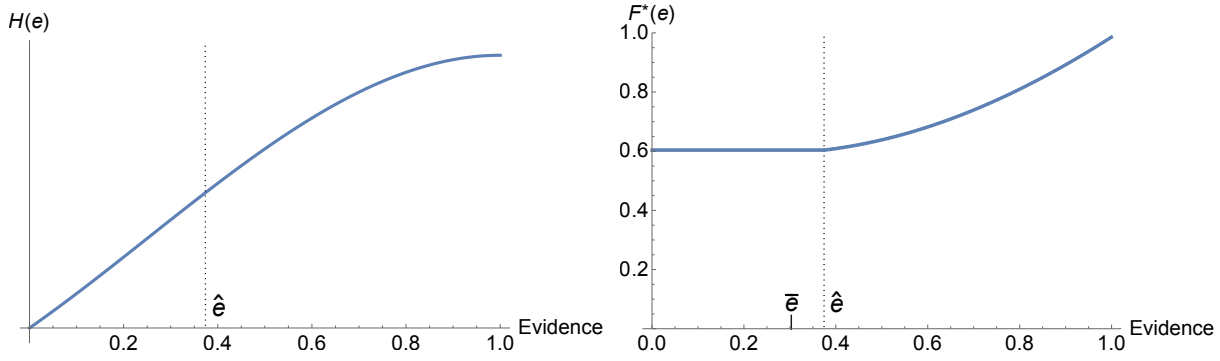
with $U = U_P^\alpha(F^*)$ as the partisan's utility given F^* and some $k > 0$.

It is then well known that the curvature of $\overline{H}$ (or the monotonicity of $\overline{h}$) captures persuasion incentives for the N types: providing information over regions where $\overline{h}$ is constant (decreasing) increases (decreases) the probability that N chooses $a = 1$. Our characterization is also in terms of $\overline{h}$ but these incentives are distorted by the fact that the investigator must also persuade P.[52]

Figure 3 presents an example of an optimal investigation. In this example, the distribution of non-partisan standards is single peaked, and so H is convex for small e,

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

[52] There are two remaining parameters in the characterization in Proposition 10: $U_P^\alpha(F^*)$ and k. These are jointly pinned down by the two constraints in (18). While an explicit expression is not always feasible, solving these two equations numerically is straightforward.



(a) CDF of non-partisan thresholds

(b) CDF of the optimal investigation

Figure 3: G is a standard logistic distribution with mean $\frac{1}{2}$, $c = \frac{1}{4}$, $q = \frac{1}{2}$, $\overline{e} = \frac{3}{10}$, and $\rho = 3$.

and concave for large e, as illustrated in the left panel. Correspondingly, the concavification of H is linear below $\hat{e}$ and equal to H above $\hat{e}$, i.e., $\overline{h}$ is constant below $\hat{e}$ and strictly decreasing above $\hat{e}$. From the right panel of Figure 3, we see that F^* provides information in a way that is consistent with N's information incentives below $\hat{e}$, but in contrast, provides some information, in a smooth way, above $\hat{e}$ at the detriment of N's outcomes. This builds on the intuitions from Proposition 6 and Corollary 2, balancing the information incentives for N and the desire to have an unpredictable investigation for P types.

F.2. Optimal Design under Ex-Post Signaling

We now 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, by Theorem 1, the investigator will always prefer the investigation in Proposition 10 to the optimal investigation under ex-post signaling. In particular, we show that under ex-post signaling the investigator does not minimize "predictability" as under ex-ante signaling.

Recall $v^\beta(e)$ is the probability of $a = 1$ as a function of the evidence given ex-post signaling. Due to the simplicity of ex-post signaling, we can explicitly derive this probability in baseline model where $\underline{r} = \overline{r}$ as

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

The main thing to note about this expression is that it is independent of the investigation F, that is, the probability of $a = 1$ is linear in the investigation. We can write the investigator's design problem as

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

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

Proposition 11. *Let $Cav(v^\beta)$ be the concavified value of v^β. There exists an optimal F with binary support if $v^\beta(\overline{e}) < Cav(v^\beta)(\overline{e})$ and an optimum with degenerate support on $\overline{e}$ if $v^\beta(\overline{e}) = Cav(v^\beta)(\overline{e})$.*

An immediate implication is that if v^β is strictly concave in e, then the uninformative investigation is uniquely optimal. Because v^β is a convex transformation of H, it is not quite sufficient for the investigator to want to withhold information from the N types. However, if the investigator is significantly harmed by providing information to the non-partisan, i.e., H is "sufficiently concave," then an uninformative investigation will be optimal under ex-post signaling.[53] Note that in these cases (and in general), the optimal investigation under ex-ante signaling provides some information; this is an immediate implication of Proposition 6. Thus, there are cases, namely those in which s types' takes $a = 1$ significantly less when given information, in which the optimal investigation under ex-ante signaling is more informative in a Blackwell sense than that under ex-post signaling.

However, the comparison can also go the other way. Because v^β is a convex transformation of G, there will be parameter specifications where the ex-post signaling optimal investigation is perfectly informative, but the investigator is harmed by providing information to non-partisans. In these cases, because concave G implies $\overline{h}$ is decreasing in e, Proposition 10 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 s types' probability of taking $a = 1$ then full information is optimal under

[53] An example is when the standards are distributed according to the standard exponential distribution.

both regimes. This means that, like under ex-ante signaling, P's behavior under ex-post signaling incentivizes the investigator to provide more information. This corollary follows the above discussion and Corollary 2 so we omit its proof.

Corollary 3. *If H is convex on $[0,1]$ then the optimal investigation is fully informative under both ex-ante and ex-post signaling.*

Proof of Proposition 10

Proof. Note that by definition $\int_S \frac{\rho q r(s) g(s)}{U+c-cF(\tilde{e}_s)} ds = \int_E \frac{\rho q h(e)}{U+c-cF(e)} de$. We solve the following relaxed version of the investigator's problem:

$$\begin{aligned} &\min_{U\geq 0, F\in\mathcal{F}} U \\ \text{subject to } &\int_E \frac{\rho q h(e)}{U+c-cF(e)} de \leq 1, \\ &\int_0^1 (1-F(e)) de \leq \overline{e}. \end{aligned} \tag{19}$$

Both constraints are convex in U and F. By Theorem 1 (Chapter 8) of Luenberger (1997), there exist multipliers $\eta, \lambda \geq 0$ such that any solution U^*, F^* to (19) will solve[54]

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

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

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

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

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

Fixing the optimal value of U^*, it is clear that the optimal investigation F^* must solve

$$\min_{F \in \mathcal{F}} \int_E \left(\frac{\eta \rho q h(e)}{U^* + c - cF(e)} - \lambda F(e) \right) de - \eta + \lambda - \lambda \bar{e}. \tag{20}$$

We have $\lambda > 0$; 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 and therefore increasing requires the use of ironing techniques to solve (20). By Theorem 3.1 of Toikka (2011), $F^*(e) = \arg\min_{x \in [0,1]} \frac{\eta \bar{h}(e) \rho q}{U^* + c - cx} - \lambda x$. The objective is differentiable, so first-order conditions are necessary. Taking the first-order condition, whenever $F^*(e) \in (0,1)$, we have

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

Letting $k = \sqrt{\frac{\eta \rho q}{c\lambda}}$, we can rewrite the above equation as $F^*(e) = \frac{U^*}{c} + 1 - k\sqrt{\bar{h}(e)}$ whenever $F^*(e) \in (0,1)$. $F^*(e) = 0$ whenever $\frac{\eta \bar{h}(e) \rho q}{c(\frac{U^*}{c} + 1)^2} - \lambda > 0$; this condition simplifies to $\frac{U^*}{c} < k\sqrt{\bar{h}(e)} - 1$. Similarly, $F^*(e) = 1$ whenever $\frac{\eta \bar{h}(e) \rho q}{c(\frac{U^*}{c})^2} - \lambda < 0$, or alternatively, when $\frac{U^*}{c} > k\sqrt{\bar{h}(e)}$. That $U^* = U_P^{\alpha}(F^*)$ follows from the fact that the first constraint in (19) holds with equality. *Q.E.D.*