Strategically Controlling Worldviews[*]

Cuimin Ba
University of Pittsburgh

Danil Dmitriev
University of Georgia

Ziqi Hang
Texas Tech University

Freddie Papazyan
Texas Tech University

March 24, 2026

Click here to see the most recent version

Abstract

This paper studies persuasion when the sender can both control the information the receiver observes and the model through which it is interpreted. Even when the receiver begins with a correctly specified model and understands the sender's strategic incentives, the sender can manipulate him and often secure her preferred action with probability one. The key mechanism highlights strong complementarity between selective information and tailored narratives. Whenever this complementarity is operative, the sender strictly outperforms what she could achieve even with commitment power. We fully characterize the sender-optimal equilibrium for a broad class of information technologies. While softer information weakens credibility, it also expands the scope for manipulative interpretation, sometimes permitting full manipulation when harder information does not. The results offer a plausible explanation for the widespread success of disinformation.

[*] The authors would like to thank Aislinn Bohren, Kevin He, Ryota Iijima, Xiao Lin, Mira Frick, Jonathan Libgober, George Mailath, Stephen Morris, Franz Ostrizek, Sulagna Dasgupta, Andrew Ferdowsian, Henrique Castro-Pires, Rahul Deb, Theo Durandard, Tristan Gagnon-Bartsch, Anne-Katrin Roesler, Yunus Aybas, Stepan Aleksenko, Natacha Spitzer, Beixi Zhou, and Omer Tamuz for insightful discussions. We are also grateful for the helpful comments from audiences at the Society for the Advancement of Economic Theory (SAET) Conference, the Stony Brook International Conference on Game Theory, the Southwest Economic Theory (SWET) Conference, the Western Economic Association International (WEAI) Annual Conference, and the Texas Economic Theory Camp, and from seminar attendees at the University of Pennsylvania, the University of Pittsburgh, and Texas Tech University.

> Certainly anyone who can make you believe absurdities can make you commit injustices.
>
> (*Certainment qui est en droit de vous rendre absurde, est en droit de vous rendre injuste.*)
>
> Voltaire, *Questions sur les Miracles*, 1765

1 Introduction

We are living in a troubling chapter of the information age, where it has become impossible to know who or what to trust. With abundant data at their disposal, strategic actors manipulate beliefs and perceptions, convincing individuals to act against their own interests. Yet the widespread success of manipulation is a puzzle for standard economic theory: if individuals rationally evaluate information and update beliefs using Bayes' rule, unreliable sources should lose credibility and play little role in decision making (Akerlof and Shiller, 2015; Stiglitz and Kosenko, 2024a,b).

We show that systematic manipulation is possible even against a sophisticated listener who initially holds a correctly specified model of the world, updates using Bayes' rule, and fully understands the speaker's strategic incentives. The key lies in what speakers actually do: they control *both* what information is communicated and how it should be interpreted, i.e., narratives. Consider an opinion leader on social media posting flashy data alongside a specious interpretation, a finance influencer recommending a falling stock while framing it as an undervalued opportunity, or a politician releasing favorable employment numbers while attributing them to their policies. While the joint use of information and narratives is pervasive in practice, the economic literature has largely studied strategic communication of information (Sobel, 2020) and persuasion through narratives (Schwartzstein and Sunderam, 2021) separately.

We develop a unifying framework that combines strategic communication and narrative persuasion. Consider the classic setting where an informed sender seeks to convince a receiver to take a risky action over a safe one. The sender privately observes a state that determines the payoff the receiver obtains from the risky action, and communicates a message using her information technology. In addition, she can propose a narrative—a model describing how the messages should be interpreted—before communication begins. To stack the environment against the sender, we consider a sophisticated receiver who starts with a correctly specified model of the sender's incentives and information technology, but is not dogmatically certain that it is correct. After observing the message, he compares how well the two models explain its occurrence and switches to the proposed model if the *Bayes factor*—the likelihood ratio of the proposed model to the initial model given the observed message—exceeds a fixed threshold (Schwartzstein and Sunderam, 2021; Ba, 2026).

We characterize when the sender can *fully manipulate* the receiver and induce the risky action with probability one, when she can *partially manipulate* him and induce the risky action more often than through strategic communication alone, and when narrative persuasion provides no additional benefit. To study how the sender's persuasiveness depends on the properties of the information technology, our analysis covers a full spectrum of information technologies, ranging from cheap talk to verifiable disclosure, as well as intermediate cases with partial verifiability. We find that narrative persuasion reverses the usual intuition from pure communication models that hard information provides a form of ex post commitment and benefits the sender more than soft information.

We begin with the cheap talk benchmark in Section 3, where the sender faces no constraints on what messages she can send. In Theorem 1, we show that with cheap talk the sender either fully manipulates the receiver or fails to manipulate at all, with the outcome determined by a simple demand-budget comparison that depends on the prior about the state and the receiver's skepticism. Full manipulation is achieved through a confidence trick: the sender claims that she will communicate like a "Bayesian persuader"—sending low messages in bad states and high messages in good states—but actually reverses this mapping, sending the high message in bad states and an *extra-high* message in good states. If the standard high message is sufficiently more likely under the proposed model than under the sender's true strategy, the receiver switches models and takes the risky action. The extra-high message, although contradicting the narrative and revealing that the sender is behaving strategically, credibly conveys that the state is indeed favorable, so the receiver again takes the risky action. Hence, full manipulation is feasible if and only if the *demand* for narrative persuasion—the probability mass of bad states—can be covered by the *budget*—the probability mass of good states—at a *price* equal to the receiver's switching threshold.

Because this condition depends only on aggregate probability mass rather than the full distribution of states, stark welfare implications follow. Full manipulation remains feasible even when the risky action offers a vanishing upside and arbitrarily severe downside compared to the safe action, resulting in extreme welfare loss for the receiver (Corollary 1). Moreover, a more favorable state distribution can make the receiver strictly worse off, since it enables manipulation that was previously infeasible (Corollary 2).

In Section 4, we generalize the analysis to *monotone interval information technologies*, a broad class that nests cheap talk, verifiable disclosure, and natural intermediate cases with partial verifiability such as noisy and coarse evidence. We characterize the constrained Bayesian persuasion solution (Proposition 1) and sender-optimal equilibrium (Proposition 2) for this full class of technologies, results which, to the best of our knowledge, are new to the literature. With narratives, the demand-budget logic extends: the sender fully manipulates whenever her narrative budget covers the residual demand after messaging, with demand now shaped by the technology (Theorem 2). When full manipulation is infeasible, message constraints introduce *partial manipulation*, a possibility entirely absent under cheap talk (Theorem 3).

One might expect harder information to limit manipulation by making it more difficult to commit fraud, but this is only partly true. Paradoxically, harder information can enable partial manipulation precisely because it restores the credibility that cheap talk lacks. On the other hand, softer information facilitates full manipulation because its flexibility reduces the demand for narrative persuasion, making it easier to cover with a fixed budget. The technology that best resolves the tension between flexibility and credibility is one that allows the sender to underreport but not overreport the state: it minimizes the demand for narratives while ensuring bad types can never mimic good ones, making it maximally persuasive in all environments (Proposition 3 and Theorem 4).

Section 5 explores several extensions to confirm the robustness of our findings. We first isolate the value of endogenous messaging by fixing the sender's messaging strategy exogenously. This reveals a fundamental tradeoff: optimizing messaging reduces demand for narrative persuasion and makes full manipulation easier to achieve, but committing to a fixed strategy allows for partial manipulation when full manipulation is infeasible (Theorem 5). Second, allowing the sender to propose multiple narratives scales up the budget linearly and further facilitates manipulation (Theorem 6). Third, an initially misspecified receiver is generally easier to manipulate, as misspecification creates additional opportunities for exploitation (Theorem 7). Finally, replacing discrete model switching with Bayesian model averaging preserves the possibility of full manipulation, although the characterization becomes less sharp (Theorem 8).

1.1 Related Literature

Our framework nests several existing models of communication and persuasion as special cases. The strategic communication literature, initiated by Crawford and Sobel (1982) and surveyed recently by Sobel (2020), studies how much a sender can influence actions through strategic information transmission. Under cheap talk, communication is uninformative when interests are sufficiently misaligned; under partially verifiable information, message constraints can restore some informativeness (Hart et al., 2017; Rappoport, 2025), though credibility remains a binding constraint. The Bayesian persuasion literature (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) bypasses these constraints through commitment and characterizes the maximal payoff the sender can achieve in a standard Bayesian framework. We show that narrative persuasion provides an alternative route to circumvent credibility limitations while surpassing the commitment benchmark. Relatedly, our paper is connected to the literature on persuasion with non-standard receivers (e.g., De Clippel and Zhang, 2022), with the key distinction that our receiver begins with a correctly specified model and yet can be fully manipulated.

This paper contributes to the growing literature on model misspecification and narrative persuasion. Ba (2026) studies whether misspecified models persist in the long term when a decision maker switches models using the same Bayes factor rule we adopt. While she analyzes a single-agent problem with exogenous models, we study a strategic setting in which the sender endogenously proposes the alternative model.[1] Several papers study narrative persuasion with endogenous models while holding the information source fixed. For example, Schwartzstein and Sunderam (2021) considers a sender who proposes a model of interpretation after observing the signal realization, allowing narratives to be tailored to realized evidence. In contrast, our framework requires the sender to propose a model ex ante rather than an interpretation ex post. Aina (2025) share this timing and study how proposing multiple competing narratives can be strategically used to make all signal realizations favorable to the sender, but our framework differs in allowing the sender to jointly control both the narrative and the messaging strategy, expanding the scope for manipulation. Bauch and Foerster (2024) study cheap-talk communication of narratives under model uncertainty, where narratives signal the senders private information about the data-generating process.

A few recent papers consider settings where a sender chooses both information and narratives, though in ways distinct from our framework. Eliaz and Spiegler (2024) study a news media setting in which outlets choose both which information to report and which narrative to present. They model narratives as causal graphs describing relationships among actions, signals, and outcomes; in our model, narratives pertain to how messages are generated, leaving the underlying decision problem unchanged. Ichihashi and Meng (2021) consider a sender who commits to a Blackwell experiment and, after the signal is realized, proposes an interpretation as in Schwartzstein and Sunderam (2021). In contrast, our sender lacks commitment power and proposes a fixed alternative model. Jain (2023) introduce a separate narrator interprets information supplied by an information provider, separating the roles of data provision and interpretation, whereas in our framework a single sender jointly controls both. Galperti (2019) studies Bayesian persuasion with exogenously given worldviews; Ko (2025) studies a related setting in which the receiver's default model treats the experiment as uninformative. Beyond these differences in model structure, our paper is the first to study how the effectiveness of narratives differs with the information technology.

[1] Relatedly, Schwartzstein and Sunderam (2024) study the exchange of models in social interactions where different social groups are endowed with exogenously given models.

The remainder of the paper is organized as follows. Section 2 introduces the framework. Section 3 analyzes the cheap talk benchmark and Section 4 extends the analysis to general information technologies. Section 5 explores extensions and Section 6 concludes. All proofs can be found in Appendix A.

2 Framework

2.1 Model

Preliminaries The players are a sender (S, she) and receiver (R, he). The receiver chooses an action $a \in \mathcal{A} = \{0, 1\}$. The safe action ($a = 0$) yields a known payoff $\omega_0 \in \mathbb{R}$, and the risky action ($a = 1$) yields an uncertain payoff equal to the state $\omega \in \Omega \subset \mathbb{R}$. The state ω is drawn by nature from a common prior $F \in \Delta\Omega$ and privately observed by the sender, which is interpreted as her type. We assume that Ω is either finite or a possibly unbounded closed interval and F admits a continuous, strictly positive density f. While the receiver's payoff depends on the chosen action and the realized state, the sender only cares about whether the risky action is taken. Specifically, $U_R(a, \omega) := (1 - a)\omega_0 + a\omega$ and $U_S(a) := a$. We focus on the non-trivial case where $\mathbb{E}[\omega] < \omega_0$ so that the receiver strictly prefers the safe action absent additional information.

Information technology The sender communicates her private information by sending a message m from a set $M(\omega) \subseteq M$, where M is the message space and $M(\omega)$ is the feasible message set in state ω. We assume M has cardinality $|M| \geq |\Omega| + 2$ so that M is rich enough to communicate any state or action recommendation. A *signal structure* is a mapping $\sigma : \Omega \to \Delta M$ from states to distributions over messages sent by the sender. Let $\bar{\Sigma}$ denote the set of all signal structures. Feasibility requires that the sender can only send messages available in the realized state, so the space of *feasible signal structures* is

$$\Sigma := \{\sigma : \text{supp}(\sigma(\cdot|\omega)) \subseteq M(\omega), \ \forall \omega \in \Omega\}. \tag{1}$$

The sender's *information technology* $\mathcal{M} := \{M(\omega)\}_{\omega \in \Omega}$ summarizes the constraints on what she can say as a function of the state, capturing the extent to which her information is hard or soft. The sender cannot commit to a signal structure in advance.

Remark 1. This setup accommodates a broad spectrum of information technologies, spanning from perfectly unverifiable information ("cheap talk"), where $M(\omega) = M$ for all ω, and perfectly verifiable information, where $M(\omega) = \{\omega, \varnothing\}$ for all ω and $\varnothing$ denotes a null message. Other choices of $M(\omega)$ naturally capture intermediate cases with partial verifiability.

Model uncertainty The receiver initially holds a correctly specified model of the decision environment under which he perfectly understands the sender's strategic incentives and information technology constraints.[2] If the receiver maintains this model, he correctly anticipates the sender's communication strategy σ in equilibrium (formally stated in Definition 1). However, the receiver is not dogmatically certain that this model is correct. At the beginning of the game, the sender can propose an alternative model $\hat{\sigma} \in \bar{\Sigma}$ describing how she will communicate.[3] This proposed model need not coincide with the true σ she ultimately uses and may even violate the feasibility constraints imposed by $\mathcal{M}$, since the proposal itself is unrestricted cheap talk. The receiver considers both models possible but does not hold a well-defined Bayesian prior over them. Instead, he chooses which model $\sigma_R \in \{\sigma, \hat{\sigma}\}$ to adopt using a model-switching rule based on *Bayes factors* (Ba, 2026; Schwartzstein and Sunderam, 2021).

Formally, let P_σ and $P_{\hat{\sigma}}$ denote the unconditional distributions of messages induced by σ and $\hat{\sigma}$. For any measurable set of messages $\tilde{M} \subset M$,

$$P_\sigma(\tilde{M}) := \int_\Omega \sigma(\tilde{M}|\omega) dF(\omega) \;\text{ and }\; P_{\hat{\sigma}}(\tilde{M}) := \int_\Omega \hat{\sigma}(\tilde{M}|\omega) dF(\omega). \tag{2}$$

Upon observing m, the receiver computes the Bayes factor $\lambda : M \to \mathbb{R}_+ \cup \{\infty\}$,

$$\lambda(m) \equiv \lambda(m|\sigma, \hat{\sigma}) := \frac{dP_{\hat{\sigma}}(m)}{dP_\sigma(m)}, \tag{3}$$

whenever the Radon-Nikodym derivative exists.[4] Intuitively, $\lambda(m)$ compares how well the two models explain the observation of message m; it is infinite when the message can arise under $\hat{\sigma}$ but not under σ. He switches to the proposed model $\hat{\sigma}$ if and only if $\lambda(m) \geq \alpha$ for some fixed *switching threshold* $\alpha > 0$. Formally,

$$\sigma_R = \sigma_R(m|\sigma, \hat{\sigma}) := \begin{cases} \hat{\sigma} & \text{if } \lambda(m|\sigma,\hat{\sigma}) \geq \alpha \\ \sigma & \text{if } \lambda(m|\sigma,\hat{\sigma}) < \alpha \end{cases}. \tag{4}$$

In other words, to convince the receiver to switch models, the message must be at least α times as likely to be sent under $\hat{\sigma}$ than under the initial model σ.

Remark 2. The threshold α measures the receiver's resistance to adopting the proposed model. If $\alpha < 1$, the receiver is a priori biased towards the proposed model and switches even when the observed message is less likely under $\hat{\sigma}$. Larger values make him more skeptical and reluctant to switch. If $\alpha = \infty$, the receiver becomes dogmatically certain

[2] We use the terms *model* and *narrative* interchangeably.

[3] We explore what happens if the sender can propose multiple alternative models in Section 5.2.

[4] We allow σ_R to be chosen arbitrarily if $\lambda(m)$ is not well-defined.

that his model is correct and never switches, and the game reduces to a standard strategic communication game in which the receiver correctly anticipates the sender's behavior.

Belief updating Given the adopted model σ_R and message m, the receiver forms a posterior belief $\pi_{\sigma_R}(\cdot|m) \in \Delta\Omega$ using Bayes' rule. For any measurable $\tilde{\Omega} \subset \Omega$,

$$\pi_{\sigma_R}(\tilde{\Omega}|m) = \frac{\int_{\tilde{\Omega}} \sigma_R(dm|\tilde{\omega})dF(\tilde{\omega})}{\int_{\Omega} \sigma_R(dm|\omega)dF(\omega)}, \tag{5}$$

whenever the denominator is positive. The receiver then chooses an optimal action $A(\pi_{\sigma_R}) \in \Delta\mathcal{A}$.

Timing The game proceeds as follows.

1. The sender proposes a model $\hat{\sigma}$ of how she will communicate to the receiver.[5]
2. Nature draws the state $\omega \sim F$, which is privately observed by the sender.
3. The sender sends a message m according to messaging strategy $\sigma(\cdot|\omega, \hat{\sigma})$.
4. Upon receiving m, the receiver decides which model to adopt, $\sigma_R \in \{\sigma, \hat{\sigma}\}$.
5. The receiver forms a posterior belief $\pi_{\sigma_R}(\cdot|m)$.
6. The receiver chooses action $a \in \mathcal{A}$, and payoffs $U_R(a, \omega)$ and $U_S(a)$ are realized.

Strategies and equilibrium The sender's persuasion strategy has two components: a *narrative strategy* $\hat{\sigma} \in \bar{\Sigma}$ describing the model she proposes and a *messaging strategy* $\sigma : \bar{\Sigma} \to \Sigma$ specifying how she actually communicates for any realized state and proposed model. Denote by $\sigma(\hat{\sigma}) \in \Sigma$ as the signal structure chosen following proposal $\hat{\sigma}$. With slight abuse of notation, we often write σ in place of $\sigma(\hat{\sigma})$, referring to the on-path signal structure induced by the equilibrium narrative $\hat{\sigma}^*$. Our equilibrium concept adapts weak Perfect Bayesian equilibrium to accommodate model-switching.

Definition 1. An *equilibrium* is a tuple $(\hat{\sigma}^*, \sigma^*, A^*, \pi^*, \sigma_R^*)$ that satisfies the following:

(i) Optimality of the narrative strategy:

$$\hat{\sigma}^* \in \arg\max_{\hat{\sigma} \in \bar{\Sigma}} \mathbb{E}_{\omega \sim F, m \sim \sigma^*(\cdot|\omega, \hat{\sigma})} \left[U_S \left(A^* \left(\pi^*_{\sigma^*_R(m|\sigma^*(\hat{\sigma}), \hat{\sigma})}(\cdot|m) \right) \right) \right]. \tag{6}$$

(ii) Optimality of the messaging strategy: for every $\omega \in \Omega$ and $\hat{\sigma}$,

$$\sigma^*(\cdot|\omega, \hat{\sigma}) \in \arg\max_{\sigma \in \Sigma} \mathbb{E}_{m \sim \sigma(\cdot|\omega, \hat{\sigma})} \left[U_S\Big(A^*\Big(\pi^*_{\sigma^*_R(m|\sigma^*(\hat{\sigma}), \hat{\sigma})}(\cdot|m) \Big) \Big) \right]. \tag{7}$$

[5] We assume the narrative is proposed before the state is realized to shut down any signaling effect.

(iii) $\sigma_R^*(m|\sigma,\hat{\sigma})$ is chosen by the model-switching rule (Eq. (4)) whenever possible for all m on the equilibrium path.[6]

(iv) $\pi_{\sigma_R}(\cdot|m) \in \Delta\Omega$ is updated by Bayes' rule (Eq. (5)) whenever possible for all m on the equilibrium path.

(v) $A^*(\pi) \in \arg\max_{a\in\mathcal{A}} \mathbb{E}_{\omega\sim\pi}\left[U_R(a,\omega)\right]$ for all $\pi \in \Delta\Omega$.

Intuitively, in equilibrium (i) the sender proposes a model to maximize her expected payoff anticipating her own equilibrium messaging strategy and the receiver's subsequent equilibrium model adoption, belief, and actions; (ii) the sender chooses her messaging strategy to maximize her expected payoff given any state and proposed model, including off-path proposals; (iii) the receiver adopts models using the Bayes factor model switching rule on path; (iv) the receiver forms posterior beliefs given any message and adopted model using Bayes rule whenever possible; (v) the receiver chooses actions to maximize his expected payoff.

Objects of interest Throughout we focus on the *sender-optimal equilibrium*, namely the equilibrium that maximizes the sender's expected payoff. This focus serves two purposes. First, it provides a natural selection criterion when multiplicity arises.[7] Second, and more importantly, it identifies the maximum gain the sender can extract from jointly controlling both the narrative and messaging strategy. With this, we can assume without loss that the receiver breaks ties in favor of the risky action when indifferent.

One central question is how much narrative persuasion expands the sender's ability to manipulate the receiver beyond what is achievable through strategic communication alone, by which we mean the benchmark in which the sender cannot propose a model and the receiver maintains his correctly specified model (i.e., a pure communication game).[8]

Definition 2. The sender *fully manipulates* the receiver if he chooses $a = 1$ with probability 1, and *partially manipulates* the receiver if he chooses $a = 1$ with probability strictly less than 1 but strictly greater than the probability attainable in any equilibrium with strategic communication alone.

[6] For conditions (iii) and (iv), "on the equilibrium path" means that $\hat{\sigma} = \hat{\sigma}^*, \sigma = \sigma^*(\hat{\sigma}^*)$, $m \in \text{supp}(\sigma^*(\hat{\sigma}^*)) \cup \text{supp}(\hat{\sigma}^*)$, and $\sigma_R = \sigma_R^*(m|\sigma^*(\hat{\sigma}^*), \hat{\sigma}^*)$. Off the equilibrium path (i.e., following a deviation $\hat{\sigma}' \neq \hat{\sigma}^*$ or $m' \notin \text{supp}(\sigma^*(\hat{\sigma}^*)) \cup \text{supp}(\hat{\sigma}^*)$), the receiver's model-switching rule and posterior beliefs are left unrestricted. Condition (i) and (ii) require only that no deviation $\hat{\sigma}'$ or m' is profitable for some specification of these off-path objects, consistent with the conventional definition of a weak PBE.

[7] As is standard in communication games under weak PBE, the flexibility of off-path beliefs generally implies non-uniqueness.

[8] While our main focus is on the incremental role of narrative persuasion relative to strategic communication, we also discuss the incremental role of optimized messaging over narrative persuasion alone in Section 5.1.

2.2 Discussion of Assumptions

Interpretation of the proposed model While the proposed model and the initial model may appear asymmetric—one described by a single $\hat{\sigma}$ and the other a contingent plan $(\hat{\sigma}, \sigma(\cdot))$—they are symmetric in the following sense: under the proposed model, the receiver believes the sender both proposes and communicates according to $\hat{\sigma}$ on and off path, as if she had commitment power, so that $\hat{\sigma}$ is a sufficient description of such a sender's strategy. The proposed model $\hat{\sigma}$ is thus a *behavioral* description of the sender's communication rule rather than a strategically justified equilibrium strategy. The receiver takes it at face value and evaluates its plausibility by comparing how well it explains the observed messages.

(In)correctly specified initial model We assume the receiver's initial model is correctly specified, which stacks the environment against the sender: the receiver begins with the most accurate understanding of the sender's strategic incentives and information technology constraints. In Section 5.3, we relax this assumption by allowing the initial model to be incorrect and show that this generally makes the receiver more vulnerable to manipulation.

Why does the receiver switch at all? Model switching is a natural response to uncertainty about the correct model. In a rich environment where the receiver may encounter different kinds of senders who truthfully follow distinct communication strategies, it is reasonable for him to remain open to alternatives. However, while he can reason carefully about the sender's incentives within a given model, he may lack the knowledge or capacity to specify and average over a large set of sender types. Instead, he compares his initial model with the salient alternative proposed by the sender. Our paper shows how this limited form of model comparison can be exploited by a strategic sender even when the receiver's initial model is correct.

Switching vs. averaging Anecdotal and experimental evidence show that people often adopt one model at a time and switch when evidence sufficiently favors an alternative, rather than maintaining continuous probabilistic beliefs over them (Barron and Fries, 2024; Aina and Schneider, 2025; Ba, 2026). This may reflect cognitive constraints on reasoning across structurally distinct models. Our switching framework delivers sharp results, with the extent of manipulation cleanly characterized by the threshold α. A more Bayesian approach would instead require the receiver to assign a positive prior weight to the proposed model before hearing any message, which is less plausible for narratives the receiver had not previously considered. Nevertheless, our main results persist qualitatively in such a framework (Section 5.4).

3 The Cheap Talk Benchmark

We begin with the cheap talk benchmark, where any message can be sent in any state ($M(\omega) = M, \forall \omega \in \Omega$). This case removes all constraints on the sender's information technology and therefore provides a natural baseline. Strikingly, although cheap talk is entirely powerless in a standard communication game, narrative persuasion makes full manipulation possible under economically intuitive conditions.

Without narrative persuasion To set up the analysis, we first establish two benchmarks and introduce the key objects that will be used throughout the paper. Consider what the sender can achieve through strategic communication alone, with or without commitment power. Any signal structure σ gives rise to *winning* and *losing* messages when the receiver believes in σ. Winning messages induce the risky action,

$$M_\sigma^+ := \left\{ m \in \operatorname{supp}(\sigma) : \mathbb{E}_{\omega \sim \pi_\sigma(\cdot|m)}[\omega] \geq \omega_0 \right\}, \tag{8}$$

while losing messages induce the safe action, $M_\sigma^- {:=} \operatorname{supp}(\sigma) \setminus M_\sigma^+$. A sender with commitment power solves a Bayesian persuasion (BP) problem to maximize her ex ante winning probability:

$$\max_{\sigma \in \Sigma} \int_\Omega \sigma(M_\sigma^+|\omega)\, dF(\omega). \tag{9}$$

The solution has the familiar cutoff structure of Kamenica and Gentzkow (2011). Define $\omega^* := \max\{\tilde{\omega} : \mathbb{E}(\omega|\omega < \tilde{\omega}) \leq \omega_0\}$. An optimal policy sends a favorable message H for $\omega > \omega^*$, an unfavorable one L for $\omega < \omega^*$, and may mix at $\omega = \omega^*$, making the receiver indifferent after H and strictly prefer the safe action after L. Intuitively, this maximizes the probability of winning by excluding the lowest states while pooling higher ones. Let p^* denote the resulting winning probability, equal to $F(\omega \geq \omega^*)$ when F is atomless.[9]

Lemma 1. *In any solution to the Bayesian persuasion problem (9), almost all types $\omega \geq \omega^*$ win and almost all types $\omega < \omega^*$ lose, yielding a winning probability of p^*.*

Without commitment, however, the familiar unraveling logic of Crawford and Sobel (1982) applies. Any winning message would be flooded by losing types, causing informative communication to collapse. As a result, babbling is the unique equilibrium, and the sender's maximum winning probability is 0. Narrative persuasion could potentially bridge the gap between the winning probability a cheap talk sender can achieve with commitment power (p^*) and without it (0), but as we will now see, this is a vast understatement.

[9] If F has an atom at ω^*, the sender sends message H at ω^* with probability $q = \frac{\mathbb{E}[\omega|\omega>\omega^*]-\omega_0 F((\omega^*,\bar{\omega}])}{(\omega_0-\omega^*)F(\{\omega^*\})}$ and message L with probability $1-q$; in this case, $p^* = F(\omega > \omega^*) + F(\{\omega^*\}) \cdot q$.

With narrative persuasion We now allow the sender to also propose a narrative $\hat{\sigma} \in \bar{\Sigma}$ alongside controlling the messaging strategy. The sender's ability to manipulate the receiver hinges on a simple demand-budget comparison.

Under messaging strategy σ, the sender's *demand for narrative persuasion* is the probability of sending a losing message,

$$D(\sigma) := \int_{\Omega} \sigma(M_{\sigma}^{-}|\omega)\, dF(\omega), \tag{10}$$

which is bounded below by $D := 1 - p^*$ by Lemma 1. The quantity D is precisely the minimum probability mass of types that must lose through messaging and therefore can only win if the receiver adopts an alternative model $\hat{\sigma}$. Conversely, given a proposed model $\hat{\sigma}$, the sender's *budget for narrative persuasion* is

$$B(\hat{\sigma}) := \int_{\Omega} \hat{\sigma}(M_{\hat{\sigma}}^{+}|\omega)\, dF(\omega), \tag{11}$$

which is bounded above by $B := p^*$ by Lemma 1. The quantity B is the maximum winning probability if the receiver adopted $\hat{\sigma}$, and thus represents the pool of types the sender can leverage in narrative persuasion. Theorem 1 shows a sharp dichotomy: the sender either fully manipulates the receiver or ends up babbling, depending on whether the budget covers the demand.

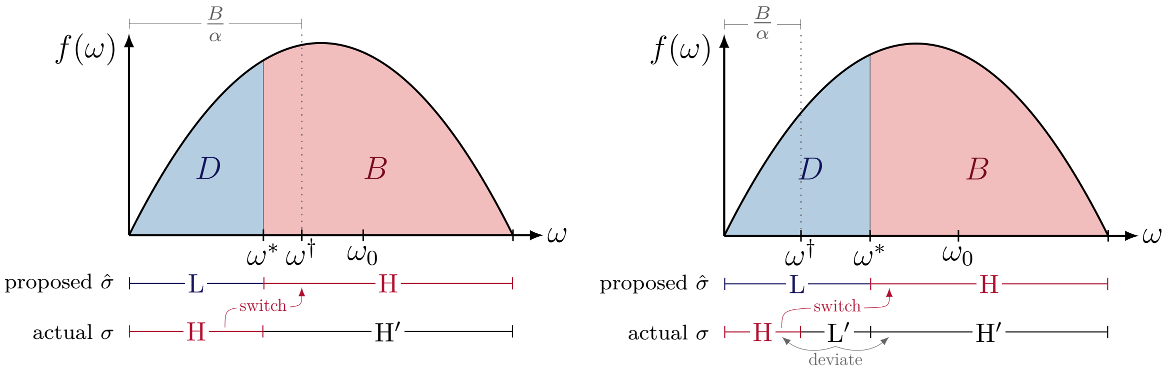
Theorem 1. *A cheap-talk sender fully manipulates in the sender-optimal equilibrium if and only if $\alpha D \leq B$, or equivalently $p^* \geq \alpha/(1+\alpha)$, and cannot manipulate otherwise.*[10]

The condition for full manipulation amounts to the narrative budget B being large enough relative to the demand D at *price* α: to convince the receiver to switch models on any message m, the sender must send m at least α times as often under $\hat{\sigma}$ than under σ, so each unit of demand costs α units of budget. The equivalent expression $p^* \geq \alpha/(1+\alpha)$ yields intuitive comparative statics. A more skeptical receiver (higher α) raises the price and tightens the budget constraint, whereas a more favorable prior (higher p^*) relaxes the constraint by simultaneously raising the budget and lowering the demand. The role of α is stark at the extremes: as $\alpha \to 0$, the price vanishes and full manipulation is always feasible regardless of the prior, while as $\alpha \to \infty$ the price becomes prohibitive and full manipulation is impossible.

The sender achieves full manipulation by running a *confidence trick*, illustrated in Figure 1a.[11] She proposes the narrative that she is a *Bayesian persuader*, sending H for $\omega \geq \omega^*$ and L for $\omega < \omega^*$, essentially bluffing by claiming to have commitment power that

[10] The result requires only that the message space M contains at least three distinct messages.

[11] This illustration focuses on the case where the state is continuously distributed to simplify exposition. The discrete case is analogous, except that the sender may mix at threshold states.



(a) Full manipulation when $\alpha D \leq B$.

(b) Manipulation breakdown when $\alpha D > B$.

Figure 1: Manipulation with cheap talk information technology.

she does not possess. In actuality she does the opposite: she sends H for $\omega \leq \omega^*$ and H' for $\omega \geq \omega^*$. The two strategies are mirror images of each other across ω^*. After observing H, the receiver switches models because H is sent with probability D under the true strategy and with probability B under the proposed model, so $\lambda(H) = B/D \geq \alpha$. He then incorrectly expects $\mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0$, which leads to the risky action even though it is objectively worse, since $\omega \leq \omega^* < \omega_0$. Types above ω^* strategically contradict their own narrative by sending H', a message never sent under $\hat{\sigma}$, so $\lambda(H') = 0$ and the receiver sticks to his original model. He then correctly expects $\mathbb{E}[\omega|\omega > \omega^*] = \omega_0$ and takes the risky action. Thus every type induces to the risky action, and since no type has a profitable deviation, this is indeed an equilibrium. The sender fully manipulates the receiver while vastly outperforming the Bayesian persuader she claims to be.

This construction also uses the narrative budget as efficiently as possible. It maximizes $\lambda(H)$ at B/D, subject to the condition that the receiver weakly prefers the risky action following any message. When $\alpha D > B$, manipulation breaks down entirely, as illustrated in Figure 1b. The previous strategy fails because the receiver no longer switches after observing H. To restore switching, the sender could instead send H only for $\omega \leq \omega^\dagger$ where $\omega^\dagger < \omega^*$ is chosen so that $\lambda(H) = \alpha$. However, this leaves a residual interval $(\omega^\dagger, \omega^*)$ of losing types who can profitably deviate to a winning message, causing any informative strategy to unravel. No other allocation resolves this problem: any message sent often enough to trigger switching exhausts the budget before covering all losing types, leaving a residual interval of deviators. Babbling is therefore the unique equilibrium.[12]

Theorem 1 reveals a novel *complementarity* between strategic messaging and narrative

[12] A babbling equilibrium exists as follows: the sender proposes $\hat{\sigma}^*$ that pools all types on a single message, say H, and plays $\sigma^*(H|\omega, \hat{\sigma}) = 1$ for all ω and all $\hat{\sigma}$ on and off path. The receiver never switches models and his posterior equals the prior, so he takes the safe action. No type has a profitable deviation as all messages lead to the safe action off path under beliefs that assign zero probability to the risky action being optimal.

persuasion. Strategic messaging without narratives yields exactly 0, while narrative persuasion alone yields at most p^*/α, both potentially far below the commitment payoff p^*. Yet when used jointly, the sender leaps all the way to payoff 1, fully manipulating a receiver who began with a correctly specified model of her strategic incentives. The mechanism is deeply synergistic. Narrative persuasion not only enables losing types to win by triggering a model switch, but also eliminates the credibility hurdle that prevents informative communication in the first place: since every type wins in equilibrium, no type has a profitable deviation, and the unraveling logic that ordinarily dooms cheap talk disappears. This in turn frees strategic messaging to operate at full force, replicating the commitment payoff and then surpassing it.

Welfare implications The characterization generates two stark welfare implications. First, full manipulation can generate *arbitrarily large* welfare loss for the receiver.

Corollary 1. *For any α and prior satisfying $\alpha D \leq B$, the sender can fully manipulate the receiver even as $\mathbb{E}(\omega)$ decreases to $-\infty$.*

We illustrate Corollary 1 with the following example.

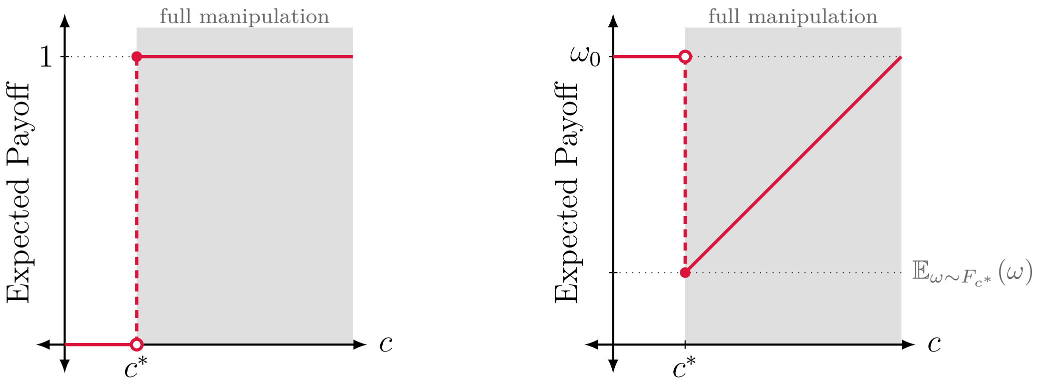
Example 1. (Picking up pennies in front of a steamroller[13]) Suppose $\alpha = 1$, $\omega_0 = 0$, and the risky action yields either $\omega = \epsilon > 0$ or a low payoff $\omega = -b < 0$ with equal probability. As $-b \to -\infty$, the risky action has a 50% chance of yielding an arbitrarily bad payoff. Nevertheless, since $p^* > 1/2 = \alpha/(1+\alpha)$, full manipulation is feasible for any b by Theorem 1. The sender can therefore always convince the receiver to take the risky action—which offers only a penny of upside (ϵ) against an arbitrarily large downside (b)—even as its expected payoff approaches $-\infty$. The same logic holds for any α: if the risky action yields $\omega_0 + \epsilon$ with probability $\alpha/(1+\alpha)$ and $-b$ otherwise, full manipulation remains feasible as $-b \to -\infty$.

Second, while the sender always benefits from a more favorable state distribution, the receiver's welfare is non-monotone. To see this, shift the entire state distribution by a constant $c \in \mathbb{R}$. Let $F_c \in \Delta\Omega$ denote the distribution with density $f_c(\omega) := f(\omega - c)$ for all $\omega \in \Omega$ while keeping ω_0 fixed and maintaining the assumption $\mathbb{E}_{\omega\sim F_c}(\omega) < \omega_0$. As c increases, the BP cutoff ω_c^* shifts up by c and p_c^* increases, since a more favorable state distribution allows more types to pool and win.

Corollary 2. *In the sender-optimal equilibrium, there exists a threshold c^* such that:*

(i) the sender's payoff is 0 for $c < c^*$ and 1 for $c \geq c^*$;

[13] "Picking up pennies in front of a steamroller" (Taleb, 2007, p. 19) describes a bet with negligible upside and catastrophic downside. Here it is taken to its logical extreme, where the pennies' worth is vanishing while the steamroller is infinitely threatening.



(a) Sender's Welfare.

(b) Receiver's Welfare.

Figure 2: Equilibrium welfare comparison as c increases.

(ii) the receiver's payoff ω_0 for $c < c^*$, drops discontinuously to $\mathbb{E}_{\omega\sim F_{c^*}}(\omega) < \omega_0$ at c^*, and increases linearly in c thereafter.

The result is illustrated in Figure 2. The sender either babbles or fully manipulates, and a better distribution only makes manipulation easier and her payoff weakly improves. For the receiver, however, a better distribution can be a curse. When the distribution is sufficiently unfavorable ($c < c^*$), the sender cannot manipulate and the receiver comfortably takes the safe action, obtaining ω_0. Once c crosses c^*, the receiver's payoff collapses. The very improvement in the state distribution makes full manipulation feasible, forcing the receiver to take the risky action. His payoff becomes the unconditional mean of the state, which is strictly below ω_0 and can be arbitrarily bad (as Example 1 illustrates). Beyond c^*, this mean increases linearly in c, so his payoff gradually recovers.

4 General Information Technology

We now generalize the analysis beyond cheap talk. Section 4.1 introduces a broad class of information technologies, Section 4.2 characterizes the sender-optimal equilibrium, and Section 4.3 examines how persuasiveness varies with the properties of the technology. We focus on a continuously distributed state throughout this section.

4.1 Monotone Interval Technologies

Following Hart et al. (2017), we take the feasible message set to be a subset of the type space, so that sending message ω' in state ω can be interpreted as type ω mimicking type ω'. We focus on a class of information technologies we call *monotone interval technologies*, which capture many standard choices while remaining analytically tractable. In state ω

the sender can claim that the state is some $\omega' \in [L(\omega), U(\omega)] \cap \Omega$. The lower and upper bounds $L : \Omega \to \bar{\mathbb{R}}$ and $U : \Omega \to \bar{\mathbb{R}}$ are weakly increasing with $L(\omega) \leq \omega \leq U(\omega)$. This captures the natural idea that higher states have access to higher claims, and that the sender can always truthfully report her state. For tractability, we assume that both L and U are piecewise C^1 and right-continuous.[14] Denote the set of all such technologies as $\mathbb{M}$.[15] If, in addition, L and U are strictly increasing on Ω, we call the technology *strictly monotone* and denote the set of such technologies by $\mathbb{M}^*$.

We focus on monotone interval technologies for two reasons. First, this class is flexible enough to capture a wide range of environments that vary in the hardness of information: the wider the interval $[L(\omega), U(\omega)]$, the more types a sender can mimic and the less verifiable her information is. We illustrate this with examples below.

Example 2. Examples of monotone interval technologies in $\mathbb{M}$ include:

(i) *Cheap talk*: $L(\omega) = \inf \Omega$ and $U(\omega) = \sup \Omega$, so $M(\omega) = \Omega, \forall \omega$.

(ii) *Perfectly verifiable disclosure*: $L(\omega) = U(\omega) = \omega$, so $M(\omega) = \{\omega\}$, $\forall \omega$.

(iii) *Disclosing lower types*: $L(\omega) = \inf \Omega$ and $U(\omega) = \omega$, so $M(\omega) = \{\omega' \in \Omega : \omega' \leq \omega\}$, $\forall \omega$. The sender can underreport the state but not overreport.

(iv) *Noisy evidence*: $L(\omega) = \omega - \epsilon$ and $U(\omega) = \omega + \epsilon$ for some $\epsilon > 0$, so $M(\omega) = [\omega - \epsilon, \omega + \epsilon]$. Here, evidence is verifiable only up to noise of size ϵ.

(v) *Coarse evidence*: there exists a partition $P = \{\Omega_1, ..., \Omega_K\}$ of Ω into left-closed right-open intervals (Ω_K is right-closed), where each cell Ω_k corresponds to a distinct piece of evidence available only to types in Ω_k. For any $\omega \in \Omega_k$, $L(\omega) = \inf \Omega_k$ and $U(\omega) = \sup \Omega_k$. The sender can mimic those who share the same piece of evidence but cannot produce evidence belonging to a different cell.

Second, monotone interval technologies yield a tractable characterization of sender-optimal equilibria, mainly owing to two key properties formalized in Lemma 2. The lemma says that the set of types who can send any given message m forms a closed interval $\Omega_{\mathbb{M}}(m)$, and that any message available to a type between two types in $\Omega_{\mathbb{M}}(m)$ is also available to at least one of those two types.

[14] Piecewise C^1 means that Ω can be partitioned into finitely many intervals on each of which L and U are continuously differentiable. Right-continuity means that $L(\omega) = \lim_{\omega' \downarrow \omega} L(\omega')$ and $U(\omega) = \lim_{\omega' \downarrow \omega} U(\omega')$ for all $\omega \in \Omega$.

[15] It is often assumed in the disclosure literature that the sender may remain silent by sending $\varnothing$ (e.g., Grossman, 1981; Milgrom, 1981). Unlike non-silence messages, $\varnothing$ is available to every type regardless of the technology, making it a universal pooling device that can dramatically change the sender's persuasiveness in our setting. Thus, we treat this case where silence is available separately in Remark 3.

Lemma 2. *For any $m \in \Omega$, $\Omega_{\mathcal{M}}(m) := \{\omega \in \Omega : m \in [L(\omega), U(\omega)]\}$ is a closed interval. Moreover, for any $\omega_1, \omega_2 \in \Omega_{\mathcal{M}}(m)$ with $\omega_1 < \omega_2$ and any message m' sent by some type $\omega \in [\omega_1, \omega_2]$, at least one of ω_1 and ω_2 can also send m'.*

4.2 Characterizing the Sender-Optimal Equilibrium

We now characterize the sender-optimal equilibrium under monotone interval signal structures, following a parallel structure as in Section 3. We first analyze what the sender can achieve through strategic communication alone, with or without commitment, and then build on these to solve the full game with narrative persuasion.

Without narrative persuasion A sender with commitment power solves a constrained Bayesian persuasion problem: she commits to a feasible signal structure $\sigma \in \Sigma$ given $\mathcal{M}$ to maximize her ex ante winning probability $\int_\Omega \sigma(M_\sigma^+|\omega)dF(\omega)$. Despite the constraints, the solution retains a clean cutoff structure familiar from Section 3.

Proposition 1. *Fix $\mathcal{M} \in \mathbb{M}$. Any solution to the constrained BP problem has a cutoff structure: there exists $\omega^*_{\mathcal{M}} \in \Omega$ such that almost all types $\omega \geq \omega^*_{\mathcal{M}}$ win and almost all types $\omega < \omega^*_{\mathcal{M}}$ lose, yielding a maximum winning probability of $p^*_{\mathcal{M}} := F(\omega \geq \omega^*_{\mathcal{M}})$.*

All solutions to the constrained BP problem share the same cutoff $\omega^*_{\mathcal{M}}$ and winning probability $p^*_{\mathcal{M}}$. We illustrate one natural construction in Fig. 3. Low types in $[\omega^*_{\mathcal{M}}, \omega_0)$ send their highest feasible message $U(\omega)$ and pool with high types in $[\omega_0, \overline{\omega})$, whose strategy is calibrated so that each pooled message induces a posterior mean exactly equal to ω_0, making the receiver just indifferent. The key feature is that high types are exhausted from ω_0 upward, since by monotonicity of U they send lower messages and are therefore accessible to more lower types. Types below $\omega^*_{\mathcal{M}}$ lose, while types above $\overline{\omega}$ win outright by sending their highest feasible message. The cutoff $\omega^*_{\mathcal{M}}$ is the lowest type for which this construction is feasible. As ω^* decreases, more low types must be pooled with high types, pushing the construction further down until it hits the lower feasibility bound $L(\omega)$. The formal construction and proof of optimality are in Section A.2.2.

Without commitment, the sender must also ensure that no type has a profitable deviation. Under cheap talk this is impossible and the sender babbles. Under a general monotone interval technology, however, message constraints can discipline deviations: a type cannot send a winning message if that message lies outside her feasible set. The key condition is whether the cutoff type can credibly separate from lower types by sending a highest feasible message unavailable to them.

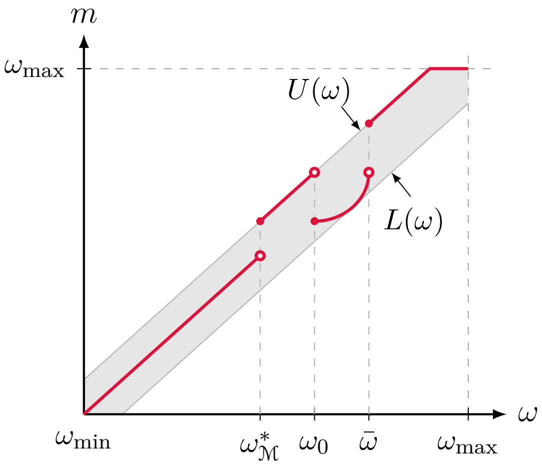


Figure 3: Solution to constrained BP problem.[16] The gray band depicts the feasible message set $[L(\omega), U(\omega)] \cap \Omega$ and the red curve shows the sender's optimal strategy.

Definition 3. Type ω is *credible* under $\mathcal{M}$ if no type below ω can send the same highest feasible message as ω, i.e., $U(\omega) > U(\omega')$ for any $\omega' < \omega$.

Right-continuity and monotonicity of U ensure that the set of credible types at or above $\omega^*_{\mathcal{M}}$ is either empty or has a well-defined minimum.

Proposition 2. *Fix $\mathcal{M} \in \mathbb{M}$. Let $\tilde{\omega}^*_{\mathcal{M}}$ be the lowest credible type at or above $\omega^*_{\mathcal{M}}$, if it exists. A sender-optimal equilibrium exists where types $\omega \geq \tilde{\omega}^*_{\mathcal{M}}$ win and types $\omega < \tilde{\omega}^*_{\mathcal{M}}$ lose, yielding winning probability $F(\omega \geq \tilde{\omega}^*_{\mathcal{M}})$, or 0 if no such type exists.*

Clearly, the sender weakly benefits from commitment power since $\tilde{\omega}^*_{\mathcal{M}} \geq \omega^*_{\mathcal{M}}$. To the best of our knowledge, this is the first explicit characterization of the BP solution and the sender-optimal equilibrium for the full class of monotone interval technologies.[17]

Example 2 (continued). We compute the BP solution and the sender-optimal equilibrium without commitment for technologies in Example 2.

(i) *Cheap talk*: $p^*_{CT} = p^*$ and $\omega^*_{CT} = \omega^*$, but $\tilde{\omega}^*_{CT}$ does not exist.

(ii) *Perfectly verifiable disclosure*: $p^*_{PV} = F(\omega \geq \omega_0)$ and $\omega^*_{PV} = \tilde{\omega}^*_{PV} = \omega_0$.

(iii) *Disclosing lower types*: $p^*_L = p^*$ and $\omega^*_L = \tilde{\omega}^*_L = \omega^*$.

(iv) *Noisy evidence*: given a uniform prior on $\Omega = [0, 1]$, $p^*_\epsilon = \min\{1 - \omega_0 + \epsilon\sqrt{2}, p^*\}$, $\omega^*_\epsilon = \max\{\omega_0 - \epsilon\sqrt{2}, \omega^*\}$, and $\tilde{\omega}^*_\epsilon = \omega^*_\epsilon$ if $\epsilon \leq 1 - \omega^*$, does not exist otherwise.[18]

[16] The figure uses the example $\Omega = [0, 1], \omega_0 = 0.6, L(\omega) = \omega - 0.1, U(\omega) = \omega + 0.1$, uniform prior, giving $\omega^*_{\mathcal{M}} = 0.6 - \frac{1}{5\sqrt{2}}$, $\overline{\omega} = 0.6 + \frac{1}{5\sqrt{2}}$, and $\sigma(\omega) = 0.7 - \sqrt{(0.6 - \omega^*)^2 - (\omega - 0.6)^2}$ for $\omega \in [\omega_0, \overline{\omega})$.

[17] If the sender can also send $\varnothing$ in all states and has commitment power, she can replicate the structure in Lemma 1 and obtain p^* by pooling high states on $\varnothing$. The sender-optimal equilibrium for the pure communication game remains unchanged, since $\varnothing$ cannot be a credible winning message.

[18] See Lemma 3 for a full derivation.

(v) *Coarse evidence*: let Ω_k be the lowest cell such that $\mathbb{E}(\omega|\omega \in \Omega_k) \geq \omega_0$. Then $\omega^*_{\mathcal{P}}$ is the lowest $\omega' \in \Omega_k$ such that $\mathbb{E}(\omega|\omega \in [\omega', \sup \Omega_k]) = \omega_0$, $p^*_{\mathcal{P}} = F(\omega \geq \omega^*_{\mathcal{P}})$, and $\tilde{\omega}^*_{\mathcal{P}} = \inf \Omega_k$.

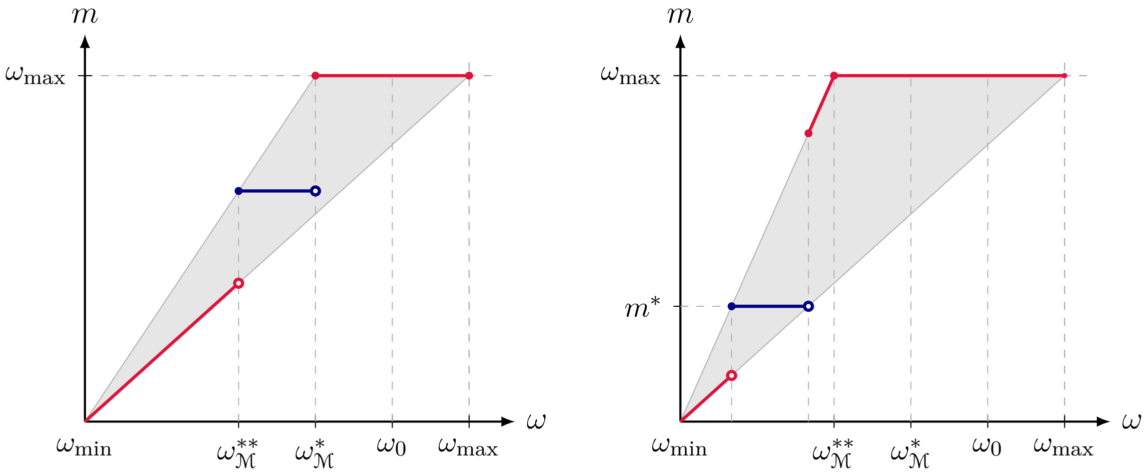
With narrative persuasion As in cheap talk, full manipulation hinges on whether the narrative budget can fully cover the demand. Since pure pooling wins at most $p^*_{\mathcal{M}}$, the narrative demand is now $D_{\mathcal{M}} := 1 - p^*_{\mathcal{M}}$. Meanwhile, the budget remains $B = p^*$ because the proposed model is unconstrained. Theorem 2 generalizes Theorem 1 accordingly. The complementarity between messaging and narratives again plays a key role: narrative persuasion removes the credibility hurdle that otherwise caps the sender's messaging payoff at $F(\omega \geq \tilde{\omega}^*_{\mathcal{M}})$, allowing strategic messaging to contribute its full commitment value $p^*_{\mathcal{M}}$.

Theorem 2 (Full Manipulation)**.** *Fix any* $\mathcal{M} \in \mathbb{M}$*. The sender* fully manipulates *in sender-optimal equilibrium if and only if* $B \geq \alpha D_{\mathcal{M}}$*, or equivalently* $p^*/\alpha \geq 1 - p^*_{\mathcal{M}}$.

A simple and versatile full manipulation strategy is as follows. The sender's actual messaging strategy replicates the constrained BP solution from Proposition 1: high types $\omega \geq \omega^*_{\mathcal{M}}$ send winning messages and low types $\omega < \omega^*_{\mathcal{M}}$ send losing messages. In the proposed model $\hat{\sigma}^*$, the sender *flips the interpretation* of losing messages: she claims that high types would randomly send exactly the messages that low types actually send. Upon observing a losing message, the receiver therefore switches models and interprets it as good news.[19] Winning messages are never sent under $\hat{\sigma}^*$, so the receiver keeps his initial model after observing them and correctly infers good news from the fact that only high types send them. This strategy is efficient: by having high types cover exactly the losing mass $D_{\mathcal{M}}$ under $\hat{\sigma}^*$ and nothing more, the sender wastes no narrative budget on types that can already win by pooling. Full manipulation is therefore feasible if and only if the budget B suffices to cover the demand $D_{\mathcal{M}}$ at price α.

When full manipulation is infeasible, some information technologies open up a new possibility absent under cheap talk: partial manipulation. Message constraints can block low types from sending winning messages, preventing unraveling and allowing the sender to win with a subset of types even when she cannot win with all. Define the set of types who can mimic all types above them as $\bar{\Omega}_{\mathcal{M}} := \{\omega \in \Omega : U(\omega) \geq \sup \Omega\}$. For example, under cheap talk $\bar{\Omega}_{CT} = \Omega$; with noisy evidence of size ϵ, $\bar{\Omega}_\epsilon = [\max \Omega - \epsilon, \max \Omega]$ if Ω is bounded above and $\bar{\Omega}_\epsilon = \varnothing$ otherwise. For Theorem 3, we restrict attention to

[19] A real-world analogue is the contrast between a momentum strategy and a contrarian strategy in financial markets: the former recommends buying when prices are rising and the latter recommends buying when prices are falling. By proposing a contrarian narrative, a sender can convince a receiver that a falling price signal is in fact a buying opportunity.



(a) Case (i): $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$.

(b) Case (ii): $\omega_{\mathcal{M}}^{**} \in \bar{\Omega}_{\mathcal{M}}$.

Figure 4: Illustration of sender-optimal equilibrium in Theorem 3. The gray band depicts $[L(\omega), U(\omega)] \cap \Omega$ and the colored curves show σ^*. In case (i), types $\omega \geq \omega_{\mathcal{M}}^{*}$ win by pooling (red) and types $\omega \in [\omega_{\mathcal{M}}^{**}, \omega_{\mathcal{M}}^{*})$ win via narrative persuasion (blue). In case (ii), types sending m^* win via narrative persuasion (blue) and all other types lose.

strictly monotone interval technologies $\mathcal{M} \in \mathbb{M}^*$ to obtain a clean characterization, as strict monotonicity ensures every type below $\bar{\Omega}_{\mathcal{M}}$ is credible.[20]

Theorem 3 (Partial and No Manipulation)**.** *Fix $\mathcal{M} \in \mathbb{M}^*$ with $B < \alpha D_{\mathcal{M}}$, and let $\omega_{\mathcal{M}}^{**}$ be defined by*

$$B = \alpha(D_{\mathcal{M}} - F(\omega < \omega_{\mathcal{M}}^{**})). \tag{12}$$

(i) If $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$, a sender-optimal equilibrium exists where types $\omega \geq \omega_{\mathcal{M}}^{**}$ win and types $\omega < \omega_{\mathcal{M}}^{**}$ lose, and the sender partially manipulates.

(ii) If $\omega_{\mathcal{M}}^{**} \in \bar{\Omega}_{\mathcal{M}}$ and $B \geq \alpha F(\Omega_{\mathcal{M}}(m^*))$ for some $m^* \in \Omega$, a sender-optimal equilibrium exists with winning probability B/α and the sender partially manipulates; otherwise all types lose and the sender cannot manipulate.

Condition Eq. (12) has a natural interpretation. Since the full demand $D_{\mathcal{M}}$ cannot be covered, the sender must concede some losing types. The cutoff $\omega_{\mathcal{M}}^{**}$ is the unique type where the narrative budget B exactly covers the residual demand $D_{\mathcal{M}} - F(\omega < \omega_{\mathcal{M}}^{**})$ at price α, assuming losing types concentrate at the bottom. Whether a credible cutoff at $\omega_{\mathcal{M}}^{**}$ can be sustained determines which instruments the sender can deploy, giving rise to two distinct regimes.

When $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$, the sender wins by combining direct pooling and narrative persuasion (Fig. 4a). Types above $\omega_{\mathcal{M}}^{*}$ replicate the constrained BP solution from Proposition 1; types

[20] Coarse evidence falls outside $\mathbb{M}^*$ and gives rise to a more complex knapsack-type problem that we leave for future work.

in $[\omega_{\mathcal{M}}^{**}, \omega_{\mathcal{M}}^{*})$ win via narrative persuasion, with the sender proposing that high types would send the very losing messages that these types actually send, flipping their interpretation from bad news to good; and types below $\omega_{\mathcal{M}}^{**}$ lose. Since $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$, it is credible and no type just below $\omega_{\mathcal{M}}^{**}$ can mimic the winning messages above, stabilizing the losing region. The complementarity between messaging and narratives is, again, stark: while the pure communication equilibrium features no manipulation in the specification of Fig. 4a, the joint use of messaging and narrative delivers a payoff of $F(\omega \geq \omega_{\mathcal{M}}^{**}) = p^*/\alpha + p_{\mathcal{M}}^*$, strictly above the commitment payoff $p_{\mathcal{M}}^*$.

When $\omega_{\mathcal{M}}^{**} \in \bar{\Omega}_{\mathcal{M}}$, the sender wins via narrative persuasion alone (Fig. 4b),. The strategy in the previous case no longer works: either $\omega_{\mathcal{M}}^{**}$ is not credible, or the winning types do not have enough distinct messages to use both pooling and narrative persuasion. This eliminates the complementarity between the two tools and makes winning via direct pooling impossible in any sender-optimal equilibrium. The sender instead finds a message m^* with a sufficiently small pool $\Omega_{\mathcal{M}}(m^*)$ that the switching budget suffices. By continuity of $F(\Omega_{\mathcal{M}}(m))$ in m, the budget can be fully exhausted by expanding the winning message interval around m^*, achieving winning probability B/α. If no such message exists, the budget is insufficient to trigger any switch and all types lose.

As in cheap talk, complementarity between messaging and narratives ensures that whenever it is at work, the sender surpasses the commitment benchmark entirely. Under full manipulation (Theorem 2), the sender achieves payoff $1 > p_{\mathcal{M}}^*$. Under partial manipulation with $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$ (Theorem 3 case (i)), the sender achieves $F(\omega \geq \omega_{\mathcal{M}}^{**}) = p_{\mathcal{M}}^* + p^*/\alpha > p_{\mathcal{M}}^*$. The exception is case (ii), where the complementarity breaks down and the sender achieves only p^*/α, which need not exceed $p_{\mathcal{M}}^*$.

Corollary 3. *Under full manipulation or partial manipulation with $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$, the sender's payoff strictly exceeds the commitment payoff $p_{\mathcal{M}}^*$.*

Remark 3. If silence is available ($\varnothing \in \mathcal{M}$), the BP solution achieves the unconstrained winning probability p^*: high types above ω^* send $\varnothing$ and win, while low types below ω^* send non-silence messages and lose, exactly as in the cheap talk benchmark. This gives $p_{\mathcal{M}}^* = p^*$ and reduces demand to $D = 1 - p^*$, so the full manipulation condition in Theorem 2 simplifies to $B \geq \alpha D$, the same as Theorem 1. Meanwhile, Theorem 3 is unchanged: since every type can send $\varnothing$, it cannot be a winning message in a partial manipulation equilibrium.

4.3 Flexibility, Credibility, and Persuasiveness

The payoffs achieved in Theorems 2 and 3 suggest a natural upper bound on what any information technology can deliver. Under full manipulation the sender achieves payoff

1, while under partial manipulation she achieves at most $p^*_{\mathcal{M}} + p^*/\alpha$. Since $p^*_{\mathcal{M}} \leq p^*$, both are bounded above by $\min\{p^*(1 + 1/\alpha), 1\}$. This bound motivates the question of which information technologies achieve it—and whether the sender can always find one that does. We now characterize how the sender's equilibrium payoff varies with the information technology and show that the payoff bound is indeed tight. Our analysis focuses on the trade-off between two key properties: *flexibility*, which determines how broadly types can pool together, and *credibility*, which determines how reliably low types can be excluded from winning pools. These properties are often in tension, and their interaction with narrative persuasion shapes the optimal choice of information technology.

Definition 4. Given two information technologies $\mathcal{M} = (L, U)$ and $\mathcal{M}' = (L', U')$ in $\mathbb{M}$, $\mathcal{M}'$ is *more flexible* than $\mathcal{M}$ if $M(\omega) \subseteq M'(\omega)$ for all $\omega \in \Omega$. In particular, $\mathcal{M}'$ is more *downward flexible* than M if $U'(\omega) = U(\omega)$ and $L'(\omega) \leq L(\omega)$ for all $\omega \in \Omega$, and more *upward flexible* if $L'(\omega) = L(\omega)$ and $U'(\omega) \geq U(\omega)$ for all $\omega \in \Omega$.

A more flexible technology benefits the sender by enabling easier pooling of good and bad types, but comes at the cost of credibility: the sender may be unable to prevent the worst types from imitating better types. When the sender can achieve full manipulation, however, this tradeoff is irrelevant: by Theorem 2, full manipulation depends only on $D_{\mathcal{M}}$, so greater flexibility can only help.

Corollary 4. *Fix $\mathcal{M} \in \mathbb{M}$ and suppose that $B \geq \alpha D_{\mathcal{M}}$. If $\mathcal{M}' \in \mathbb{M}$ is more flexible than $\mathcal{M}$, the sender can fully manipulate the receiver under $\mathcal{M}'$.*

If full manipulation is infeasible, Theorem 3 suggests that flexibility alleviates the demand $D_{\mathcal{M}}$ but may jeopardize the credibility of type $\omega^{**}_{\mathcal{M}}$. We show that the two kinds of flexibility have disparate effects. Downward flexibility allows high types to pool with more low types, with that decision in the hands of the high types, and therefore expands pooling capacity without sacrificing credibility (for any type). We can show that there exists a sender-optimal equilibrium under $\mathcal{M}$ that still remains feasible and incentive-compatible under any more downward flexibility $\mathcal{M}'$, so downward flexibility always weakly benefits the sender. In contrast, upward flexibility allows low types to pool with more high types, with that decision in the hands of the low types, and therefore typically comes at the cost of credibility. By giving low types access to messages previously exclusive to high types, it can violate the incentive-compatibility constraints that sustain the equilibrium. We illustrate both effects with the noisy evidence technology before stating the general characterization.

Example 2 (continued). Consider a noisy evidence technology $\mathcal{M} = (L, U)$ with $L(\omega) = \omega - \epsilon$ and $U(\omega) = \omega + \epsilon$, where $\epsilon > 0$ controls the degree of noise. Assume Ω

is bounded with $\omega_{\max} := \max \Omega$. A larger ϵ allows the sender to claim a wider range of states and gives her more flexibility to pool types but also allowing worse types to mimic better ones, eroding credibility. The highest credible type is $\inf \bar{\Omega}_{\mathcal{M}} = \omega_{\max} - \epsilon$, which falls as ϵ increases.

As ϵ increases from a small value, the sender initially benefits: greater pooling capacity reduces $\omega^*_{\mathcal{M}}$ and expands the set of types that win via direct pooling, raising the sender's payoff. However, credibility is simultaneously being eroded. Once ϵ is large enough that the partial manipulation cutoff $\omega^{**}_{\mathcal{M}}$ meets the highest credible type $\omega_{\max} - \epsilon$, the sender can no longer credibly exclude the worse types from imitating high types, and her payoff falls sharply. Depending on the prior, there may or may not be an intermediate flat region: if the pooling gains are exhausted before credibility is destroyed, the sender's payoff plateaus at its maximum before falling; otherwise it rises and falls without ever reaching that plateau.

This example illustrates a general pattern. Define ω^{**} as the partial manipulation cutoff that would prevail if the technology offered maximum downward flexibility, i.e., the unique type satisfying $B = \alpha(D_{\mathcal{M}} - F(\omega < \omega^{**}))$ when $D_{\mathcal{M}} = D = 1 - p^*$. This is the best cutoff achievable by any technology in $\mathbb{M}^*$ and serves as the benchmark against which upward flexibility is evaluated in the following proposition.

Proposition 3. *Consider two information technologies $\mathcal{M}, \mathcal{M}' \in \mathbb{M}^*$. In any sender-optimal equilibrium:*

(i) More downward flexible $\mathcal{M}'$ weakly benefits the sender.

(ii) More upward flexible $\mathcal{M}'$ strictly benefits the sender if and only if $\omega^{**} \notin \bar{\Omega}_{\mathcal{M}'}$ and $\exists\, \omega \in \Omega$ with $L(\omega) > U(\omega^*)$ under $\mathcal{M}$; strictly harms her if and only if $\omega^{**} \notin \bar{\Omega}_{\mathcal{M}}$ but $\omega^{**} \in \bar{\Omega}_{\mathcal{M}'}$, or $\omega^{**} \in \bar{\Omega}_{\mathcal{M}}$ and the condition of Theorem 3 case (ii) holds under $\mathcal{M}$ but not $\mathcal{M}'$; and leaves her payoff unchanged otherwise.

The noisy evidence example above illustrates the first two scenarios of Case (ii). The third scenario where $\omega^{**} \in \bar{\Omega}_{\mathcal{M}}$ and upward flexibility destroys the narrative persuasion condition—arises when the technology is already so flexible that winning under the initial model σ is infeasible, and further upward flexibility shrinks $\Omega(m)$ that can be covered by the narrative budget, eliminating narrative persuasion as well.

The proposition also reveals how narrative persuasion changes the comparative statics of flexibility. Without narratives, as a technology becomes more downward flexible, the sender's payoff increases from $F(\omega \geq \omega_0)$ toward p^* and then falls to 0 once credibility is destroyed. With narrative persuasion, the entire payoff frontier shifts up: the payoff increases from $F(\omega \geq \omega_0) + p^*/\alpha$ toward $p^* + p^*/\alpha$, stays flat once the unconstrained optimum is reached, and then falls—possibly to p^*/α first then 0, since narrative persuasion alone can still rescue a fraction p^*/α of types even after credibility is destroyed. The payoff reaches 0 only when even narrative persuasion fails, i.e., when no message pool is small enough to be covered by the narrative budget. Narrative persuasion thus not only raises the ceiling of what the sender can achieve, but also raises the floor, making the sender strictly better off at every level of flexibility where it operates.

Having shown how flexibility and credibility interact, we now characterize information technologies that are *maximally persuasive*—those that achieve the payoff bound $\min\{p^*(1 + 1/\alpha), 1\}$.

Theorem 4 (Maximally Persuasive Technologies)**.** *Fix $\mathcal{M} = (L, U) \in \mathbb{M}^*$.*

(i) If $\alpha D \leq B$, $\mathcal{M}$ is maximally persuasive if and only if ω^* shares a message with every $\omega' > \omega^*$ in Ω. The sender's payoff is 1.

(ii) If $\alpha D > B$, $\mathcal{M}$ is maximally persuasive if and only if: (a) ω^* shares a message with every $\omega' > \omega^*$ in Ω; and (b) $\omega^{**} \notin \bar{\Omega}_{\mathcal{M}}$. The sender's payoff is $p^* + p^*/\alpha$.

The two conditions have a natural economic interpretation. Condition (a)—that ω^* shares a message with every type above it—is the ability to *tell a good story*: the sender must be able to craft a winning message consistent with any high state, which is necessary for the constrained BP solution to operate at full capacity and minimize the residual demand that narrative persuasion must cover. Condition (b)—that $\omega^{**} \notin \bar{\Omega}_{\mathcal{M}}$—is the ability to *selectively withhold*: the worst types cannot mimic the winning message, credibly excluding bad news from the winning pool and preventing unraveling. Perhaps surprisingly, credibility of just one type suffices—without commitment power, credibility of the single critical type ω^{**} is exactly what substitutes for it.

Corollary 5. *$\mathcal{M} = (L, U) \in \mathbb{M}^*$ is maximally persuasive for all (F, ω_0, α) iff (a) any two types $\omega < \omega'$ in Ω share a feasible message, and (b) $U(\omega) < \sup \Omega$ for all $\omega \in \mathrm{int}(\Omega)$.*

The theorem and corollary together shed light on the persuasiveness of the technologies in Example 2. Cheap talk is maximally persuasive when $\alpha D \leq B$ since every type shares a message with every other type, but fails entirely when $\alpha D > B$ since no type is credible. Perfectly verifiable disclosure supports partial manipulation and therefore outperforms cheap talk when $\alpha D > B$, but is not maximally persuasive since ω^* cannot pool with any type above it. Noisy evidence can be maximally persuasive depending on the environment: for intermediate levels of ϵ both conditions of the theorem hold, but for large ϵ credibility is destroyed. The disclosing lower types technology satisfies both conditions of the corollary—$M(\omega) = [\inf \Omega, \omega]$ so any two types always share $\inf \Omega$, and $U(\omega) = \omega < \sup \Omega$ for all interior types—and is therefore maximally persuasive in all environments. These comparisons reveal a class of technologies that are maximally persuasive regardless of the environment or whether narratives are available. Outside this class, however, narrative persuasion can reverse technology rankings: for a given environment (F, ω_0, α), a softer technology can strictly outperform a harder one with narratives while being strictly worse without them, since narrative persuasion compensates for reduced credibility in a way that pure strategic communication cannot.

5 Extensions

We explore four extensions of the baseline model: comparing with a framework where the messaging strategy is exogenous, allowing the sender to propose multiple narratives, considering an initially misspecified receiver, and allowing the receiver to perform Bayesian updating over models rather than switching. For simplicity we focus on the cheap talk benchmark throughout, though the results extend naturally to general information technologies.

5.1 Exogenous vs. Endogenous Messaging

We isolate the value of endogenous messaging by fixing the sender's actual messaging strategy at $\sigma \in \Sigma$ exogenously and allowing her to choose only a narrative $\hat{\sigma} \in \bar{\Sigma}$. Under any fixed σ, the winning messages M_σ^+ and losing messages M_σ^- are determined, along with the winning probability $1 - D(\sigma) = P_\sigma(M_\sigma^+)$ and losing probability $D(\sigma)$. Narrative persuasion can improve the sender's payoff only with an alternative model under which losing messages under σ are sufficiently likely to trigger switching. Since the budget for narrative persuasion remains at $B = p^*$, the sender can rescue a subset $S \subseteq M_\sigma^-$ of losing messages as long as $P_\sigma(S) \le p^*/\alpha$. If $D(\sigma) \le p^*/\alpha$, all losing messages can be converted and the sender fully manipulates; otherwise she rescues as much losing mass as the budget allows.

Theorem 5. *Fix any exogenous messaging strategy σ. The sender-optimal equilibrium sees full manipulation if and only if $D(\sigma) \le p^*/\alpha$. Otherwise, the sender's winning probability in the sender-optimal equilibrium is given by*

$$1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : \; P_\sigma(S) \le p^*/\alpha} P_\sigma(S).$$

Theorem 5 clarifies the value of endogenous messaging.[21] Endogenous messaging

[21] A related paper is Aina (2025), which also fixes a default model and characterizes the set of posteriors that are achievable through narrative persuasion. By imposing a decision problem for the receiver, we

allows the sender to optimally choose σ and reduce $D(\sigma)$, the probability of losing types that can only win through narrative persuasion. Hence, when $D(\sigma) > D$, full manipulation is harder to achieve. At one extreme, if σ pools on a single losing message, then $D(\sigma) = 1$ and narrative persuasion is entirely ineffective. Interestingly, fixing an exogenous messaging strategy can itself be advantageous. In the baseline model, the sender's equilibrium messaging strategy must survive deviations, which under cheap talk forces babbling whenever full manipulation is infeasible. Fixing σ commits the sender to this messaging strategy, which prevents unraveling and allows partial manipulation even when full manipulation is infeasible.

5.2 Cheap Talk with Multiple Proposed Models

We now allow the sender to propose K narratives simultaneously, $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}^1, \ldots, \hat{\sigma}^K) \in \bar{\Sigma}^K$. After observing message m, the receiver computes the Bayes factor $\lambda_k(m) := P_{\hat{\sigma}^k}(m)/P_\sigma(m)$ for each proposed model, and switches to a maximizer if $\max_{k \le K} \lambda_k(m) \ge \alpha$, with ties broken in the sender's favor.

Theorem 6. *Fix $K \in \mathbb{N}$. Under cheap talk, in the sender-optimal equilibrium, the sender achieves full manipulation if and only if $KB \ge \alpha D$, or equivalently $p^* \ge \alpha/(K + \alpha)$, and cannot manipulate otherwise.*

Relative to Theorem 1, each additional proposed model contributes $B = p^*$ units to the narrative budget, expanding the total budget from B to KB. The optimal strategy generalizes the confidence trick of Section 3: the sender uses $K + 1$ messages $H', H^1, ..., H^K$, where H' is sent by high types above ω^* and never triggers model switching, and each H^k is sent by a $1/K$ fraction of the low types below ω^* and is designed to trigger switching to model $\hat{\sigma}^k$. Since each $\hat{\sigma}^k$ allocates its full budget B to H^k, the total budget across multiple models is KB. Multiple narratives thus relax the threshold for full manipulation, but do not alter the bang–bang nature of equilibrium outcomes under cheap talk.

5.3 Cheap Talk with an Initially Misspecified Model

We now relax the assumption that the receiver's default model is correctly specified, which as we will see makes the receiver more vulnerable to manipulation in most cases. We assume the receiver believes the sender holds a fixed default model $\sigma_0 \in \bar{\Sigma}$ regardless of the proposed $\hat{\sigma}$—as if he believes the sender were behavioral—and the sender knows

instead ask how often the posterior crosses a threshold to induce the risky action. This allows us to sharply characterize the value of endogenous messaging.

this default model. Let $M^+_{\sigma_0} := \{m : \mathbb{E}_{\omega \sim \pi_{\sigma_0}(\cdot|m)}(\omega) \geq \omega_0\}$ denote the winning messages under σ_0. The sender proposes $\hat{\sigma} \in \bar{\Sigma}$ and then chooses a true messaging strategy σ, neither of which needs to coincide with σ_0. After observing a message m, the receiver compares σ_0 and $\hat{\sigma}$ via the Bayes factor $\lambda(m|\sigma_0, \hat{\sigma}) = P_{\hat{\sigma}}(m)/P_{\sigma_0}(m)$ and switches if and only if $\lambda(m|\sigma_0, \hat{\sigma}) \geq \alpha$. Since model switching is now evaluated relative to σ_0 rather than the sender's true strategy, the sender can exploit messages that are either interpreted favorably under σ_0 or sufficiently unlikely under it to trigger model switching.

Theorem 7. *Fix a default model σ_0. In the sender-optimal equilibrium:*

(i) If $M^+_{\sigma_0} \neq \emptyset$, the sender achieves full manipulation.

(ii) If $M^+_{\sigma_0} = \emptyset$, then the sender achieves full manipulation if and only if there exists a message $m \in M$ with $P_{\sigma_0}(m) \leq p^*/\alpha$, and cannot manipulate otherwise.

In case (i), the receiver's default model already assigns a winning interpretation to some messages, so the sender simply pools all types on it and induces the risky action without any model switching. In case (ii), the default model is pessimistic as no message is interpreted favorably under σ_0, so manipulation must rely entirely on inducing model switching. Full manipulation is then feasible if and only if some message m has probability lower than p^*/α under σ_0. If so, the sender can propose an alternative model under which m is a good signal, and in reality sends m at every state. The rarity of m under σ_0 ensures that the receiver switches.

The result highlights a key contrast with the correctly specified benchmark. With correct beliefs, manipulation requires the global demand–budget constraint $\alpha D \leq B$ to hold. Under misspecification, this collapses to a local condition that σ_0 contains at least one winning or sufficiently rare message. This fails only for a small set of default models, such as when the receiver presumes the sender is babbling and pools all types on a single message.

5.4 Bayesian Updating over Models

We now consider an alternative to the baseline's discrete model-switching rule. Rather than switching entirely to the proposed model when the Bayes factor exceeds a threshold, the receiver assigns prior probability $\beta_0 \in (0, 1)$ to the proposed model $\hat{\sigma}$ and $1 - \beta_0$ to the correctly specified model σ, and updates these weights continuously to $\beta(m)$ by Bayes' rule after observing message m. Perhaps counterintuitively, full manipulation remains feasible even though the receiver's prior over states would lead him to take the safe action absent any message—and can even be feasible when the receiver's prior assigns more credibility to the true model than the proposed one ($\beta_0 < 1/2$).

Theorem 8. *The sender achieves full manipulation in the sender-optimal equilibrium if and only if* $\beta_0 \geq \bar{\beta}_0$*, where*

$$\bar{\beta}_0 := \frac{P(\omega < \omega^*)\Big(\omega_0 - \mathbb{E}[\omega \mid \omega < \omega^*]\Big)}{P(\omega > \omega_0)\Big(\mathbb{E}[\omega \mid \omega > \omega_0] - \omega_0\Big) + P(\omega < \omega^*)\Big(\omega_0 - \mathbb{E}[\omega \mid \omega < \omega^*]\Big)}. \tag{13}$$

Otherwise, the sender cannot manipulate.

The sender achieves full manipulation with the following strategy. She sends H' for types above ω^* under σ and never proposes H' under $\hat{\sigma}$, so that the receiver fully believes the correctly specified model after H' and correctly infers he should take the risky action. For message H, losing types below ω^* send it under σ, but under $\hat{\sigma}$ the sender proposes only the best types above ω_0 send H. Unlike in Section 3 where adding more types to H under $\hat{\sigma}$ helps exhaust the narrative budget, here adding types in $[\omega^*, \omega_0)$ under $\hat{\sigma}$ would only drag down the Bayesian-averaged posterior mean. The resulting average crosses ω_0 if and only if the aggregate upside of states above ω_0 under the proposed model is large enough, relative to the aggregate downside of states below ω^* under the true model, once weighted by the receiver's prior beliefs over the two models.

6 Conclusion

Disinformation continues to rampantly undermine political processes, economic outcomes, and public discourse in markets and societies around the world. While it is natural to expect that informed, skeptical audiences who understand a speaker's motives would be immune, our results show that even such receivers can be systematically manipulated. The source of this vulnerability is the joint control of messaging and narratives: a sender who chooses both what information to provide and suggests the lens through which it is interpreted can fully manipulate a receiver who begins with a correct understanding of the world, exploiting the very openness to alternative viewpoints that would be reasonable in non-adversarial settings.

Although this vulnerability is robust, our analysis precisely characterizes when manipulation succeeds and when it fails. The relevant conditions are in principle levers for policy design: the receiver's skepticism, the hardness of information, the flexibility-credibility balance of the technology. Our results suggest that the two most prominent countermeasures are insufficient: fact-checking addresses only the information itself while leaving the sender's narrative unaddressed, and media literacy campaigns cannot protect receivers who are already correctly specified yet still remain vulnerable. Promising directions for future work include designing effective policy interventions, such as introducing reputation concerns in repeated interactions or leveraging competition among multiple senders to discipline manipulative behavior.

References

Aina, Chiara, "Tailored Stories," Working Paper, 2025.

_ **and Florian H. Schneider**, "Weighting Models," *Mimeo*, 2025.

Akerlof, George A and Robert J Shiller, *Phishing for phools: The economics of manipulation and deception*, Princeton University Press, 2015.

Ba, Cuimin, "Robust Misspecified Models," *American Economic Review*, 2026. Forthcoming.

Barron, Kai and Tilman Fries, "Narrative persuasion," WZB Discussion Paper SP II 2023-301r, Berlin 2024. Additional information: January 2023 (revised April 2024).

Bauch, Gerrit and Manuel Foerster, "Strategic communication of narratives," *arXiv preprint arXiv:2410.23259*, 2024.

Bergemann, Dirk and Stephen Morris, "Information design: A unified perspective," *Journal of Economic Literature*, 2019, *57* (1), 44–95.

Clippel, Geoffroy De and Xu Zhang, "Non-bayesian persuasion," *Journal of Political Economy*, 2022, *130* (10), 2594–2642.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1431–1451.

Eliaz, Kfir and Ran Spiegler, "News media as suppliers of narratives (and information)," *arXiv preprint arXiv:2403.09155*, 2024.

Galperti, Simone, "Persuasion: The Art of Changing Worldviews," *American Economic Review*, March 2019, *109* (3), 996–1031.

Grossman, Sanford J, "The informational role of warranties and private disclosure about product quality," *The Journal of Law and Economics*, 1981, *24* (3), 461–483.

Hart, Sergiu, Ilan Kremer, and Motty Perry, "Evidence games: Truth and commitment," *American Economic Review*, 2017, *107* (3), 690–713.

Ichihashi, Shota and Delong Meng, "The design and interpretation of information," *Available at SSRN 3966003*, 2021.

Jain, Atulya, "Informing agents amidst biased narratives," *Job Mark. Pap., HEC Paris, Paris*, 2023.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, October 2011, *101* (6), 2590–2615.

Ko, Hyeonggyun, "Persuasion in the Long Run: When history matters," *arXiv preprint arXiv:2508.01662*, 2025.

Milgrom, Paul R, "Good news and bad news: Representation theorems and applications," *The Bell Journal of Economics*, 1981, pp. 380–391.

Rappoport, Daniel, "Evidence and skepticism in verifiable disclosure games," *Theoretical Economics*, 2025, *20* (4), 1213–1246.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

_ and _ , *Sharing Models to Interpret Data*, Harvard Business School, 2024.

Sobel, Joel, "Signaling games," in "Complex social and behavioral systems: Game theory and agent-based models," Springer, 2020, pp. 251–268.

Stiglitz, Joseph E and Andrew Kosenko, "Robust theory and fragile practice: Information in a world of disinformation Part 1: Indirect communication," 2024.

_ and _ , "Robust theory and fragile practice: Information in a world of disinformation Part 2: Direct communication," 2024.

Taleb, Nassim Nicholas, *The Black Swan: The Impact of the Highly Improbable*, New York: Random House, 2007.

Voltaire, *Œuvres Complètes de Voltaire 60D : Collection des Lettres Sur les Miracles. Ecrites a Genève, et a Neufchatel*, Voltaire Foundation, 2018. ProQuest Ebook Central.

A Proofs

A.1 Proofs for Section 3

A.1.1 Proof of Lemma 1

Proof. This immediately follows from Kamenica and Gentzkow (2011). □

A.1.2 Proof of Theorem 1

Proof. Full Manipulation. We first show that full manipulation is supported in equilibrium when $\alpha D \leq B$. We first focus on the case where ω has an atomless distribution and then allow the case where its distribution is discrete. Throughout the proof let L, H, and H' denote three distinct messages in M.

We construct the sender's strategy as follows. She proposes that she will use an optimal Bayesian Persuasion policy, $\hat{\sigma}^* = \sigma_{BP}$ where

$$\sigma_{BP}(H|\omega) = 1 \ \forall \omega \geq \omega^*, \ \ \sigma_{BP}(L|\omega) = 1 \ \forall \omega < \omega^*, \tag{14}$$

where $\omega^* \equiv \max\{\tilde{\omega} : \mathbb{E}[\omega|\omega < \tilde{\omega}] \leq \omega_0\}$ and $p^* \equiv F(\omega \geq \omega^*)$. By construction $B = p^*$. Given any $\hat{\sigma}$, let $\sigma^*(\cdot|\cdot, \hat{\sigma})$ be defined as follows.

$$\sigma^*(H|\omega, \hat{\sigma}) = 1 \ \forall \omega < \omega^*, \ \ \sigma^*(H'|\omega, \hat{\sigma}) = 1 \ \forall \omega \geq \omega^*. \tag{15}$$

Let the receiver's equilibrium action strategy be $A^*(\pi) = \mathbb{1}\{\mathbb{E}_{\omega\sim\pi}[\omega] \geq \omega_0\}$, which satisfies part (v) of Definition 1. We now consider the Receiver's behavior on the equilibrium path following $\hat{\sigma}^*$ and $\sigma^*(\cdot|\cdot, \hat{\sigma}^*)$. Message $m = H$ induces him to adopt the sender's proposed model $\sigma^*_R(H) = \hat{\sigma}^*$ since

$$\lambda(m) = \frac{F(\omega \geq \omega^*)}{F(\omega < \omega^*)} = \frac{p^*}{1 - p^*} \geq \alpha,$$

which induces him to form the following posterior belief and expectation in equilibrium[22]

$$\pi_{\sigma^*_R}(\tilde{\Omega}|H) = \frac{\int_{\tilde{\Omega}} \mathbb{1}\{\omega \geq \omega^*\} dF(\omega)}{F(\omega \geq \omega^*)} \ \forall \tilde{\Omega} \in \mathcal{B}(\Omega) \ \Rightarrow \mathbb{E}_{\omega\sim\pi_{\sigma^*_R}(\cdot|H)}[\omega] = \mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0.$$

and thus he chooses the risky action $a = 1$ after observing $m = H$ on the equilibrium path. Upon observing $m = H'$, the receiver retains his original model $\sigma^*_R(H') = \sigma^*(\cdot|\cdot, \hat{\sigma}^*)$

[22] $\mathcal{B}(\Omega)$ denotes the Borel sets of Ω.

since this message is not sent with positive probability under $\hat{\sigma}^*$, i.e.,

$$P_{\hat{\sigma}^*}(H') = 0 < P_{\sigma^*(\cdot|\cdot,\hat{\sigma}^*)}(H') \Rightarrow \lambda(H') = 0.$$

This induces him to form the following posterior belief and expectation in equilibrium

$$\pi_{\sigma_R^*}(\tilde{\Omega}|H') = \frac{\int_{\tilde{\Omega}} \mathbb{1}\{\omega \geq \omega^*\} dF(\omega)}{F(\omega \geq \omega^*)} \; \forall \tilde{\Omega} \in \mathcal{B}(\Omega) \; \Rightarrow \mathbb{E}_{\omega \sim \pi_{\sigma_R^*}(\cdot|H')}[\omega] = \mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0,$$

and thus he chooses the risky action $a = 1$ after observing $m = H'$ on the equilibrium path. Since the sender attains their maximum payoff with probability 1 after every message $m \in \{H, H'\}$ they send with positive probability, it then clearly follows that full manipulation is supported in equilibrium since for any fixed specification of the receiver's model choice and beliefs off the equilibrium path, neither the sender nor the receiver has a profitable deviation, no matter how optimistic the receiver's off-path beliefs are.

The proof in the case where the state ω is discrete is nearly identical. In this case it is possible that $\mathbb{E}[\omega|\omega \geq \omega^*] > \omega_0$. If so, σ_{BP} is then modified at the state $\omega_-^* := \max\{\omega : \omega < \omega^*\}$ just below the cutoff ω^* where $\sigma_{BP}(H|\omega_-^*) = q$ and $\sigma_{BP}(L|\omega_-^*) = 1 - q$ such that

$$\frac{\mathbb{E}[\omega|\omega > \omega_-^*] + q\omega_-^* F(\{\omega_-^*\})}{F((\omega_-^*, \bar{\omega}]) + qF(\{\omega_-^*\})} = \omega_0 \Leftrightarrow q = \frac{\mathbb{E}[\omega|\omega > \omega_-^*] - \omega_0 F(\omega > \omega_-^*)}{(\omega_0 - \omega_-^*)F(\{\omega_-^*\})}, \tag{16}$$

The actual strategy $\sigma^*(\cdot|\omega, \hat{\sigma})$ is defined as in Eq. (15), with the modification that at ω_-^* the sender sends H with probability q and H' with probability $1 - q$, where q is as above. The proof then proceeds just as before.

No Manipulation. We first show that no manipulation is possible in equilibrium when $\alpha D > B$. First, we prove that if $\alpha D > B$ then any $\hat{\sigma}$ and σ must have

$$\underbrace{\int_\Omega \sigma(m \in M_\sigma^+ \text{ and } \lambda(m) < \alpha|\omega)\, dF(\omega)}_{S_1} + \underbrace{\int_\Omega \sigma(m \in M_{\hat{\sigma}}^+ \text{ and } \lambda(m) \geq \alpha|\omega)\, dF(\omega)}_{S_2} < 1.$$

Notice that S_1 (resp. S_2) is the probability mass of messages after which the receiver does not switch (resp. does switch) models induce the receiver to weakly prefer the risky action but do not induce him to switch models. Thus, if $S_1 + S_2 < 1$, there must exist a message sent with strictly positive probability that induces the receiver to strictly prefer the safe action and result in a loss for the sender. First, $S_1 \leq 1 - D$ since

$$S_1 \leq \int_\Omega \sigma(M_\sigma^+|\omega) dF(\omega) \equiv 1 - D(\sigma) \leq 1 - D.$$

Second, $S_2 \leq B/\alpha$ since

$$S_2 \leq \frac{1}{\alpha}\int_\Omega \hat{\sigma}(M_{\hat{\sigma}}^+|\omega)dF(\omega) = \frac{B(\hat{\sigma})}{\alpha} \leq \frac{B}{\alpha}$$

where the first inequality follows from the fact that

$$\lambda(m) \geq \alpha \ \forall m \in M_{\hat{\sigma}}^+ \Leftrightarrow \underbrace{\int_\Omega \sigma(M_{\hat{\sigma}}^+|\omega)dF(\omega)}_{P_\sigma(M_{\hat{\sigma}}^+)} \leq \frac{1}{\alpha}\underbrace{\int_\Omega \hat{\sigma}(M_{\hat{\sigma}}^+|\omega)dF(\omega)}_{P_{\hat{\sigma}}(M_{\hat{\sigma}}^+)\equiv B(\hat{\sigma})}.$$

Therefore, since $\alpha \cdot (1 - B) = \alpha D > B \Leftrightarrow B < \frac{\alpha}{1+\alpha}$, it follows that

$$S_1 + S_2 \leq 1 - D + \frac{B}{\alpha} = B + \frac{B}{\alpha} = B \cdot \frac{1+\alpha}{\alpha} < 1.$$

Thus, for any $\hat{\sigma}$ and σ, the sender must lose on a strictly positive mass of states. Since under cheap talk any losing type can send any on-path winning message, for any candidate equilibrium profile under which the sender *wins* on a strictly positive mass of states, a profitable deviation must exist. Therefore, the sender must lose with probability one in equilibrium when $\alpha D > B$. That is, she only babbles, devolving into her equilibrium behavior in the traditional case without narrative persuasion.

For completeness, we construct such an equilibrium profile. Suppose the sender proposes that they pool on a single message $\hat{\sigma}^*(H|\omega) = 1 \ \forall\omega$; she also pools $\sigma^*(H|\omega, \hat{\sigma}) = 1$ after any proposal $\hat{\sigma}$ on or off the equilibrium path and any state realization $\omega \in \Omega$. Let A^* be as above. Off the equilibrium path, have the receiver never switch models and set their belief to the prior. On the equilibrium path the receiver switches models if and only if $\alpha \leq 1$, and in any case his posterior belief is equal to the prior by Bayes rule. The receiver chooses the safe action on and off the equilibrium path and no player has a profitable deviation. □

A.1.3 Proof of Corollary 1

Proof. This immediately follows from Example 1. This example shows that there exists a distribution $F \in \Delta\Omega$ that satisfies $\alpha D \leq B$ ($\Leftrightarrow p^* \leq \frac{\alpha}{1+\alpha}$) even as $\mathbb{E}_{\omega\sim F}[\omega] =\rightarrow -\infty$.

Another such distribution in the case of a continuous state has density

$$f(\omega) = \begin{cases} 0 & \text{, if } \omega \in (-\infty, -b - \frac{\varepsilon}{2}) \\ \frac{1}{(1+\alpha)\varepsilon} - \delta & \text{, if } \omega \in [-b - \frac{\varepsilon}{2}, -b + \frac{\varepsilon}{2}] \\ \frac{\delta}{b-\varepsilon/2} & \text{, if } \omega \in (-b + \frac{\varepsilon}{2}, 0) \\ \frac{\alpha}{(1+\alpha)\varepsilon} & \text{, if } \omega \in [0, \varepsilon] \\ 0 & \text{, if } \omega \in (\varepsilon, \infty) \end{cases} \tag{17}$$

where $\alpha, \varepsilon, \delta > 0$ and $-b < -\varepsilon/2$. □

A.1.4 Proof of Corollary 2

Proof. Consider the family $\{F_c\}_{c\in\mathbb{R}}$ with $f_c(\omega) = f(\omega - c)$ and support $\{\omega + c : \omega \in \Omega\}$, so $\mathbb{E}_{\omega\sim F_c}[\omega] = \mathbb{E}_{\omega\sim F}[\omega] + c$ while ω_0 is fixed. The cutoff ω_c^* satisfies $\mathbb{E}_{\omega\sim F_c}[\omega|\omega \geq \omega_c^*]$ which is equivalent to $\mathbb{E}_{\omega\sim F}[\omega|\omega \geq \omega_c^* + c] = \omega_0 + c$ whose left (resp. right) hand side is non-decreasing (resp. decreasing) in c. It then follows that ω_c^* is decreasing in c while $p_c^* = F_c([\omega_c^*, \infty))$ is increasing with $p_c^* \to 0$ as $c \to -\infty$ and $p_c^* \to 1$ as $c \to \infty$. Therefore by continuity, $\exists! c^* \in \mathbb{R}$ such that $p_{c^*}^* = \alpha/(1+\alpha)$ and $\text{sgn}(c - c^*) = \text{sgn}(p_{c^*}^* - \alpha/(1+\alpha))$.[23] Thus when $c < c^*$ the sender cannot manipulate the receiver to ever take the risky action, so the sender obtains 0 while the receiver obtains ω_0 in expectation. When $c \geq c^*$ the sender fully manipulates the receiver to take the risky action, so the sender obtains 1 while the receiver obtains $\mathbb{E}_{\omega\sim F_c}[\omega] = \mathbb{E}_{\omega\sim F}[\omega] + c$ in expectation, which increases linearly in c.

□

A.2 Proofs for Section 4

A.2.1 Proof of Lemma 2

Proof. We first show $\Omega_{\mathcal{M}}(m)$ is a closed interval. Let $\omega_a := \inf\{\omega : U(\omega) \geq m\}$ and $\omega_b := \sup\{\omega : L(\omega) \leq m\}$. Since U is right-continuous and weakly increasing, $U(\omega_a) \geq m$, so $\omega_a \in \Omega_{\mathcal{M}}(m)$. Since L is right-continuous and weakly increasing, $L(\omega_b) \leq m$, so $\omega_b \in \Omega_{\mathcal{M}}(m)$. For any $\omega \in [\omega_a, \omega_b]$, monotonicity gives $L(\omega) \leq L(\omega_b) \leq m$ and $U(\omega) \geq U(\omega_a) \geq m$, so $\omega \in \Omega_{\mathcal{M}}(m)$. Hence $\Omega_{\mathcal{M}}(m) = [\omega_a, \omega_b]$ is a closed interval.

For the second part, suppose $\omega \in [\omega_1, \omega_2] \subseteq \Omega_{\mathcal{M}}(m)$ can send m', so $m' \in [L(\omega), U(\omega)]$. Since L and U are weakly increasing, $L(\omega_1) \leq L(\omega) \leq m' \leq U(\omega) \leq U(\omega_2)$. If $m' \leq U(\omega_1)$, then $m' \in [L(\omega_1), U(\omega_1)]$ so ω_1 can send m'. Otherwise $m' > U(\omega_1)$,

[23] If F contains atoms, c^* may not be unique.

which implies $m' \geq L(\omega_2)$ since $\omega_1, \omega_2 \in \Omega_{\mathbb{M}}(m)$ means their feasible sets overlap at m, so $m' \in [L(\omega_2), U(\omega_2)]$ and ω_2 can send m'.

□

A.2.2 Proof of Proposition 1

Proof. The proof proceeds in three steps: we define a candidate optimal type cutoff ω^* and the corresponding signal structure σ^*, then show that no feasible strategy can make the sender better off in equilibrium, and finally show the interval structure.

Step 1. Define $\bar{U}_0 := \sup_{\omega<\omega_0} U(\omega)$. For any candidate cutoff $\tilde{\omega} \in \Omega$, we construct a *saving strategy* $\sigma_{\tilde{\omega}} : [\omega_0, \bar{\omega}(\tilde{\omega})) \to [U(\tilde{\omega}), \bar{U}_0)$ as follows, where $\bar{\omega}(\tilde{\omega})$ is defined implicitly below. We discuss strictly increasing and flat segments of U separately.

On any interval $[\omega_a, \omega_b) \subset [\tilde{\omega}, \omega_0)$ where U is strictly increasing, each message $m \in [U(\omega_a), U(\omega_b))$ is sent by exactly one low type $\omega_1(m) := U^{-1}(m) \in [\omega_a, \omega_b)$ and one high type $\omega_2(m) := \sigma_{\tilde{\omega}}^{-1}(m) \in [\omega_0, \bar{\omega}(\tilde{\omega}))$. The posterior mean condition $\mathbb{E}[\omega|m] = \omega_0$ requires:

$$\frac{\omega_1(m) \cdot \frac{f(\omega_1(m))}{U'(\omega_1(m))} + \omega_2(m) \cdot \frac{f(\omega_2(m))}{\sigma_{\tilde{\omega}}'(\omega_2(m))}}{\frac{f(\omega_1(m))}{U'(\omega_1(m))} + \frac{f(\omega_2(m))}{\sigma_{\tilde{\omega}}'(\omega_2(m))}} = \omega_0. \tag{18}$$

Rearranging (18) and substituting $\omega = \omega_2(m)$, so that $m = \sigma_{\tilde{\omega}}(\omega)$ and $\omega_1(m) = U^{-1}(\sigma_{\tilde{\omega}}(\omega))$, yields the following first-order ODE for $\sigma_{\tilde{\omega}}$ on the corresponding subinterval of $[\omega_0, \bar{\omega}(\tilde{\omega})]$:

$$\sigma_{\tilde{\omega}}'(\omega) = \frac{(\omega - \omega_0) f(\omega) U'(U^{-1}(\sigma_{\tilde{\omega}}(\omega)))}{(\omega_0 - U^{-1}(\sigma_{\tilde{\omega}}(\omega))) f(U^{-1}(\sigma_{\tilde{\omega}}(\omega)))}. \tag{19}$$

By the piecewise C^1 assumption, the right-hand side of (19) is locally Lipschitz in $\sigma_{\tilde{\omega}}(\omega)$ on each smooth piece, so a unique solution exists locally by the Picard–Lindelöf theorem.

On any interval $[\omega_a, \omega_b) \subset [\tilde{\omega}, \omega_0)$ where U is flat, i.e., $U(\omega) = \bar{m}$ for all $\omega \in [\omega_a, \omega_b)$, all types in $[\omega_a, \omega_b)$ send the same message $\bar{m}$. In this case the posterior mean condition becomes an algebraic equation: there exists a unique interval $[\omega_2^a, \omega_2^b) \subset [\omega_0, \bar{\omega}(\tilde{\omega}))$ of high types pooling on $\bar{m}$, determined by:

$$\frac{\int_{\omega_a}^{\omega_b} \omega f(\omega)\, d\omega + \int_{\omega_2^a}^{\omega_2^b} \omega f(\omega)\, d\omega}{\int_{\omega_a}^{\omega_b} f(\omega)\, d\omega + \int_{\omega_2^a}^{\omega_2^b} f(\omega)\, d\omega} = \omega_0. \tag{20}$$

Equation (20) has a unique solution $[\omega_2^a, \omega_2^b)$ given the mass of low types $\int_{\omega_a}^{\omega_b} f(\omega)\, d\omega$, since the left-hand side is strictly increasing in ω_2^b given ω_2^a. On this interval, $\sigma_{\tilde{\omega}}(\omega) = \bar{m}$ for all $\omega \in [\omega_2^a, \omega_2^b)$.

The piecewise C^1 assumption implies that $[\tilde{\omega}, \omega_0)$ can be partitioned into finitely many half-open pieces $[\omega_a^k, \omega_b^k)$, $k = 1, \ldots, K$, each either strictly increasing or flat, with $\omega_a^1 = \tilde{\omega}$ and $\omega_b^K = \omega_0$. For each piece k, let $[\alpha^k, \beta^k] \subset [\omega_0, \bar{\omega}(\tilde{\omega})]$ denote the interval of high types saving piece k, with $\alpha^1 = \omega_0$ and $\alpha^k = \beta^{k-1}$ for $k \geq 2$ ensuring continuity. The initial condition for each piece is $\sigma_{\tilde{\omega}}(\alpha^k) = U(\omega_a^k)$, with β^k determined by $\sigma_{\tilde{\omega}}(\beta^k) = U(\omega_b^k)$. Each piece is solved using (19) or (20) as appropriate. The upper boundary is then $\bar{\omega}(\tilde{\omega}) := \beta^K$, which satisfies $\lim_{\omega \to^+ \bar{\omega}(\tilde{\omega})} \sigma_{\tilde{\omega}}(\omega) = \lim_{\omega \to^+ \omega_b^K} U(\omega) = \bar{U}_0$ by definition of $\bar{U}_0$.

The saving strategy $\sigma_{\tilde{\omega}}$ is feasible only if $\sigma_{\tilde{\omega}}(\omega) \geq L(\omega)$ for all $\omega \in [\omega_0, \bar{\omega}(\tilde{\omega})]$. As $\tilde{\omega}$ approaches $\sup \Omega$, the interval $[\tilde{\omega}, \omega_0)$ becomes empty, no pooling is required, and feasibility holds trivially. By the piecewise C^1 assumption, the solution $\sigma_{\tilde{\omega}}(\omega)$ varies continuously in $\tilde{\omega}$ by the continuous dependence of ODE solutions on initial conditions (applied piecewise and joined at boundary points), so the feasibility constraint $\sigma_{\tilde{\omega}}(\omega) \geq L(\omega)$ varies continuously in $\tilde{\omega}$. Therefore define:

$$\omega_{\mathcal{M}}^* := \inf \left\{ \tilde{\omega} \in \Omega : \sigma_{\tilde{\omega}}(\omega) \geq L(\omega) \; \forall \, \omega \in [\omega_0, \bar{\omega}(\tilde{\omega})] \right\},$$

which is well-defined and attained by continuity.

Given $\omega_{\mathcal{M}}^*$, define signal structure σ^* by:

(i) $\sigma^*(\omega) = \omega$ for all $\omega < \omega_{\mathcal{M}}^*$.

(ii) $\sigma^*(\omega) = U(\omega)$ for all $\omega \in [\omega_{\mathcal{M}}^*, \omega_0)$.

(iii) $\sigma^*(\omega) = \sigma_{\omega_{\mathcal{M}}^*}(\omega)$ for all $\omega \in [\omega_0, \bar{\omega})$, where $\bar{\omega} := \bar{\omega}(\omega_{\mathcal{M}}^*)$.

(iv) $\sigma^*(\omega) = U(\omega)$ for all $\omega \geq \bar{\omega}$.

By construction σ^* is feasible and all types $\omega \geq \omega_{\mathcal{M}}^*$ win under σ^*.

Step 2. We show that no feasible $\sigma' \in \Sigma_{\mathcal{M}}$ can make all types in $[\tilde{\omega}, \sup \Omega]$ win for any $\tilde{\omega} < \omega_{\mathcal{M}}^*$. Suppose for contradiction that there exists $\tilde{\omega} < \omega_{\mathcal{M}}^*$ and a feasible $\sigma' \in \Sigma_{\mathcal{M}}$ such that all types $\omega \geq \tilde{\omega}$ win under σ'. Since $\tilde{\omega} < \omega_{\mathcal{M}}^*$, the saving strategy $\sigma_{\tilde{\omega}}$ is infeasible, so there exists $\hat{\omega} \in [\tilde{\omega}, \omega_0)$ such that $U(\hat{\omega}) < L(\omega_2(U(\hat{\omega})))$, where $\omega_2(U(\hat{\omega}))$ is the high type paired with $U(\hat{\omega})$ by the ODE. Define the deficit of $[\tilde{\omega}, \hat{\omega}]$ and surplus of $[\omega_0, \omega_2(U(\hat{\omega}))]$ as:

$$\mathcal{D} := \int_{\tilde{\omega}}^{\hat{\omega}} (\omega_0 - \omega) f(\omega) \, d\omega, \qquad \mathcal{S} := \int_{\omega_0}^{\omega_2(U(\hat{\omega}))} (\omega - \omega_0) f(\omega) \, d\omega.$$

Since the ODE satisfies (18) with equality at every $m \in [U(\tilde{\omega}), U(\hat{\omega})]$, integrating over m gives $\mathcal{S} = \mathcal{D}$.

By monotonicity of L, every $\omega > \omega_2(U(\hat{\omega}))$ satisfies $L(\omega) \geq L(\omega_2(U(\hat{\omega}))) > U(\hat{\omega})$ and cannot send any message in $[U(\tilde{\omega}), U(\hat{\omega})]$. Therefore the feasible saver types for messages in $[U(\tilde{\omega}), U(\hat{\omega})]$ under any feasible strategy are restricted to $\{\omega \in [\omega_0, \omega_2(U(\hat{\omega}))] : L(\omega) \leq U(\hat{\omega})\}$, a strict subset of $[\omega_0, \omega_2(U(\hat{\omega}))]$, so the total feasible surplus satisfies:

$$\mathcal{S}' := \int_{\{\omega \in [\omega_0, \omega_2(U(\hat{\omega}))] : L(\omega) \leq U(\hat{\omega})\}} (\omega - \omega_0) f(\omega)\, d\omega < \mathcal{S} = \mathcal{D}. \tag{21}$$

But for σ' to make all types in $[\tilde{\omega}, \hat{\omega}]$ win, each low type $\omega \in [\tilde{\omega}, \hat{\omega}]$ must send some message $m(\omega) \leq U(\omega) \leq U(\hat{\omega})$ and pool with feasible saver types to achieve $\mathbb{E}[\omega | m(\omega)] \geq \omega_0$. Integrating over all $\omega \in [\tilde{\omega}, \hat{\omega}]$ and using the fact that saver types contributing to messages in $[U(\tilde{\omega}), U(\hat{\omega})]$ must come from $\mathcal{S}'$ gives $\mathcal{S}' \geq \mathcal{D}$, contradicting (21). Therefore no feasible $\sigma' \in \Sigma_{\mathcal{M}}$ achieves a winning probability strictly greater than $p^*_{\mathcal{M}} = F(\omega \geq \omega^*_{\mathcal{M}})$, and σ^* is optimal. When F is atomless, any signal structure that coincides with σ^* almost everywhere is also optimal.

Step 3. It remains to show that any optimal solution must have the cutoff structure. Suppose σ' is optimal with winning probability $p^*_{\mathcal{M}}$, but there exist positive-measure sets $A \subset [\omega^*_{\mathcal{M}}, \sup \Omega)$ that lose and $B \subset (-\infty, \omega^*_{\mathcal{M}})$ that win under σ', with $F(A) = F(B) > 0$. For any winning message m sent by types in B under σ', the posterior mean condition requires:

$$\frac{\int_B \omega \cdot \sigma'(m|\omega) f(\omega)\, d\omega + \int_{H(m)} \omega \cdot \sigma'(m|\omega) f(\omega)\, d\omega}{\int_B \sigma'(m|\omega) f(\omega)\, d\omega + \int_{H(m)} \sigma'(m|\omega) f(\omega)\, d\omega} = \omega_0,$$

where $H(m)$ denotes the high types sending m under σ'. Since $\omega > \omega^*_{\mathcal{M}} > \omega_0$ for all $\omega \in A$ and $\omega < \omega^*_{\mathcal{M}} < \omega_0$ for all $\omega \in B$, replacing B with A in message m gives:

$$\frac{\int_A \omega \cdot \sigma'(m|\omega) f(\omega)\, d\omega + \int_{H(m)} \omega \cdot \sigma'(m|\omega) f(\omega)\, d\omega}{\int_A \sigma'(m|\omega) f(\omega)\, d\omega + \int_{H(m)} \sigma'(m|\omega) f(\omega)\, d\omega} > \omega_0,$$

so the high types in $H(m)$ are no longer fully needed to cover message m. The freed high types satisfy $U(\omega) \geq U(\omega_0) > \bar{U}_0 \geq U(\omega')$ for any $\omega' \in [\omega^*_{\mathcal{M}}, \omega_0)$ by monotonicity of U, so they can send messages in $[U(\omega^*_{\mathcal{M}}), \bar{U}_0]$ that are feasible for low types in $[\omega^*_{\mathcal{M}}, \omega_0)$ and distinct from messages sent by types in A. By the surplus-deficit accounting in Step 1, the freed surplus $\int_{H(m)} (\omega - \omega_0) f(\omega)\, d\omega > 0$ is sufficient to cover a positive-measure set of additional losing types in $(-\infty, \omega^*_{\mathcal{M}})$ via the saving strategy $\sigma_{\omega^*_{\mathcal{M}}}$, strictly increasing the winning probability above $p^*_{\mathcal{M}}$. Therefore almost all types above $\omega^*_{\mathcal{M}}$ win and almost all types below lose in any optimal solution.

□

A.2.3 Proof of Proposition 2

Proof. We first handle the case where no credible type at or above $\omega^*_{\mathcal{M}}$ exists, in which case babbling is the unique equilibrium. Any winning message m must be sent by some type $\omega \geq \omega^*_{\mathcal{M}}$, but since no such type is credible, every winning message is available to some lower type who would profitably deviate to it. Hence no informative equilibrium exists and the sender's winning probability is 0. Now suppose $\tilde{\omega}^*_{\mathcal{M}}$ exists. We construct an equilibrium and show it is sender-optimal. We consider two cases.

Case 1: $\tilde{\omega}^*_{\mathcal{M}} < \omega_0$. Take the signal structure $\sigma_{\omega^*_{\mathcal{M}}}$ constructed in the proof of Proposition 1 and modify it as follows. Any type ω that previously sent a message $m \geq U(\tilde{\omega}^*_{\mathcal{M}})$ under $\sigma_{\omega^*_{\mathcal{M}}}$ continues to send the same message. Any type ω that previously sent a message $m < U(\tilde{\omega}^*_{\mathcal{M}})$ under $\sigma_{\omega^*_{\mathcal{M}}}$ instead sends $U(\omega)$. Call this modified strategy σ^*.

By construction, types $\omega \in [\tilde{\omega}^*_{\mathcal{M}}, \omega_0)$ send $U(\omega) \geq U(\tilde{\omega}^*_{\mathcal{M}})$ and remain in the same pools that include the same high types they were pooling with as under $\sigma_{\omega^*_{\mathcal{M}}}$, so the posterior mean is at leastω_0 and they win. All types $\omega \geq \omega_0$ that previously sent messages $m \geq U(\tilde{\omega}^*_{\mathcal{M}})$ continue to win as before. Types $\omega \geq \omega_0$ that previously sent messages $m < U(\tilde{\omega}^*_{\mathcal{M}})$ now send $U(\omega) > \bar{U}_0$, a message sent only by types above ω_0, so $\mathbb{E}[\omega | U(\omega)] > \omega_0$ and they win outright. Thus all types $\omega \geq \tilde{\omega}^*_{\mathcal{M}}$ win, yielding winning probability $F(\omega \geq \tilde{\omega}^*_{\mathcal{M}})$.

No type $\omega < \tilde{\omega}^*_{\mathcal{M}}$ has a profitable deviation: the lowest message used by any winning type under σ^* is $U(\tilde{\omega}^*_{\mathcal{M}})$, and by credibility of $\tilde{\omega}^*_{\mathcal{M}}$, $U(\omega) < U(\tilde{\omega}^*_{\mathcal{M}})$ for all $\omega < \tilde{\omega}^*_{\mathcal{M}}$, so no winning message lies in the feasible set of any losing type.

Case 2: $\tilde{\omega}^*_{\mathcal{M}} \geq \omega_0$. Define σ^* as follows: (i) types $\omega < \tilde{\omega}^*_{\mathcal{M}}$: send ω and lose; (ii) Types $\omega \geq \tilde{\omega}^*_{\mathcal{M}}$: send $U(\omega)$. Since $\tilde{\omega}^*_{\mathcal{M}} \geq \omega_0$, all winning types have $\omega \geq \omega_0$, so $\mathbb{E}[\omega | U(\omega)] \geq \omega_0$ and the receiver takes the risky action. By credibility, types below $\tilde{\omega}^*_{\mathcal{M}}$ cannot send $U(\tilde{\omega}^*_{\mathcal{M}})$ or higher, so no profitable deviation exists.

In both Case 1 and Case 2, the winning probability is $F(\omega \geq \tilde{\omega}^*_{\mathcal{M}})$. No equilibrium can achieve strictly higher: any equilibrium with cutoff $\hat{\omega} > \tilde{\omega}^*_{\mathcal{M}}$ would require $\hat{\omega}$ to be credible at or above $\omega^*_{\mathcal{M}}$, contradicting the definition of $\tilde{\omega}^*_{\mathcal{M}}$ as the lowest such type. Any cutoff $\hat{\omega} < \omega^*_{\mathcal{M}}$ would require winning probability $F(\omega \geq \hat{\omega}) > p^*_{\mathcal{M}}$, contradicting Proposition 1. □

A.2.4 Statement and Proof of Lemma 3

Lemma 3. *Under noisy evidence $\mathcal{M}_\epsilon$ with uniform prior on $\Omega = [0, 1]$, the constrained BP cutoff is $\omega^*_\epsilon = \max\{\omega_0 - \epsilon\sqrt{2},\ \omega^*\}$, the maximum winning probability is $p^*_\epsilon = \min\{1 - \omega_0 + \epsilon\sqrt{2},\ p^*\}$, and $\tilde{\omega}^*_\epsilon = \omega^*_\epsilon$ if $\omega^* \leq 1 - \epsilon$, and does not exist if $\omega^* > 1 - \epsilon$.*

Proof. Assume $\omega^*_\epsilon > \epsilon$ and $\bar{\omega} < 1 - \epsilon$ so boundary effects do not bind; we verify this ex post. The saving strategy $\sigma(\omega)$ satisfies the ODE from Section A.2.2, which with $f(\omega) = 1$, $U'(\omega) = 1$, and $U^{-1}(\sigma(\omega)) = \sigma(\omega) - \epsilon$ reduces to:

$$\sigma'(\omega) = \frac{\omega - \omega_0}{\omega_0 + \epsilon - \sigma(\omega)}.$$

Separating variables and integrating gives $(\sigma(\omega) - (\omega_0 + \epsilon))^2 + (\omega - \omega_0)^2 = R^2$, a circle of radius R centered at $(\omega_0, \omega_0 + \epsilon)$. Applying the initial condition $\sigma(\omega_0) = \omega_\epsilon^* + \epsilon$ gives $R^2 = (\omega_\epsilon^* - \omega_0)^2$, and taking the lower branch:

$$\sigma(\omega) = \omega_0 + \epsilon - \sqrt{(\omega_\epsilon^* - \omega_0)^2 - (\omega - \omega_0)^2}.$$

The strategy reaches $\bar{U}_0 = \omega_0 + \epsilon$ when the square root vanishes, giving $\bar{\omega} = 2\omega_0 - \omega_\epsilon^*$ by symmetry. The feasibility constraint $\sigma(\omega) \geq \omega - \epsilon$ requires:

$$(\omega_\epsilon^* - \omega_0)^2 \leq (\omega_0 + 2\epsilon - \omega)^2 - (\omega - \omega_0)^2 = 4\epsilon(\omega_0 + \epsilon - \omega),$$

which is tightest at $\omega = \omega_0 + \epsilon$ (noting $\bar{\omega} \geq \omega_0 + \epsilon$ since $\epsilon(\sqrt{2} - 1) \geq 0$), giving $(\omega_\epsilon^* - \omega_0)^2 \leq 2\epsilon^2$. The lowest feasible cutoff is thus $\omega_\epsilon^* = \omega_0 - \epsilon\sqrt{2}$, provided $\omega_0 - \epsilon\sqrt{2} \geq \omega^*$; otherwise the technology achieves the cheap talk ceiling and $\omega_\epsilon^* = \omega^*$. Hence $\omega_\epsilon^* = \max\{\omega_0 - \epsilon\sqrt{2}, \omega^*\}$ and $p_\epsilon^* = \min\{1 - \omega_0 + \epsilon\sqrt{2}, p^*\}$.

For credibility, since $U(\omega) = \min(\omega + \epsilon, 1)$, every type $\omega \leq 1 - \epsilon$ satisfies $U(\omega) = \omega + \epsilon > U(\omega') = \omega' + \epsilon$ for all $\omega' < \omega$, and is therefore credible. Types $\omega > 1 - \epsilon$ have $U(\omega) = 1 = U(1 - \epsilon)$ and are not credible. Thus $\tilde{\omega}_\epsilon^* = \omega_\epsilon^*$ when $\omega^* \leq 1 - \epsilon$, and does not exist when $\omega^* > 1 - \epsilon$.

Finally, $\omega_\epsilon^* \geq \omega^* > \epsilon$ and $\bar{\omega} = 2\omega_0 - \omega_\epsilon^* \leq 2\omega_0 - \omega^* < 1 - \epsilon$ for ϵ small enough that $\omega^* > \epsilon$ and $2\omega_0 - \omega^* < 1 - \epsilon$, confirming the boundary assumption. □

A.2.5 Proof of Theorem 2

Proof. We first construct an equilibrium achieving full manipulation when $B \geq \alpha D_{\mathcal{M}}$, then show full manipulation is impossible when $B < \alpha D_{\mathcal{M}}$.

Sufficiency. Let σ^* be the constrained BP solution from Proposition 1: all types $\omega \geq \omega_{\mathcal{M}}^*$ send winning messages and all types $\omega < \omega_{\mathcal{M}}^*$ send losing messages that equal to their true type, with $P_{\sigma^*}(M^-) = D_{\mathcal{M}}$ where $M^- = \{\omega \in \Omega : \omega \leq \omega_{\mathcal{M}}^*\}$ denotes the support of losing messages under σ^*.

Construct the proposed model $\hat{\sigma}^*$ as follows. For each $m \in M^-$, set

$$P_{\hat{\sigma}^*}(m) = \frac{p^*}{D_{\mathcal{M}}} \cdot P_{\sigma^*}(m),$$

and set $P_{\hat{\sigma}^*}(m) = 0$ for every winning message $m \notin M^-$. This is achieved by having all types $\omega \geq \omega^*$ randomize over M^- with probability proportional to P_{σ^*}, using up the full budget p^*. In addition, have all types $\omega < \omega^*$ send a single message $m' \in M \setminus \Omega$. This is feasible since $\hat{\sigma}^* \in \bar{\Sigma}$ is unconstrained and M has cardinality at least $|\Omega| + 2$.

We verify that every on-path message induces the risky action. First, losing messages $m \in M^-$ induce the risky action by triggering a switch. By construction, $\lambda(m) = P_{\hat{\sigma}^*}(m)/P_{\sigma^*}(m) = B/D_{\mathcal{M}} \geq \alpha$, so the receiver switches to $\hat{\sigma}^*$. Since under $\hat{\sigma}^*$ every high type $\omega \geq \omega^*$ sends each $m \in M^-$ with the same relative probability, the posterior under $\hat{\sigma}^*$ after any $m \in M^-$ is the distribution of ω conditional on $\omega \geq \omega^*$. Therefore

$$\mathbb{E}_{\hat{\sigma}^*}[\omega|m] = \mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0.$$

Hence the receiver takes the risky action after any $m \in M^-$. Second, winning messages $m \notin M^-$ induce the risky action by revealing the default model is correct. Since $P_{\hat{\sigma}^*}(m) = 0$, we have $\lambda(m) = 0 < \alpha$, so the receiver retains σ^*. Under σ^*, any $m \notin M^-$ is a winning message, so $\mathbb{E}_{\sigma^*}[\omega|m] \geq \omega_0$ by construction of the constrained BP solution. Hence the receiver takes the risky action.

Since every on-path message induces the risky action, the sender attains payoff 1 with probability 1. No type has a profitable deviation because any off-path message can be assigned beliefs under which the receiver takes the safe action.

Necessity. Suppose $B < \alpha D_{\mathcal{M}}$. Consider any proposed model $\hat{\sigma} \in \bar{\Sigma}$ and any feasible messaging strategy $\sigma \in \Sigma$. We show that

$$\underbrace{\int_\Omega \sigma(m \in M_\sigma^+ \text{ and } \lambda(m) < \alpha|\omega)\, dF(\omega)}_{S_1} + \underbrace{\int_\Omega \sigma(m \in M_{\hat{\sigma}}^+ \text{ and } \lambda(m) \geq \alpha|\omega)\, dF(\omega)}_{S_2} < 1, \tag{22}$$

which implies full manipulation is impossible since $S_1 + S_2 < 1$ means some message is sent with positive probability that induces the safe action. First, $S_1 \leq 1 - D_{\mathcal{M}}$ since

$$S_1 \leq \int_\Omega \sigma(M_\sigma^+|\omega)\, dF(\omega) := 1 - D(\sigma) \leq 1 - D_{\mathcal{M}},$$

where the last inequality uses the fact that $\sigma \in \Sigma$ is feasible under $\mathcal{M}$, so Proposition 1 bounds its winning probability by $p_{\mathcal{M}}^* = 1 - D_{\mathcal{M}}$. Second, $S_2 \leq B/\alpha$ since

$$S_2 \leq \frac{1}{\alpha} \int_\Omega \hat{\sigma}(M_{\hat{\sigma}}^+|\omega)\, dF(\omega) := \frac{B(\hat{\sigma})}{\alpha} \leq \frac{B}{\alpha},$$

where the first inequality follows from

$$\lambda(m) \geq \alpha \ \forall m \in M_{\hat{\sigma}}^{+} \Leftrightarrow \underbrace{\int_{\Omega} \sigma(M_{\hat{\sigma}}^{+}|\omega)\, dF(\omega)}_{P_{\sigma}(M_{\hat{\sigma}}^{+})} \leq \frac{1}{\alpha} \underbrace{\int_{\Omega} \hat{\sigma}(M_{\hat{\sigma}}^{+}|\omega)\, dF(\omega)}_{P_{\hat{\sigma}}(M_{\hat{\sigma}}^{+}):=B(\hat{\sigma})},$$

and the last inequality uses the fact that $\hat{\sigma} \in \bar{\Sigma}$ is unconstrained, so $B(\hat{\sigma}) \leq p^* = B$ by Lemma 1. Therefore

$$S_1 + S_2 \leq 1 - D_{\mathcal{M}} + \frac{B}{\alpha} = p_{\mathcal{M}}^* + \frac{p^*}{\alpha} < 1,$$

since $B < \alpha D_{\mathcal{M}}$. Hence $S_1 + S_2 < 1$ for any $(\hat{\sigma}, \sigma)$, so full manipulation is impossible. □

A.2.6 Proof of Theorem 3

Proof. Part (i): $\omega_{\mathcal{M}}^{**} \notin \bar{\Omega}_{\mathcal{M}}$. We first show that an equilibrium exists such that types $\omega \geq \omega_{\mathcal{M}}^{**}$ win. Since $\mathcal{M} \in \mathbb{M}^*$, U is strictly increasing, so $U(\omega_{\mathcal{M}}^{**}) > \omega_{\mathcal{M}}^{**}$ and types in $[\omega_{\mathcal{M}}^{**}, \omega_{\mathcal{M}}^{*})$ can send any feasible message in $[\omega_{\mathcal{M}}^{**}, U(\omega_{\mathcal{M}}^{*}))$, which is non-empty since $\omega_{\mathcal{M}}^{**} < \omega_{\mathcal{M}}^{*} \leq U(\omega_{\mathcal{M}}^{*})$. Define σ^* as follows. For types $\omega \geq \omega_{\mathcal{M}}^{*}$, let σ^* be the constrained BP solution from Proposition 1. For types $\omega < \omega_{\mathcal{M}}^{**}$, let them send their true type ω; collect these messages in $M^{\text{lose}} := \{m : m < \omega_{\mathcal{M}}^{**}\}$. For types $\omega \in [\omega_{\mathcal{M}}^{**}, \omega_{\mathcal{M}}^{*})$, let them send any feasible message in $[\omega_{\mathcal{M}}^{**}, U(\omega_{\mathcal{M}}^{*}))$; denote the resulting message pool by M^-, so $P_{\sigma^*}(M^-) = D_{\mathcal{M}} - F(\omega < \omega_{\mathcal{M}}^{**})$. Note that M^-, the winning messages, and the messages sent by types below $\omega_{\mathcal{M}}^{**}$ are pairwise disjoint.

Construct $\hat{\sigma}^*$ as follows. For each $m \in M^-$, set

$$P_{\hat{\sigma}^*}(m) = \frac{p^*}{D_{\mathcal{M}} - F(\omega < \omega_{\mathcal{M}}^{**})} \cdot P_{\sigma^*}(m),$$

and set $P_{\hat{\sigma}^*}(m) = 0$ for every other message. This is achieved by having all types $\omega \geq \omega^*$ randomize over M^- with probability proportional to P_{σ^*}, using up the full budget p^*, and having all types $\omega < \omega^*$ send a single message $m' \in M \setminus \Omega$. This is feasible since $\hat{\sigma}^* \in \bar{\Sigma}$ is unconstrained and M has cardinality at least $|\Omega| + 2$.

We verify that every on-path message induces the correct action.

Messages $m \in M^-$. By construction,

$$\lambda(m) = \frac{P_{\hat{\sigma}^*}(m)}{P_{\sigma^*}(m)} = \frac{p^*}{D_{\mathcal{M}} - F(\omega < \omega_{\mathcal{M}}^{**})} \geq \alpha,$$

where the inequality follows from Eq. (12). Hence the receiver switches to $\hat{\sigma}^*$. Since every type $\omega \geq \omega^*$ sends each $m \in M^-$ with the same relative probability under $\hat{\sigma}^*$, the posterior after any $m \in M^-$ is the distribution of ω conditional on $\omega \geq \omega^*$, so

$$\mathbb{E}[\omega|m, \hat{\sigma}^*] = \mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0,$$

and the receiver takes the risky action.

Winning messages $m \notin M^- \cup M^{lose}$. Since $P_{\hat{\sigma}^*}(m) = 0$, we have $\lambda(m) = 0 < \alpha$, so the receiver retains σ^*. These are winning messages by the constrained BP solution, so $\mathbb{E}[\omega|m, \sigma^*] \geq \omega_0$ and the receiver takes the risky action.

Messages $m \in M^{lose}$. Since $P_{\hat{\sigma}^*}(m) = 0$, we have $\lambda(m) = 0 < \alpha$, so the receiver retains σ^*. Types below $\omega^{**}_{\mathcal{M}}$ send their true type, as in the constrained BP solution. Since $\omega^{**}_{\mathcal{M}} \leq \omega^*_{\mathcal{M}} \leq \omega^*$, these types are a subset of the losing types in the constrained BP solution, and removing types in $[\omega^{**}_{\mathcal{M}}, \omega^*_{\mathcal{M}})$ from their pools only lowers the posterior mean, so $\mathbb{E}[\omega|m, \sigma^*] < \omega_0$ and the receiver takes the safe action.

All types $\omega \geq \omega^{**}_{\mathcal{M}}$ win. No type $\omega < \omega^{**}_{\mathcal{M}}$ has a profitable deviation: since U is strictly increasing, $U(\omega) < U(\omega^{**}_{\mathcal{M}})$ for all $\omega < \omega^{**}_{\mathcal{M}}$, so no type below $\omega^{**}_{\mathcal{M}}$ can send any message in M^- or any winning message.

We now show that the constructed equilibrium is sender-optimal. Consider any proposed model $\hat{\sigma} \in \bar{\Sigma}$ and any feasible $\sigma \in \Sigma$. Define S_1 and S_2 as in the proof of Theorem 2 (Eq. (22)). Since $\sigma \in \Sigma$ is feasible, $S_1 \leq p^*_{\mathcal{M}} = 1 - D_{\mathcal{M}}$. Since $\hat{\sigma} \in \bar{\Sigma}$ is unconstrained, $S_2 \leq p^*/\alpha = B/\alpha$. Therefore

$$S_1 + S_2 \leq 1 - D_{\mathcal{M}} + \frac{B}{\alpha} = 1 - D_{\mathcal{M}} + D_{\mathcal{M}} - F(\omega < \omega^{**}_{\mathcal{M}}) = F(\omega \geq \omega^{**}_{\mathcal{M}}),$$

where the second equality uses Eq. (12). Since the construction achieves exactly $F(\omega \geq \omega^{**}_{\mathcal{M}})$, it is sender-optimal.

Part (ii): $\omega^{**}_{\mathcal{M}} \in \bar{\Omega}_{\mathcal{M}}$. In this case Ω must be bounded above since otherwise $\bar{\Omega}_{\mathcal{M}} = \emptyset$. We first show $S_1 = 0$ in any sender-optimal equilibrium $(\hat{\sigma}^*, \sigma^*)$.

Since $\mathcal{M} \in \mathbb{M}^*$, L is strictly increasing, so for any $\omega, \omega' \in \bar{\Omega}_{\mathcal{M}}$ with $\omega' < \omega$, $M(\omega) \subseteq M(\omega')$ since $U(\omega) = U(\omega') \geq \sup \Omega$ and $L(\omega) \geq L(\omega')$. Therefore if any type $\omega^\dagger \in \bar{\Omega}_{\mathcal{M}}$ loses, all types above $\omega^\dagger$ in $\bar{\Omega}_{\mathcal{M}}$ must also lose since they cannot access any winning message that $\omega^\dagger$ cannot. Let $\omega^\dagger$ be the infimum of losing types in $\bar{\Omega}_{\mathcal{M}}$, which exists since $S_1 + S_2 \leq 1 - D_{\mathcal{M}} + B/\alpha < 1$ (where the strict inequality uses $B < \alpha D_{\mathcal{M}}$) implies not all types win.

Suppose for contradiction that some type $\omega' \in \bar{\Omega}_{\mathcal{M}}$ wins via pooling on some message m under $(\hat{\sigma}^*, \sigma^*)$. By the monotonicity of feasible message sets, all types $\omega'' \in \bar{\Omega}_{\mathcal{M}}$ with $\omega'' < \omega'$ can also send m and must also win. The pool on m must contain some type above ω_0 for the posterior mean to reach ω_0. Since ω' wins while all types above $\omega^\dagger$ lose, we have $\omega_0 < \omega^\dagger$. Now construct $(\hat{\sigma}', \sigma')$ coinciding with $(\hat{\sigma}^*, \sigma^*)$ everywhere except that σ' has all losing types in $\bar{\Omega}_{\mathbb{M}}$ pool on $m' = \max \Omega$. Since $\omega_0 < \omega^\dagger \leq \max \Omega$, these types have posterior mean above ω_0 and win outright. Note that $m' = \max \Omega$ must have been a losing message under $(\hat{\sigma}^*, \sigma^*)$ since types above $\omega^\dagger$ lose. If $\hat{\sigma}^*$ assigns positive probability to m', replace it with some message outside Ω in $\hat{\sigma}'$. This construction is a valid equilibrium where the set of winning types strictly expands while no losing type outside $\bar{\Omega}_{\mathbb{M}}$ can send $\max \Omega$, contradicting sender-optimality. Hence no type in $\bar{\Omega}_{\mathbb{M}}$ wins via pooling.

Now suppose all types in $\bar{\Omega}_{\mathbb{M}}$ do not win via pooling while some type $\hat{\omega} \notin \bar{\Omega}_{\mathbb{M}}$ wins via pooling. The pool must contain some type above ω_0, so $\omega_0 < \inf \bar{\Omega}_{\mathbb{M}}$. By the same argument, all losing types in $\bar{\Omega}_{\mathbb{M}}$ satisfy $\omega > \omega_0$ and can pool on $\max \Omega$, a message not previously used by any winning type. Constructing $(\hat{\sigma}', \sigma')$ where these types pool on $\max \Omega$ and replacing any use of $\max \Omega$ in $\hat{\sigma}^*$ with a message outside Ω strictly increases the sender's payoff, contradicting sender-optimality. Hence $S_1 = 0$ in any sender-optimal equilibrium.

Sub-case (a): We show that an equilibrium exists achieving winning probability B/α when $B \geq \alpha F(\Omega_{\mathbb{M}}(m^*))$ for some $m^* \in \Omega$. Since $L(\omega) \leq \omega \leq U(\omega)$ for all $\omega \in \Omega$, for any $m \in \Omega$ the boundary points $\omega_a(m) := \inf\{\omega : U(\omega) \geq m\}$ and $\omega_b(m) := \sup\{\omega : L(\omega) \leq m\}$ are well-defined and satisfy $\Omega_{\mathbb{M}}(m) = [\omega_a(m), \omega_b(m)]$. Since U and L are right-continuous and strictly increasing in $\mathbb{M}^*$, $\omega_a(m)$ is non-increasing and $\omega_b(m)$ is non-decreasing in m, both right-continuous. Since F is atomless, $F(\Omega_{\mathbb{M}}(m)) = F(\omega_b(m)) - F(\omega_a(m))$ is continuous in m.

As we expand the interval of messages $[m', m'']$ around m^* both upward and downward, the winning region $\bigcup_{m' \leq m \leq m''} \Omega_{\mathbb{M}}(m)$ grows since $\omega_a(m')$ decreases as m' decreases and $\omega_b(m'')$ increases as m'' increases by strict monotonicity of L and U. Hence $F(\bigcup_{m' \leq m \leq m''} \Omega_{\mathbb{M}}(m))$ increases continuously from $F(\Omega_{\mathbb{M}}(m^*))$. The switching budget needed, $\alpha \cdot F(\bigcup_{m' \leq m \leq m''} \Omega_{\mathbb{M}}(m))$, also increases continuously from $\alpha F(\Omega_{\mathbb{M}}(m^*)) \leq B$. By the intermediate value theorem, we can expand $[m', m'']$ until $\alpha \cdot F(\bigcup_{m' \leq m \leq m''} \Omega_{\mathbb{M}}(m)) = B$, giving a winning region $W := \bigcup_{m' \leq m \leq m''} \Omega_{\mathbb{M}}(m)$ with $F(W) = B/\alpha$.

Now construct σ^* by having all types in W send their respective message $m \in [m', m'']$ and all other types send a feasible message outside of $[m', m'']$. Construct $\hat{\sigma}^*$ by having all types $\omega \geq \omega^*$ randomize over $[m', m'']$ proportionally to P_{σ^*}, using up the full budget p^*, and having all types $\omega < \omega^*$ send $m' \in M \setminus \Omega$. Then for each $m \in [m', m'']$,

$$\lambda(m) = \frac{P_{\hat{\sigma}^*}(m)}{P_{\sigma^*}(m)} = \frac{p^* \cdot P_{\sigma^*}(m)/F(W)}{P_{\sigma^*}(m)} = \frac{p^*}{F(W)} = \frac{B}{B/\alpha} = \alpha,$$

so the receiver switches after any $m \in [m', m'']$. Since $\hat{\sigma}^*$ assigns messages in $[m', m'']$ only to types $\omega \geq \omega^*$, the posterior after any $m \in [m', m'']$ is the distribution of ω

conditional on $\omega \geq \omega^*$, so $\mathbb{E}[\omega|m, \hat{\sigma}^*] = \mathbb{E}[\omega|\omega \geq \omega^*] = \omega_0$ and the receiver takes the risky action. All types in W win, achieving winning probability $F(W) = B/\alpha$.

Since $S_1 = 0$ and $S_2 \leq B/\alpha$, and the construction achieves exactly B/α, it is sender-optimal.

Sub-case (b): We now show that all types lose when $B < \alpha F(\Omega_{\mathcal{M}}(m))$ for all $m \in \Omega$. Since $S_1 = 0$, it suffices to show $S_2 = 0$. Suppose for contradiction that some message m triggers switching, so $\lambda(m) \geq \alpha$ and $P_\sigma(m) \leq P_{\hat{\sigma}}(m)/\alpha \leq B/\alpha < F(\Omega_{\mathcal{M}}(m))$. Since F is atomless, there exists a positive-measure set of types in $\Omega_{\mathcal{M}}(m)$ that do not send m. Since $S_1 = 0$ these types lose, but they can profitably deviate to m and trigger switching, contradicting equilibrium. Therefore $S_2 = 0$ and all types lose.

□

A.2.7 Proof of Corollary 4

Proof. Since $B \geq \alpha D_{\mathcal{M}}$, under $\mathcal{M}$ there exists a sender-optimal equilibrium in which the sender fully manipulates the receiver. Note that both the alternative and the true strategies remain feasible for the sender under $\mathcal{M}'$, since $\mathcal{M}'$ is more flexible than $\mathcal{M}$. Since those strategies achieve the best possible payoff for the sender (namely, risky action with probability 1), the same sender-optimal equilibrium remains an equilibrium under $\mathcal{M}'$. □

A.2.8 Proof of Proposition 3

Proof. **Case (i):** Suppose that the sender-optimal equilibrium under $\mathcal{M}$ is given by $(\hat{\sigma}, \sigma)$. This strategy is still feasible under $\mathcal{M}'$ since $\mathcal{M}$ is more downward flexible, so the sender is able to achieve the same ex-ante payoff under $\mathcal{M}'$, provided that no incentive-compatibility constraints are broken.

Note that an implication of Theorem 3 is that the sender's interim payoff in a sender-optimal equilibrium is increasing in the sender's type. If two types $\omega_1 < \omega_2$ earn different interim payoffs in the sender-optimal equilibrium under $\mathcal{M}$, $V(\omega_1) < V(\omega_2)$, then ω_1 must be unable to send the message that type ω_2 is sending in equilibrium, i.e. $\sigma^*(\omega_2) > U(\omega_1)$. Increasing downward flexibility cannot give ω_2 an option to imitate ω_1. Therefore, allowing greater downward flexibility does not break any incentive-compatibility constraints. In fact, it may increase $p^*_{\mathcal{M}}$ by allowing some underutilized high types to pool with bad types. Thus, the sender's ex-ante payoff is weakly greater under $\mathcal{M}'$ than under $\mathcal{M}$.

Case (ii): Since $\mathcal{M} \in \mathbb{M}^*$ and $F(\bar{\Omega}_{\mathcal{M}'}) < p^* + p^*/\alpha$, it follows from Theorem 3 that the optimal strategy $(\sigma_{\mathcal{M}}, \hat{\sigma}_{\mathcal{M}})$ under both of them takes the following reduced form: $\omega \in [\omega^*_{\mathcal{M}}, \sup \Omega]$ win by pooling using σ, $\omega \in [\omega^{**}_{\mathcal{M}}, \omega^*_{\mathcal{M}})$ win by narrative persuasion using

$\hat{\sigma}_{\mathcal{M}}$, and $\omega \in [\inf \Omega, \omega^{**}_{\mathcal{M}})$ lose. That applies to $\mathcal{M}'$ as well.

As long as $\omega^*_{\mathcal{M}}$ and $\omega^{**}_{\mathcal{M}}$ exist (which is ensured by $F(\bar{\Omega}_{\mathcal{M}}) \leq F(\bar{\Omega}_{\mathcal{M}'}) < p^* + p^*/\alpha$), the sender is able to save exactly p^*/α mass of types using narrative persuasion and $p^*_{\mathcal{M}}$ using pooling under σ. Therefore, for a more upward-flexible technology $\mathcal{M}' = (L', U')$ to offer the sender a strictly higher payoff, it must be that $p^*_{\mathcal{M}'} > p^*_{\mathcal{M}}$, which is equivalent to $\omega^*_{\mathcal{M}} > \omega^*_{\mathcal{M}'}$. That is the case if and only if ω^* cannot pool with $\sup \Omega$ under $\mathcal{M}$; otherwise, we would have $\omega^*_{\mathcal{M}} = \omega^*$.

There are only two ways for the sender's equilibrium payoff to be lower under $\mathcal{M}'$ than under $\mathcal{M}$. If the sender is successfully manipulating the receiver under $\mathcal{M}$, they cannot be successfully manipulating the receiver under $\mathcal{M}'$, since that would imply $p^*_{\mathcal{M}'} \geq p^*_{\mathcal{M}}$ and the same mass of types that are saved via narrative persuasion. That may occur if $\mathcal{M}'$ allows types at or below ω^{**} to imitate any types above them (i.e.,$\omega^{**} \in \bar{\Omega}_{\mathcal{M}}$), in which case we would have $p^*_{\mathcal{M}'} = 0$. The necessary and sufficient condition for that to occur is $F(\bar{\Omega}_{\mathcal{M}'}) \geq p^* + p^*/\alpha$.

Alternatively, the sender may not save no types with direct pooling under $\mathcal{M}$ while still saving a mass p^*/α of types via narrative persuasion. Recall that this happens if and only if $F(\bar{\Omega}_{\mathcal{M}}) \geq p^* + p^*/\alpha$ and there exists $m \in \Omega$ such that $F(\Omega_{\mathcal{M}}(m)) \leq p^*/\alpha$. In this case, the only way for $\mathcal{M}'$ to perform strictly worse than $\mathcal{M}$ is for there to be no message $m \in \Omega$ such that $F(\Omega_{\mathcal{M}'}(m)) \leq p^*/\alpha$. The sender is unable to pool or succeed via narrative persuasion, so her equilibrium payoff drops from p^*/α under $\mathcal{M}$ to 0 under $\mathcal{M}'$.

The sender's equilibrium payoff is the same under $\mathcal{M}$ and $\mathcal{M}'$ if and only if the two technologies have the same $p^*_{\mathcal{M}} = p^*_{\mathcal{M}'}$ (since they already save the same mass of types via narrative persuasion). Since $\mathcal{M}'$ is more upward flexible than $\mathcal{M}$, $U'(\omega^*_{\mathcal{M}}) > U(\omega_{\mathcal{M}})$, which would lead to $\omega_{\mathcal{M}'} > \omega_{\mathcal{M}}$ unless we are in one of two subcases: either $\omega^*_{\mathcal{M}} = \omega^*$ or $\omega^*_{\mathcal{M}}$ does not exist due to $F(\bar{\Omega}_{\mathcal{M}}) \geq p^* + p^*/\alpha$. The first subcase is equivalent to $U(\omega^*) \geq L(\sup \Omega) = L'(\sup \Omega)$, and we additionally require $F(\bar{\Omega}_{\mathcal{M}'}) < p^* + p^*/\alpha$ because otherwise $\omega^*_{\mathcal{M}'}$ would not exist (i.e., the sender would be unable to successfully pool under σ). The second subcase also requires that $\mathcal{M}$ and $\mathcal{M}'$ either both get the mass p^*/α of types to win via narrative persuasion, or they both fail. The two conditions are equivalent to there existing $m \in \Omega$ such that $F(\Omega_{\mathcal{M}'}(m)) \leq p^*/\alpha$ and there not existing $m \in \Omega$ such that $F(\Omega_{\mathcal{M}}(m)) \leq p^*/\alpha$, respectively. □

A.2.9 Proof of Theorem 4

Proof. We will prove the theorem case-by-case.

Case (i)**.** Since $\alpha D \leq B$, by Theorem 1 there exists a cheap-talk strategy $(\hat{\sigma}^{CT}, \sigma^{CT})$ such that the types $\omega \in [\inf \Omega, \omega^*)$ successfully manipulate the receiver by inducing a model switch, and the types $\omega \in [\omega^*, \sup \Omega]$ win by pooling among themselves. That gives the sender an equilibrium payoff of 1. Since that is the highest feasible payoff, all sender-optimal technologies must give the sender the same payoff.

Consider an arbitrary technology $\mathcal{M} = (L, U) \in \mathbb{M}^*$. For it to be sender-optimal, there must exist a strategy $(\hat{\sigma}, \sigma)$ that attains the same payoff as $(\hat{\sigma}^{CT}, \sigma^{CT})$. For a strategy $\hat{\sigma}$ to allow the types $\omega \in [\inf \Omega, \omega^*)$ to win through narrative persuasion, we can replicate $\hat{\sigma}^{CT}$ by mapping each message sent by a type $\omega \in [\omega^*, \sup \Omega]$ under $\hat{\sigma}^{CT}$ into feasible messages of types $[\inf \Omega, \omega^*)$ under $\mathcal{M}$. Since $\hat{\sigma}$ is unrestricted, this replicated mapping is straightforward.

Whether $\mathcal{M}$ is sender-optimal or not then comes down to the existence of a strategy σ that pools together types $\omega \in [\omega^*, \sup \Omega]$. It is both necessary and sufficient that there exists $\hat{m} \in \Omega$ such that $[\omega^*, \sup \Omega] \subseteq \Omega_{\mathcal{M}}(\hat{m})$. The condition is necessary because ω^* is defined as $\mathbb{E}[\omega | \omega \geq \omega^*] = \omega_0$; if ω^* cannot be pooled with $\sup \Omega$, then it is impossible to find a message m such that $\mathbb{E}[\omega | \omega \in [\omega^*, \sup \Omega_{\mathcal{M}}(m)]) \geq \omega_0$. The condition is also sufficient, since we can set $\sigma(\omega) = \hat{m}$ for all $\omega \in [\omega^*, \sup \Omega]$.

Case (ii)**.** If $\alpha D > B$, the best payoff the Sender would be able to achieve under commitment is $p^* + p^*/\alpha$. A BP-optimal strategy σ^{BP} pools the types $\omega \in [\omega^*, \sup \Omega]$ together, and there exists an alternative strategy $\hat{\sigma}$ that uses the types in $[\omega^*, \sup \Omega]$ to allow the types in $[\omega^{**}, \omega^*)$ to win via narrative persuasion. Recall that ω^{**} is defined as $\alpha F([\omega^{**}, \omega^*)) = B$.

Consider an arbitrary technology $\mathcal{M} = (L, U) \in \mathbb{M}^*$. For it to be sender-optimal, there must exist a strategy $(\hat{\sigma}, \sigma)$ that attains the payoff $p^* + p^*/\alpha$. By Theorem 3, such a strategy will split the types into $[\inf \Omega, \omega^{**}_{\mathcal{M}})$ that lose, $[\omega^{**}_{\mathcal{M}}, \omega^*_{\mathcal{M}})$ that win via narrative persuasion, and $[\omega^*_{\mathcal{M}}, \sup \Omega]$ that win via pooling.

First, we will prove that the conditions in the theorem's statement are sufficient. If $U(\omega^*) \geq L(\sup \Omega)$ and $F(\bar{\Omega}_{\mathcal{M}}) < p^* + p^*/\alpha$, then for all $\omega \in [\omega^*, \sup \Omega]$ we can set $\sigma(\omega) = U(\omega^*)$. By definition of ω^*, we have $\mathbb{E}[\omega | m = U(\omega^*)] = \omega_0$. Note that $F(\bar{\Omega}_{\mathcal{M}}) < p^* + p^*/\alpha$ ensures that the types below ω^{**} cannot send $U(\omega^*)$. We now need to create $\hat{\sigma}$ such that it allows the types in $[\omega^{**}, \omega^*)$ to trick the receiver with model persuasion. Since ω^{**} is defined as $B = \alpha(D_{\mathcal{M}} - F(\omega < \omega^{**}_{\mathcal{M}})$, such $\hat{\sigma}$ exists as long as the type ω^{**} is credible under $\mathcal{M}$. Since $U(\omega)$ is strictly increasing and $U(\omega) < \sup \Omega$ (by the condition $F(\bar{\Omega}_{\mathcal{M}}) < p^* + p^*/\alpha$), it follows that ω^{**} is credible under $\mathcal{M}$, and so we can separate that type from the types in $[\inf \Omega, \omega^{**})$.

Second, we will prove that the conditions in the theorem's statement are necessary. If $U(\omega^*) < L(\sup \Omega)$, we cannot find a message to pool ω^* and $\sup \Omega$ together, which implies that the best pooling strategy under $\mathcal{M}$ will achieve a winning probability of less than p^*, and thus there is no strategy $(\sigma, \hat{\sigma})$ that achieves the sender-optimal payoff of $p^* + p^*/\alpha$. If $F(\bar{\Omega}_{\mathcal{M}}) \geq p^* + p^*/\alpha$ (meaning $\omega^{**} \in \bar{\Omega}_{\mathcal{M}}$), it is impossible to separate some types below ω^{**} from the types pooling on $m = U(\omega^*) = \sup \Omega$. Thus, there is no pooling strategy σ that achieves a positive winning probability, so the sender once again is unable to achieve her optimal payoff of $p^* + p^*/\alpha$. □

A.2.10 Analysis with finite Ω

A.3 Proofs for Section 5

A.3.1 Proof of Theorem 5

Proof. Fix any exogenous cheap-talk messaging strategy σ. Under σ, the sender induces the risky action with probability $B(\sigma) = P_\sigma(M_\sigma^+) = 1 - D(\sigma)$, so narrative persuasion can improve her payoff only by converting some on-path losing messages into winning ones.

Necessity. Let $S \subseteq M_\sigma^-$ be any subset of on-path losing messages that are rescued in equilibrium. For each $m \in S$, switching requires $\lambda(m) = \frac{P_{\hat{\sigma}}(m)}{P_\sigma(m)} \geq \alpha$, hence $P_\sigma(m) \leq \frac{1}{\alpha} P_{\hat{\sigma}}(m)$. Summing over $m \in S$ gives $P_\sigma(S) \leq \frac{1}{\alpha} P_{\hat{\sigma}}(S)$. Moreover, any rescued message must be winning under the proposed model, so $S \subseteq M_{\hat{\sigma}}^+$. Therefore $P_{\hat{\sigma}}(S) \leq P_{\hat{\sigma}}(M_{\hat{\sigma}}^+) \leq p^*$, where p^* is the maximal winning probability in the cheap-talk Bayesian persuasion benchmark. It follows that $P_\sigma(S) \leq \frac{p^*}{\alpha}$. Thus any subset of losing messages that is rescued in equilibrium must satisfy $P_\sigma(S) \leq \frac{p^*}{\alpha}$. Consequently, in any equilibrium the sender's winning probability is at most $1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : P_\sigma(S) \leq p^*/\alpha} P_\sigma(S)$.

Sufficiency. Fix any subset $S \subseteq M_\sigma^-$ such that $P_\sigma(S) \leq \frac{p^*}{\alpha}$. If $P_\sigma(S) = 0$, the claim is immediate. Let $r := P_\sigma(S) > 0$. Then $\alpha r \leq p^*$. By Lemma 1, there exists a signal structure with a winning message occurring with probability p^*. Since $\alpha r \leq p^*$, by randomization we can refine this structure to obtain signal h such that $P(h) = \alpha r$ and $\mathbb{E}[\omega | h] \geq \omega_0$.

Let q be the conditional distribution of messages in S under the true model, and choose some off-path message $L \notin \text{supp}(\sigma)$. Construct a proposed model $\hat{\sigma}$ as follows: whenever the winning signal h is realized, instead of reporting h directly, draw a message from S according to q; all other realizations are mapped into L. Then for each $m \in S$,

$$P_{\hat{\sigma}}(m) = P_{\hat{\sigma}}(h)\, q(m) = \alpha r \cdot \frac{P_\sigma(m)}{r} = \alpha P_\sigma(m), \tag{23}$$

so $\lambda(m) = \frac{P_{\hat{\sigma}}(m)}{P_\sigma(m)} = \alpha$. Hence every message in S triggers switching.

Moreover, under the proposed model, any message in S can arise only through the winning signal h. Therefore, after observing any message $m \in S$, the receiver infers that

h occurred, and hence $\mathbb{E}_{\hat{\sigma}}[\omega|m] = \mathbb{E}[\omega|h] \geq \omega_0$. Thus every message in S is winning under the proposed model.

For any on-path message $m \notin S$, we have $P_{\hat{\sigma}}(m) = 0$ since L is off path under σ. Hence if $m \in M_\sigma^+$, then $\lambda(m) = 0 < \alpha$, so the receiver sticks with the true model and still takes the risky action. Therefore the sender induces the risky action with probability at least $1 - D(\sigma) + P_\sigma(S)$.

Since this construction works for every subset $S \subseteq M_\sigma^-$ satisfying $P_\sigma(S) \leq \frac{p^*}{\alpha}$, the sender's winning probability is at least $1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : P_\sigma(S) \leq p^*/\alpha} P_\sigma(S)$. Combining this lower bound with the upper bound established in the necessity part yields

$$\sup \Pr(\text{risky action}) = 1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : P_\sigma(S) \leq p^*/\alpha} P_\sigma(S). \tag{24}$$

If $D(\sigma) \leq \frac{p^*}{\alpha}$, then $S = M_\sigma^-$ is feasible, and therefore

$$1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : P_\sigma(S) \leq p^*/\alpha} P_\sigma(S) = 1 - D(\sigma) + P_\sigma(M_\sigma^-) = 1. \tag{25}$$

Hence full manipulation is achievable. If instead $D(\sigma) > \frac{p^*}{\alpha}$, then no feasible subset $S \subseteq M_\sigma^-$ can satisfy $P_\sigma(S) = D(\sigma)$, so

$$1 - D(\sigma) + \sup_{S \subseteq M_\sigma^- : P_\sigma(S) \leq p^*/\alpha} P_\sigma(S) < 1, \tag{26}$$

and full manipulation is impossible. □

A.3.2 Proof of Theorem 6

Proof. We first show if $KB \geq \alpha D$, full manipulation is achievable. Note that $KB \geq \alpha D \iff Kp^* \geq \alpha(1 - p^*) \iff p^* \geq \frac{\alpha}{K+\alpha}$. Since $KB \geq \alpha D$, we can choose numbers $d_1, \ldots, d_K \geq 0$ such that $d_k \leq \frac{B}{\alpha}$ for all k, $\sum_{k=1}^K d_k = D$. If $D > 0$, define $\beta_k \equiv d_k/D$, so that $\beta_k \geq 0$ and $\sum_{k=1}^K \beta_k = 1$. Let $H_1, \ldots, H_K, H', L_1, \ldots, L_K$ be distinct elements of M. For each k, let $\hat{\sigma}^k$ be a relabeling of the optimal Bayesian persuasion policy from Lemma 1, with winning message H_k and losing message L_k. Then $P_{\hat{\sigma}^k}(H_k) = B$ and $\mathbb{E}_{\hat{\sigma}^k}[\omega|H_k] \geq \omega_0$.

Now define the sender's actual messaging strategy σ^* as follows. For every $\omega > \omega^*$, let $\sigma^*(H'|\omega) = 1$, and at $\omega = \omega^*$, if necessary, use the same mixing probability as in the BP benchmark so that the posterior mean after H' is exactly ω_0. For every $\omega < \omega^*$, let $\sigma^*(H_k|\omega) = \beta_k$ for each $k = 1, \ldots, K$. Hence $P_{\sigma^*}(H_k) = \beta_k D = d_k \leq \frac{B}{\alpha}$. Therefore, for every k with $d_k > 0$, $\lambda_k(H_k) = \frac{P_{\hat{\sigma}^k}(H_k)}{P_{\sigma^*}(H_k)} = \frac{B}{d_k} \geq \alpha$. So after observing H_k, the receiver switches to model $\hat{\sigma}^k$. Since H_k is the winning message under the BP-optimal policy $\hat{\sigma}^k$, we have $\mathbb{E}_{\hat{\sigma}^k}[\omega|H_k] \geq \omega_0$, and hence the receiver chooses the risky action.

Next consider message H'. Since $P_{\hat{\sigma}^k}(H') = 0$ for every k, observing H' does not trigger switching to any proposed model. Under the actual strategy σ^*, however, H' is sent only by states above the cutoff, with mixing at ω^* if necessary as in the BP benchmark, so $\mathbb{E}_{\sigma^*}[\omega|H'] = \mathbb{E}[\omega|\omega \geq \omega^*] \geq \omega_0$, and the receiver again chooses the risky action. Thus every on-path message induces the risky action, and the sender obtains payoff 1 with probability 1. Since 1 is the maximal feasible payoff, full manipulation is supported in the sender-optimal equilibrium.

Assume now $KB < \alpha D$. Consider any equilibrium with actual messaging strategy σ and proposed models $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}^1, \ldots, \hat{\sigma}^K)$. Let W_0 be the set of messages after which the receiver does not switch and chooses the risky action, and for each $k = 1, \ldots, K$, let W_k be the set of messages after which the receiver switches to model k and chooses the risky action. The sender's winning probability is $P_\sigma(W_0) + \sum_{k=1}^K P_\sigma(W_k)$. Since $W_0 \subseteq M_\sigma^+$, $P_\sigma(W_0) \leq P_\sigma(M_\sigma^+) \leq B$, where the last inequality follows because $B = p^*$ is the maximal winning probability in the cheap-talk Bayesian persuasion benchmark.

For each k and each message $m \in W_k$, we have $\lambda_k(m) \geq \alpha$, so $P_\sigma(m) \leq \frac{1}{\alpha} P_{\hat{\sigma}^k}(m)$. Integrating over W_k yields $P_\sigma(W_k) \leq \frac{1}{\alpha} P_{\hat{\sigma}^k}(W_k)$. Moreover, because the receiver chooses the risky action after switching to model k on messages in W_k, we have $W_k \subseteq M_{\hat{\sigma}^k}^+$, and therefore $P_{\hat{\sigma}^k}(W_k) \leq P_{\hat{\sigma}^k}(M_{\hat{\sigma}^k}^+) \leq B$. Hence $P_\sigma(W_k) \leq \frac{B}{\alpha}$ for each k. Summing across k gives $P_\sigma(W_0) + \sum_{k=1}^K P_\sigma(W_k) \leq B + \frac{KB}{\alpha}$. If $KB < \alpha D = \alpha(1 - B)$, then $B + \frac{KB}{\alpha} < 1$. So full manipulation is impossible.

Suppose toward a contradiction that there exists an equilibrium in which the sender wins with strictly positive probability. Then there exists an on-path message m^+ after which the receiver chooses the risky action. Since full manipulation is impossible, the sender's winning probability is strictly less than 1. Hence there must exist some state ω from which, with positive probability, an on-path message m^- is sent that induces the safe action. Under cheap talk, every message is feasible in every state, so at state ω the sender can profitably deviate and send m^+ instead. After observing m^+, the receiver's continuation behavior is the equilibrium behavior assigned to m^+, which yields the risky action and sender payoff 1, whereas sending m^- yields sender payoff 0. This contradicts optimality of the sender's messaging strategy. Therefore, when $KB < \alpha D$, the sender cannot manipulate. □

A.3.3 Proof of Theorem 7

Proof. Part (i). Suppose $M_{\sigma_0}^+ \neq \varnothing$, and let $m^+ \in M_{\sigma_0}^+$. Consider the following strategy profile. The sender proposes $\hat{\sigma}^* = \sigma_0$ and sends message m^+ in every state, i.e., $\sigma^*(m^+|\omega, \hat{\sigma}^*) = 1 \ \forall \omega \in \Omega$. Since $\hat{\sigma}^* = \sigma_0$, the receiver interprets m^+ using σ_0 regardless of whether he formally switches. By definition of $m^+ \in M^+_{\sigma_0}$, he therefore chooses the risky action after observing m^+. Thus the sender obtains payoff 1 in every state, so full manipulation is achievable.

Part (ii): sufficiency. Now suppose $M^+_{\sigma_0} = \varnothing$, and suppose there exists a message $m^* \in M$ such that $P_{\sigma_0}(m^*) \leq \frac{p^*}{\alpha}$. We show that full manipulation is achievable.

Let L be another message distinct from m^*. Let $\hat{\sigma}^*$ be a relabeling of the optimal Bayesian persuasion policy from Lemma 1, with winning message m^* and losing message L. Then $P_{\hat{\sigma}^*}(m^*) = p^*$ and $\mathbb{E}_{\hat{\sigma}^*}[\omega|m^*] \geq \omega_0$. Now let the sender's actual messaging strategy pool on m^* in every state: $\sigma^*(m^*|\omega, \hat{\sigma}^*) = 1$ $\forall \omega \in \Omega$. Upon observing m^*, the receiver compares σ_0 and $\hat{\sigma}^*$ and obtains

$$\lambda(m^*|\sigma_0, \hat{\sigma}^*) = \frac{P_{\hat{\sigma}^*}(m^*)}{P_{\sigma_0}(m^*)} = \frac{p^*}{P_{\sigma_0}(m^*)} \geq \alpha. \tag{27}$$

Hence he switches to $\hat{\sigma}^*$. Since m^* is the winning message under the BP-optimal policy $\hat{\sigma}^*$, we have $\mathbb{E}_{\hat{\sigma}^*}[\omega|m^*] \geq \omega_0$. Therefore the receiver chooses the risky action after m^*, and since m^* is sent in every state, the sender obtains payoff 1 with probability 1. Thus full manipulation is achievable.

Part (ii): necessity. Still suppose $M^+_{\sigma_0} = \varnothing$, and suppose instead that $P_{\sigma_0}(m) > \frac{p^*}{\alpha}$ $\forall m \in M$. We show that no manipulation is possible.

Consider any equilibrium with proposed model $\hat{\sigma}$ and actual messaging strategy σ. Since $M^+_{\sigma_0} = \varnothing$, if the receiver does not switch after observing a message m, then he chooses the safe action. Hence any message that induces the risky action must do so after switching to $\hat{\sigma}$. Suppose there exists a message m that induces the risky action in equilibrium. Then, by the previous observation, the receiver must switch to $\hat{\sigma}$ after observing m, and therefore $m \in M^+_{\hat{\sigma}}$. Hence $P_{\hat{\sigma}}(m) \leq P_{\hat{\sigma}}(M^+_{\hat{\sigma}}) \leq p^*$. Therefore

$$\lambda(m|\sigma_0, \hat{\sigma}) = \frac{P_{\hat{\sigma}}(m)}{P_{\sigma_0}(m)} < \frac{p^*}{p^*/\alpha} = \alpha. \tag{28}$$

So the receiver does not switch after observing m, a contradiction. Hence no message can induce the risky action. It follows that under any equilibrium the receiver always chooses the safe action, so the sender's payoff is 0 with probability 1. Therefore no manipulation is possible. □

A.3.4 Proof of Theorem 8

Proof. Let $\mu_L := \mathbb{E}[\omega|\omega < \omega^*]$, $\mu_H := \mathbb{E}[\omega|\omega > \omega_0]$, $p_L := P(\omega < \omega^*)$, and $p_H := P(\omega > \omega_0)$. After observing message m, the receiver assigns posterior probability

$\beta(m) = \frac{\beta_0 P_{\hat{\sigma}}(m)}{\beta_0 P_{\hat{\sigma}}(m)+(1-\beta_0)P_\sigma(m)}$ to the proposed model. He therefore takes the risky action after m if and only if

$$\beta(m)\mu_{\hat{\sigma}}(m) + (1-\beta(m))\mu_\sigma(m) \geq \omega_0. \tag{29}$$

Necessity. For the sender to manipulate, some positive-measure set of types $\omega < \omega^*$ must be induced to take the risky action. Since the game is cheap talk, any type $\omega < \omega^*$ can imitate any on-path winning message. Hence if some types $\omega < \omega^*$ induced the risky action while others did not, the latter would have a profitable deviation to an on-path winning message. Therefore any manipulation must induce the risky action for all types $\omega < \omega^*$.

Among all constructions with this property, the posterior mean on a rescued low-type message is maximized by pooling the low types on a single message H and, under the proposed model $\hat{\sigma}$, assigning H only to the highest states $\omega > \omega_0$. Under this most favorable construction, $P_\sigma(H) = p_L$, $P_{\hat{\sigma}}(H) = p_H$, $\mu_\sigma(H) = \mu_L$, and $\mu_{\hat{\sigma}}(H) = \mu_H$, so the highest posterior mean attainable after H is $\frac{\beta_0 p_H \mu_H + (1-\beta_0)p_L\mu_L}{\beta_0 p_H + (1-\beta_0)p_L}$.

For (29) to hold after H, it is therefore necessary that $\frac{\beta_0 p_H \mu_H + (1-\beta_0)p_L\mu_L}{\beta_0 p_H + (1-\beta_0)p_L} \geq \omega_0$, or equivalently, $\beta_0 p_H(\mu_H - \omega_0) \geq (1-\beta_0)p_L(\omega_0 - \mu_L)$. Rearranging yields $\beta_0 \geq \bar{\beta}_0$.

Sufficiency. Suppose $\beta_0 \geq \bar{\beta}_0$. Define the messaging strategy σ and proposed model $\hat{\sigma}$ using three messages $\{H, H', \tilde{H}\}$ as follows: $\sigma(H|\omega) = 1$ for $\omega < \omega^*$, $\sigma(H'|\omega) = 1$ for $\omega \geq \omega^*$, $\hat{\sigma}(H|\omega) = 1$ for $\omega > \omega_0$, and $\hat{\sigma}(\tilde{H}|\omega) = 1$ for $\omega \leq \omega_0$.

After H. We have $P_\sigma(H) = p_L$, $P_{\hat{\sigma}}(H) = p_H$, $\mu_\sigma(H) = \mu_L$, and $\mu_{\hat{\sigma}}(H) = \mu_H$, so the receiver's posterior mean after H is $\frac{\beta_0 p_H \mu_H + (1-\beta_0)p_L\mu_L}{\beta_0 p_H + (1-\beta_0)p_L}$, which is at least ω_0 if and only if $\beta_0 \geq \bar{\beta}_0$.

After H'. Since $P_{\hat{\sigma}}(H') = 0$, we have $\beta(H') = 0$, so the receiver places full weight on the correctly specified model. Under σ, message H' is sent only by types $\omega > \omega^*$ and, at $\omega = \omega^*$ if necessary, with the same mixing probability as in the Bayesian persuasion benchmark, so $\mu_\sigma(H') = \omega_0$. By the tie-breaking rule, the receiver takes the risky action after H'.

Thus every type induces the risky action, no type has a profitable deviation, and the sender achieves full manipulation when $\beta_0 \geq \bar{\beta}_0$. □