Signaling by Taking Stands*

Andrew McClellan† and Daniel Rappoport‡

January 18, 2026

Abstract

A decision maker (DM) cares about his reputation for being "good," i.e., seeking the "correct" decision, as opposed to being "bad," i.e., having a fixed partisan agenda. The correct action depends on publicly revealed evidence as well as the DM's privately known standard. While the DM will take his decision after the evidence is realized, he 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 standard 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 available evidence as opposed to the potentially biased agenda of the decision maker. While the decision is only taken after the evidence is made public, in many contexts the decision maker has an opportunity to "take a stand," or state their intentions, before this evidence is revealed. Our paper explores how such "ex-ante" signaling efforts affect outcomes. To make this concrete, consider the following examples.

*We thank Ian Ball, 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 for helpful comments and suggestions.

†University of Chicago, Booth School of Business. Email: andrew.mcclellan@chicagobooth.edu

‡Georgetown University. Email: dr1097@georgetown.edu

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., want 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]

An initial question in such settings is how informative can communication be prior to the revelation of evidence, e.g., how much can politicians distinguish their standards during an investigation? Our main focus though, is how such communication affects incentives and outcomes. Would we expect that Republican senators who indicate conditions for impeachment up front convict more or less than those who wait and see? Would more drugs be approved with prespecified efficacy standards as opposed to case-by-case approval decisions? 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. Conversely, 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

[4] Pinker, Steven (2014) "The Trouble With Harvard" *The New Republic*, 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. Preregistration is often discussed in terms of its incentive effects on authors (see Orzach (2025) who studies the effect on reputationally concerned authors). Our paper speaks to the impact that preregistration would have on the decisions of 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 6.2, 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 chosen action.

The DM's standard can be interpreted in two ways. Our leading interpretation is that the standard represents private non-verifiable information about the "right" evidence threshold, e.g., FDA officials have specific expertise about the current state of treatment for the experimental drug's targeted disease. An alternative interpretation is that the standard represents 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. Key to both of these is that the standard is interpreted as transitory because it is specific to the current issue or decision. In contrast, partisanship is interpreted as persistent. Thus, whether the DM is expected to be aligned with the public on future decisions, and hence his reputation value in our model, only depends on beliefs about his partisanship.[8] 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.[9] Non-partisans may have different ideology or information for the current issue, but they are still interested in the objective evidence on that issue.

Our first result shows that, at the decision stage, each non-partisan type uses a standard-dependent threshold, i.e., takes $a = 1$ if and only if the evidence is higher than some $\tilde{e}_s$. We then 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 threshold contingent plan with respect to the realized evidence, e.g., stating "I will approve if the evidence meets ... standard," with the non-partisan DM with higher standards committing to higher thresholds. Conversely, ex-post signaling can be interpreted as 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. In all equilibria, reputation concerns induce the partisan to mimic

[8] In Subsection 6.1 we allow for the DM's reputation value to depend on beliefs about his standard and show that our results continue to hold as long as the DM is still sufficiently interested in signaling non-partisanship.

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

the non-partisan, and the non-partisan to exhibit "political correctness" by taking $a = 1$ at some evidence realizations where they would prefer $a = 0$ in the absence of reputation concerns.

This characterization shows how communication prior to the realization of evidence can effectively change the signaling space for the DM. Under ex-post signaling, the DM signals only with his action choice following the evidence—$a \in \{0, 1\}$—while under ex-ante signaling, the DM signals with his entire contingent plan over evidence—$x : E \to \{0, 1\}$. Our main goal is to compare signaling incentives across these different equilibria, and their associated outcomes. Theorem 1 shows that ex-ante signaling has the highest probability of $a = 1$ across all equilibria, 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"; government 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, obtaining a higher reputation requires taking a different action with probability one. More specifically, suppose that the DM chooses between using two evidence thresholds $t_1 < t_2$ for taking $a = 1$. From an ex-ante perspective, declaring an intention to use t_1 is relatively inexpensive (as compared to t_2) as these thresholds do not require different decisions following "non-pivotal" evidence $e \notin [t_1, t_2)$. Conversely, credibly distinguishing that the DM is using t_1 rather than t_2 under ex-post signaling is only possible conditional on evidence realizations $e \in [t_1, t_2)$, and requires taking $a = 1$ with probability one. This implies that the reputational compensation for using t_1 must be larger at such evidence realizations under ex-post signaling than under ex-ante signaling. We show that generating this larger reputational difference requires the DM to use the lower threshold less often, i.e., take $a = 1$ less often, under ex-post signaling.

Following Theorem 1 we discuss how the main tradeoff hinges on two key types of uncertainty in our model: uncertainty over (i) which contingent plan will be used by the non-partisan DM, and (ii) whether these contingent plans will be distinguishable ex-post, i.e., the presence of non-pivotal evidence. These points can be seen in the simple intuition above. The first type of uncertainty leads to multiple evidence thresholds being used (i.e., t_1 and t_2 above) while the second type of uncertainty makes using lower thresholds cheaper from an ex-ante perspective, (i.e., because t_1 takes the same actions as t_2 for non-pivotal evidence $e \notin [t_1, t_2)$).

The main forces behind our result do not rely on many of the specific details assumed in our baseline model. To deliver this point, we generalize our model to one with arbitrary material preferences for the DM and a larger action space in Subsection 6.1.[10] We use this model to identify relatively weak conditions on preferences that guarantee the two key types of uncertainty mentioned above and thereby preserve our main results. More specifically, we find that, as in our baseline model, the DM acts most against his partisan interests under ex-ante signaling, as compared to all other equilibria.

Ex-ante signaling and ex-post signaling are focal because their outcomes have simple implementations. Ex-post signaling outcomes, which deliver 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 deliver 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 Proposition 9, we show that "intermediate outcomes" can be implemented by regulating how much information about evidence is revealed before or after the DM commits to a contingent plan.[11]

Such implementations can be used to change the equilibrium probability of $a = 1$, both conditional and unconditional on evidence. This is our central focus for several reasons. First, from a positive perspective, the rate of taking $a = 1$ is the only observed outcome in applications where the standard is never publicly revealed. Second, our results help us identify who benefits from changes in the timing and explains why interested parties may lobby or push for different communication protocols. For example, in

[10] We preserve some specific assumptions of our baseline model such as the fact that there is a single partisan type with the lowest reputation. As we discuss in Subsection 6.1, this assumption buys significant tractability. In Appendix F we show that our main takeaway holds in a version of our model where the DM has continuous degree of partisanship and thereby violates this assumption.

[11] One could also ask which outcomes will be implemented in the absence of these institutional interventions. Proposition 8 shows that ex-ante signaling outcomes are uniquely selected under a natural refinement. The intuition is that taking a clear stand is the best way to signal responsiveness to evidence rather than to one's idiosyncratic agenda.

the context of merger approvals, our results say that firms will prefer transparent criteria (ex-ante signaling) if the approval agency is possibly adversarial, and prefer opacity (ex-post signaling) when the approval agency is possibly aligned. A natural alternative design objective is social welfare, where the measure is informed by our leading interpretation, i.e., where s is a private signal about the "correct" evidence threshold from the observer's perspective. In Subsection 6.4, we compare ex-ante signaling and ex-post signaling on this social welfare measure. We discuss how this comparison turns on a familiar tradeoff in signaling models between the DM's level of bias versus his tendency toward political correctness.

In Section 5 we investigate how outcomes change with parameters of our model under our focal equilibrium of ex-ante signaling. We highlight two such comparative statics. Proposition 4 and Proposition 5 identify conditions under which a mean-preserving spread in standards, or respectively a mean-preserving contraction in evidence, decreases the probability of $a = 1$ under ex-ante signaling. At a high level, the intuition for these results are two sides of the same coin. That is, a spread in standards or a contraction in evidence makes the private information of the DM more important in decision making relative to the public evidence. This greater uncertainty in the non-partisan's decisions gives the partisan more latitude in mimicking them, which he then in turn uses to decrease the probability of $a = 1$.

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 robustness results, equilibrium selection, alternative commitment and timing assumptions, and welfare comparisons.

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)). Our DM's preferences are closest to those in Morris (2001).[12] He studies an informed sender who seeks a reputation for being respon-

[12] 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.

sive to the state—similar to our non-partisan—rather than having a state-independent preference—similar to our partisan.[13] 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). Like Bénabou and Tirole (2006), Esteban and Ray (2006), and Frankel and Kartik (2019), we allow for multidimensional types of the DM—namely preference heterogeneity of the DM beyond the dimension on which he seeks reputation. As in these papers, this multidimensionality allows for intermediate reputations in our equilibria through "plausible deniability," i.e., the DM could take a low action because he is a partisan, or because he is non-partisan but with a high standard. As we discuss further after Theorem 1, this multidimensionality is also key for how communication impacts equilibrium outcomes in our framework. Frankel and Kartik (2022) and Ball (2025), among others bring a design perspective to such settings, studying how to design scoring systems in the presence of strategic manipulation.[14,15]

The role of communication in our model is to permit the use of an "expanded signaling space": from the ex-post perspective, the DM simply makes a binary-action choice, but from the ex-ante perspective, the DM chooses over contingent plans. The reason that communication facilitates signaling over this larger space is because a decision-relevant state—namely, evidence—is revealed in between our communication and decision stages.[16] This differentiates our use of communication from that of previous work that introduces cheap talk as a coordination device between players with some degree of shared interests (see Farrell and Rabin (1996) for a survey).[17]

[13] 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.

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

[15] 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 over a payoff-relevant state and subsequent costly signaling, but considers different material and reputation incentives of the DM and has private, rather than public, revelation of the state.

[16] Cheap talk is only one way to deliver this enriched signaling space. More directly, we show in Subsection 6.3 that endowing the DM with public commitment over their contingent plans implements the same outcomes.

[17] Relatedly, Kartik (2009) studies a model with simultaneous cheap talk and costly signaling where

Some papers have studied cheap-talk/signaling prior to the revelation of an unknown state. For example, Ottaviani and Sorensen (2006a), Ottaviani and Sorensen (2006b) ask how an expert makes predictive cheap-talk statements about an ex-ante unknown state given that his reputation is assessed after the revelation of this state. Chen (2012) studies a Crawford and Sobel (1982)-like cheap-talk model and asks whether welfare is higher if the sender communicates before or after a public signal about the state is revealed.[18] In contrast to our setting, these papers do not feature an action choice after the state is revealed. This means that signaling over contingent plans does not arise in these papers, and so the main comparison in our paper and the forces behind it are also absent.

One interpretation of our two focal equilibria is that the DM is transparent about his entire contingent plan under ex-ante signaling while he only reveals his realized action choice under ex-post signaling. Thus, our main comparison can be interpreted as saying that "endogenous" transparency about an agent's contingent plan leads to fewer partisan actions. Other career-concern models have studied how exogenous changes in transparency about other features affect outcomes. For example, Prat (2005) shows how observing an expert's recommendation in addition to whether they were correct can harm outcomes.[19] Levy (2007) shows transparency about the individual votes of committee members makes them less likely to vote for the prior-optimal decision. Both of these papers focus on tradeoffs around herding on the prior-optimal decision.

2. Model

Overview There are two players: a decision maker (DM) and an inactive Bayesian observer. The DM chooses an action $a \in A = \{0, 1\}$. His preferences over this action depend on his privately known type $\theta \in \Theta$ and evidence $e \in E \subseteq \mathbb{R}$ which is distributed according to a commonly known CDF F (which we will refer to as an investigation) and independent of the DM's type θ. We assume that $\int_E e dF(e)$ is finite. The DM also values his reputation in the eyes of the observer.

cheap talk is used as an imperfect substitute to resolve coordination issues given that all costly signals have been exhausted by other types.

[18] Daley and Green (2014) considers a costly signaling model where the public observes an independent signal about the agent's ability in addition to his action choice.

[19] 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.

The game takes place in two stages. In the initial communication stage, the DM observes his private type but not the evidence, and sends a cheap-talk message $m \in M$ to the observer, where M is some sufficiently large metrizable space.[20] 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 $\theta \in \{N, P\} \times S$. 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 a non-partisan is

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

We refer to the first component of the payoff that depends (directly) on the action as the **material payoff** and $\rho\nu$ as the **reputation payoff**. N types have higher material utility from $a = 1$ relative to $a = 0$ if (i) the evidence is more convincing (e is higher), or (ii) their standards are lower (s is lower). 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. The reputation payoff captures the DM's desire to appear as a non-partisan to the observer, and the weight $\rho > 0$ parameterizes the strength of these reputation concerns.[21]

We provide additional discussion at the end of this section, but in our leading interpretation the DM's standard represents a private non-verifiable signal about the "correct" evidence threshold for action $a = 1$ from the observer's perspective.[22] The non-

[20] 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.

[21] The fact that all DM types place an identical weight ρ on reputation is a normalization; we could allow for the weight ρ to depend on θ without changing the results.

[22] Under this interpretation, the assumption that s and e are independent fits applications where they represent different dimensions of a decision. For example, s could capture a representative's private information about her constituents' idiosyncratic preference for a policy, whereas e represents a cost estimate at the federal level; in the merger-approval context, s could capture the regulator's internal projection of market outcomes (e.g., the impact of increased market concentration), while e represents the degree of synergies of a particular merger being proposed. However, in some cases where s and e are both information about whether $a = 1$ is a good decision, it is natural that they may be correlated. While we expect our forces would also play a role with this kind of correlation, independence makes the analysis signif-

partisan's material utility then represents his alignment with the interests of the observer.[23] In contrast, the P type does not care about taking the "right" decision, and as a result, neither the standard nor evidence enter his material utility. Thus, under our leading interpretation, the reputational concerns represent a desire to portray alignment with the observer.

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 $S \equiv \mathrm{Supp}(G)$. 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 are 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 on-path 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 $\zeta : \Theta \times M \times E \to \Delta(A)$, 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.[24]

2. ν_2 is obtained from ζ using Bayes rule with prior $\nu_1(\cdot|m)$.[25]

icantly more tractable and more clearly illustrates these forces. In this way, we follow other studies like Levy (2007) that also make an independence assumption about different aspects of a decision for similar reasons.

[23] Under this interpretation, the specification in (1) assumes "complete" alignment, i.e., that the non-partisan's material utility is also that of the observer, but this is purely for expositional compactness. Indeed, in Subsection 6.4 we allow the non-partisan to have arbitrary positive weights on observer welfare and his partisan bias.

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

[25] For all Borel $\hat{\Theta} \subseteq \mathrm{Supp}(\nu_1(\cdot|m))$, $\int_{\hat{\Theta}} \zeta(a|\theta, m, e)d\nu_1(\theta|m) = \nu_2(\hat{\Theta}|m, e, a) \int_{\Theta} \zeta(a|\theta, m, e)d\nu_1(\theta|m)$ and $\mathrm{Supp}(\nu_2(\cdot|m, e, a)) \subseteq \mathrm{Supp}(\nu_1(\cdot|m))$.

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

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 been sent, where the type space is Θ_m.[26] 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. We provide the formal definition of D1 in the context of our game in Appendix A.[27]

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,[28] $v^{\mathcal{E}}(e, F)$ 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

[26] 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.

[27] Without this refinement, the game admits many unnatural equilibria—e.g., for bounded evidence and standards and high ρ, there is an equilibrium in which all DM types pool on $a = 0$, with a choice of $a = 1$ leading the observer to believe the DM must be a P type with probability one, even though s types may have stronger preferences for $a = 1$ than P. The D1 refinement allows us to rule out such equilibria. Moreover, weaker refinements, like the intuitive criterion, would not rule out such equilibria due to the high reputation concerns laid out in Assumption 1. A full reputation would induce all types to deviate to any off-path action, and so the intuitive criterion does not eliminate any types from possible off-path beliefs.

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

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 the P type 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 $a = 0$ 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 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. Our model captures the idea that non-partisans may be willing to take the right action even if it means a large loss of reputation, e.g., if s or e is extreme. Instead, as mentioned in our literature review, our preferences (but not our game) mirror those in Morris (2001) in that good types' preferences depend on the "state"—evidence and standards, in our model—whereas bad types' preferences do not. 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.

We note here that the structure of the DM's preferences and binary action space is made for fit with the applications and for ease of exposition. In Subsection 6.1 we show that our main results hold in an extension with arbitrary material preferences and any number of finite actions.

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 many applications. Indeed, our leading interpretation is that the standard is private information about the "correct" evidence threshold which is only relevant for the decision at hand. In this case, it is natural that obtaining information that $a = 1$ is probably a bad choice, i.e., a high standard, should not carry a lower reputation. However, in some settings the

DM may 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. Our generalization in Subsection 6.1 allows for reputational payoffs to depend on the observer's belief about the DM's standards.

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 6.2 shows that endowing the DM with commitment uniquely implements the same outcomes as this salient cheap-talk equilibrium.

3. Equilibrium Characterization

Our main results in Section 4 show how equilibrium outcomes change when the DM "takes stands," i.e., communicates informatively prior to the revelation of evidence. Toward this end, we begin with an equilibrium characterization result in Lemma 1. We then show in Lemma 2 that equilibrium outcomes can be categorized by the amount of information transmitted about the N's standard at the communication stage.

It will be useful to make statements in terms of induced mappings from evidence to actions, i.e., contingent plans $x \in \mathcal{X} \equiv \{x' : E \to A\}$. 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 first and third 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 specified in Assumption 1. If a message is only sent by P, then it receives the lowest possible reputation, and high reputation incentives induce P to deviate away from this message. If a message is never sent by P, then it receives the highest possible reputation, and high reputation incentives induce P to deviate towards this message. The reason for why, after sending message m, P mixes over the actions chosen by s types who also send m is analogous.

The second point states that the s type's realized contingent plan (almost surely) follows the fixed rule $x_s(e)$.[29] To avoid probability one caveats, going forward we focus on equilibria where N's actions correspond with $x_s(e)$ *everywhere*, i.e., for all $e \in E,\ s \in S$. 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 $a = 0$.

To provide intuition for point 2, consider the case in which both actions are on path following some evidence realization e. 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.[30]

Given this argument, an equivalent characterization of x_s is that it chooses $a = 1$ for

[29] 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.

[30] 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 $s \leq e + c$) for all $s \in \Theta_m$, i.e., actions are consistent with x_s.

evidence realizations where the s type loses less material utility from this choice than the P type, and chooses $a = 0$ otherwise. Because of the separability across evidence realizations, this induces a contingent plan that also maximizes the material utility difference from an ex-ante perspective, i.e.,

$$x_s \in \arg\max_{x \in \mathcal{X}} \int_E (e - s + c)x(e)dF(e). \tag{2}$$

This perspective will be useful in the following analysis.

We next categorize the set of equilibrium outcomes by how much information N communicates about his standard at the communication stage. Accordingly, define an **information structure on standards** (ISS) as a Bayes-plausible distribution of posteriors over standards $\Lambda \in \Delta(\Delta(S))$.[31] Given an equilibrium $\mathcal{E}$ with communication strategy $\sigma : \Theta \to \Delta(M)$, we say the ISS Λ is **induced** by $\mathcal{E}$ if for each Borel $H \subset \Delta(S)$, $\Lambda(H) = \int_M \int_S \mathbb{1}\big(\nu_1(\cdot|m, \theta \in S) \in H\big) d\sigma(m|s)dG(s)$. That is, if the observer knew the DM were an N type, the communication strategy σ induces interim beliefs over S that are distributed according to Λ.

Lemma 2. *For each ISS Λ, there exists an equilibrium that induces Λ. Moreover, the set of equilibria that induce Λ admit a unique equilibrium outcome.*

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

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

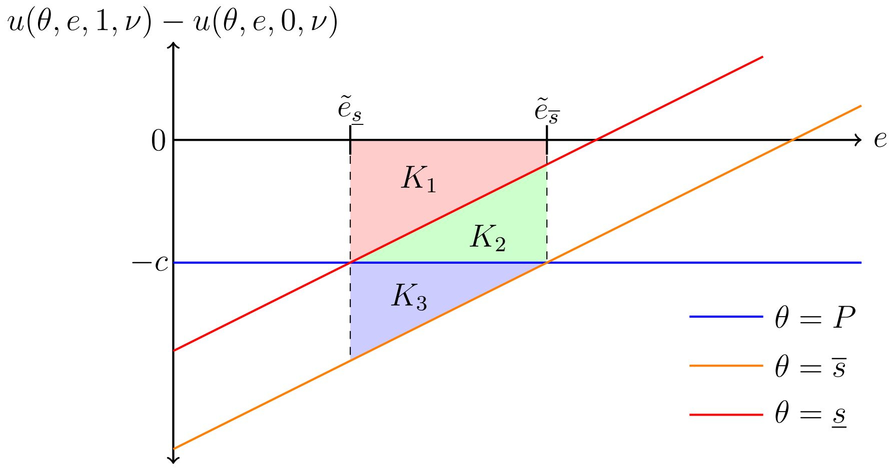

Figure 1: Material utility difference between $a = 1$ and $a = 0$ as a function of evidence.

mixes over all messages sent by the s types at the communication stage and all on-path follow up actions 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. The central idea comes from the characterization in (2): if an s type considers some deviation to sending an alternative message and follow up contingent plan $\tilde{x}$, the fact that x_s maximizes the material utility difference between s and P across all contingent plans means that P will have a higher value for this deviation. That P does not find such a deviation profitable then implies neither does s.

Figure 1 provides further illustration. 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 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 area 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 area 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.

4. The Effects of Informative Stands

In light of Lemma 2, we refer to equilibria by their associated ISS. We begin our analysis by comparing the two salient extreme cases of equilibria that we define 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 and, consistent with Lemma 1, P positively mixes over all m_s to ensure his own indifference. 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 P and N types behave identically after sending this message. That is, the equilibrium has no **residual strategic uncertainty**: there does not exist a positive probability set of m, e for which more than one action is on-path after message m and evidence e is realized.[32] Since the threshold $\tilde{e}_s$ is increasing in s, P takes $a = 1$ more frequently when declaring a lower standard and therefore must be compensated with a higher reputation.

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 what the evidence is." Ex-post signaling admits residual strategic uncertainty under the weak condition that there exist two types s', s'' such that $F(\tilde{e}_{s'}) \neq F(\tilde{e}_{s''})$. The distribution F only impacts equilibrium outcomes conditional on an evidence realization e through changes in the messaging strategies (and therefore the updated belief about the DM's type before evidence is realized). Because messages are uninformative under ex-post signaling, a distinctive feature of this equilibrium is that 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. In effect, these equilibria differ in the timing of when signaling occurs, before or after evidence is revealed.

[32] Formally, a is on-path after m, e if $\mathbb{P}^{\mathcal{E}}(a|m, e) = \int_\Theta \zeta(|\theta, m, e) d\nu_2(\theta|m) > 0$, so an equilibrium $\mathcal{E}$ has residual strategic uncertainty if $\int_M \int_E \mathbb{1}(\mathbb{P}^{\mathcal{E}}(a = 1|m, e) \in (0, 1)) dF(e)(q d\Sigma_N(m) + (1 - q) d\sigma^{\mathcal{E}}(m|P)) > 0$.

4.1. Binary Standards

To gain some intuition for how the timing of signaling shapes equilibrium outcomes, it is useful to compare outcomes in these two extreme equilibria in a simple binary standard version of our model. Suppose G is supported on two types $\underline{s} < \overline{s}$ and F has full support on $\mathbb{R}$. 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 $e \in [\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$.

Under ex-ante signaling, P will mix between declaring himself to be $\underline{s}$ and $\overline{s}$, and follow through with a threshold for taking $a = 1$ of $\tilde{e}_{\underline{s}}$ and $\tilde{e}_{\overline{s}}$ respectively. Thus, the probability that P takes $a = 1$ after pivotal evidence is realized is the probability that he mimics the $\underline{s}$ type at the communication stage, which is pinned down by P's indifference across messages. In order to be indifferent, the difference in reputational payoff from declaring $\underline{s}$ and $\overline{s}$ must be equal to the difference in probability (weighted by c) with which $\underline{s}$ takes $a = 1$ relative to $\overline{s}$, which is given by the probability that pivotal evidence is realized:

$$\rho(\text{Ex-Ante Signaling Reputational Diff.}) = c(F(\tilde{e}_{\overline{s}}) - F(\tilde{e}_{\underline{s}})).$$

Under ex-post signaling, every DM type chooses the same message at the communication stage. Given a pivotal evidence realization at the decision stage, P chooses whether to mimic $\underline{s}$ by taking $a = 1$ or to mimic $\overline{s}$ by taking $a = 0$. To leave him indifferent, the reputational payoff difference from mimicking $\underline{s}$ and $\overline{s}$ must be equal to the relative cost of taking $a = 1$, namely c:

$$\rho\big(\text{Ex-Post Signaling Reputational Diff.}\big) = c.$$

Thus, the difference in reputations must be larger under ex-post signaling than 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, ex-ante signaling leads to a higher rate of $a = 1$ than ex-post signaling.

Figure 2 illustrates this logic. The left graph shows the probability that P takes $a = 1$ as a function of the evidence threshold used for taking $a = 1$. As shown in Lemma 1, the only two thresholds used in this binary case will be $\tilde{e}_{\underline{s}}$ and $\tilde{e}_{\overline{s}}$. The graph shows this probability for ex-ante signaling (before evidence is realized) and ex-post conditional on

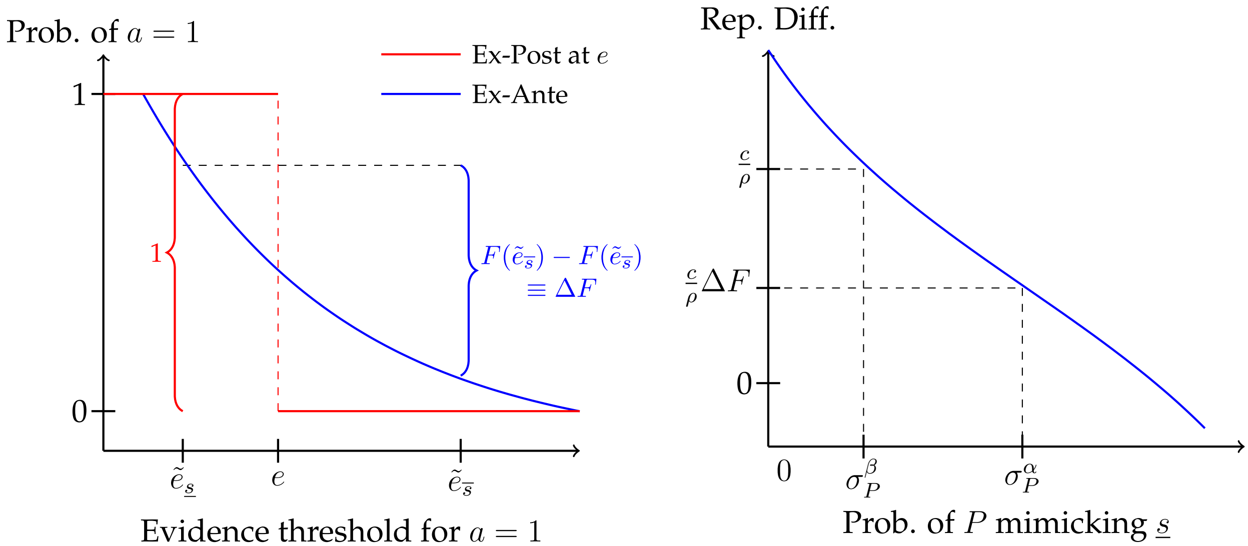

Figure 2: The left figure plots the probability of $a = 1$ as a function of the evidence threshold used, where the probability for ex-ante signaling is taken before evidence is realized while ex-post is taken after pivotal evidence $e \in (\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$ is realized. The bracketed distances represent the associated differences in probability of taking $a = 1$ as a consequence of using $\tilde{e}_{\underline{s}}$ rather than $\tilde{e}_{\overline{s}}$. The right figure plots the reputation difference between mimicking $\underline{s}$ and $\overline{s}$ as a function of the probability P mimics $\underline{s}$, where σ_P^α and σ_P^β denote this probability under ex-ante and ex-post signaling respectively.

e being realized, where $e \in (\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$. Because the ex-post probability conditions on e being realized, using any threshold below e is equivalent to choosing $a = 1$ with probability one and any threshold above e is equivalent to choosing $a = 0$ with probability one, leading to a step function. In contrast, because the ex-ante probability takes the expectation over the evidence, it is continuously decreasing in the threshold chosen. The right graph maps the corresponding reputational differences needed to make P indifferent between mimicking $\underline{s}$ and $\overline{s}$ under ex-ante signaling and ex-post signaling (conditional on evidence e), namely $\frac{c}{\rho}(F(\tilde{e}_{\overline{s}}) - F(\tilde{e}_{\underline{s}}))$ and $\frac{c}{\rho}$ respectively, into the equilibrium probability that P chooses to mimic $\underline{s}$ and choose $a = 1$ at $e \in [\tilde{e}_{\underline{s}}, \tilde{e}_{\overline{s}})$. Because the reputational difference between mimicking $\underline{s}$ and $\overline{s}$ is decreasing in the probability that P mimics $\underline{s}$, P's equilibrium probability of mimicking $\underline{s}$ is lower under ex-post than ex-ante signaling.

4.2. Comparisons Across Equilibria

How far does this intuition that committing up-front increases the rate of taking the action extend? Will it hold for richer type spaces or across equilibria other than ex-post? Our main result answers these questions. First, we introduce a technical condition. We say there is **mild agreement** if for every pair $s', s'' \in S$, $x_{s'}(e) = x_{s''}(e)$ on a positive probability set of 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 P preferred $a = 1$ instead (i.e., $c < 0$), 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 6.2 and Subsection 6.3 elaborate, showing how ex-ante signaling outcomes arise uniquely for variations of the current model.

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

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

The high-level intuition behind these results is as follows: the DM is willing to promise more ex-ante because he is only called to act on this promise for a subset of evidence realizations while still enjoying the higher reputation for all evidence realizations. As in our binary-standards analysis, under ex-ante signaling, when P claims to be the lower standard $\underline{s}$ instead of the higher one $\overline{s}$, he obtains the higher reputation regardless of the evidence realization, but only suffers the material loss upon the realization of pivotal evidence where $x_{\underline{s}}$ and $x_{\overline{s}}$ take different actions. Thus, the reputational gain from claiming the lower standard ex-ante only needs to compensate P for this relatively "small"

[33] Another natural comparison to look at is that of the probability of $a = 1$ across intermediate equilibria, not just with ex-ante signaling. Because these equilibria are less salient, and the analysis is more technical we defer this question to Appendix G. There, we show that for any two ISS that are Blackwell comparable, the more Blackwell informative ISS has a higher probability of $a = 1$ for sufficiently high ρ.

average loss. In contrast, under ex-post signaling, $x_{\underline{s}}$ and $x_{\overline{s}}$ can only be distinguished after pivotal evidence realizations, upon which, mimicking $\underline{s}$ requires P to take $a = 1$ with probability one. As a result, the required reputational compensation must be larger, which implies that P takes $a = 1$ less often.

This intuition relies on two kinds of uncertainty in our model. First, uncertainty in the standard leads to uncertainty in the contingent plan that will be used by the non-partisan type, i.e., x_s depends on s. Second, due to the uncertainty over the evidence at the communication stage, there is corresponding uncertainty about whether two such contingent plans will be distinguishable ex-post, i.e., choose different actions following the evidence realization. The importance of these is clear in our binary-standards analysis. If the first type of uncertainty were absent, then $x_{\underline{s}}$ and $x_{\overline{s}}$ would coincide and all equilibria would be equivalent. If the second type of uncertainty were absent, then $x_{\underline{s}}(e) \neq x_{\overline{s}}(e)\ \forall e \in \text{Supp}(F)$ and again the two equilibria would coincide. This is because the cost of mimicking $x_{\underline{s}}$ relative to $x_{\overline{s}}$ is only lower ex-ante because $x_{\underline{s}}$ and $x_{\overline{s}}$ take the same action for some "non-pivotal" evidence realizations. If evidence is degenerate, then one of these types of uncertainty is absent. Indeed, if there is no uncertainty over evidence, then it is as if communication happens after evidence is revealed at the decision stage, and so it will have no impact on equilibrium outcomes.

Guaranteeing these two types of uncertainty is the purpose of the strictness conditions in Theorem 1: residual strategic uncertainty guarantees that there exist two x_s contingent plans that admit pivotal evidence realizations, and mild agreement guarantees that every two x_s admit non-pivotal evidence realizations.[34] These conditions are relatively lenient; for example, both hold whenever F has full support on $\mathbb{R}$.

The main intuition above does not depend on many specifics of our model. In Subsection 6.1, we generalize the characterization of the equilibrium set and Theorem 1 to a setting with arbitrary preferences and actions, and discuss how the two types of uncertainty mentioned above remain central to this comparison.

4.3. Extending the Binary-Standards Intuition

We next discuss how to extend the logic from Subsection 4.1 with binary standards to establish Proposition 1 and Theorem 1. The first extension turns out to be straightforward. In our proof of Proposition 1, we show that for any evidence e and distribution over standards G, there exists an alternative distribution over standards $\tilde{G}$ with binary

[34] 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., $1 - F(\tilde{e}_s)$).

support that delivers the same probability of $a = 1$ conditional on evidence e under both ex-post and ex-ante signaling. The result then follows from the logic in Subsection 4.1 concerning binary standards.

We next show how Proposition 1 can be used to establish Theorem 1. This step relies on the following 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 also dictates how P's equilibrium expected reputation changes. 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.

Combining Lemma 3 and Proposition 1 gives that P's utility is lower under ex-ante signaling than ex-post signaling. In addition, Lemma 3 implies that establishing Theorem 1 is equivalent to establishing that P's utility is lower under ex-ante signaling than any other equilibrium—namely, $U_P^\alpha \leq U_P^{\mathcal{E}}$. Toward this end, fix an arbitrary equilibrium $\mathcal{E}$ with associated interim belief ν_1. Let $U_P^{\alpha,m}$ and $U_P^{\beta,m}$ respectively be the ex-ante and ex-post signaling utilities for P under prior $\nu_1(\cdot|m)$ after on-path message m. Note that P's indifference over messages implies that his expected utility under $\mathcal{E}$ is equal to his expected utility conditional on sending an on-path message m, which itself is equal to the ex-post signaling equilibrium utility with prior equal to the interim belief $\nu_1(\cdot|m)$, i.e., $U_P^{\beta,m} = U_P^{\mathcal{E}}$ for every on-path m. Using the comparison between ex-post and ex-ante signaling in Proposition 1, we then obtain that P's expected utility conditional on sending message m under $\mathcal{E}$ is higher than under ex-ante signaling when the prior is given by the interim belief $\nu_1(\cdot|m)$—namely, $U_P^{\alpha,m} \leq U_P^{\beta,m} = U_P^{\mathcal{E}}$.

Now consider an alternative messaging strategy in which the DM selects both an "initial" message m according to the original equilibrium strategy under $\mathcal{E}$ along with a "follow-up" message m_s according to the ex-ante signaling equilibrium given prior

$\nu_1(\cdot|m)$. In addition, given that the observer forms beliefs knowing this alternative messaging strategy, the induced utility for P after initial message m is then $U_P^{\alpha,m}$ regardless of which follow up message m_s is sent. This of course may not be an equilibrium as P may no longer be indifferent across messages—i.e., $U_P^{\alpha,m}$ need not equal $U_P^{\alpha,m'}$ for on-path m, m'. To reestablish equilibrium, P must adjust his mixing probabilities to equalize these utilities. Since every s type sends a different message, the resulting equilibrium is ex-ante signaling and the associated utility for P is U_P^{α}. We show that P's utility following a message is decreasing in the rate he sends it. Thus, the resulting U_P^{α} after P equalizes utilities must be less than $\max_m U_P^{\alpha,m}$. Because we have already shown that $U_P^{\alpha,m} \leq U_P^{\mathcal{E}}$ for all m, our desired conclusion that U_P^{α} is less than $U_P^{\mathcal{E}}$ then follows.[35]

4.4. Comparing the DM's Utility

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

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

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

Given 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. The fact that each s type's contingent plan is constant across equilibria also implies that their preference across equilibria only depends on their expected reputation. The second point of Corollary 1 then implies that s types' expected reputation must be lowest under ex-ante signaling, which in turn, by Bayes plausibility, implies that P obtains his highest reputation under ex-ante signaling. At a high level, this means P and N types "separate" the least under ex-ante signaling as compared with

[35] The proof in Appendix C follows from Theorem 2 for our more general model considered in Subsection 6.1. As we describe at the end of that section, for the first step comparing ex-ante and ex-post signaling, that proof uses an alternative approach which is more easily adapted to that setting.

other equilibria. This is consistent with our main intuition which says that the cost of signaling that one is a "good" type, i.e., taking $a = 1$ more frequently, is lower from the ex-ante perspective.

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 6.2.

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.[36]

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.*

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

[36] 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.

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 results concern 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 leading interpretation, 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 other contexts, they can be interpreted as an increase in polarization on the current issue, 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 $[\underline{e}, \overline{e}]$.[37]

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 get a sense of the intuition, consider the binary-standards analysis from Subsection 4.1 where the evidence distribution has full support on the reals. An MPS in the standards distribution corresponds to moving $\underline{s}$ down and $\overline{s}$ up. Such a change induces more pivotal evidence realizations where $\underline{s}$ and $\overline{s}$ take different actions, and thereby increases P's cost of mimicking $\underline{s}$ relative to $\overline{s}$. As a result P sends message $m_{\overline{s}}$ more frequently and thereby takes $a = 1$ less frequently.

Recall that given message m_s sent at the communication stage, the probability of $a = 1$ at the decision stage is $1 - F(s - c)$. This means that, even fixing an ex-ante signaling messaging strategy, an MPS in G can affect the probability of $a = 1$ depending on the curvature of F. The regularity condition that $\frac{f}{\rho q + c(1-F)}$ is increasing guarantees that this mechanical effect does not overwhelm the former strategic effect. 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

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

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 $a=1$. If there is a large amount of mass within a small interval, then $1-F(e)$ will decrease "quickly" over this interval and $\frac{h(e)}{\rho q + c(1-F(e))}$ will be increasing in e. 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.[38] The 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

[38] In an earlier working version of this paper (McClellan and Rappoport (2024)), we fully characterize optimal investigations in this problem when the evidence is a 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; we show that this is not a feature of the optimal investigation under ex-post signaling, for which an uninformative investigation may be optimal.

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.[39] 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.

6. Extensions

6.1. General Preferences

We now extend the model to allow for more general material and reputation preferences as well as a larger set of actions.[40,41] Let A be a finite set and E a measurable space.[42] We let $w_\theta : A \times E \to \mathbb{R}$ be θ's material payoffs and assume each has finite expectation with respect to e for any action strategy. Finally, we allow for more general reputation payoffs: for some $r : \Theta \to \mathbb{R}_+$, where we normalize $r(P) = 0$.[43] We let $\rho \int_\Theta r(\theta) d\nu(\theta)$ be the reputation payoff when the public holds beliefs $\nu \in \Delta(\Theta)$. The utility for each DM type is again the sum of their material and reputational payoffs.

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

[40] While binary actions capture the most important aspects of the DM's choices in our motivating examples, richer action choices are also natural. A university may decide not only whether to admit a student, but how much financial aid to give and what programs to allow them into. The FTC may decide not only to let a merger go through or not, but to require the merging companies to divest in certain markets.

[41] Unless otherwise specified, our subsequent extensions will use the preferences and binary action space from our baseline model.

[42] A natural case not explicitly covered by the model is one where there is uncertainty about whether a decision needs to be taken at all. However, this could easily be incorporated without changing the results by allowing the set of available actions A to depend on e. We omit such dependence for ease of notation.

[43] Given that we are placing no restrictions on the DM's preferences, P can simply be defined as the type with the lowest reputation, which we then normalize to zero.

Let $\underline{r} \equiv \inf_{s\in S} r(s)$, $\overline{r} \equiv \sup_{s\in S} r(s)$, $\underline{w}_P \equiv \inf_{e\in E,\, a\in A} w_P(a,e)$ and $\overline{w}_P \equiv \sup_{e\in E, a\in A} w_P(a,e)$. To avoid uninteresting cases, we assume $\underline{w}_P < \overline{w}_P$. We then adapt Assumption 1 as follows.[44]

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

The role of Assumption 2 is identical to that of Assumption 1 in our baseline model: it ensures that neither the P type nor any s type will ever fully reveal themselves in equilibrium. The main differences from Assumption 1 are twofold. First, under the general form of P's utility, we now need the reputational incentives to compensate for his maximal utility change from switching actions. Second, we need the difference between $\underline{r}$ and $\overline{r}$ to be not too large relative to the difference between $\underline{r}$ and $r(P) = 0$ so that the benefit of appearing as any s type is attractive enough to invalidate full separation. We discuss how our methods use this assumption and what would change without it later in the section.

Our next assumption places some mild technical conditions on preferences and evidence, the first of which is equivalent in our baseline model to assuming F and G are atomless and the second ensures that material indifferences over actions is a probability zero event for P.

Assumption 3. *Both* $\{e, s : \; |\arg\max_{a\in A} (w_s(a,e) - w_P(a,e))| > 1\}$ *and* $\{e : \; \exists a, a' \in A,\; w_P(a,e) = w_P(a',e)\}$ *occur with probability zero.*

Redefine x_s by $x_s(e) = \arg\max_{a\in A} (w_s(a,e) - w_P(a,e))$;[45] this function maximizes the material utility difference between s and P at each e, and is the analogue of (2) in the baseline model. In Appendix B, we prove that the equilibrium characterization in Lemma 1 and Lemma 2 continues to hold under this definition of x_s; we omit restating these lemmas as the statements would be unchanged. That is, (i) each s type uses contingent plan x_s under any equilibrium, and P positively mixes over all messages sent by the N types at the communication stage, and then mixes over the set of x_s for s types who sent the same message at the decision stage, and (ii) every ISS is associated with a unique equilibrium outcome. In particular, under ex-ante signaling—still denoted as equilibrium α—each s type sends message m_s and the P type positively mixes over all these m_s at the communication stage. The DM then follows up message m_s with contingent plan x_s regardless of his type. That is, as in our baseline model, ex-ante signaling

[44] Our baseline model takes $\overline{r} = \underline{r} = 1$ and $\underline{w}_P = -c < 0 = \overline{w}_P$, in which case Assumption 2 is identical to Assumption 1.

[45] For e in which the $\arg\max$ is not unique, take any arbitrary selection from it.

allows the DM to effectively commit to a contingent plan and features no residual strategic uncertainty. We extend our main results to this setting:

Theorem 2. *For any $\mathcal{E}$ and F, the following hold:*

1. $U_P^\alpha(F) \leq U_P^{\mathcal{E}}(F)$,
2. $\mathbb{E}^\alpha[w_P(a,e)|\theta = P] \leq \mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta = P]$,
3. *If* $w_P(a,e) = -ca$ *for some* $c > 0$, *then* $\mathbb{E}^{\mathcal{E}}[a] \leq \mathbb{E}^\alpha[a]$,

with strict inequalities under mild-agreement if $\mathcal{E}$ has residual strategic uncertainty.

As in our baseline model, P is worst off (point 1) and acts the most against his material self-interest (point 2) under ex-ante signaling; because s type behavior is fixed across equilibria, this means that the DM unconditionally acts most against his partisan interest in ex-ante signaling. Given that we allow for arbitrary preferences for P, point 2 does not immediately translate into a ranking of the expected action. However, point 3 establishes that this ranking does hold when P has a constant disutility from increasing his action, as in our baseline model.

This theorem not only illustrates the robustness of our main results, but is also useful for understanding what drives our equilibrium comparison. It highlights that our results are driven not by the specifics of material payoffs in the baseline model, but instead by the two types of uncertainty discussed after Theorem 1: uncertainty over what contingent plan N will follow—namely, x_s—and uncertainty about whether different x_s will be distinguishable from each other ex-post. Both kinds of uncertainty are guaranteed if there is mild agreement and residual strategic uncertainty.[46] These two conditions are relatively weak—for example, residual strategic uncertainty holds whenever the DM's action cannot be perfectly predicted after the evidence is revealed, something that often arises in our motivating examples.

In our baseline model, uncertainty about what contingent plan will N will choose arises due to uncertainty about what the "right" action is conditional on evidence. More specifically, there are evidence realizations for which some s types prefer $a = 1$ more relative to the P type, while others prefer $a = 0$ more relative to the P type. While the

[46] While residual strategic uncertainty is a necessary condition for the comparison to be strict, mild agreement is simply a sufficient condition. Other conditions would also suffice—e.g., if $w_P(a,e)$ is non-constant in e for each $a \in A$—to establish a strict comparison when residual strategic uncertainty is present.

presence of such uncertainty clearly arises in a much richer class of preferences than those in our baseline model, it is not true for all preferences. For example, consider Spence (1973)-like preferences—i.e., $A \subset \mathbb{R}$ with $w_s(a, e) - w_s(a', e) > w_P(a, e) - w_P(a', e)$ for all s and $a > a'$—where the "bad" P type is identified by his maximal distaste for higher actions relative to reputation. In this case, $x_s(e) = \max\ A$ for all s, e, i.e., there is no uncertainty about what action maximizes signaling value. For such preferences, all types pool on this single contingent plan. While such preferences are captured by the theorem, the lack of residual strategic uncertainty leads all equilibria to have the same outcomes, i.e., communication has no value.

However, uncertainty about x_s is not sufficient for communication to impact outcomes. As we discuss following Theorem 1, if the evidence distribution is degenerate at $\tilde{e}$, then the second type of uncertainty is absent. Simply, for any two contingent plans x_{s_1} and x_{s_2}, there is no uncertainty about whether they will be ex-post distinguishable, i.e., they are either never distinguishable because $x_{s_1}(\tilde{e}) = x_{s_2}(\tilde{e})$, or always distinguishable because $x_{s_1}(\tilde{e}) \neq x_{s_2}(\tilde{e})$. More generally, this can be the case if there are multiple evidence realizations, but s types always disagree on the action that maximizes their utility difference relative to the P type for any evidence realization, i.e., $\forall s_1, s_2 \in S$, $x_{s_1}(e) \neq x_{s_2}(e)\ \forall e \in E$. Mild agreement rules this out.

As mentioned, Assumption 2 ensures that appearing as the P type is sufficiently worse than appearing as any other type so that neither s types nor the P type can reveal themselves following any on-path message-action combination. This is the main feature we preserve from our baseline model as it buys significant tractability, namely, it ensures there is a single type whose incentives pin down equilibrium outcomes (i.e., P is indifferent over messages and actions). However, we believe that our central intuition described in Section 4 does not depend on this assumption. To illustrate this, in Appendix F we analyze a version of our baseline model in which the DM has a continuous degree of partisanship and binary standards, and so no type is significantly worse in the eyes of the observer than other types, i.e., Assumption 2 is violated. Nonetheless, we show that the same forces as in our baseline model arise and lead to a higher probability of $a = 1$ under ex-ante signaling than ex-post signaling.

Our proof of Theorem 2 compares P's utility across equilibria and then uses a generalization of Lemma 3 to establish that the same comparison holds for P's material utility across equilibria. The idea behind the argument for the former point, i.e., point (1) in Theorem 2, is as follows. The key step is to show the comparison between ex-ante and ex-post signaling, i.e., that $U_P^{\alpha}(F) \leq U_P^{\beta}(F)$. The comparison to other equilibria then follows from the same logic as described in Subsection 4.3. Towards this end, we make two observations. Let δ_e denotes the degenerate distribution on e. First, note that all equilibria yield equivalent outcomes when F is degenerate, there is no difference between the decision stage and the communication stage. In particular $U_P^\alpha(\delta_e) = U_P^\beta(\delta_e)\ \forall e \in E$. Second, observe that the ex-post signaling utility $U_P^\beta(F)$ is linear in F owing to the fact that $v^\beta(e)$ is independent of F. These points imply

$$U_P^\beta(F) = \int_E U_P^\beta(\delta_e)dF(e) = \int_E U_P^\alpha(\delta_e)dF(e).$$

Thus, the comparison that $U_P^\alpha(F) \leq U_P^\beta(F)$ holds if $\int_E U_P^\alpha(\delta_e)dF(e) \geq U_P^\alpha(F)$ for any F. We establish this in the next lemma.

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

The high level idea is that P is hurt by uncertainty over the evidence as it makes his "targeting" less effective; when the evidence is close to certain, he can claim standards that are just out of reach of the likely evidence. This proof approach to comparing ex-ante and ex-post partisan utility is an alternative one to that developed for Proposition 1. We use this one for Theorem 2 as it is simpler to adapt to the current general model and its intermediate results are useful in other results (e.g., that in Subsection 6.3).

6.2. 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 Appendix E, 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 Appendix E, 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 observer belief $\nu \in \Delta(\Theta)$ is given by $u(\theta, e, a, \nu) + \varepsilon a$ where ε is a random variable that is mean zero, 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. 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 P type than it is for N types. The reason is that N will only take ε into account for evidence realizations where $e - s$ 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 the optional commitment x_s goes unused, then it will be given a full reputation, which is not possible in equilibrium given our high ρ assumption.

This result 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.[47] 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.

6.3. Timing of Evidence Disclosure

In our model, 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 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. Evidence e is an unknown state with prior F. Prior to the com-

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

munication stage, the DM and observer see an initial public signal about e that induces a belief $\pi \in \Delta(E)$ over e with $\pi \sim \mu \in \Delta(\Delta(E))$. Evidence is revealed as before at the decision stage, only now players assign a belief π instead of F. The preferences of the DM are the same as in Section 2 with only the final evidence e being payoff relevant. We maintain the focus on ex-ante signaling equilibria in each subgame following the realization of the signal, and so our results also apply to the commitment model.

By varying the distribution μ, we can span various timing of evidence distributions. When the initial signal is uninformative, i.e., μ is degenerate on $\pi = F$, information is "fully back-loaded" until after the DM communicates, in which case equilibrium outcomes correspond to those under ex-ante signaling in our baseline model. When signals are fully informative, i.e., μ places weight only on degenerate beliefs over E, information is "fully front-loaded" to before the DM communicates, in which case equilibrium outcomes correspond to those under ex-post signaling in our baseline model. That is, even though we focus on the ex-ante signaling equilibrium conditional on π, front-loading disclosure generates ex-post signaling outcomes due to the fact that when the evidence distribution is degenerate, ex-ante signaling and ex-post signaling are identical. More broadly, when the initial signal is more informative, this corresponds to more of the evidence being revealed up-front. Our next result examines how the timing of information disclosure impacts the actions taken.

Proposition 9. *Take information structures over evidence $\mu, \tilde{\mu} \in \Delta(\Delta(E))$ such that $\tilde{\mu}$ is more informative about $e \in E$ than μ in the Blackwell sense. Then the probability of $a = 1$ is higher under μ than $\tilde{\mu}$. In particular, when information is fully front-loaded, the probability of $a = 1$ is minimized, and when information is fully back-loaded, the probability of $a = 1$ is maximized.*

This result follows from the convexity of $U_P^{\alpha}(\cdot)$. Thus, delaying evidence disclosure hurts P and leads to a higher probability of taking $a = 1$.

6.4. Welfare Comparisons Across Equilibria

Our main results compare the probability of $a = 1$ across equilibria—both conditional and unconditional on evidence. This is an important outcome *per se*: in many natural applications, it the only publicly observed outcome. We have also argued that it is an important design objective, e.g., our results imply that a designer who prefers $a = 1$ will favor ex-ante signaling and can implement it through the methods we describe in Subsection 6.2. One case where the designer's choice is not so obvious is when they aim to maximize social welfare, where the measure of welfare is informed by our leading interpretation. That is, the standard s is a private signal representing the "correct" evidence standard for taking $a = 1$. In this section, we follow this interpretation and define social welfare given evidence e, standard s, and action $a \in \{0, 1\}$ as $(e - s)a$. Our question is which equilibrium has higher social welfare: ex-ante signaling or ex-post signaling?

To give more descriptive results, we also generalize the material utility of the N type to $(\lambda_1(e - s) - \lambda_2 c)a$ for some $\lambda_1 > 0$ and $\lambda_2 \geq 0$. This allows us to talk about the N type's degree of alignment with social objectives—λ_1—and their degree of bias—λ_2.[48] P's material utility remains $-ca$, so it is still the case that, in the absence of reputation concerns, for any such λ_1, λ_2 the N type will take actions closer to the social objective.[49] These generalizations fit into the framework of Subsection 6.1, which means our equilibrium characterization and comparison continue to hold. Indeed, we focused on the specification in our baseline model (i.e., $\lambda_1 = 1$ and $\lambda_2 = 0$) to be most parsimonious, but the analysis and interpretations for this generalized version are largely similar. In particular, the specifications where $\lambda_2 = 1$ are particularly natural, as they capture the case where the non-partisan has the same incentives towards the agenda of the partisan, but also cares about social welfare. For example, the partisan bias may be induced by external incentives for all DM types such as by a bribe for taking $a = 0$.

Because P's preferences do not depend on the standard, our equilibrium analysis is silent on how his strategy depends on the standard.[50] However, this dependence clearly affects welfare. For this extension, we focus on the selection where the P type uses a monotonic strategy, i.e., his probability of $a = 1$ is decreasing in the standard given every evidence realization and every message.[51] We defer the formalization of this monotonicity selection, as well as the expression of expected welfare to Appendix E. To guarantee strict comparisons between equilibria, we assume that F has full support over $\mathbb{R}$. To ease exposition, we also assume that G has compact support and admits a positive density.

Proposition 10.

1. *Ex-ante signaling has higher welfare than ex-post signaling if $\lambda_2 \geq 1$, or if $\lambda_2 < 1$ and λ_1*

[48] These payoffs are defined relative to reputational payoffs and so we cannot normalize one of the values of λ_i as this would change the weight on reputation for N. Indeed, taking $\lambda_1 \to \infty$ when $\lambda_2 = 0$ is equivalent to taking the weight on reputation for N to zero (holding fixed that for P).

[49] P's material utility is maximized over contingent plans by choosing $x(e) = 0$ for all e. In contrast, the s type's material utility is maximized by the threshold contingent plan $\mathbb{1}(e \geq s + \frac{\lambda_2}{\lambda_1} c)$ which always takes $a = 0$ when society prefers it, and sometimes takes $a = 1$ when society prefers it.

[50] Recall that as a convention, P observes the standard s, even though this assumption does not affect the analysis for our previous results as it is payoff irrelevant for him.

[51] Indeed, these outcomes represent the unique limit of equilibria in the version of the model where P's material utility is given by $(\gamma(e - s) - c)a$, which fits into our framework of Subsection 6.1, as $\gamma \to 0$.

is sufficiently large.

2. *If $\lambda_2 < 1$, then ex-post signaling has higher welfare than ex-ante signaling if ρ is sufficiently large or if λ_1 is sufficiently small.*

To see the logic for these comparisons, note that Proposition 1 says that ex-ante signaling increases the probability of $a = 1$ for every evidence realization.[52] In our monotonic selection, this corresponds to the P type using contingent plans whose threshold evidence level for taking $a = 1$ is shifted downwards under ex-ante signaling relative to ex-post signaling for each standard s.

Whether this downshift in thresholds is welfare improving depends on which thresholds the N type is using, i.e., which thresholds the P type is shifting towards under ex-ante signaling. Our analysis in Subsection 6.1 shows that each (N, s) type uses evidence threshold $s - \frac{1-\lambda_2}{\lambda_1}c$. Thus, when $\lambda_2 \geq 1$, despite being lower than that used by the (P, s) type, the (N, s) type's thresholds are still (weakly) higher than the socially optimal threshold due to their bias. Thus, the (P, s) type's thresholds moving closer to the (N, s) type's thresholds under ex-ante signaling also represents a movement toward society's preferred thresholds.

When $\lambda_2 < 1$, political correctness in the N's types strategy causes them to "overshoot" the socially optimal threshold and take $a = 1$ too frequently (as compared to the socially optimal rule). The (P, s) type mimicking the (N, s) type's threshold more closely under under ex-ante signaling then also has the potential to similarly overshoot and be worse for social welfare than ex-post signaling. As λ_1 becomes large, the high social alignment implies that the (N, s) types choose evidence thresholds very close to s, i.e., very close to the social optimum, and so ex-ante signaling is still preferred as it will not overshoot the optimal threshold. In contrast, when λ_1 is small, the low alignment relative to reputation concerns causes extreme political correctness on behalf of the N types, and so ex-post signaling has higher welfare. Lastly, this overshooting problem becomes worse as ρ increases because the high reputation concerns cause the (P, s) type's threshold choices to move closer to that of the (N, s) types, becoming identical in the limit as $\rho \to \infty$.

[52] We adapt this proposition for the current altered N preferences in the Appendix E.

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.

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. (2025). Scoring strategic agents. *American Economic Journal: Microeconomics*, 17(1):97–129.

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

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.

Farrell, J. and Rabin, M. (1996). Cheap talk. *Journal of Economic perspectives*, 10(3):103–118.

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.

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

McClellan, A. and Rappoport, D. (2024). Signaling good faith by taking stands.

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

Orzach, R. (2025). Self-censorship of hard information.

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. *American 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.

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

We prove Lemma 1, Lemma 2 and Theorem 2 under the more general preferences (and corresponding assumptions) of Subsection 6.1. Theorem 1 immediately follows. Proofs for the rest of the results maintain the assumptions of our baseline model.

A. Preliminaries

We begin by defining some useful notation. Given an equilibrium $\mathcal{E}$, let $q_m \equiv \nu_1(S|m)$ be the interim belief that $\theta \in S$ type after message m is sent, $q(m,e,a) \equiv \nu_2(S|m,e,a)$ be the posterior belief that $\theta \in S$ after message m, action a and evidence e. The reputational payoff is $R(m,e,a) \equiv \int_\Theta r(\theta) d\nu_2(\theta|m,e,a)$. Let $A^*_\theta(m,e) \equiv \arg\max_a \; u(\theta,e,a,R(m,e,a))$. 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 $U^{\mathcal{E}}_\theta(F)$ in the proofs below and for those from Section 3 as it is held fixed. Let $U^{\mathcal{E}}_{\theta,m}$ be the equilibrium expected utility to θ from sending message m.

Our first result uses Assumption 2 to place bounds on the reputations that may arise in equilibrium. We say a is *off-path* after m, e if a is played with probability zero after m, e.[53] 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 $a \in A$, $q_m \leq \frac{\rho q \bar{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}} < 1$ and $q(m,e,a) < 1$. For all $m \in M^*_P$ and $a \in A^*_P(m,e)$, $q_m > 0$ and $q(m,e,a) > 0$.*

Proof. First, we show that $U^{\mathcal{E}}_{P,m} \in [\underline{w}_P + \rho q_m \underline{r}, \overline{w}_P + \rho q_m \bar{r}]$ 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 $\sum_{a \in A} q(m,e,a)\zeta(a|P,m,e)$, is at most q_m. Using $R(m,e,a) \leq \bar{r} q(m,e,a)$, we then have

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

[53] 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 $\theta \in \text{Supp}(\Theta_m)$.

For each 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 $U^{\mathcal{E}}_{P,m} \geq \int_E (w_P(\overline{a}_e, e) + \rho R(m, \overline{a}_e, e))dF(e)$. Using $R(m, e, \overline{a}_e) \geq q(m, e, \overline{a}_e)\underline{r} \geq q_m\underline{r}$, we then have $\underline{w}_P + \rho q_m \underline{r} \leq U^{\mathcal{E}}_{P,m}$.

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 $U^{\mathcal{E}}_{P,m} \leq \overline{w}_P + \rho q_m \overline{r} \leq \overline{w}_P + \rho q \overline{r}$ 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 $U^{\mathcal{E}}_{P,m'} \geq \underline{w}_P + \rho q_{m'}\underline{r} \geq \underline{w}_P + \rho q \underline{r}$.

Next, for any $m \in M$, we show $q_m \leq \frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}} < 1$. Using $\underline{w}_P + \rho q_m \underline{r} \leq U^{\mathcal{E}}_{P,m} \leq U^{\mathcal{E}}_P \leq \overline{w}_P + \rho q \overline{r}$, we have $q_m \leq \frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}}$. If $\frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}} \geq 1$, then $\rho \leq \frac{\overline{w}_P - \underline{w}_P}{\underline{r} - q\overline{r}}$, because $\underline{r} - q\overline{r} > 0$ per Assumption 2. Using the same assumption, $\rho \geq \frac{(\overline{w}_P - \underline{w}_P)(\underline{r} + \overline{r})}{\underline{r}^2 - q\overline{r}^2}$, so $\frac{\overline{w}_P - \underline{w}_P}{\underline{r} - q\overline{r}} \geq \frac{(\overline{w}_P - \underline{w}_P)(\underline{r} + \overline{r})}{\underline{r}^2 - q\overline{r}^2}$, which simplifies to $0 \geq \overline{r}\underline{r}(1-q)$, a contradiction. Similarly, for any $m \in M^*_P$, using $\underline{w}_P + \rho q \underline{r} \leq U^{\mathcal{E}}_P = U^{\mathcal{E}}_{P,m} \leq \overline{w}_P + \rho q_m \overline{r}$, we have $\rho q_m \geq \frac{\rho q \underline{r} + \underline{w}_P - \overline{w}_P}{\overline{r}}$. By Assumption 2, $\rho \geq \frac{(\overline{w}_P - \underline{w}_P)(\underline{r} + \overline{r})}{q\underline{r}^2} > \frac{\overline{w}_P - \underline{w}_P}{q\underline{r}}$, so $\rho q \underline{r} + \underline{w}_P - \overline{w}_P > 0$, 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, by Bayes plausibility, there exists $a' \in A^*_P(m, e)$ with $a' \neq a$ and $R(m, e, a') \leq q_m \overline{r}$. For P not to have a profitable deviation to choose a, it must be that $w_P(a', e) + \rho q_m \overline{r} \geq w_P(a, e) + \rho \underline{r}$, which implies $q_m \geq \frac{w_P(a,e) - w_P(a',e) + \rho \underline{r}}{\rho \overline{r}} \geq \frac{\rho \underline{r} + \underline{w}_P - \overline{w}_P}{\rho \overline{r}}$. Combining this inequality with $q_m \leq \frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}}$ and simplifying, we conclude that $\rho \leq \frac{(\overline{w}_P - \underline{w}_P)(\overline{r} + \underline{r})}{\underline{r}^2 - q\overline{r}^2}$, a contradiction of Assumption 2. Therefore, $q(m, e, a) < 1$.

Next, take $m \in M^*_P, e \in E$ and, for the sake of contradiction, suppose $q(m, e, a) = 0$ for some $a \in A^*_P(m, e)$. P's utility from taking action a is then $w_P(a, e)$. Take a' such that $q(m, e, a') \geq q_m$ (such an a' exists by Bayes plausibility), so $R(m, e, a') \geq q(m, e, a')\underline{r} \geq q_m \underline{r}$. Optimality of a for P implies $w_P(a', e) + \rho q_m \underline{r} \leq w_P(a, e)$; simplifying, we get $\rho q_m \leq \frac{w_P(a,e) - w_P(a',e)}{\underline{r}} \leq \frac{\overline{w}_P - \underline{w}_P}{\underline{r}}$. Combining this inequality with $\rho q_m \geq \frac{\rho q \underline{r} - (\overline{w}_P - \underline{w}_P)}{\overline{r}}$ and simplifying, we have $\rho \leq \frac{(\overline{w}_P - \underline{w}_P)(\overline{r} + \underline{r})}{q\underline{r}^2}$, a contradiction of Assumption 2. Therefore, $q(m, e, a) > 0$. *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 and let a' be an on-path action. Let

$U^*_m(\theta, e)$ be the equilibrium utility from type θ sending message m after evidence e is realized. 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

$$\begin{aligned}&\{\nu \in \Delta(\Theta_m):\ u\big(\theta'', e, a, \int_{\Theta_m} r(\theta) d\nu(\theta)\big) \geq U^*_m(\theta'', e)\} \\ &\subsetneq \{\nu \in \Delta(\Theta_m):\ u\big(\theta', e, a, \int_{\Theta_m} r(\theta) d\nu(\theta)\big) > U^*_m(\theta', e)\}.\end{aligned} \tag{3}$$

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. Define $\Delta_s(a, e) \equiv w_s(a, e) - w_P(a, e)$, which implies that $x_s(e) \in \arg\max_a\ \Delta_s(a, e)$.

Lemma 6. *Take any $m \in M$ such that $q_m > 0$, $e \in E$ and for each $s \in S_m$, let $a'_s \in A^*_s(m, e)$. Let a be an off-path action following m, e. If $\Delta_s(a, e) > \Delta_s(a'_s, e)$ for some $s \in S_m$, then D1 requires $q(m, e, a) = 1$. If $\Delta_s(a, e) < \Delta_s(a'_s, e)$ for all $s \in S_m$, then D1 requires $q(m, e, a) = 0$.*

Proof. Let a be an off-path action following m, e. By $q_m > 0$, $S_m \neq \emptyset$. By Bayes plausibility, there exists on-path $a' \in \mathrm{Supp}(\zeta(\cdot|P, m, e))$ (which implies $a' \in A^*_P(m, e)$) for which $q(m, e, a') \leq q_m$, so $R(m, e, a') \leq q_m \overline{r}$. We note that

$$\begin{aligned}&\{\nu \in \Delta(\Theta_m):\ u(P, e, a, \int_{\Theta_m} r(\theta) d\nu(\theta)) > U^*_m(P, e)\} \neq \emptyset \\ &\iff \rho(\max_{s \in S_m} r(s) - R(m, e, a')) > w_P(a', e) - w_P(a, e).\end{aligned} \tag{4}$$

The last inequality of (4) holds if $\rho(\underline{r} - q_m \overline{r}) > \overline{w}_P - \underline{w}_P$, or equivalently $q_m < \frac{\rho \underline{r} - (\overline{w}_P - \underline{w}_P)}{\rho \overline{r}}$, which holds because, by Lemma 5, $q_m < \frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}}$ and $\frac{\rho q \overline{r} + \overline{w}_P - \underline{w}_P}{\rho \underline{r}} < \frac{\rho \underline{r} - (\overline{w}_P - \underline{w}_P)}{\rho \overline{r}}$ by $\rho \geq \frac{(\overline{w}_P - \underline{w}_P)(\underline{r} + \overline{r})}{\underline{r}^2 - q \overline{r}^2}$ (Assumption 2). Thus, the set of beliefs that could induce P to deviate to an off-path action is non-empty.

Next, we argue that $A^*_s(m, e) \subseteq A^*_P(m, e)$ for $s \in S_m$. For the sake of contradiction, suppose there exists $a \in A^*_s(m, e) \backslash A^*_P(m, e)$. Then a is off path, as otherwise $q(m, e, a) = 1$, a contradiction of Lemma 5. In order for $q(m, e, a) < 1$, (3) must not hold when $\theta' = s, \theta'' = P$, which simplifies to

$$U^*_m(P, e) - w_P(a, e) \leq U^*_m(s, e) - w_s(a, e).$$

Note that $U^*_m(s, e) = w_s(a, e) + \rho R(m, e, a)$ and $U^*_m(P, e) > w_P(a, e) + \rho R(m, e, a)$ (by

$a \in A_s^*(m,e)$ and $a \notin A_P^*(m,e)$ respectively). Plugging these into the above inequality and simplifying yields a contradiction.

Take any $s \in S_m$ and $a'_s \in A_s^*(m,e)$. The above argument implies $a'_s \in A_P^*(m,e)$, so $U_m^*(P,e) = u(P,e,a'_s,R(m,e,a'_s))$. D1 requires $\nu_2(P|m,e,a) = 0$ (which implies $q(m,e,a) = 1$) if (3) holds for $\theta'' = P$ and some $\theta' \in S_m$, which simplifies to $\Delta_s(a,e) > \Delta_s(a'_s,e)$ for some $s \in S_m$. Similarly, D1 requires $\nu_2(S_m|m,e,a) = 0$ (which implies $q(m,e,a) = 0$) if (3) holds for $\theta' = P$ and all $\theta'' \in S_m$, which simplifies to $\Delta_s(a'_s,e) > \Delta_s(a,e)$ for all $s \in S_m$ (with $a'_s \in A_s^*(m,e)$). *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, $M' \subseteq M_P^*$ and 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 $m \in M_P^*$, $s \in S_m$ and e such that $|\arg\max_a \Delta_s(a,e)| = 1$. For the sake of contradiction, suppose there exists $a \in A_s^*(m,e)$ with $a \neq x_s(e)$. Then s (weakly) prefers a over $x_s(e)$, so

$$w_s(a,e) + \rho R(m,e,a) \geq w_s(x_s(e),e) + \rho R(m,e,x_s(e)). \tag{5}$$

P (weakly) prefers $x_s(e)$ to a if and only if

$$w_P(x_s(e),e) + \rho R(m,e,x_s(e)) \geq w_P(a,e) + \rho R(m,e,a). \tag{6}$$

Adding (5) to (6) and simplifying yields $\Delta_s(a,e) \geq \Delta_s(x_s(e),e)$, a contradiction of the definition of $x_s(e)$ and uniqueness of $\arg\max_a \Delta_s(a,e)$. Therefore, (6) cannot hold, which implies $x_s(e) \notin A_P^*(m,e)$. If $x_s(e)$ is off-path, then, by Lemma 6, $q(m,e,x_s(e)) = 1$, which contradicts Lemma 5. Thus, $x_s(e)$ must be on-path. But then $\zeta(x_s(e)|P,m,e) = 0$ by $x_s(e) \notin A_P^*(m,e)$, which implies $q(m,e,x_s(e)) = 1$, again a contradiction. We conclude that $A_s^*(m,e) = \{x_s(e)\}$, which implies $\zeta(x_s(e)|s,m,e) = 1$.

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 $|\arg\max_a \ \Delta_s(a, e)| = 1$, which by Assumption 3 is a probability one set. Therefore, $\int_E \int_S \int_M \zeta(x_s(e)|s, m, e) d\sigma(m|s) dG(s) dF(e) = 1$.

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 $\tilde{M} \subseteq M$.

Adopting the convention that $\frac{0}{0} = 0$, define

$$\tilde{R}_a(e, z; \tilde{q}, \tilde{G}) \equiv \frac{\tilde{q} \int_S r(s) \mathbb{1}(a = x_s(e)) d\tilde{G}(s)}{\tilde{q} \int_S \mathbb{1}(a = x_s(e)) d\tilde{G}(s) + (1 - \tilde{q}) z}.$$

Let $\overline{q} = \frac{\rho \underline{r} + \underline{w}_P - \overline{w}_P}{\rho \overline{r}}$, which is strictly greater than zero by Assumption 2, and $\tilde{A}(e; \tilde{G}) \equiv \{a \in A : \ \int_S \mathbb{1}(a = x_s(e)) d\tilde{G}(s) > 0\}$. For $\tilde{q} \in (0, \overline{q})$ and $\tilde{G}$ a CDF over S, define $z_a(\cdot; \tilde{q}, \tilde{G})$ in the following way (for notational convenience, we will drop dependence on $\tilde{G}$ in $\tilde{R}, \tilde{A}, \tilde{z}$ where no confusion is caused). If $a \notin \tilde{A}(e)$, set $z_a(e; \tilde{q}) = 0$. If $\{a\} = \tilde{A}(e)$, set $z_a(e; \tilde{q}) = 1$. If $\tilde{A}(e) = \{a_1, ..., a_N\}$, define $z_a(e; \tilde{q})$ for $a \in \tilde{A}(e)$ such that $\sum_{a \in \tilde{A}(e)} z_a(e; \tilde{q}) = 1$ and

$$w_P(a_1, e) + \rho \tilde{R}_{a_1}(e, z_{a_1}(e; \tilde{q}); \tilde{q}) = ... = w_P(a_N, e) + \rho \tilde{R}_{a_N}(e, z_{a_N}(e; \tilde{q}); \tilde{q}). \tag{7}$$

We now show such a unique solution $\{z_{a_n}\}_{n=1}^N$ to above equations exists and that $z_{a_n} \in (0, 1)$ for all n. For $k \in \mathbb{R}$, define $z_n(k)$ such that $w_P(a_n, e) + \rho \tilde{R}_{a_n}(e, z_n(k); \tilde{q}) = k$ if a solution exists; if $k > w_P(a_n, e) + \rho \tilde{R}_{a_n}(e, 0; \tilde{q})$, set $z_n(k) = 0$; if $k < w_P(a_n, e) + \rho \tilde{R}_{a_n}(e, 1; \tilde{q})$, set $z_n(k) = 1$; because $\tilde{R}_{a_n}(e, z; \tilde{q})$ is continuous and strictly decreasing in z, for all other k there exists a unique $z_n(k) \in (0, 1)$ that solves our equation. In addition, $z_n(k)$ is continuous and decreasing in k, strictly so when $z_n(k) \in (0, 1)$. For $k = \underline{w}_P$, we have

$z_n(k) = 1$ for all n while for $k = \overline{w}_P + \rho\overline{r}$, $z_n(k) = 0$ for all n. Therefore, we can find a unique value k such that $\sum_{n=1}^{N} z_n(k) = 1$; call this $\tilde{k}$. Given our definition of $z_n(k)$, we have that (7) holds (taking $z_{a_n}(e;\tilde{q}) = z_n(k)$) if $z_n(\tilde{k}) \in (0,1)$ for all $n = 1, ..., N$. For the sake of contradiction, suppose $z_n(\tilde{k}) = 0$ for some n; without loss, assume $n = 1$. Then there exists n' with $z_{n'}(\tilde{k}) > 0$ and $\tilde{R}_{a_{n'}}(e, z_{n'}(\tilde{k}); \tilde{q}) < \tilde{q}\overline{r}$. We then have $w_P(a_1, e) + \rho\tilde{R}_{a_1}(e, 0; \tilde{q}) \leq \tilde{k} \leq w_P(a_{n'}, e) + \rho\tilde{R}_{a_{n'}}(e, z_{n'}(\tilde{k}); \tilde{q})$, which itself implies $\underline{w}_P + \rho\underline{r} \leq \overline{w}_P + \rho\tilde{q}\overline{r}$. Simplifying this last inequality yields $\tilde{q} \geq \frac{\rho\underline{r} + \underline{w}_P - \overline{w}_P}{\rho\overline{r}} = \overline{q}$, a contradiction. Therefore, $z_n(\tilde{k}) > 0$ for all n, which by $\sum_{n=1}^{N} z_n(k) = 1$, implies $z_n(k) < 1$ for all n. Therefore, (7) holds. Because $\tilde{R}_a(e, z; \tilde{q}, \tilde{G})$ is continuous in $z, \tilde{q}$ for all e and a, the solution $z_a(e;\tilde{q})$ is also continuous in $\tilde{q}$.

With some abuse of notation, let $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ be equal to $\tilde{R}_a(e, z_a(e; \tilde{q}, \tilde{G}); \tilde{q}, \tilde{G})$. For any $a \in \{x_s(e)\}_{s \in \mathrm{Supp}(\tilde{G})}$, we have that $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ is continuous in $\tilde{q}$ because $\tilde{R}_a(e, z; \tilde{q}, \tilde{G})$ is continuous in z and $\tilde{q}$ and $\tilde{z}_a(e;\tilde{q})$ is continuous in $\tilde{q}$. We also argue that $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ is strictly increasing in $\tilde{q}$ on $(0, \overline{q})$. Take $\tilde{q}_1 < \tilde{q}_2 < \overline{q}$ and suppose $\tilde{R}_a(e; \tilde{q}_1, \tilde{G}) \geq \tilde{R}_a(e; \tilde{q}_2, \tilde{G})$ for some $a \in \tilde{A}(e; \tilde{G})$. This inequality, along with fact that $w_P(a', e) + \rho\tilde{R}_{a'}(e; \tilde{q}, \tilde{G})$ is constant across a' for all $a' \in \tilde{A}(e; \tilde{G})$, implies that $\tilde{R}_{a'}(e; \tilde{q}_1, \tilde{G}) \geq \tilde{R}_{a'}(e; \tilde{q}_2, \tilde{G})$ for all $a' \in \tilde{A}(e; \tilde{G})$. Because $\tilde{R}_{a'}(e, z; \tilde{q}, \tilde{G})$ is strictly increasing in $\tilde{q}$ and strictly decreasing in z, this inequality implies that $z_{a'}(e; \tilde{q}_1, \tilde{G}) < z_{a'}(e; \tilde{q}_2, \tilde{G})$ for all $a' \in \tilde{A}(e; \tilde{G})$, which is a contradiction as both sides of this inequality must sum to 1 over $a' \in \tilde{A}(e; \tilde{G})$. Therefore, $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ is strictly increasing in $\tilde{q}$ on $(0, \overline{q})$.

For an arbitrary $\tilde{q} \in (0, \overline{q})$ and CDF $\tilde{G}$ on S, define for an arbitrary $a \in \tilde{A}(e; \tilde{G})$

$$J(e; \tilde{q}, \tilde{G}) = w_P(a, e) + \rho\tilde{R}_a(e; \tilde{q}, \tilde{G})).$$

Because $w_P(a', e) + \rho\tilde{R}_{a'}(e; \tilde{q}, \tilde{G}))$ is the same for all $a' \in \tilde{A}(e; \tilde{G})$, the value of $J(e; \tilde{q}, \tilde{G})$ is the same for any selection of $a \in \tilde{A}(e; \tilde{G})$. Given our constructed strategy, $J(e; \tilde{q}, \tilde{G})$ 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

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

Our next result gives some properties of $\overline{J}$.

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

Proof. That $\tilde{R}_a(e; \tilde{q}, \tilde{G})$ is continuous and strictly increasing in $\tilde{q}$ then implies that $\overline{J}(\tilde{q}; \tilde{G})$

is as well. We now show $J(e;\tilde{q},\tilde{G}) \in [\underline{w}_P + \rho\tilde{q}\underline{r}, \overline{w}_P + \rho\tilde{q}\overline{r}]$ (which immediately implies $\overline{J}$ respects the same bounds). When $|\tilde{A}(e)| = 1$ there is no updating of beliefs, so $R_a(e;\tilde{q},\tilde{G}) = \tilde{q}\int_S r(s)d\tilde{G}(s) \in [\tilde{q}\underline{r}, \tilde{q}\overline{r}]$ for $a \in \tilde{A}(e)$ and these bounds hold for $J(e;\tilde{q},\tilde{G})$. Take e such that $|\tilde{A}(e)| > 1$. By Bayes plausibility, there exists $a, a' \in \tilde{A}(e)$ such that

$$\tilde{R}_a(e;\tilde{q},\tilde{G}) \geq \tilde{q}\int_S r(s)d\tilde{G}(s) \geq \tilde{R}_{a'}(e;\tilde{q},\tilde{G}).$$

These inequalities imply $\tilde{R}_a(e;\tilde{q},\tilde{G}) \geq \tilde{q}\underline{r}$ and $\tilde{R}_{a'}(e;\tilde{q},\tilde{G}) \leq \tilde{q}\overline{r}$. By definition of $z_a(e;\tilde{q},\tilde{G})$, we have $J(e;\tilde{q},\tilde{G}) = w_P(x_s(e),e) + \rho\tilde{R}_{x_s(e)}(e;\tilde{q},\tilde{G})$ for all $s \in \text{Supp}(\tilde{G})$. Our desired bounds then follow. ■

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 $\hat{M} \subseteq M$. These correspond to the rate that P uses messages relative to N types. 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 about θ 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, $\rho > \frac{(\overline{w}_P - \underline{w}_P)(\overline{r}+\underline{r})}{\underline{r}^2 - q\overline{r}^2}$, which implies $\underline{w}_P + \rho\overline{q}\underline{r} > \overline{w}_P + \rho q\overline{r}$. Similarly, $\rho > \frac{(\overline{w}_P - \underline{w}_P)(\underline{r}+\overline{r})}{q\underline{r}^2}$ implies $\overline{w}_P < \underline{w}_P + \rho q\underline{r}$. Using these bounds and the bounds on $\overline{J}$ from Claim 1, we have

$$\begin{aligned} \lim_{t \downarrow \underline{t}} \overline{J}(\varphi(t), G_\nu) &\geq \lim_{t \downarrow \underline{t}} \underline{w}_P + \rho\varphi(t)\underline{r} = \underline{w}_P + \rho\overline{q}\underline{r} > \overline{w}_P + \rho q\overline{r}, \\ \lim_{t \to \infty} \overline{J}(\varphi(t), G_\nu) &\leq \lim_{t \to \infty} \overline{w}_P + \rho\varphi(t)\overline{r} = \overline{w}_P < \underline{w}_P + \rho q\underline{r}. \end{aligned} \tag{8}$$

For $U \in [\underline{w}_P + \rho q\underline{r}, \overline{w}_P + \rho q\overline{r}]$, define $\psi^*(U; m_\nu)$ to be the value of t such that $U = \overline{J}(\varphi(t), G_\nu)$—that is, $\psi^*(U; m_\nu)$ is the rate at which P must send message m_ν in order for his associated expected payoffs at the decision stage, given indifference across N's used actions (i.e., $\overline{J}$) to be equal to U given the the updated belief that the DM is type N is $\varphi(\psi^*(U; m_\nu))$. We note that such a t exists and is unique follows from (8) and the fact that $\overline{J}(\cdot, G_\nu)$ is continuous and strictly increasing. Because, in addition, $\varphi(\cdot)$ is continuous and strictly decreasing, $\psi^*(U; m_\nu)$ is continuous and strictly decreasing in U. By (8), this implies $\psi^*(U; m_\nu) > \underline{t}$ for all $U \in [\underline{w}_P + \rho q\underline{r}, \overline{w}_P + \rho q\overline{r}]$ (and hence $\varphi(\psi^*(U; m_\nu)) \in (0, \overline{q})$).

Claim 2. *There exists a unique $U^* \in [\underline{w}_P + \rho q\underline{r}, \overline{w}_P + \rho q\overline{r}]$ such that $1 = \int_{M_\Lambda} \psi^*(U^*; m_\nu)d\Sigma_N(m_\nu)$. Moreover, $\underline{w}_P + \rho\varphi(\psi^*(U^*; m_\nu))\underline{r} > \overline{w}_P$ 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 $U = \underline{w}_P + \rho q \underline{r}$. Because $\overline{J}(\tilde{q}; G_\nu) \geq \underline{w}_P + \rho \tilde{q} \underline{r}$ for all $\tilde{q} \in (0, \overline{q})$, we have

$$\underline{w}_P + \rho q \underline{r} = U = \overline{J}(\varphi(\psi^*(U; m_\nu)), G_\nu) \geq \underline{w}_P + \rho \varphi(\psi^*(U; m_\nu))\underline{r}.$$

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

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

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 $\psi^*(U; m_\nu)$ is continuous and strictly decreasing in U, there exists a unique $U^* \in [\underline{w}_P + \rho q \underline{r}, \overline{w}_P + \rho q \overline{r}]$ 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

$$\underline{w}_P + \rho q \underline{r} \leq \underline{w}_P + \rho \varphi(\psi^*(U^*; m_{\nu'}))\underline{r} \leq U^* \leq \overline{w}_P + \rho \varphi(\psi^*(U^*; m_\nu))\overline{r},$$

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

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, $\sigma(m_\nu|s) = 0 \ \forall s \notin \mathrm{Supp}(\nu)$. 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

$$\zeta(a|s,m,e)=\begin{cases}\mathbb{1}(a=x_s(e)) & \text{if } m=m_\nu\in M_\Lambda,\ s\in S_m,\\ \frac{\mathbb{1}\left(a\in\arg\max_{a'}\ u(s,e,a',\tilde{R}_{a'}(e;\nu_1(S|m_\nu),G_\nu))\right)}{|\arg\max_{a'}\ u(s,e,a',\tilde{R}_{a'}(e;\nu_1(S|m_\nu),G_\nu))\big)|} & \text{if } m=m_\nu\in M_\Lambda,\ s\notin S_m,\\ \frac{\mathbb{1}\left(a\in\arg\max_{a'}\ w_s(a',e)\right)}{|\arg\max_{a'}\ w_s(a',e)|} & \text{if } m\notin M_\Lambda,\end{cases}$$

$$\zeta(a|P,m,e)=\begin{cases}z_a(e;\varphi(\psi^*(U^*;m_\nu),G_\nu) & \text{if } m=m_\nu\in M_\Lambda,\\ \frac{\mathbb{1}(a\in\arg\max_{a'}\ w_P(a',e))}{|\arg\max_{a'}\ w_P(a',e)|} & \text{if } m\notin M_\Lambda.\end{cases}$$

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 any $a\in\arg\max_{a'}\ w_P(a',e)$ is clearly optimal; thus, P has no incentive to deviate after such m. Take $m_\nu\in M_\Lambda$. P has no incentive to deviate to actions among $\tilde{A}(e)$ after e by construction of z_a. Any deviation to $a\notin\tilde{A}(e)$ is unprofitable as such a deviation leads to a utility of at most $\overline{w}_P$ while P's payoff from the proposed equilibrium is at least $\underline{w}_P+\rho\varphi(\psi^*(U^*;m_\nu)\underline{r}$, which is greater than $\overline{w}_P$. 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 $U^*\geq\underline{w}_P+\rho q\underline{r}$, strictly prefers the expected utility of U^* from any $m_\nu\in M_\Lambda$ to the expected utility from sending $m\notin M_\Lambda$, which is at most $\overline{w}_P$.

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 from the definition of ζ because, for such m, the definition of ζ only puts weight on actions that maximize the material payoff, which maximizes the utility as the reputations are equal to zero for $m\notin M_\Lambda$. Take an arbitrary $m_\nu\in M_\Lambda,\ s\in S_{m_\nu}$ and e. Take any $a'\neq x_s(e)$. By the definition of x_s, $\Delta_s(x_s(e),e)\geq\Delta_s(a',e)$. By the definition of z and Claim 2, $x_s(e)\in\text{Supp}(\zeta(\cdot|P,m_\nu,e))$, so P weakly prefers $x_s(e)$ to a', namely $w_P(x_s(e),e)+\rho R(m_\nu,e,x_s(e))\geq w_P(a',e)+\rho R(m_\nu,e,a')$. If s has a strict incentive to deviate to a', then $w_s(x_s(e), e) + \rho R(m_\nu, e, x_s(e)) < w_s(a', e) + \rho R(m_\nu, e, a')$. Subtracting the inequality for P and simplifying, we get $\Delta_s(x_s(e), e) < \Delta_s(a', e)$, 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

$$\int_E (w_s(x'(e), e) + \rho R(m', x'(e), e)) dF(e) > \int_E (w_s(x_s(e), e) + \rho R(m_\nu, x_s(e), e)) dF(e). \quad (9)$$

By our earlier arguments, P (weakly) prefers to send m_ν and follow with x_s than send m' and follow with x':

$$\int_E (w_P(x_s(e), e) + \rho R(m_\nu, x_s(e), e)) dF(e) \geq \int_E (w_P(x'(e), e) + \rho R(m', x'(e), e)) dF(e). \quad (10)$$

Adding (9) and (10) and simplifying yields $\int_E \Delta_s(x'(e), e) dF(e) > \int_E \Delta_s(x_s(e), e) dF(e)$, a contradiction of $x_s \in \arg\max_{x \in \mathcal{X}} \int_E \Delta_s(x(e), e) dF(e)$. 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 $\nu_1(P|m) = 1$.[54] Take $m = m_\nu \in M_\Lambda$. Suppose there is an off-path action a after e. For the sake of contradiction, suppose $\nu(P|m, e, a) = 1$ is not compatible with D1. Then there must exist some $s \in S_m$ for which (3) holds. Taking such an s, because $U^*_m(s, e) = u(s, e, x_s(e), R(m, e, x_s(e))$ (and similarly for P), (3) simplifies to $\Delta_s(a, e) > \Delta_s(x_s(e), e)$, a contradiction of the definition of $x_s(e)$. ■

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 $\psi^*(U^*; m)$ defines the unique messaging strategy that leaves P indifferent across messages $m_\nu \in M_\Lambda$ and $\{z_a(e; \varphi(\psi^*(U^*; m_\nu)), G_\nu)\}_{a \in A}$ gives the unique 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 mes-

[54] 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.

sages; 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 $m \in M_s^*$, $U^* = \int_E (w_P(x_s(e), e) + \rho R(m, e, x_s(e)))dF(e)$, so s's equilibrium utility is $\int_E w_s(x_s(e), e) + \rho R(m, e, x_s(e)))dF(e) = \int_E (w_s(x_s(e), e) - w_P(x_s(e), e))dF(e) + U^*$. 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 of Main Results

Lemma 3 follows immediately from the next lemma. Here we relate P's equilibrium utility to the expectation of $w_P(a, e)$ given the equilibrium distribution over actions and evidence from all DM types:

$$\mathbb{E}^{\mathcal{E}}[w_P(a, e)] = \int_E \int_\Theta \int_M \sum_{a \in A} w_P(a, e)\zeta^{\mathcal{E}}(a|\theta, m, e)d\sigma^{\mathcal{E}}(m|\theta)d\nu(\theta)dF(e),$$

so, consistent with the notation in Theorem 2, $\mathbb{E}^{\mathcal{E}}[w_P(a, e)|\theta = P]$ is P's expected material utility under equilibrium $\mathcal{E}$.

Lemma 7.
For every equilibrium $\mathcal{E}$, $\mathbb{E}^{\mathcal{E}}[w_P(a, e)] = U_P^{\mathcal{E}}(F) - \rho\mathbb{E}[r(\theta)]$.

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 $\theta \in \Theta_m$:

$$U_P^{\mathcal{E}}(F) = \int_E \left(\sum_{a \in A} (w_P(a, e) + \rho R(m, e, a))\zeta^{\mathcal{E}}(a|\theta, m, e) \right) dF(e).$$

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

$$U_P^{\mathcal{E}}(F) = \mathbb{E}^{\mathcal{E}}[w_P(a, e)|m] + \rho\mathbb{E}^{\mathcal{E}}[r(\theta)|m].$$

Taking the ex-ante expectation of both sides over messages in M_P^* (which is a probability one set under $\sigma^{\mathcal{E}}(\cdot|P)$ and $\Sigma_N^{\mathcal{E}}(\cdot)$ by Lemma 1) and again applying the law of iterated expectations then yields[55]

$$U_P^{\mathcal{E}}(F) = \mathbb{E}^{\mathcal{E}}[w_P(a,e)] + \rho\mathbb{E}[r(\theta)].$$

Rearranging terms then yields our desired result. *Q.E.D.*

Next, we give a lemma that characterizes the partisan's value under ex-ante signaling.

Lemma 8. *$U_P^{\alpha}(F)$ is the unique solution U to*

$$1 = \int_S \frac{\rho q r(s)}{U - \int_E w_P(x_s(e),e)dF(e)} dG(s).$$

Proof. 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. Thus, their reputation after m_s, 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 $\int_E w_P(x_s(e),e)dF(e)$ and so his equilibrium expected utility is given by

$$\begin{aligned} U_P^{\alpha}(F) &= \int_E w_P(x_s(e),e)dF(e) + \rho\frac{qr(s)}{q+(1-q)\psi(m_s)} \ \forall s \in S \\ \iff q + (1-q)\psi(m_s) &= \frac{\rho q r(s)}{U_P^{\alpha}(F) - \int_E w_P(x_s(e),e)dF(e)} \ \forall s \in S. \end{aligned}$$

Noting that $\sigma(m_s|s') = \mathbb{1}(s'=s)$ gives $\int_S \psi(m_s)dG(s) = \int_S \psi(m_s)d\Sigma_N(m_s) = \int_M d\sigma(m|P) = 1$. Taking the expectation over both sides of the last line of the display above with respect to s, we have

$$1 = \int_S \frac{\rho q r(s)}{U_P^{\alpha}(F) - \int_E w_P(x_s(e),e)dF(e)} dG(s). \tag{11}$$

Note that the right-hand side of (11) is strictly decreasing in $U_P^{\alpha}(F)$ and so it must be the unique solution to our desired equation. *Q.E.D.*

[55] We can drop dependence on $\mathcal{E}$ in the expectation for $r(\theta)$ as this is equal across all $\mathcal{E}$.

Proof of Lemma 4

Proof. 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 Lemma 8, we then have

$$\begin{aligned}
&\int_S \frac{\rho q r(s) dG(s)}{U_P^\alpha(F_\lambda) - \int_E w_P(x_s(e), e) dF_\lambda(e)} \qquad (12)\\
&= \lambda \int_S \frac{\rho q r(s) dG(s)}{U_P^\alpha(F_1) - \int_E w_P(x_s(e), e) dF_1(e)} + (1-\lambda) \int_S \frac{\rho q r(s) dG(s)}{U_P^\alpha(F_2) - \int_E w_P(x_s(e), e) dF_2(e)}\\
&\geq \int_S \frac{\rho q r(s) dG(s)}{\lambda U_P^\alpha(F_1) + (1-\lambda) U_P^\alpha(F_2) - \lambda \int_E w_P(x_s(e), e) dF_1(e) - (1-\lambda) \int_E w_P(x_s(e), e) dF_2(e)}\\
&= \int_S \frac{\rho q r(s) dG(s)}{\lambda U_P^\alpha(F_1) + (1-\lambda) U_P^\alpha(F_2) - \int_E w_P(x_s(e), e) dF_\lambda(e)}.
\end{aligned}$$

Thus, (12) 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 2

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

Proof. First, we show that $U_P^\alpha(\delta_e) = U_P^\beta(\delta_e)$ for all e. By Lemma 8, it suffices to show $\int_S \frac{\rho q r(s)}{U_P^\beta(\delta_e) - w_P(x_s(e), e)} dG(s) = 1$. Let z_a be the probability under ex-post signaling that P selects a when $F = \delta_e$. If $a \in \{a' : \int_S \mathbb{1}(a' = x_s(e)) dG(s) > 0\}$, then a is an optimal action for P, so $U_P^\beta(\delta_e) = w_P(a, e) + \frac{\rho q \int_S r(s) \mathbb{1}(a = x_s(e)) dG(s)}{q \int_S \mathbb{1}(a = x_s(e)) dG(s) + (1-q) z_a}$, which implies

$$q \int_S \mathbb{1}(a = x_s(e)) dG(s) + (1-q) z_a = \frac{\rho q \int_S \mathbb{1}(a = x_s(e)) r(s) dG(s)}{U_P^\beta(\delta_e) - w_P(a, e)}.$$

This same equation trivially holds for $a \notin \{a' : \int_S \mathbb{1}(a = x_s(e)) dG(s) > 0\}$ as both the left and right-hand side above are zero. Adding these equations together across all a, we have

$$1 = \sum_{a \in A} \frac{\rho q \int_S r(s) \mathbb{1}(a = x_s(e)) dG(s)}{U_P^\beta(\delta_e) - w_P(a, e)} = \int_S \frac{\rho q r(s)}{U_P^\beta(\delta_e) - w_P(x_s(e), e)} dG(s).$$

Thus, $U_P^\alpha(\delta_e) = U_P^\beta(\delta_e)$.

As discussed in the text, $U_P^\beta(F)$ is linear in F, so Lemma 4 implies

$$U_P^\alpha(F) \leq \int_E U_P^\alpha(\delta_e)dF(e) = \int_E U_P^\beta(\delta_e)dF(e) = U_P^\beta(F). \tag{13}$$

Note that the inequality in (12) is strict if and only if there exists a positive probability set $S' \subseteq S$ such that $U_P^\alpha(F_1) - \int_E w_P(x_s(e),e)dF_1(e) \neq U_P^\alpha(F_2) - \int_E w_P(x_s(e),e)dF_2(e)$ 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 (13), we are taking a convex combination over δ_e, so the inequality is strict if there exists a positive probability set $S' \subseteq S$ such that, for each $s \in S'$, $U_P^\alpha(\delta_e) - w_P(x_s(e),e)$ is not constant in e on a probability one set of e.

For the sake of contradiction, suppose no such S' existed when mild agreement holds and β has residual strategic uncertainty. Then there exists a probability one set S'' such that, for each $s \in S''$, such that $U_P^\alpha(\delta_e) - w_P(x_s(e),e)$ is constant in e on a probability one set E'' of e. This implies that $U_P^\alpha(\delta_e) - U_P^\alpha(\delta_{e'}) = w_P(x_s(e),e) - w_P(x_s(e'),e')$ for all $e, e' \in E''$ and all $s \in S''$.

Take any $s, s' \in S''$. By mild agreement, there exists $e \in E''$ such that $w_P(x_s(e),e) = w_P(x_{s'}(e),e)$. Therefore, for any other $e' \in E''$, because $U_P^\alpha(\delta_e) - U_P^\alpha(\delta_{e'}) = w_P(x_s(e),e) - w_P(x_s(e'),e') = w_P(x_{s'}(e),e) - w_P(x_{s'}(e'),e')$, we have $w_P(x_s(e'),e') = w_P(x_{s'}(e'),e')$. By Assumption 3, this equality implies $x_s(e') = x_{s'}(e')$ for a probability one set of e'. This implies only one action is chosen on-path in β with probability one, a contradiction of residual strategic uncertainty. Therefore, the inequality in (13) holds strictly under mild agreement when β has residual strategic uncertainty. ■

We now turn to the proof of Theorem 2; note that Theorem 1 follows immediately from this result.

Proof. Take any equilibrium $\mathcal{E}$ with strategies $\{\sigma(\cdot|\theta)\}_{\theta\in\Theta}$. Recall that $\Sigma_N(\cdot) = \int_S \sigma(\cdot|s)dG(s)$. 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 $\int_S \frac{\rho q_m r(s)}{U - \int_E w_P(x_s(e),e)dF(e)}dG_m(s) = 1$.[56] Given this definition of $U_P^{\alpha,m}(F)$, the arguments from the proof of Claim 4 to show $U_P^\alpha(\delta_e) = U_P^\beta(\delta_e)$ then also imply that $U_P^{\alpha,m}(\delta_e) = U_P^{\beta,m}(\delta_e)$.

[56] That a unique solution exists follows from the following arguments. Given the bounds on q_m in Lemma 5 and Assumption 2, it is straightforward to show that $\rho q_m \underline{r} > \overline{w}_P - \underline{w}_P$ which implies $\int_S \frac{\rho q_m r(s)}{U - \int_E w_P(x_s(e),e)dF(e)}dG_m(s) > 1$ when $U = \overline{w}_P$. Because $\int_S \frac{\rho q_m r(s)}{U - \int_E w_P(x_s(e),e)dF(e)}dG_m(s)$ is strictly decreasing in U on $[\overline{w}_P, \infty)$ with a limit of zero as $U \to \infty$, a unique solution to

By the arguments made in Lemma 4, $U_P^{\alpha,m}(\cdot)$ is convex and so,[57] for all $m \in M_P^*$,

$$U_P^{\alpha,m}(F) \leq \int_E U_P^{\alpha,m}(\delta_e)dF(e) = \int_E U_P^{\beta,m}(\delta_e)dF(e) = U_P^{\beta,m}(F) = U_P^{\mathcal{E}}(F). \tag{14}$$

For the sake of contradiction, suppose $U_P^{\alpha}(F) > U_P^{\mathcal{E}}(F)$. Then, by (14), $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$ 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 $\int_S \frac{\rho q_m r(s)}{U_P^{\alpha,m}(F) - \int_E w_P(x_s(e),e)dF(e)} dG_m(s) = 1$ to get

$$\begin{aligned} 1 &= \int_{M_P^*} \left[\int_S \frac{\rho q_m r(s) dG_m(s)}{U_P^{\alpha,m}(F) - \int_E w_P(x_s(e),e)dF(e)} \right] (qd\Sigma_N(m) + (1-q)d\sigma(m|P)) \\ &= \int_S \int_{M_P^*} \frac{\rho q r(s)}{U_P^{\alpha,m}(F) - \int_E w_P(x_s(e),e)dF(e)} d\sigma(m|s)dG(s), \\ &> \int_S \int_{M_P^*} \frac{\rho q r(s)}{U_P^{\alpha}(F) - \int_E w_P(x_s(e),e)dF(e)} d\sigma(m|s)dG(s) \\ &= \int_S \frac{\rho q r(s)}{U_P^{\alpha}(F) - \int_E w_P(x_s(e),e)dF(e)} dG(s). \end{aligned} \tag{15}$$

where the second equality follows from Bayes rule and the inequality from $U_P^{\alpha}(F) > U_P^{\alpha,m}(F)$. But the equations contradict (11). Therefore, $U_P^{\alpha}(F) \leq U_P^{\mathcal{E}}(F)$.

For the sake of contradiction, suppose there is mild agreement and residual strategic uncertainty in $\mathcal{E}$ and $U_P^{\alpha}(F) = U_P^{\mathcal{E}}(F)$. By Claim 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 (14), $U_P^{\alpha}(F) \geq U_P^{\alpha,m}(F)$ with strict inequality for a positive probability set of messages. The same arguments as above in (15) lead to a contradiction. This completes the proof of first part of the theorem.

To establish second part of the theorem, we first note that $U_P^{\alpha}(F) \leq U_P^{\mathcal{E}}(F)$ and Lemma 7 imply $\mathbb{E}^{\alpha}[w_P(a,e)] \leq \mathbb{E}^{\mathcal{E}}[w_P(a,e)]$ (all inequalities below will hold strictly under residual strategic uncertainty for $\mathcal{E}$ and mild agreement). Next, we observe that, by Lemma 1, $\mathbb{E}^{\alpha}[w_P(a,e)|\theta \in S] = \mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta \in S]$ as each s type follows the same

---

$\int_S \frac{\rho q_m r(s)}{U - \int_E w_P(x_s(e),e)dF(e)} dG_m(s) = 1$ exists.

[57] 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 $\int_S \frac{\rho q r(s) dG(s)}{U - \int_E w_P(x_s(e),e)dF(e)} = 1$, and so apply to $U_P^{\alpha,m}$ as well.

contingent plan x_s in both α and $\mathcal{E}$. We then have

$$\begin{aligned}\mathbb{E}^{\alpha}[w_P(a,e)] &= q\mathbb{E}^{\alpha}[w_P(a,e)|\theta\in S]+(1-q)\mathbb{E}^{\alpha}[w_P(a,e)|\theta=P]\\ &= q\mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta\in S]+(1-q)\mathbb{E}^{\alpha}[w_P(a,e)|\theta=P], \text{ and}\\ \mathbb{E}^{\mathcal{E}}[w_P(a,e)] &= q\mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta\in S]+(1-q)\mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta=P].\end{aligned}$$

That $\mathbb{E}^{\alpha}[w_P(a,e)]\le\mathbb{E}^{\mathcal{E}}[w_P(a,e)]$ then implies $\mathbb{E}^{\alpha}[w_P(a,e)|\theta=P]\le\mathbb{E}^{\mathcal{E}}[w_P(a,e)|\theta=P]$.

Finally, the third part of the theorem follows from the fact that, as noted above, $\mathbb{E}^{\alpha}[w_P(a,e)]\le\mathbb{E}^{\mathcal{E}}[w_P(a,e)]$, which when $w_P(a,e)=-c$, is equal to $\mathbb{E}^{\alpha}[-ca]\le\mathbb{E}^{\mathcal{E}}[-ca]$. Dividing both sides by c yields the result. Again, strict inequalities will hold under residual strategic uncertainty for $\mathcal{E}$ and mild agreement. *Q.E.D.*

Following this point, everything below uses the baseline model unless otherwise specified. We also note that Proposition 1, which we prove below, does not rely on Theorem 1 or Theorem 2, but instead only uses the equilibrium characterization of Lemma 1 and Lemma 2, and the characterization of $U_P^{\alpha}(F)$ in Lemma 8.

Proof of Proposition 1

Proof. Take any $e\in E$. For notational simplicity, we drop dependence of v^{α} on F where no confusion is caused. The proof is immediate if $G(e+c)=0$ or if $G(e+c)=1$ as Lemma 1 implies $v^{\alpha}(e)=v^{\beta}(e)=0$, or $v^{\alpha}(e)=v^{\beta}(e)=1$ respectively. Suppose $G(e+c)\in(0,1)$. Note that $v^{\alpha}(e)=\int_S\mathbb{1}(e\ge\tilde{e}_s)(qdG(s)+(1-q)d\sigma(m_s|P))$. 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 8, i.e., it is the relative rate at which P mimics the s type under ex-ante signaling. 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 8, $q+(1-q)\psi(m_s)=\frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)}$, so $v^{\alpha}(e)=\int_{-\infty}^{e+c}\frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)}dG(s)$.

Consider the ex-post signaling equilibrium where there is a single message m_0 sent by all DM types at the communication stage. Let $R^{\beta}(e,a)\equiv\nu_2^{\beta}(S|m_0,e,a)$ be the ex-post signaling reputation at the decision stage after evidence e and action a. P's indifference condition is $\rho R^{\beta}(e,1)-c=\rho R^{\beta}(e,0)$. Since the N type takes action $a=1$ after evidence e with probability $G(e+c)$ by Lemma 1, and $v^{\beta}(e)$ is the probability of $a=1$ given e, Bayes rule gives that $R^{\beta}(e,1)=\frac{qG(e+c)}{v^{\beta}(e)}$ and $R^{\beta}(e,0)=\frac{q(1-G(e+c))}{1-v^{\beta}(e)}$. Thus, putting these together gives that $v^{\beta}(e)$ is the unique solution to

$$\frac{\rho qG(e+c)}{v^{\beta}(e)}-c=\frac{\rho q(1-G(e+c))}{1-v^{\beta}(e)}. \tag{16}$$

This means $v^{\beta}(\cdot)$ does not depend on F and only depends on G through the value of $G(e+c)$.

Suppose F is atomless; we will deal with atoms later via a continuity argument. Next, we will construct another distribution over standards $\tilde{G}$ that delivers the same probability of $a=1$ at e under ex-post and ex-ante signaling as G. In addition, we ensure it delivers the same ex-ante signaling utility to P (taking expectation over e). Towards this end, let $s_1 < e+c$ be defined by the value of s such that $\frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)}G(e+c) = \int_{-\infty}^{e+c} \frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_{\tilde{s}})} dG(\tilde{s})$. To see that such an s exists, note that, as F is assumed to be atomless, $\frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)}$ is a continuous function of s. Existence of s_1 then follows by the intermediate value theorem. Similarly, define $s_2 > e+c$ such that $\int_{e+c}^{\infty} \frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)} dG(s) = \frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_{s_2})}(1-G(e+c))$. Let $\tilde{G}$ be a binary-support distribution that places mass $G(e+c)$ on s_1 and mass $1-G(e+c)$ on s_2.

We claim that $\tilde{G}$ generates an ex-ante signaling utility for P equal to $\tilde{U}_P^{\alpha}(F) = U_P^{\alpha}(F)$. This follows from the fact that $\tilde{U}_P^{\alpha}(F)$ is the unique solution U to

$$1 = \int_S \frac{\rho q}{U+c-cF(\tilde{e}_s)} d\tilde{G}(s) = \frac{\rho q}{U+c-cF(\tilde{e}_{s_1})}G(e+c) + \frac{\rho q}{U+c-cF(\tilde{e}_{s_2})}(1-G(e+c)),$$

which is solved by $U_P^{\alpha}(F)$ by construction of s_1, s_2 and the fact that $1 = \int_S \frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)} dG(s)$. Note that $\tilde{G}$ delivers the same probability of $a=1$ under ex-post signaling as $\tilde{G}(e+c) = G(e+c)$. Moreover, the ex-ante signaling probability of $a=1$ under under $\tilde{G}$ is $\int_{-\infty}^{e+c} \frac{\rho q}{\tilde{U}_P^{\alpha}(F)+c-cF(\tilde{e}_s)} d\tilde{G}(s) = \frac{\rho q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_{s_1})}G(e+c) = v^{\alpha}(e)$ by construction.

Given the above argument, proving that $v^{\alpha}(e) \geq v^{\beta}(e) \forall e \in E$ for any $\tilde{G}$ with binary support is sufficient to prove the proposition. This follows from the logic in Subsection 4.1, which we formalize here. Let $\mathrm{Supp}(\tilde{G}) = \{s_1, s_2\}$ where $s_1 \leq s_2$. By Lemma 1, since s_i uses contingent plan x_{s_i} and the P type mixes only over these at the decision stage, $v^{\alpha}(e) = v^{\beta}(e) = 1\ \forall e \geq s_2 - c$ and $v^{\alpha}(e) = v^{\beta}(e) = 0\ \forall e < s_1 - c$.

The only remaining case is $e \in [s_1 - c, s_2 - c)$. Let $R^{\alpha}(m_{s_i}) \equiv \nu_1^{\alpha}(S|m_i)$ be the ex-ante equilibrium reputations after message m_i at the communication stage. P's indifference condition is given by $\rho R^{\alpha}(m_{s_1}) - c\left(F(s_2-c) - F(s_1-c)\right) = \rho R^{\alpha}(m_{s_2})$. Since the DM uses x_{s_i} under ex-ante signaling after sending m_{s_i}, and $e \in [s_1-c, s_2-c)$, the probability of $a=1$ given e, i.e., $v^{\alpha}(e)$, is the probability the DM sends m_{s_1} at the communication stage. The non-partisan sends m_{s_1} if and only if they have standard s_1, i.e., with probability $\tilde{G}(s_1)$. Thus, using Bayes rule gives $R^{\alpha}(m_{s_1}) = \frac{q\tilde{G}(s_1)}{v^{\alpha}(e)}$ and $R^{\alpha}(m_{s_2}) = \frac{q(1-\tilde{G}(s_1))}{1-v^{\alpha}(e)}$. With some rearranging, substituting these into the indifference condition and reproducing the analogous condition in (16) under ex-post signaling for this parameterization gives

$$\frac{\rho q\tilde{G}(s_1)}{v^\alpha(e)} - \frac{\rho q(1-\tilde{G}(s_1))}{1-v^\alpha(e)} = c\left(F(s_2-c) - F(s_1-c)\right),$$
$$\frac{\rho q\tilde{G}(s_1)}{v^\beta(e)} - \frac{\rho q(1-\tilde{G}(s_1))}{1-v^\beta(e)} = c$$

respectively. But the left-hand side of the first and second line are the same decreasing function of $v^\alpha(e)$ and $v^\beta(e)$ respectively. Since the right-hand side of the first line is less than that of the second line, this means that $v^\alpha(e) \geq v^\beta(e)$.

Finally, suppose F has a mass point; since we assumed that one of F or G is atomless, this implies G is atomless. Let $\{F_n\}_{n=0}^\infty$ be a sequence of atomless distributions that converge in distribution to F. As v^β is independent of F, our desired ranking follows from the case of atomless F if $\lim_{n\to\infty} v^\alpha(e, F_n) = v^\alpha(e, F)$. Because G is atomless, for each U, $\lim_{n\to\infty} \frac{\rho q}{U+c-c\tilde{F}_n(\tilde{e}_s)} = \frac{\rho q}{U+c-c\tilde{F}(\tilde{e}_s)}$ for each s such that $F(\cdot)$ is continuous at $\tilde{e}_s$, i.e., for a probability one set of s. Thus, $\lim_{n\to\infty} \int_S \frac{\rho q}{U+c-cF_n(\tilde{e}_s)} dG(s) = \int_S \frac{\rho q}{U+c-cF(\tilde{e}_s)} dG(s)$ for each U. This implies

$$\begin{aligned}\int_S \frac{\rho q}{U_P^\alpha(F)+c-cF(\tilde{e}_s)} dG(s) = 1 &= \lim_{n\to\infty} \int_S \frac{\rho q}{U_P^\alpha(F_n)+c-cF_n(\tilde{e}_s)} dG(s) \\ &= \int_S \frac{\rho q}{\lim_{n\to\infty} U_P^\alpha(F_n)+c-cF(\tilde{e}_s)} dG(s),\end{aligned}$$

so $\lim_{n\to\infty} U_P^\alpha(F_n) = U_P^\alpha(F)$. This then implies that

$$\begin{aligned}\lim_{n\to\infty} v^\alpha(e, F_n) &= \lim_{n\to\infty} \int_{-\infty}^{e+c} \frac{\rho q}{U_P^\alpha(F_n)+c-cF_n(\tilde{e}_s)} dG(s) \\ &= \int_{-\infty}^{e+c} \frac{\rho q}{U_P^\alpha(F)+c-cF(\tilde{e}_s)} dG(s) \\ &= v^\alpha(e, F).\end{aligned}$$

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 $s \in$

$\mathrm{Supp}(\nu_1(\cdot|m))$ and $s \in \mathrm{Supp}(\nu_1'(\cdot|m'))$ 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

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

where R and R' are the equilibrium reputation functions in $\mathcal{E}$ and $\mathcal{E}'$ respectively. Thus, $U_s^{\mathcal{E}}(F) - U_s^{\mathcal{E}'}(F) = \rho\int_E(R(m,e,x_s(e)) - R'(m',e,x_s(e)))dF(e)$ 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 $s \in \mathrm{Supp}(\nu_1(\cdot|m))$ 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 baseline 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 (11) with respect to ρ, we have

$$-\frac{dU_P^{\alpha}(F)}{d\rho}\int_S \frac{\rho q}{(U_P^{\alpha}(F)+c-cF(\tilde{e}_s))^2}dG(s) + \int_S \frac{q}{U_P^{\alpha}(F)+c-cF(\tilde{e}_s)}dG(s) = 0. \quad (17)$$

By (11), $\int_S \frac{q}{U_P^\alpha(F)+c-cF(\tilde{e}_s)}dG(s) = \frac{1}{\rho}$. Substituting this into (17) and simplifying, we get

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

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 (11). 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

$$\begin{aligned}
1 = \int_S \frac{\rho q}{U_2+c-cF(\tilde{e}_s)}dG_2(s) &> \int_S \frac{\rho q}{U_1+c-cF(\tilde{e}_s)}dG_2(s) \\
&\geq \int_S \frac{\rho q}{U_1+c-cF(\tilde{e}_s)}dG_1(s).
\end{aligned} \tag{18}$$

These inequalities hold because the integrand is increasing in s and G_2 FOSD G_1. But (18) 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 $\frac{h(e)}{(U_P^\alpha(F_\lambda)+c(1-F_\lambda(e)))^2}$ 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 $1 = \int_E \frac{\rho q h(e)}{U_P^\alpha(F_\lambda)+c(1-F_\lambda(e))}de$. Taking the derivative of both sides with respect to λ and rearranging yields

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

The second equality follows from integration by parts, and the inequality follows from the fact that $\frac{h(e)}{(U_P^\alpha(F_\lambda)+c(1-F_\lambda(e)))^2}$ 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 $\varepsilon > 0$, and $\delta \in (0,\Delta)$. 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 higher 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 $d(x,x') = \int_E |x(e)-x'(e)|dF(e)$.[58] 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[59] with $\mathrm{Supp}(\nu(\cdot|x)) \subseteq \{\theta : x \in \mathrm{Supp}(\xi(\cdot|\theta))\}$ if $\{\theta : x \in \mathrm{Supp}(\xi(\cdot|\theta))\} \neq \emptyset$,
2. $\xi(\mathcal{X}_\theta^*|\theta) = 1$ where $\mathcal{X}_\theta^* \equiv \arg\max_{x\in\mathcal{X}} \int_E u(\theta, e, x(e), \nu(S))dF(e)$.

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 $\Theta' \subseteq \Theta$

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

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

such that, for each $\theta'' \notin \Theta'$, there exists $\theta' \in \Theta'$ such that

$$\{\nu \in \Delta(\Theta) : \int_E u\big(\theta'', e, x(e), \nu(S)\big) dF(e) > U_{\theta''}\}$$
$$\subsetneq \{\nu \in \Delta(\Theta) : \int_E u\big(\theta', e, x(e), \nu(S)\big) dF(e) > U_{\theta'}\}.$$

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

Proof of Proposition 7

Proof. Throughout, given an equilibrium (ξ, ν), we denote the reputation payoff for x as $R(x)$. 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 $\frac{d\xi(x_s|P)}{dG(s)} = \frac{q}{1-q}[\frac{\rho}{U_P^\alpha(F)+c-cF(\tilde{e}_s)} - 1]$.[60] Set equilibrium beliefs $\nu(s|x) = \frac{U_P^\alpha(F)+c-cF(\tilde{e}_s)}{\rho}$, $\nu(P|x) = 1 - \nu(s|x)$ 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 the fact that ex-ante signaling is an equilibrium, no type wants to deviate from their equilibrium strategy to another $x \in \cup_{s\in S} \mathcal{X}_s$ (which is equivalent under ex-ante signaling to deviating to another $m_{s'}$ and following it up with $x_{s'}$) and no type has an incentive to deviation to $x \notin \cup_{s\in S} \mathcal{X}_s$ as such a deviation leads to $R(x) = 0$ (which is equivalent in ex-ante signaling to send message $m \notin \{m_s\}_{s\in S}$ and then following action plan x). The fact that these are not profitable deviations under ex-ante signaling implies they are not profitable deviations here.

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

$$\{\nu \in \Delta(\Theta) : -c\int_E x(e)dF(e) + \rho\nu(S) \geq U_P^\alpha(F)\}$$
$$\subsetneq \{\nu \in \Delta(\Theta) : \int_E (e-s)x(e) + \rho\nu(S) \geq \int_E (e-s)x_s(e) + \rho R(x_s)\}.$$

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

The left-hand side above is non-empty.[61] 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 $\mathcal{X}_s$ (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 $R(x) = 1$. By Bayes' plausibility, there must exist $x' \in \mathcal{X}_P^*$ such that $R(x') \leq q$. For x' to be in $\mathcal{X}_P^*$, we must have

$$-c\int_E x'(e)dF(e) + \rho R(x') \geq -c\int_E x(e)dF(e) + \rho R(x). \tag{19}$$

Using our bounds on $R(x)$ and $R(x')$, this implies $-c+\rho \leq \rho q$, or $\rho \leq \frac{c}{1-q}$, a contradiction of Assumption 1. ■

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: cl$(\mathcal{X}_P^*) \cap \mathcal{X}_s \neq \emptyset$ and cl$(\mathcal{X}_P^*) \cap \mathcal{X}_s = \emptyset$ where cl$(\mathcal{X}_P^*)$ is the closure of $\mathcal{X}_P^*$.

In the first case, where 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':

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

[61] Take $x' \in \{x_s\}_{s\in S}$ such that $\nu(S|x') \leq q$ (such an x' exists by Bayes plausibility), which implies $U_P^\alpha(F) \leq \rho R(x') \leq \rho q$. If the set on the left-hand side was empty, then $\rho - c\int_E x(e)dF(e) < U_P^\alpha(F) \leq \rho q \overline{r}$ or $\rho < \frac{c\int_E x(e)dF(e)}{1-q} \leq \frac{c}{1-q}$, which contradicts Assumption 1.

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

$$\begin{aligned}\{\nu \in \Delta(\Theta) :\ -c\int_E x_s(e)dF(e) + \rho\nu(S) > -c\int_E x'(e)dF(e) + \rho R(x')\} \\ \subsetneq \{\nu \in \Delta(\Theta) :\ \int_E (e-s)x_s(e) + \rho\nu(S) > \int_E (e-s)x(e)dF(e) + \rho R(x)\}.\end{aligned} \tag{20}$$

By analogous arguments to those in Step 1, the left-hand side of (20) 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, (20) holds if

$$\begin{aligned}\{\nu \in \Delta(\Theta) :\ -c\int_E x_s(e)dF(e) + \rho\nu(S) > -c\int_E x(e)dF(e) + \rho R(x)\} \\ \subsetneq \{\nu \in \Delta(\Theta) :\ \int_E (e-s)x_s(e) + \rho\nu(S) > \int_E (e-s)x(e)dF(e) + \rho R(x)\}.\end{aligned}$$

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 $R(x') \geq q$, in which case P can achieve a utility of $-c\int_E x'(e)dF(e) + \rho R(x') \geq \rho q - c > 0 \geq -c\int_E x(e)dF(e)$. 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 $\zeta : \Theta \times M \times E \times [-\delta, \delta] \to \Delta(A)$ 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 $\mathrm{Supp}(\nu_1(\cdot|x)) \subseteq \{\theta : x \in \mathrm{Supp}(\sigma(\cdot|\theta))\}$ if $\{\theta : x \in \mathrm{Supp}(\sigma(\cdot|\theta))\} \neq \emptyset$,
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
$$A^*_{\theta,m,e,\varepsilon} \equiv \arg\max_{a \in A} \; u(\theta, e, a, \nu_2(S|m, e, a)) + \varepsilon a,$$
4. For each θ, $\sigma(\mathcal{Y}^*_\theta|\theta) = 1$ where
$$\mathcal{Y}^*_\theta \equiv \arg\max_{y \in M \cup \mathcal{X}} \int_E \Big[\mathbb{1}(y \in M) \big\{ \int_{-\delta}^{\delta} \big(\max_{a \in A} u(\theta, e, a, \nu_2(S|y, a, e)) + \varepsilon a \big) d\lambda(\varepsilon) \big\} \\ + \mathbb{1}(y \in \mathcal{X}) u(\theta, e, y(e), \nu_1(S|y)) \Big] dF(e),$$

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

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, $U^\alpha_P(F) \geq \rho q - c$. 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 $\rho q - c > 0$ by $\rho > \frac{c}{q}$ and $\rho q - c > -c + \delta$ by $\rho > \frac{\delta}{q}$. Now consider $s \in S$. By $U_P^\alpha(F) = -c(1 - F(\tilde{e}_s)) + \rho R(x_s)$, we have $\rho R(x_s) \geq U_P^\alpha(F) + c(1 - F(\tilde{e}_s)) \geq \rho q - cF(\tilde{e}_s) \geq \rho q - c$. 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 $\rho q - c \geq \max\{c, \delta\}$, or $\rho \geq \frac{\max\{\delta + c, 2c\}}{q}$. This follows by Assumption 1 if $c > \delta$ and that $\rho \geq 2 \max\{\frac{\delta}{q}, \frac{\delta}{1-q}\}$ 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 zero. Thus, P attains a maximum utility of $\max\{-c + \delta, 0\}$ from sending m. However, the P type can attain at least $\rho q - c - \delta$ by mimicking some s type whose expected equilibrium reputation is at least q (such s exist by Bayes plausibility). For mimicking m to be optimal for P, we must have $\rho q - c - \delta \geq \max\{-c + \delta, 0\}$. But this contradicts either $\rho > \frac{2\delta}{q}$ or $\rho > \frac{2c}{q}$.

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 equal to one 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. If $y \in M$, then the expected equilibrium reputation for P following y is at most q (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. 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 $R(m, e, a) = \nu_2((S, [-\delta, \delta])|m, e, a)$ and $R(x) = \nu_1(S|x)$. 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

$$\int_E \int_{-\delta}^{\delta} \big(\max\{e - s + \varepsilon + \rho R(m,e,1), \rho R(m,e,0)\} - (e - s + \varepsilon)x_s(e) - \rho R(x_s)\big) d\lambda(\varepsilon) dF(e), \tag{21}$$

and for P, it is given by

$$\int_E \int_{-\delta}^{\delta} \big(\max\{-c + \varepsilon + \rho R(m,e,1), \rho R(m,e,0)\} - (-c + \varepsilon)x_s(e) - \rho R(x_s)\big) d\lambda(\varepsilon) dF(e). \tag{22}$$

Notice that the integrand for s in (21) is weakly less than the integrand for P in (22) for every e, ε and so this comparison holds for the expressions themselves.

Suppose, for the sake of contradiction, that (21) is strictly less than (22). First, consider that the commitment $x_s \in \mathcal{Y}_P^*$. Since $m \in \mathcal{Y}_P^*$, (22) is zero and so by hypothesis (21) 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 $x_s \notin \mathrm{Supp}(\sigma(\cdot|s))$ for all $s \in S$. The strict inequality holds between (21) and (22) for every $R(x_s)$. If $R(x_s) = 1$, then because P's equilibrium payoff is less than $\rho q + \max\{\delta - c, 0\}$, (22) is less than $\rho q + \max\{\delta - c, 0\} - (\rho - c) < 0$. If $R(x_s) = 0$ then because P's equilibrium payoff is greater than $\rho q - c - \delta$, (22) is greater than $\rho q - c - \delta > 0$. 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 $\nu_1(P|x_s) = 0$, $R(x_s) = 1$ and P's utility from x_s is at least $\rho - c$. His equilibrium utility is again bounded above by $\rho q + \max\{\delta - c, 0\}$ which we have already shown is less than $\rho - c$. Thus, P prefer x_s, a contradiction.

Therefore, (21) is equal to (22). Because the integrand in (21) is weakly point wise lower than (22) 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 (21) is equal to (22), 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 $R(x) = 1$, so P's expected utility from x is at least $\rho - c$. 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 ρq, so his expected utility from y is at most $\rho q + \max\{-c+\delta, 0\}$, which is less than $\rho - c$ as argued above. Thus, $\nu(S|x) < 1$.

If $x \in \operatorname{Supp}(\sigma(\cdot|s))$ for some $s \in S$, then $x \in \operatorname{Supp}(\sigma(\cdot|P))$; otherwise, $\nu(S|x) = 1$. By analogous arguments as in Proposition 7, $\operatorname{Supp}(\sigma(\cdot|s)) \cap \mathcal{X} \subseteq \mathcal{X}_s$. 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 $-c(1 - F(\tilde{e}_s)) + \rho\nu_1(S|y)$. 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. Let $\mathbb{P}_{\hat{\mu}}(a = 1)$ be the equilibrium probability of $a = 1$ under an arbitrary $\hat{\mu}$ and let F_π be the CDF associated with a belief $\pi \in \Delta(E)$. Under any $\hat{\mu}$, conditional on a realization of π, the subsequent equilibrium is identical to the ex-ante signaling equilibrium in our baseline model when the distribution of evidence is F_π. Therefore, P's expected utility conditional on π is $U_P^\alpha(F_\pi)$ and, by Lemma 3, $\mathbb{P}_{\hat{\mu}}(a = 1|\pi) = \frac{1}{c}(\rho q - U_P^\alpha(F_\pi))$.

We then have

$$\mathbb{P}_{\hat{\mu}}(a = 1) = \int_{\Delta(E)} \mathbb{P}_{\hat{\mu}}(a = 1|\pi) d\hat{\mu}(\pi) = \frac{1}{c}\left[\rho q - \int_{\Delta(E)} U_P^\alpha(F_\pi) d\hat{\mu}(\pi)\right].$$

Therefore, we have $\mathbb{P}_\mu(a = 1) \geq \mathbb{P}_{\tilde{\mu}}(a = 1)$ if and only if

$$\int_{\Delta(E)} U_P^\alpha(F_{\tilde{\pi}}) d\tilde{\mu}(\tilde{\pi}) \geq \int_{\Delta(E)} U_P^\alpha(F_\pi) d\mu(\pi). \tag{23}$$

Let $\mathcal{B}(\Delta(E))$ be the Borel subsets of $\Delta(E)$. That $\tilde{\mu}$ being more informative than μ

is equivalent to $\tilde{\mu}$ being a mean-preserving spread of μ—namely, that there exists a Markov kernel $\kappa : \Delta(E) \times \mathcal{B}(\Delta(E)) \to [0,1]$ such that, $\tilde{\mu}(A) = \int_{\Delta(E)} \kappa(\pi, A)d\mu(\pi)$ for all $A \in \mathcal{B}(\mathbb{R})$ and $\pi = \int_{\Delta(E)} \tilde{\pi}\kappa(\pi, d\tilde{\pi})$ for μ-almost all π. This last point implies that $F_\pi = \int_{\Delta(E)} F_{\tilde{\pi}}\kappa(\pi, d\tilde{\pi})$. Our desired inequality in (23) then follows from the fact that

$$\begin{aligned}\int_{\Delta(E)} U_P^\alpha(F_{\tilde{\pi}})d\tilde{\mu}(\tilde{\pi}) &= \int_{\Delta(E)} \left[\int_{\Delta(E)} U_P^\alpha(F_{\tilde{\pi}})\kappa(\pi, d\tilde{\pi})\right] d\mu(\pi) \\ &\geq \int_{\Delta(E)} U_P^\alpha\big(\int_{\Delta(E)} F_{\tilde{\pi}}\kappa(\pi, d\tilde{\pi})\big)d\mu(\pi) \\ &= \int_{\Delta(E)} U_P^\alpha(F_\pi)d\mu(\pi),\end{aligned}$$

where the inequality follows by convexity of $U_P^\alpha(\cdot)$. *Q.E.D.*

Welfare Comparison

As described in the text, we impose a tie-breaking assumption on how P mixes given his observation of s. To formulate such a tie-breaking rule formally, we need to consider the expanded type space $\Theta = \{P, N\} \times S$ that also identifies for P what standard is realized. As in our main model, the communication stage strategy is $\sigma : \Theta \to \Delta(M)$ and the decision stage strategy is $\zeta : \Theta \times M \times E \to \Delta(A)$. Welfare in equilibrium $\mathcal{E}$ is given by

$$\int_\Theta \int_M \int_E (e - s_\theta)\zeta^{\mathcal{E}}(1|\theta, m, e)dF(e)d\sigma^{\mathcal{E}}(m|\theta)d\nu_0(\theta).$$

where s_θ is the standard associated with θ.

By the characterization of equilibria for Theorem 2, we know that (N, s) types will choose to follow the contingent plan x_s that maximizes the difference between $w_s(a, e) = (\lambda_1(e - s) - \lambda_2 c)a$ and $w_P(a, e) = -ca$, which amounts to taking $a = 1$ if and only if $\lambda_1(e - s) - \lambda_2 c \geq -c$. Thus, (N, s) uses the contingent plan $x_s(e) = \mathbb{1}(e \geq s - \frac{1-\lambda_2}{\lambda_1}c)$. Note that, if $\lambda_2 \geq 1$, then (N, s) uses a threshold that is weakly higher than the social optimal threshold of s. However, for any value of λ_2, this threshold converges to the optimal one as $\lambda_1 \to \infty$. Otherwise, (N, s) uses a threshold that is too low relative to the social optimal one.

For ex-ante signaling, we describe the tie-breaking by a mimicking function $\tau^\alpha : S \to S$ such that (P, s) chooses to send the message $m_{\tau^\alpha(s)}$ sent by $(N, \tau^\alpha(s))$. Letting $\sigma^\alpha(\{m_{s'}\}_{s' \leq s}|P)$ be the ex-ante signaling probability of P sending a message identified with an (N, s') type for $s' \leq s$, $\tau^\alpha(s)$ is identified by $G(s) = \sigma^\alpha(\{m_{s'}\}_{s' \leq \tau^\alpha(s)}|P)$. Similarly, we describe the tie-breaking under ex-post signaling by a mimicking function $\tau^\beta : S \to S$ so that type (P, s) mimics the strategy of type $(N, \tau^\beta(s))$. This function is pinned down by the fact that the probability that P takes $a = 1$ at e is $G(\tau^{\beta,-1}(e + \frac{1-\lambda_2}{\lambda_1}c))$, which must be equal to $\zeta^\beta(1|P, m_0, e)$.

Next, we turn to the proof of Proposition 10.

Proof. Let $\underline{s} = \min \mathrm{Supp}(G)$ and $\overline{s} = \max \mathrm{Supp}(G)$. We argue next that $\tau^\alpha(s) > s \in (\underline{s}, \overline{s})$. We know that $\tau^\alpha(\underline{s}) = \underline{s}$ and $\tau^\alpha(\overline{s}) = \overline{s}$.[62] Let $t^\alpha(s) = \frac{d\tau^\alpha(s)}{ds}$. The reputation from sending the message m_s (and then following contingent plan x_s) is given by

$$\frac{qg(s)}{qg(s) + (1-q)g(\tau^{\alpha,-1}(s))\frac{1}{t(\tau^{\alpha-1}(s))}}.$$

The reputation at $\underline{s}$ must be strictly greater than q, which implies $t(\underline{s}) > 1$. Similarly, the reputation at $\overline{s}$ must be strictly less than q, so $t(\overline{s}) < 1$. Thus, for s sufficiently close to $\overline{s}$ and $\underline{s}$, we have $\tau^\alpha(s) > s$. Suppose, for the sake of contradiction, that $\tau^\alpha(s) \leq s$ for some $s \in (\underline{s}, \overline{s})$. Then there exists $s' \leq s$ such that $\tau^\alpha(s') = s'$ and $t(s') \leq 1$, which implies that the reputation at s' is weakly less than q. By the fact that $\tau(\tilde{s}) > \tilde{s}$ for $\tilde{s}$ close enough to $\overline{s}$, there must exist $s'' > s$ such that $\tau(s'') = s''$ and $t(s'') > 1$. But this implies that the reputation at s'' is strictly greater than q, contradicting the fact that reputations must be decreasing in the threshold type announced (this follows immediately from P's indifference condition). Therefore, $\tau^\alpha(s) > s$ for all $s \in (\underline{s}, \overline{s})$.

Next we argue that $\tau^\alpha \leq \tau^\beta$. Note that the comparison in Proposition 1, i.e., $v^\alpha(e) \geq v^\beta(e)\ \forall e$, continues to hold for these modified N preferences. To see this, note that the N type's strategy is the same across equilibria and so we only need to establish the comparison for the P type. The P type's strategy in ex-ante and ex-post signaling only depends on the distribution over contingent plans used by the N type. The current model, where (N, s) uses the contingent plans $\mathbb{1}(e \geq s - \frac{1-\lambda_2}{\lambda_1}c)$, induces exactly the same distribution over N's contingent plans as in our baseline model (where (N, s) uses contingent plan $\mathbb{1}(e \geq s - c)$) when the distribution over s is given by $\tilde{G}$ defined by $\tilde{G}(e + c) = G(e + \frac{1-\lambda_2}{\lambda_1}c)$. Thus, the conclusion of Proposition 1 goes through.

Take an arbitrary e. The cutoff (N, s) type taking $a = 1$ is given by $s = e + \frac{1-\lambda_2}{\lambda_1}c$. The fact that the probability of $a = 1$ is higher for each e under ex-ante as compared to

[62] This follows from the fact that the support of P's strategy is equal to $\{m_s\}_{s \in S}$ and the monotonic selection we impose for P's mixing strategy.

ex-post signaling implies that the probability P mimics a type below the pivotal s type i.e., $s' < e + \frac{1-\lambda_2}{\lambda_1}c$, under ex-ante signaling is higher than under ex-post, namely that $\tau^{\alpha,-1}(e + \frac{1-\lambda_2}{\lambda_1}c) \geq \tau^{\beta,-1}(e + \frac{1-\lambda_2}{\lambda_1}c)$. Varying e, this implies $\tau^{\alpha} \leq \tau^{\beta}$.

Now we compare welfare under ex-ante and ex-post signaling. Because s types' decision-stage strategies are fixed across equilibria, the comparison depends only on whether P's strategy generates higher welfare. We can write the welfare when P follows strategy $\tau(s)$ as $\int_S \int_E (e-s)x_{\tau(s)}(e)dF(e)dG(s)$. Therefore, the difference in welfare when $\theta = P$ under ex-ante and ex-post signaling is

$$\begin{aligned}
&\int_S \int_E (e-s)x_{\tau^{\alpha}(s)}(e)dF(e)dG(s) - \int_S \int_E (e-s)x_{\tau^{\beta}(s)}(e)dF(e)dG(s) \\
&= \int_S \int_{\tau^{\alpha}(s)-\frac{1-\lambda_2}{\lambda_1}c}^{\tau^{\beta}(s)-\frac{1-\lambda_2}{\lambda_1}c} (e-s)dF(e)dG(s). \qquad (24)
\end{aligned}$$

That this is positive whenever $\lambda_2 \geq 1$ follows from the observation that, for all s, we have $e - s \geq 0$ for all $e \in [\tau^{\alpha}(s) - \frac{1-\lambda_2}{\lambda_1}c, \tau^{\beta}(s) - \frac{1-\lambda_2}{\lambda_1}c]$ as $\tau^{\alpha}(s) - \frac{1-\lambda_2}{\lambda_1}c - s \geq \tau^{\alpha}(s) - s \geq 0$.

Next, consider the case when $\lambda_1 \to \infty$; we will make dependence on λ_1 explicit in $\tau^{\alpha}, \tau^{\beta}$. In the limit, each s type chooses the contingent plan that takes $a = 1$ whenever $e \geq s$. This is equivalent to our baseline model, where each s type chooses the contingent plan taking $a = 1$ whenever $e \geq s - c$, with the distribution of s given by $\tilde{G}$ defined by $\tilde{G}(e + c) = G(e)$. Because F has full support, ex-post signaling (under $\tilde{G}$ in our baseline model) has residual strategic uncertainty and mild agreement holds, so $V^{\alpha}(F) > V^{\beta}(F)$. This then implies that P must take $a = 1$ strictly more often under ex-ante signaling, so the $\lim_{\lambda_1 \to \infty} \tau^{\alpha}(s; \lambda_1) < \lim_{\lambda_1 \to \infty} \tau^{\beta}(s; \lambda_1)$ for a positive probability set of s. For each s, the associated welfare difference satisfies $\lim_{\lambda_1 \to \infty} \int_{\tau^{\alpha}(s;\lambda_1)-\frac{1-\lambda_2}{\lambda_1}c}^{\tau^{\beta}(s;\lambda_1)-\frac{1-\lambda_2}{\lambda_1}c} (e-s)dF(e) \geq 0$, where the inequality follows from the fact that $\tau^{\alpha}(s; \lambda_1) \geq s$ for all λ_1, so the lower bound of the integral converges to something weakly above s. Moreover, this integral is strictly positive in the limit at s for which $\lim_{\lambda_1 \to \infty} \tau^{\beta}(s; \lambda_1) > \lim_{\lambda_1 \to \infty} \tau^{\alpha}(s; \lambda_1)$. Therefore, we can find a λ_1 large enough such that the last line in (24) is strictly positive.

Next, consider the case in which $\lambda_2 < 1$. Note that, if $\tau^{\beta}(s) - \frac{1-\lambda_2}{\lambda_1}c < s$ for all s, then $e - s < 0$ for all $e \in [\tau^{\alpha}(s) - \frac{1-\lambda_2}{\lambda_1}c, \tau^{\beta}(s) - \frac{1-\lambda_2}{\lambda_1}c]$. This implies that the last line in (24) is strictly negative, and ex-post signaling yields higher welfare than ex-ante signaling in this case. Thus, for the case of high ρ or low λ_1, it suffices to show $\tau^{\beta}(s) - \frac{1-\lambda_2}{\lambda_1}c < s$ for all s.

Consider the case of ρ sufficiently large. As $\rho \to \infty$, realized reputations under ex-post signaling for $a = 1$ and $a = 0$ when on-path must converge to q. If both are on-path, then P's indifference requires arbitrarily close reputations as ρ becomes large. If only one action is on path, there is no update, and the associated reputation is exactly q. Under ex-post signaling, the reputation following on-path action $a = 0$ and $a = 1$ after evidence e are respectively given by

$$\frac{qG(e + \frac{1-\lambda_2}{\lambda_1}c)}{qG(e + \frac{1-\lambda_2}{\lambda_1}c) + (1-q)G(\tau^{\beta,-1}(e + \frac{1-\lambda_2}{\lambda_1}c))}, \text{ and}$$
$$\frac{q(1 - G(e + \frac{1-\lambda_2}{\lambda_1}c))}{q(1 - G(e + \frac{1-\lambda_2}{\lambda_1}c)) + (1-q)(1 - G(\tau^{\beta,-1}(e + \frac{1-\lambda_2}{\lambda_1}c)))}.$$

Thus, the fact that both of these must converge to q whenever on-path implies that $\tau^\beta(s)$ converges to s as $\rho \to \infty$. We can then take ρ large enough that $\tau^\beta(s) - \frac{1-\lambda_2}{\lambda_1}c < s$ for all s. Finally, consider the case of λ_1 sufficiently small. Because $\tau^\beta(s) \leq \overline{s}$ for all s, we can take λ_1 sufficiently small such that $\tau^\beta(s) - \frac{1-\lambda_2}{\lambda_1}c < s$ for all s. *Q.E.D.*

F. Continuous Degree of Partisanship

As mentioned, throughout our analysis, we maintain the assumption that the reputation for the partisan is sufficiently bad. This ensures that the partisan must be present in any on-path realization of outcomes, a fact which lends significant tractability to the model. However, we believe our main forces do not rely on this assumption. To make this point, we study our binary-standards version of the baseline model from Subsection 4.1 with the change that the DM has a continuous degree of partisanship.

The only modification to our baseline model of Section 2 is to the type space and preferences of the DM. Our timing and game remains the same, and we maintain our focus on equilibria satisfying the D1 refinement laid out in Appendix A. The type is given by $\theta = (\lambda, s)$ where $s \in \{s_1, s_2\}$ (with $s_1 < s_2$) and $\lambda \in [0, 1]$. We assume λ and s are independent, with G_λ and G_s being the marginal CDFs, and g_s being the associated PMF for the standard distribution. We assume G_λ has full support on $[0, 1]$ and, for simplicity, that it admits a density. Given type (λ, s), evidence $e \in E$, action $a \in \{0, 1\}$, and belief over partisanship $\nu \in \Delta([0, 1])$, the DM's utility is now given by

$$u((\lambda, s), e, a, \nu) = (\lambda(e - s) - (1-\lambda)c)a + \rho\mathbb{E}[\lambda|\lambda \sim \nu].$$

Our baseline model (restricted to binary standards) is this generalized framework where

λ is supported on $\{0,1\}$. Thus, the current framework allows the DM to have intermediate degrees of partisanship. In our leading interpretation, as λ increases, the DM puts more weight on the socially-optimal decision—as determined by the private signal s and the evidence e—and less weight on his partisan bias. Accordingly, the reputation concerns now take the form of maximizing the public's expectation of λ. This model does not fit into the generalization in Subsection 6.1 because it violates Assumption 2: there is no "gap" in reputation value between the reputation commanded by the worst type and that of other types.

This setting is more complex than our baseline model in multiple ways. In particular there is no sense in which the communication strategy of s types—namely, types $(1,s_1)$ and $(1,s_2)$—pin down equilibrium outcomes, which makes equilibrium characterization more challenging. We instead focus only on comparing ex-ante and ex-post signaling equilibria, which we now define as equilibria with "effective commitment" and "no communication" respectively. More formally, we say that an equilibrium features ex-post signaling if the communication stage is uninformative, or, without loss, every DM type sends the same message. We say that an equilibrium features ex-ante signaling if, after each message sent, there is no residual strategic uncertainty, i.e., the distributions over actions following each evidence realization is degenerate. One additional complication, is that unlike in our baseline model, there are multiple equilibrium outcomes that feature ex-ante signaling and ex-post signaling respectively. Nonetheless, we obtain the following result.

Proposition 11. *There exists $\overline{\rho}$ such that, for $\rho \geq \overline{\rho}$, there exists an ex-ante signaling equilibrium that yields a (weakly) higher probability of $a=1$ for each realization of e than any ex-post equilibrium.*

The ex-ante signaling equilibrium relevant to the proposition can be described as follows. There are two on-path messages at the decision stage, m_1 and m_2, interpreted as claiming standard s_1 and s_2 respectively. Message m_i is followed up by contingent plan $x_{s_i}(e)=\mathbb{1}(e \geq s_i - c)$ with probability one. This means the DM effectively commits at the communication stage and this equilibrium does in fact feature ex-ante signaling. The communication stage strategy is described by a threshold $\lambda^\alpha \in (0,1)$ such that the DM of type (λ, s) sends m_1 if and only if $\lambda \geq \lambda^\alpha$ and $s = s_1$.[63] The threshold type (λ^α, s_1), is indifferent between both messages, where, just like in our baseline model, he receives

[63] One use of the high ρ assumption is to guarantee that such a threshold exists, i.e. that a full reputation for m_1 would incur a deviation relative to the prior reputation under m_2.

a lower material utility under m_1, but is compensated by its higher reputation.[64]

We make the comparison to ex-post signaling equilibria for each evidence realization. As in our main analysis, for $e < s_1 - c$, every ex-post signaling equilibrium has $a = 0$ being taken with probability one. The reason is that, if $a = 1$ were on-path, it would be most preferred by types with low λ for these evidence realizations which implies that $a = 1$ would have a lower reputation than $a = 0$. But this is at odds with the fact that $a = 1$ is costly for these low λ types. Similarly, there exists an ex-post signaling equilibrium such that $a = 1$ is always taken for $e \geq s_2 - c$.

Thus, as in Subsection 4.1, the comparison turns on pivotal evidence realizations $e \in [s_1 - c, s_2 - c)$. For each such realization under ex-post signaling, there is a threshold type $\lambda^\beta(e) \in (0, 1)$ such that type (λ, s) takes $a = 1$ if and only if $\lambda \geq \lambda^\beta(e)$ and $s = s_1$. We show that there exists an ex-ante signaling equilibrium as above with $\lambda^\alpha \leq \lambda^\beta(e)\ \forall e \in E$. The intuition is the same as for our main results. When the threshold type λ^α considers sending m_1 instead of m_2 under ex-ante signaling, he gets the reputation benefit always but only has to take the action for pivotal evidence realizations $e \in (s_1 - c, s_2 - c)$. In contrast, in order for such a threshold type $\lambda^\beta(e)$ to pool with the same higher set of λ under ex-post signaling after evidence e, he must take $a = 1$ with probability one. Thus, mimicking the lower standard is more costly under ex-post signaling, and so the associated threshold must shift upward to increase the compensating reputational gain. As a result, the DM takes $a = 1$ more frequently under ex-ante signaling than ex-post signaling. Figure 3 illustrates this comparison in a specific example.

Proof. Throughout the proof, we will construct ρ_1, ρ_2, ρ_3, and take $\overline{\rho} > \max\{\rho_1, \rho_2, \rho_3\}$.

The types $(0, s_1)$ and $(0, s_2)$ have the same material utility of $-ca$ for all $e \in E$ and so we refer to this pair as "a $\theta = c$ type". Let $\mathbb{E}[\lambda] \equiv \mu_\lambda$.

Claim 6. *The material utility difference between $a = 1$ and $a = 0$ given evidence e is strictly increasing (decreasing) in λ for $\theta = (\lambda, s_1)$ if $e > (<) s_1 - c$, and strictly increasing (decreasing) in λ for $\theta = (\lambda, s_2)$ if $e > (<) s_2 - c$.*

Proof. The material utility difference between $a = 1$ and $a = 0$, is given by $\lambda(e - s) - (1 - \lambda)c$. Taking the derivative in λ gives $e - s + c$. The claim follows. ■

[64] One difference is that the threshold (λ^α, s_1) has evidence dependent material utility, so the lower utility for the lower threshold is an equilibrium consequence of it having a higher reputation, rather than the indifferent P type intrinsically disliking lower threshold contingent plans as in our baseline model.

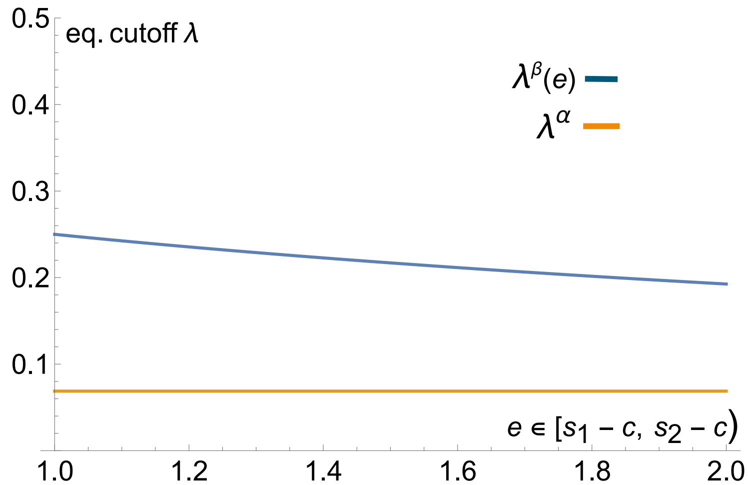

Figure 3: This plots the thresholds under ex-ante signaling and under ex-post signaling for different realizations of pivotal evidence in an example where $s_1 = 2$, $s_2 = 3$, $c = 1$, $\rho = 5$, F is uniform on $[0, 5]$, G_s is uniform on $\{s_1, s_2\}$, and G_λ is uniform on $[0, 1]$.

First, consider an arbitrary ex-post signaling equilibrium and let $R^\beta(e, a)$ be the reputation after e and a. By Claim 6, for $e < s_2 - c$, $\theta = (1, s_2)$ has the largest material utility difference between taking $a = 0$ and $a = 1$. If $a = 0$ is off-path, then D1 implies $R^\beta(e, 0) = 1$, whereas $a = 1$ obtains the prior reputation, i.e., $R^\beta(e, 1) = \mu_\lambda$. Thus, a positive probability set of types near c (i.e., low λ) would strictly prefer to $a = 0$ because it has strictly higher material and reputational payoff. Therefore, $a = 0$ must be on-path for $e < s_2 - c$.

Now define $\rho_1 \equiv \max\{\frac{c+s_2-s_1}{\mu_\lambda}, \frac{c+s_2-s_1}{1-\mu_\lambda}\}$. By Claim 6, for $e > s_1 - c$, $\theta = (1, s_1)$ has the largest material utility difference between $a = 1$ and $a = 0$. If $a = 1$ is off-path, then $a = 0$ gets the prior reputation, i.e., $R^\beta(e, 0) = \mu_\lambda$, and D1 implies that $R^\beta(e, 1) = 1$. But then, a positive probability set of types near c would strictly prefer $a = 1$ to $a = 0$ as $-c + \rho > \rho\mu_\lambda$ by $\rho > \rho_1$. Therefore, $a = 1$ must be on-path for $e > s_1 - c$.

We now argue that $a = 0$ must be chosen with probability one for $e < s_1 - c$. Suppose not, so that $a = 1$ is chosen with positive probability. By Claim 6, the material utility difference between $a = 1$ and $a = 0$ is decreasing in λ for both s_1 and s_2, and so the set of types choosing $a = 1$ must be given by, for some λ_1 and λ_2, $\{(\lambda, s_1) : \lambda \le \lambda_1\} \cup \{(\lambda, s_2) : \lambda \le \lambda_2\}$. Because both actions are chosen with positive probability, $\lambda_1 > 0$ and $\lambda_2 < 1$. This yields

$$R^\beta(e, 1) = \frac{\sum_{i=1}^2 g_s(s_i) \int_0^{\lambda_i} \tilde{\lambda} dG_\lambda(\tilde{\lambda})}{\sum_{i=1}^2 g_s(s_i) G_\lambda(\lambda_i)} < \mu_\lambda < \frac{\sum_{i=1}^2 g_s(s_i) \int_{\lambda_i}^1 \tilde{\lambda} dG_\lambda(\tilde{\lambda})}{\sum_{i=1}^2 g_s(s_i)(1 - G_\lambda(\lambda_i))} = R^\beta(e, 0). \tag{25}$$

But then a positive probability set of types near c would have a strict incentive to deviate to $a = 0$ as it yields a strictly higher material and reputational payoff. Therefore, we must have $a = 0$ chosen with probability one when $e < s_1 - c$.

Next, consider $e \in [s_1 - c, s_2 - c)$. As established above, both $a = 1$ and $a = 0$ must be taken with positive probability. By Claim 6, either there exists λ_2 such that $a = 0$ is taken if and only if $\theta \in \{(\lambda, s_2) : \ \lambda > \lambda_2\}$, or there exists λ_1 such that $a = 1$ is taken if and only if $\theta \in \{(\lambda, s_1) : \ \lambda > \lambda_1\}$. In the former case,

$$R^\beta(e, 0) = \frac{\int_{\lambda_2}^{1} \tilde{\lambda} dG_\lambda(\tilde{\lambda})}{1 - G_\lambda(\lambda_2} > \mu_\lambda > \frac{g_s(s_1) \int_0^1 \tilde{\lambda} dG_\lambda(\tilde{\lambda}) + g_s(s_2) \int_0^{\lambda_2} \tilde{\lambda} dG_\lambda(\tilde{\lambda})}{g_s(s_1) + g_s(s_2) G_\lambda(\lambda_2} = R^\beta(e, 1).$$

But this means that a positive probability set of types near c would choose $a = 0$ as it has both a strictly higher material payoff and reputational payoff. This contradicts the fact that type c and all types sufficiently close to c choose $a = 1$ under this candidate equilibrium. Thus, only the latter case, with threshold type (λ_1, s_1), can be part of an equilibrium.

Define $R_+(\tilde{\lambda}) \equiv \mathbb{E}[\lambda | \lambda \geq \tilde{\lambda}]$ and $R_-(\tilde{\lambda}) \equiv \mathbb{E}[\lambda | s = s_2, \text{ or } \lambda \leq \tilde{\lambda} \text{ and } s = s_1]$. Observe that by definition, $R_+(0) = R_-(0) = \mu_\lambda$, and $R_+(\tilde{\lambda}) > \mu_\lambda > R_-(\tilde{\lambda})\ \forall \tilde{\lambda} \in (0, 1]$. In addition, since the distribution G_λ is atomless, R_+ and R_- are continuous. The cutoff type in any ex-post signaling equilibrium must solve the following indifference equation across actions:

$$(1 - \lambda_1)c + \lambda_1(s_1 - e) = \rho(R_+(\lambda_1) - R_-(\lambda_1)). \tag{26}$$

There may be many solutions to this equation; because our goal is to show that an ex-ante signaling equilibrium has a higher probability of $a = 1$ than ex-post signaling for any evidence e, we focus on the equilibrium with the highest probability of $a = 1$, i.e., the lowest solution to (26). Moreover, this solution exists and has $1 > \lambda^\beta(e) > 0$. To see this, note that the LHS and RHS of (26) at $\lambda_1 = 0$ are, respectively, c and 0, in which case so the LHS is greater than the RHS, and at $\lambda_1 = 1$ are, respectively, $\lambda_1(s_1 - e) < c$ and $\rho(1 - \mu_\lambda)$, in which case the LHS is less than the RHS by $\rho > \rho_1$. Note also that $\lambda^\beta(e)$ is strictly decreasing in $e \in [s_1 - c, s_2 - c)$. This is because the LHS of (26) is strictly decreasing in e for any $\lambda_1 > 0$, but as mentioned above, for all e the LHS is greater than the RHS at $\lambda_1 = 0$.

Next, let us consider ex-ante signaling. We first specify the requisite bound on ρ so that our construction is valid. Define $\overline{F} \equiv F(s_2 - c) - F(s_1 - c)$. Let $\lambda_1(\rho)$ be the lowest solution $\tilde{\lambda} \in [0, 1]$ to

$$c\overline{F} = \rho(R_+(\tilde{\lambda}) - R_-(\tilde{\lambda})). \tag{27}$$

A solution to (27) exists and is strictly decreasing in ρ with $\lim_{\rho \to \infty} \lambda_1(\rho) = 0$. To see this, first note that the RHS of (27) at $\tilde{\lambda} = 0$ is 0, in which case the LHS is greater than the RHS, and at $\tilde{\lambda} = 1$ is $\rho(1 - \mu_\lambda) > c$ by $\rho > \rho_1$, in which case the LHS is less than the RHS. The limit and monotonicity of $\lambda_1(\rho)$ follow from the fact that the RHS strictly increases, and unboundedly so, in ρ for any $\tilde{\lambda} > 0$. Let $\lambda_2(\rho)$ be the highest solution $\tilde{\lambda} \in [0, 1]$ to

$$s_2 - s_1 + c = \rho(R_-(\tilde{\lambda}) - \tilde{\lambda}). \tag{28}$$

Such a solution exists and is strictly increasing in ρ. This is because the RHS of (28) at $\tilde{\lambda} = 0$ is $\rho\mu_\lambda > s_2 - s_1 + c$ by $\rho > \rho_1$ and at $\tilde{\lambda} = 1$ is $\rho(\mu_\lambda - 1) < 0$. The monotonicity follows from the RHS being strictly increasing in ρ and, because $\lambda_2(\rho)$ is the highest solution, the RHS must be strictly decreasing in $\tilde{\lambda}$ at $\lambda_2(\rho)$. Thus, let ρ_2 be the solution to $\lambda_1(\rho_2) = \lambda_2(\rho_2) \equiv \hat{\lambda}$. We take $\rho > \rho_2$, which implies that $c\overline{F} < \rho(R_+(\hat{\lambda}) - R_-(\hat{\lambda}))$ and, for all $\lambda \leq \hat{\lambda}$, we have $s_2 - s_1 + c < \rho(R_-(\tilde{\lambda}) - \tilde{\lambda})$.

We will construct an equilibrium that leads to a strictly higher rate of taking the action than under ex-post signaling for each realization of e. To do so, we specify that the DM sends one of two messages, m_1, m_2 at the communication stage. For types who send m_i, they follow contingent plan x_{s_i}. Define type λ^α as the lowest solution $\tilde{\lambda} \in [0, 1]$ to the following indifference condition:

$$\begin{aligned}
&\int_E (\lambda^\alpha(e - s_1) - (1 - \tilde{\lambda})c)x_{s_1}(e)dF(e) + \rho R_+(\tilde{\lambda}) \\
&= \int_E (\tilde{\lambda}(e - s_1) - (1 - \tilde{\lambda})c)x_{s_2}(e)dF(e) + \rho R_-(\tilde{\lambda}). \\
\iff &\int_{s_1 - c}^{s_2 - c} (\tilde{\lambda}(s_1 - e) + (1 - \tilde{\lambda})c)dF(e) = \rho(R_+(\tilde{\lambda}) - R_-(\tilde{\lambda})).
\end{aligned} \tag{29}$$

We observe that a such a solution exists and has $\lambda^\alpha < \hat{\lambda}$. To see this, note that at $\tilde{\lambda} = \hat{\lambda}$ the LHS and RHS of (29) evaluate to $\int_{s_1 - c}^{s_2 - c}(\hat{\lambda}(s_1 - e) + (1 - \hat{\lambda})c)dF(e) \leq c\overline{F}$ and $\rho(R_+(\hat{\lambda}) - R_-(\hat{\lambda}))$ respectively, so, as shown above, the LHS is less than the RHS because $\rho > \rho_2$. At $\tilde{\lambda} = 0$, the LHS and RHS of (29) evaluate to $c\overline{F}$ and 0 respectively, so that the LHS is greater than the RHS. Existence of a solution follows from continuity of both sides with respect to $\tilde{\lambda}$.

We specify strategies such that $\{(\lambda, s_1) : \lambda \geq \lambda^\alpha\}$ choose m_1 (and follow contingent plan x_{s_1}) and all other types choose m_2 (and follow up with contingent plan x_{s_2}). Any other message is assigned a reputation of zero. Any deviation from x_{s_1} after m_1 yields a reputation of λ^α; a deviation from x_{s_2} after m_2 yields a reputation of zero for $e \notin (s_1 - c, s_2 - c)$ and a reputation of λ^α for $e \in (s_1 - c, s_2 - c)$.

We first show that the that these reputations satisfy D1. First consider the decision stage following m_1. By Claim 6, for $e \geq s_1 - c$, the material utility difference between $a = 1$, which is on-path, and $a = 0$, which is off-path, is minimized at (λ^α, s_1) for $\theta \in \Theta_{m_1}$. Similarly, for $e < s_1 - c$ the material utility difference between $a = 1$, which is off-path, and $a = 0$, which is on-path, is maximized again at (λ^α, s_1) for $\theta \in \Theta_{m_1}$. Now consider the decision stage following m_2. For $e < s_1 - c$, the material utility difference between $a = 1$, which is off-path, and $a = 0$, which is on-path, is maximized again at the c type for $\theta \in \Theta_{m_2}$. If $e \in [s_1 - c, s_2 - c)$ this material utility is maximized at (λ^α, s_1) for $\theta \in \Theta_{m_2}$. Finally, if $e \geq s_2 - c$, then the material utility difference between $a = 1$, which is on-path, and $a = 0$, which is off-path, is minimized at the c type for $\theta \in \Theta_{m_2}$. This means that all off-path reputations are consistent with D1.

Next, we argue that this is an equilibrium. We begin by arguing that all types, on or off path, will follow contingent plan x_{s_i} after m_i. Pick an arbitrary e. Consider a type (λ, s) after sending m_i. If $e \geq s_i - c$, a deviation to $a = 0$ earns payoff of at most $\rho\lambda^\alpha$ while choosing $a = 1$ yields a payoff of at least $\lambda(e - s) - (1 - \lambda)c + \rho R_-(\lambda^\alpha)$. This deviation is most tempting when $i = 1$, $e = s_1 - c$, $s = s_2$, and $\lambda = 1$, in which case the payoff from $a = 1$ yields at least $s_1 - s_2 - c + \rho R_-(\lambda^\alpha)$. So, this deviation is unprofitable if

$$s_2 - s_1 + c \leq \rho(R_-(\lambda^\alpha) - \lambda^\alpha). \tag{30}$$

But this is guaranteed by the fact that $\lambda^\alpha < \hat{\lambda}$ and $\rho > \rho_2$. Now suppose $e \leq s_i - c$, a deviation to $a = 1$ earns payoff of at most $\lambda(e - s) - (1 - \lambda)c + \rho\lambda^\alpha$ while choosing $a = 0$ yields a payoff of at least $\rho R_-(\lambda^\alpha)$. This deviation is most tempting when $i = 2$, $e = s_2 - c$, $s = s_1$, and $\lambda = 1$, in which case the payoff from $a = 1$ yields at least $s_2 - s_1 - c + \rho\lambda^\alpha$. So, this deviation is unprofitable if

$$s_2 - s_1 - c \leq \rho(R_-(\lambda^\alpha) - \lambda^\alpha).$$

But this is implied by the previous inequality in (30).

Finally, we argue that no type has an incentive to deviate at the messaging stage.

No type has incentive to deviate to $m \notin \{m_1, m_2\}$ as they can always achieve a strictly higher utility by sending message m_1 and then following the strategy that maximizes their material utility (and which will achieve a positive reputation for some e). Thus, we only need to check they have no incentive to deviate to another message in $\{m_1, m_2\}$. For type (λ, s), the expected utility difference between m_1 and m_2 is

$$\begin{aligned}
&\int_E (\lambda(e-s) - (1-\lambda)c)x_{s_1}(e)dF(e) + \rho R_+(\lambda^\alpha) \\
&\qquad - \int_E (\lambda(e-s) - (1-\lambda)c)x_{s_2}(e)dF(e) - \rho R_-(\lambda^\alpha) \\
&= \int_{s_1-c}^{s_2-c} (e - s_1 + c)(\lambda - \lambda^\alpha)dF(e) - (s - s_1)\overline{F}\lambda.
\end{aligned}$$

If $s = s_1$, the the above expression is positive (i.e., (λ, s) will prefer m_1 to m_2) if and only if $\lambda \geq \lambda^\alpha$. If $s = s_2$, then the integral in the last line is bounded above by $(s_2 - s_1)(\lambda - \lambda^\alpha)\overline{F} < (s_2 - s_1)\lambda\overline{F}$, so (λ, s_2) will prefer m_2.

Finally, we compare the probability of $a = 1$ under ex-ante and ex-post signaling. As in our main text, denote the probability of $a = 1$ under an ex-ante and ex-post signaling equilibrium following evidence e as $v^\alpha(e)$ and $v^\beta(e)$ respectively. Given our analysis, the probability of $a = 1$ under our constructed ex-ante signaling equilibrium and the highest probability of $a = 1$ that is consistent with an ex-post signaling equilibrium is no higher than,[65] respectively

$$v^\alpha(e) = \begin{cases} 0 & \text{if } e < s_1 - c, \\ \big(1 - G_\lambda(\lambda^\alpha)\big)g_s(s_1) & \text{if } e \in [s_1 - c, s_2 - c), \\ 1 & \text{if } e > s_2 - c. \end{cases}$$

$$v^\beta(e) = \begin{cases} 0 & \text{if } e < s_1 - c, \\ \big(1 - G_\lambda(\lambda^\beta(e))\big)g_s(s_1) & \text{if } e \in [s_1 - c, s_2 - c), \\ 1 & \text{if } e > s_2 - c. \end{cases}$$

Thus, the result holds if $\lambda^\beta(e) > \lambda^\alpha$ for all $e \in [s_1 - c, s_2 - c)$. Since we established that $\lambda^\beta(e)$ is decreasing in e, it is sufficient to show that $\lim_{e \uparrow s_2 - c} \lambda^\beta(e) \equiv \lambda^\beta > \lambda^\alpha$.

We first introduce the relevant bound on ρ. Define $\overline{\lambda} \equiv \min\{(\frac{(1-\overline{F})c}{s_2 - s_1}, 1\}$, and $\rho_3 \equiv$

[65] It can be shown that an ex-post signaling equilibrium with $v^\beta(e)$ described in the display exists, but we omit this because existence is not necessary for our comparison.

$\frac{c-\overline{\lambda}(s_2-s_1)}{R_+(\overline{\lambda})-R_-(\overline{\lambda})}$. We take $\rho > \rho_3$. By taking the limit of the indifference condition for ex-post signaling as e approaches s_2-c from below, we get that λ^β is the lowest solution $\lambda \in [0,1]$ to

$$c - \lambda(s_2 - s_1) = \rho(R_+(\lambda) - R_-(\lambda)). \tag{31}$$

The definition of ρ_3 implies that the LHS of (31) is less than the RHS at $\lambda = \overline{\lambda}$ and $\rho > \rho_3$. Because, at $\lambda = 0$, the LHS is greater than the RHS, we have $\lambda^\beta < \overline{\lambda}$. We will next show that the LHS of (29) is less than the LHS of (31) for $\lambda < \overline{\lambda}$. That is,

$$\int_{s_1-c}^{s_2-c} (\tilde{\lambda}(s_1 - e) + (1 - \tilde{\lambda})c)dF(e) < c - \lambda(s_2 - s_1).$$

Maximizing the integral over F with mass $\overline{F}$ on $[s_1 - c, s_2 - c]$ yields an upper bound of $c\overline{F}$. Therefore, the above inequality holds if $c\overline{F} < c-\lambda(s_2-s_1)$, which is the case if $\lambda < \overline{\lambda}$. But this means that, at $\tilde{\lambda} = \lambda^\beta$, the LHS of (29) is less than the RHS since $\lambda^\beta < \overline{\lambda}$. Because λ^α was defined as the lowest solution to (29), both sides are continuous in $\tilde{\lambda}$, and the LHS and RHS are $c\overline{F}$ and 0 respectively at $\tilde{\lambda} = 0$, this means that $\lambda^\alpha < \lambda^\beta$. *Q.E.D.*

G. Comparing Other Equilibria

Theorem 1 highlights how the most informative ISS maximizes the probability of $a = 1$. While it is natural to ask if a similar comparative static holds with respect to informativeness of the ISS outside of the extreme case, it is difficult in general to compare equilibria as many of the nice properties held by ex-ante and ex-post signaling no longer hold for "intermediate" equilibria. Under our current bound on ρ (Assumption 1) in the baseline model, we can find "non-standard" examples where the Blackwell more informative ISS has a lower probability of $a = 1$.[66] Instead, to make such comparisons, we will focus on the case of large ρ which restores some degree of tractability to the analysis and, as we show below, is sufficient to generate a comparison across intermediate equilibria.

Proposition 12. *Let $\Lambda, \tilde{\Lambda}$ be ISSs such that $\tilde{\Lambda}$ is Blackwell more informative than Λ. Then there exists $\overline{\rho}$ such that the equilibrium that induces $\tilde{\Lambda}$ has a higher probability of $a = 1$ than that for Λ for all $\rho \geq \overline{\rho}$.*

[66] For example, such examples must have sufficiently "multi-modal" evidence distributions which are concentrated on multiple far away evidence realizations.

Proof. To make notation simpler, we will normalize the DM's payoffs by ρ, so the weight on reputation is equal to one and the material payoffs for P are $-\delta ca$. Taking $\rho \to \infty$ is equivalent to taking $\delta \to 0$. The proposition will follow if we can show that the ranking of the probability of $a = 1$ holds for all sufficiently low δ.

We begin by considering the equilibrium corresponding to Λ. Let $U(\delta)$ be P's expected utility in the equilibrium that induces Λ at δ. Let G_ν be the CDF of s for belief $\nu \in \Delta(S)$. Without loss, assume let the set of messages chosen by N be $\mathrm{Supp}(\Lambda)$, with messages m_ν identified by the $\nu \in \Delta(S)$ they induce. Let $q_\nu(\delta)$ be interim probability of $\theta \in S$ given m_ν. Let $z_\nu(\delta)$ be the Radon-Nikodym derivative of P's strategy relative to the set of s types sending m_ν, so that $q_\nu(\delta) = \frac{q}{q+(1-q)z_\nu(\delta)}$.

Let $U_\nu(e;\delta)$ be P's equilibrium utility after a message inducing ν and evidence e with $U_\nu(\delta) \equiv \int_E U_\nu(e;\delta)dF(e)$. Note that because P is indifferent across all messages sent by N in any equilibrium, we have $U(\delta) = U_\nu(\delta)$ for every $\nu \in \mathrm{Supp}(\Lambda)$. As shown in the proof of Claim 4, this utility is pinned down by

$$1 = \int_S \frac{q_\nu(\delta)}{U_\nu(e;\delta) + \delta c x_s(e)} dG_\nu(s). \tag{32}$$

Next, we verify that all the appropriate objects are all sufficiently differentiable in δ.

Claim 7. *$z_\nu(\delta), q_\nu(\delta)$ and $U_\nu(e;\delta)$ are twice differentiable in δ.*

Proof. Let $\tilde{q}(z) = \frac{q}{q+(1-q)z}$. Under ex-post signaling, when evidence e is realized and the prior on $\theta \in S$ is q' and the distribution over s is G_ν, the realized utility for P is given by the solution U to $1 = \int_S \frac{q'}{U+c\delta x_s(e)} dG_\nu(s)$, which is given by $\frac{q'-c\delta+\sqrt{(q'+c\delta)^2-4q'c\delta G_\nu(e+c)}}{2}$. For an arbitrary $U \in [-c\delta + q, q]$, define $\tilde{z}_\nu(U;\delta)$ by the unique z that solves $U = \int_E \frac{\tilde{q}(z)-c\delta+\sqrt{(\tilde{q}(z)+c\delta)^2-4\tilde{q}(z)c\delta G_\nu(e+c)}}{2} dF(e)$; uniqueness follows by the monotonicity of the term in the integral with respect to z. By the implicit function theorem, $\tilde{z}_\nu(U;\delta)$ is twice differentiable in U and δ. Note that $U(\delta)$ is given by the unique U that solves $\int_{\Delta(S)} \tilde{z}_\nu(U;\delta)d\Lambda(\nu) = 1$. By the implicit function theorem, $U(\delta)$ is twice-differentiable in δ. This then implies that $z_\nu(\delta) = \tilde{z}_\nu(U(\delta);\delta)$ is twice differentiable in δ. This also implies that $q_\nu(\delta) = \frac{q}{q+(1-q)z_\nu(\delta)}$ is differentiable in δ as well and, by the characterization of $U_\nu(e;\delta)$ in (32) and an application of the implicit function theorem, that $U_\nu(e;\delta)$ is twice differentiable as well. ∎

Taking the derivative with respect to δ (letting $U'_\nu(e;\delta) = \frac{dU_\nu(e;\delta)}{d\delta}$), we get

$$0 = \int_S \left[\frac{q'_\nu(\delta)}{U_\nu(e;\delta) + \delta c x_s(e)} - \frac{q_\nu(\delta)(U'_\nu(e;\delta) + c x_s(e))}{(U_\nu(e;\delta) + \delta c x_s(e))^2} \right] dG_\nu(s). \tag{33}$$

It is easy to check to that $q_\nu(0) = q = U_\nu(e;0)$. That $q_\nu(0) = q$ implies $q + (1-q)z_\nu(0) = 1$. Evaluating (33) at $\delta = 0$ and taking the expectation over both sides with respect to e yields

$$U'_\nu(0) \equiv \int_E U'_\nu(e;0) dF(e) = -q(1-q)z'_\nu(0) - c\int_E G_\nu(e+c) dF(e).$$

$U'(0) = U'_\nu(0)$ is the same for all $\nu \in \mathrm{Supp}(\Lambda)$ by the fact P is indifferent across messages m_ν.

Note that, because the N and P's strategy is absolutely continuous, $\int_{\Delta(S)} z_\nu(\delta) d\Lambda(\nu) = 1$, so $\int_{\Delta(S)} z'_\nu(\delta) d\Lambda(\nu) = 0$. Let $\overline{w} \equiv -c\int_E G(e+c)dF(e)$. Let us take the expectation over m_ν for both sides above. Using the facts that $\int_{\Delta(S)} G_\nu(s) d\Lambda(\nu) = G(s)$ and $U'(0) = \int_{\Delta(S)} U'_\nu(0) d\Lambda(\nu)$ by $U'(0) = U'_\nu(0)$ for all $\nu \in \mathrm{Supp}(\Lambda)$, we have

$$U'(0) = -c\int_{\Delta(S)} \int_E G_\nu(e+c) dF(e) d\Lambda(\nu) = \overline{w}.$$

Note that this implies $q'_\nu(0) = \overline{w} + \int_E c G_\nu(e+c) dF(e)$.

Taking the derivative of (33) with respect to δ, we have

$$\begin{aligned} 0 = \int_S \Bigg[& \frac{q''_\nu(\delta)}{U_\nu(e;\delta) + \delta c x_s(e)} - \frac{2q'_\nu(\delta)(U'_\nu(e;\delta) + c x_s(e))}{(U_\nu(e;\delta) + \delta c x_s(e))^2} \\ & - \frac{q_\nu(\delta) U''_\nu(e;\delta)}{(U_\nu(e;\delta) + \delta c x_s(e))^2} + \frac{2q_\nu(\delta)(U'_\nu(e;\delta) + c x_s(e))^2}{(U_\nu(e;\delta) + \delta c x_s(e))^3} \Bigg] dG_\nu(s). \end{aligned}$$

Evaluating this at $\delta = 0$ and solving for $U''_\nu(e;0)$, we have

$$U''_\nu(e;0) = q''_\nu(0) + \frac{2}{q} \int_S \Big[- q'_\nu(0)(U'_\nu(e;0) + c x_s(e)) + (U'_\nu(e;0) + c x_s(e))^2 \Big] dG_\nu(s).$$

Using the fact that $U'_\nu(e;0) = q'_\nu(0) - cG_\nu(e+c)$, we have

$$\begin{aligned}
U''_\nu(e;0) &= q''_\nu(0) + \frac{2}{q}\int_S \Big[- q'_\nu(0)(q'_\nu(0) - cG_\nu(e+c) + cx_s(e)) \\
&\qquad + (q'_\nu(0) - cG_\nu(e+c) + cx_s(e))^2 \Big] dG_\nu(s) \\
&= q''_\nu(0) + \frac{2}{q}\int_S [q'_\nu(0)c(x_s(e) - G_\nu(e+c)) + c^2(x_s(e) - G_\nu(e+c))^2] dG_\nu(s) \\
&= q''_\nu(0) + \frac{2}{q}[q'_\nu(0)c(\int_S x_s(e)dG_\nu(s) - G_\nu(e+c)) + \int_S c^2(x_s(e) - G_\nu(e+c))^2 dG_\nu(s)] \\
&= q''_\nu(0) + \frac{2}{q}\int_S c^2(x_s(e) - G_\nu(e+c))^2 dG_\nu(s) \\
&= -q(1-q)z''_\nu(0) + 2q((1-q)z'_\nu(0))^2 + \frac{2}{q}\int_S c^2(x_s(e) - G_\nu(e+c))^2 dG_\nu(s) \\
&= -q(1-q)z''_\nu(0) + \frac{2}{q}(c\int_E G_\nu(e+c)dF(e) + \overline{w})^2 + \frac{2}{q}\int_S c^2(x_s(e) - G_\nu(e+c))^2 dG_\nu(s)
\end{aligned}$$

Taking the expectation over both sides with respect to e, we have

$$\begin{aligned}
U''_\nu(0) &= -q(1-q)z''_\nu(0) + \frac{2}{q}\Big[(c\int_E G_\nu(e'+c)dF(e') + \overline{w})^2 \\
&\qquad + \int_E\int_S c^2(x_s(e) - G_\nu(e+c))^2 dG_\nu(s)dF(e)\Big] \\
&= -q(1-q)z''_\nu(0) + \frac{2}{q}\Big[c^2\Big\{(\int_E G_\nu(e+c)dF(e))^2 - \int_E G_\nu(e+c)^2 dF(e)\Big\} \\
&\qquad + \overline{w}^2 + 2\overline{w}c(\int_E G_\nu(e+c)dF(e)) \\
&\qquad + \int_E\int_S c^2 x_s(e)^2 dG_\nu(s)dF(e)\Big].
\end{aligned}$$

Let $\mathbb{V}(\nu) \equiv c^2\big[\int_E G_\nu(e+c)^2 dF(e) - (\int_E G_\nu(e+c)dF(e))^2\big] = \frac{c^2}{2}\int_E\int_E (G_\nu(e+c) - G_\nu(e'+c))^2 dF(e)dF(e')$ be the variance of $cG_\nu(e+c)$. Taking expectation over ν with respect to Λ and using $\int_{\Delta(S)} z''_\nu(0)d\Lambda(\nu) = 0$ and $U''(0) = \int_{\Delta(S)} U''_\nu(0)d\Lambda(\nu)$, we have

$$U''(0) = \frac{2}{q}\Big[\int_{\Delta(S)}\Big\{ - \mathbb{V}(\nu)\Big\}d\Lambda(\nu) + \int_E\int_S c^2 x_s(e)^2 dG(s)dF(e) - \overline{w}^2\Big]$$

Let $\tilde{U}$ be P's utility under $\tilde{\Lambda}$. By Lemma 3, it suffices to show that $\tilde{U}(\delta) < U(\delta)$ for sufficiently small δ. Taking a Taylor approximation of $U, \tilde{U}$, we then have

$$U(\delta) - \tilde{U}(\delta) = \frac{1}{q}\bigg[\int_{\Delta(S)} \mathbb{V}(\nu)d\tilde{\Lambda}(\nu) - \int_{\Delta(S)} \mathbb{V}(\nu)d\Lambda(\nu)\bigg]\delta^2 + o(\delta^3).$$

Because $\tilde{\Lambda}$ is more informative than Λ, we can represent $\tilde{\Lambda}$ by drawing a belief ν according to Λ and then applying an mean-preserving spread of ν—namely, there exists a κ such that $\tilde{\Lambda}(\cdot) = \int_{\Delta(S)} \kappa(\cdot|\nu)d\Lambda(\nu)$. That the term in brackets is weakly positive, which follows by noting that $\mathbb{V}(\nu)$ is convex in the distribution ν and then applying Jensen's inequality. Because the variance function is strictly convex, the term in brackets is an equality if and only if, for a probability one set of ν under Λ, κ maps ν to distributions ν' (with probability one) such that $G_{\nu'}(e + c) = G_\nu(e + c)$ on a probability one set of e (according to F). But this says that, with probability one, $\tilde{\Lambda}$ only provides additional information on s that indistinguishable in terms of their chosen contingent plans x_s. It is clear that such information will not impact P's equilibrium expected utility, in which case $U(\delta) = \tilde{U}(\delta)$. In all other cases, we have that the term in brackets above is strictly positive, which implies that for sufficiently small δ, we have $U(\delta) > \tilde{U}(\delta)$. *Q.E.D.*