Bayesian Persuasion with Lie Detection[*]

Weicheng Min[†] Florian Ederer[‡]

December 10, 2025

Abstract

How does lie detection constrain the potential for one person to persuade another to change her action? We consider a model of persuasion in which the Receiver can probabilistically detect lies. We show that under full commitment, the Sender lies more when the lie detection probability increases. When this probability is sufficiently small, the Sender's and the Receiver's equilibrium payoffs are unaffected by the presence of lie detection because the Sender simply compensates by lying more. When lie detection is sufficiently accurate, the Sender still lies more, sometimes seemingly against her own interest, to partially neutralize lie detection, but improvements in the lie detection technology strictly harm the Sender and benefit the Receiver. Under partial commitment, these strategic forces remain intact, and stronger lie detection can even harm the Receiver. Our model's main insights continue to hold qualitatively under several extensions, including general state, message, and action spaces, and different detection technologies.

JEL Codes: D83, D82, K40, D72, M31

Keywords: communication, lie detection, lying, constrained information design

[*] We are particularly grateful to Andrew Little for inspiring our initial analysis. We also thank Nageeb Ali, Joshua Gans, Scott Gehlbach, Matt Gentzkow, Philippe Jehiel, Navin Kartik, Kohei Kawamura, Elliot Lipnowski, Igor Letina, Barry Nalebuff, Jacopo Perego, Larry Samuelson, Julia Simon-Kerr, Joel Sobel, Alex Tabarrok, Satoru Takahashi, Ian Turner, and Akhil Vohra for helpful comments.

[†] Antai College of Economics and Management, Shanghai Jiao Tong University, w.min@sjtu.edu.cn

[‡] Boston University Questrom School of Business, CEPR, ECGI, and NBER, florian.ederer@gmail.com

1 Introduction

Lies are a pervasive feature of communication, even when that communication is subject to intense public and media scrutiny. For example, during his tenure as U.S. president, Donald Trump made over 20,000 false or misleading claims.[1] However, such lies are also often detectable. Monitoring and fact-checking should constrain how much license a sender of communication has when making false statements. Interestingly, however, in the face of increased fact-checking and media focus, Trump's rate of lying increased rather than decreased—a development that runs counter to this intuition.[2]

Lies and misinformation have also become particularly widespread on social media. Half of U.S. adults use social media for news consumption (Liedke and Wang, 2023), and lies that spread on social media can disrupt elections (Allcott and Gentzkow, 2017; Aral, 2021) and lead receivers of such lies to make bad health choices (Naeem et al., 2021). Recognizing this problem, social media platforms such as Facebook and X (formerly Twitter) have instituted fact-checking features (e.g., Meta's use of the International Fact-Checking Network and X's Community Notes) aimed at detecting and labeling false content and improving information accuracy on their respective platforms. However, a recent report showed that X's false content detection feature has failed to address harmful misinformation and has not effectively deterred accounts from "disseminating debunked claims and gaining more followers" (Kao and Bengani, 2023) and, in the political context, has "minimal effects on candidate evaluations or vote choice" (Nyhan et al., 2020).

This paper analyzes how lie detection and a sender's commitment power jointly affect communication strategies and payoffs. We show that a sender may optimally choose to lie more frequently when it is more likely that her false statements will be flagged as lies, and this behavior renders lie detection ineffective for the receiver for a large set of parameter values. As a clean benchmark that allows us to isolate the strategic forces introduced by probabilistic lie detection, we adopt a Bayesian persuasion framework (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011) as our

[1] See https://www.washingtonpost.com/politics/2020/07/13/president-trump-has-made-more-than-20000-false-or-misleading-claims/ for a comprehensive analysis of the increasing lack of veracity of Trump's statements.

[2] Former New York Representative George Santos also repeatedly lied to Congress and the media about his family history, educational attainments, and professional experience. Although his many lies were detected and publicly documented in the media, he continued lying and was ultimately expelled from Congress.

baseline model, in which the sender has full commitment. In Section 4, following Lipnowski et al. (2022) and Min (2021), we extend this framework to allow for partial commitment by assuming that the Sender's commitment binds probabilistically.

The central innovation of our model—that lies are sometimes detectable—is a natural assumption for many applications of (Bayesian) persuasion and communication, including political campaigns, courts, advertising, expert advice, lobbying, and financial disclosure. For example, in a court case, facts may surface that contradict the statements of a plaintiff, defendant, or witness and affect the judge's or the jury's verdict.[3] Similarly, a pre-sale inspection of a product may reveal that the seller has misrepresented some of the product's features, which in turn may influence the buyer's purchase decision. Or, as in our examples above, politicians and social media accounts may continue to peddle lies to their followers even when these lies are detected and exposed by journalists or platform fact-checking features.

In our model, a Sender and a Receiver engage in one round of communication. The Sender observes a binary state of nature and commits to a messaging strategy. We assume that the message space equals the state space and define a lie as a message that differs from the true state of nature. If the Sender tells a lie, it is flagged as such with some probability. The Receiver observes both the message and the lie detection outcome and then takes an action. Whereas the Sender prefers the Receiver to take the "favorable" action regardless of the state of nature, the Receiver wants to match the action to the underlying state. Finally, the payoffs are realized for both parties.

Our model delivers the following set of results. First, the Sender lies more frequently when the lie detection technology improves. Second, as long as the lie detection probability is sufficiently small, the equilibrium payoffs of both players are unaffected by the lie detection technology because the Sender simply compensates by lying more frequently in the unfavorable state of nature by claiming that the state is favorable. That is to say, the lie detection technology changes the Sender's messaging strategy but does not impact the payoff of either player. Third, when the lie detection technology is sufficiently reliable, any further increase in the lie detection probability

[3] Courts also focus on demeanor (e.g., facial expression, tone of voice, body language, gaze) as a lie detection tool, but the effectiveness of this policy is not supported by scientific studies (Simon-Kerr, 2020).

causes counterintuitive lying where the Sender lies more frequently in the favorable state of nature, seemingly against her own interest. The Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability.

A simple example illustrates the central intuition of our model. Suppose that politicians (Sender) always want war, but war is not always good for voters (Receiver) who make the ultimate decision of supporting a war. If voters can never detect a lie we obtain the canonical Bayesian persuasion outcome. The politicians always say "war is good" when war is good, but when war is bad they sometimes say "war is bad" and sometimes say "war is good" (i.e., they sometimes lie). In equilibrium, politicians tell the truth just often enough that voters are indifferent between following the politicians' advice that war is good and ignoring them completely. The politicians are better off because relative to truth-telling, occasional lying results in war not only when war is good but sometimes even when it is bad.

Now assume that there is a technology that detects lies with some (low) probability (e.g., occasional fact-checking). Holding all else equal, voters would be better off because they can detect some lies of politicians saying "war is good" when in fact war is bad. However, due to the additional information generated by the lie detection technology, voters now strictly prefer to support a war when the politicians claim "war is good" without it being detected as a lie. Thus, politicians now have the incentive to report "war is good" more frequently when war is actually bad (i.e., they lie more often) to restore the voters' indifference condition. The voters' threat point of ignoring the politicians altogether has not changed, and so in the new equilibrium, their expected utility is still the same.

We also study the impact of lie detection in partial commitment settings that bridge Bayesian persuasion and cheap talk. We show that as long as the Sender's commitment power is sufficiently high, the main forces identified in the baseline model remain: weak lie detection can still be completely neutralized by strategic messaging by the Sender, and sufficiently accurate lie detection continues to generate counterintuitive lying by the Sender and to benefit the Receiver. Even with lower commitment power, the Sender can still partially neutralize lie detection. Commitment therefore affects the intensity, but not the nature, of the strategic responses to lie detection.

Moreover, under partial commitment, improvements in lie detection may sometimes even hurt the Receiver.

Our framework is sufficiently tractable to analyze a number of extensions. First, our main insights also hold in more general persuasion environments with richer state, message, and action spaces. Specifically, with a larger state or message space, both players' payoffs are completely independent of the lie detection technology, whereas with a larger action space, our baseline results about the players' equilibrium payoffs continue to hold if and only if the prior is sufficiently low or high. Second, we consider alternative detection technologies such as lie detection with false alarms, truth detection, and state detection, showing that the central insights of our model continue to hold. Third, we analyze the (nontrivial) case in which the default action coincides with the Sender's preferred action and show that the main results are analogous to those in the baseline model.

Our paper contributes to the study of constrained information design (Doval and Skreta, 2018; Kamenica et al., 2021; Ball and Espín-Sánchez, 2022). One of the key assumptions in the information design literature is that the information designer (i.e., the Sender in our setting) can commit and flexibly choose any information structure. In reality, however, the designer may not be able to commit to all information structures, and the exact nature of the constraints depends on the particular application. Our paper studies one such constraint (i.e., lie detection) that imposes realistic limits on the power of the designer and analyzes the optimal design problem under this constraint. Despite its simplicity, this constraint is different from constraints previously considered in the literature. Among them, Tsakas and Tsakas (2021) and Le Treust and Tomala (2019) are closest to our paper. They allow imperfect communication by introducing purely exogenous noise to the messages of the Sender. Thus, the Receiver obtains less information if the Sender's strategy is held fixed. In contrast, lie detection is endogenous as it depends on the message, and the Receiver obtains more information if the Sender's strategy is held fixed. The distinction is also apparent in their results. Whereas Tsakas and Tsakas (2021) show that the Sender may be strictly better off as the noise decreases, we show that the Sender can never be strictly better off as the lie detection technology improves. Matyskova and Montes (2023) consider a related setting in which the Receiver can gather outside information that in turn decreases the Sender's power to persuade.

Two recent papers (Balbuzanov, 2019; Dziuda and Salas, 2018) specifically investigate the role of lie detection in communication. We follow their analysis and assume that messages have literal meanings. The most significant difference with respect to our paper lies in the communication protocol. In both papers, the communication game takes the form of cheap talk (Crawford and Sobel, 1982) rather than Bayesian persuasion.[4] Although it is debatable whether the extreme cases of full commitment (as in Bayesian persuasion) or no commitment (as in cheap talk) constitute more plausible assumptions about real-life communication settings, our baseline model is an important step toward studying communication games with lie detection. Furthermore, in our extension with partial commitment, we show that our insights are not limited to the extreme case of full commitment. Due to the difference in the communication protocol, a large number of equilibria arise in the two papers, making the comparative statics difficult. Dziuda and Salas (2018) impose two assumptions on off-path beliefs and consider a special environment to guarantee the uniqueness of the (informative) equilibrium. Balbuzanov (2019) does not consider comparative statics and instead focuses on the existence of a fully revealing equilibrium.

Related theoretical work on lying in communication games includes Kartik et al. (2007) and Kartik (2009), who do not consider lie detection but instead introduce an exogenous cost of lying tied to the size of the lie in a cheap talk setting. They find that most types inflate their messages, but only up to a point. In contrast to our results, they obtain full information revelation for some or all types depending on the bounds of the type and message space. Guo and Shmaya (2021) considers a communication protocol in which the message space is over the distribution of states, and the Sender incurs a miscalibration cost if a message differs from the induced posterior of the message in equilibrium. They show that when this cost is sufficiently high, the Sender can obtain her commitment payoff. In contrast, if the Sender in our model loses all commitment power, she

[4] Jehiel (2021) considers a setting with two rounds of communication à la Crawford and Sobel (1982) but includes the innovative feature that a Sender who lied in the first period cannot remember the exact lies that she told. However, the potential inconsistency of messages never arises in any pure strategy equilibrium. As a result, no lies are ever detected in equilibrium. In Perez-Richet and Skreta (2022) the Sender can falsify inputs in the experiment rather than lie about outputs as in our model. Levkun (2022) considers the role of strategic fact-checking in communication. Finally, in the cheap talk model of Morris (2001) reputational incentives can lead a sender who has the same preferences as the receiver (unlike in our setting), to lie and be more “politically correct” in order to enhance her reputation to report truthfully in future periods.

cannot obtain the commitment payoff for any lie detection probability. Sobel (2020) adopts a more abstract approach and clarifies the relationship between lying and deception in a general framework. The definition of lying in his paper is informally consistent with ours.

In the domain of political science, Luo and Rozenas (2018, 2021) consider Bayesian persuasion with lying, yet with a different approach and a different definition of lies. In their models, the Sender does not have full commitment power and lies by misreporting the signal realization she observes. In contrast, in our model the Sender has full commitment power and lies by committing to a strategy that is not fully truth-telling. Moreover, as mentioned earlier, our paper can be viewed as a standard Bayesian persuasion problem with additional constraints on the set of feasible information structures. However, no such constraint is imposed in those papers. Furthermore, in contrast to our findings they show that the Sender lies only if the probability of lie detection is intermediate. In a slightly different vein, Gehlbach et al. (2022) analyze how improvements that benefit the Sender (e.g., censorship and propaganda) impact communication under Bayesian persuasion. In contrast, we focus on an improvement in the Receiver's communication technology.

Finally, a large and growing experimental literature (Gneezy, 2005; Hurkens and Kartik, 2009; Sánchez-Pagés and Vorsatz, 2009; Ederer and Fehr, 2017; Gneezy et al., 2018) examines lying in a variety of communication games. Most closely related to our work is Fréchette et al. (2022) who investigate models of cheap talk, information disclosure, and Bayesian persuasion in a unified experimental framework. Their experiments provide general support for the strategic rationale behind the role of commitment and, more specifically, for the Bayesian persuasion model of Kamenica and Gentzkow (2011).

2 Model

Consider the following simple model of Bayesian persuasion in the presence of lie detection.

Timing and Strategies: Let $\omega \in \{0,1\}$ denote the state of the world and $\Pr(\omega = 1) = \mu \in (0,1)$. The Sender ($S$, she) sends a message $m \in \{0,1\}$ to the Receiver (R, he). We assume that the Sender has full commitment power, as is common in the Bayesian persuasion framework.[5] Specifically, the

[5] For a detailed discussion and relaxation of this assumption, see Min (2021), Fréchette et al. (2022), Lipnowski

strategy of the Sender is a mapping $\sigma : \{0,1\} \longrightarrow \Delta(\{0,1\})$. The Receiver observes the message m together with a lie detection outcome $d \in \{lie, \neg lie\}$, and then takes an action $a \in \{0,1\}$. The exact nature of the lie detection technology is specified below. The strategy of the Receiver is a mapping $a : \{0,1\} \times \{lie, \neg lie\} \longrightarrow \Delta(\{0,1\})$. Denote the Sender's and Receiver's strategy space by Σ and A, respectively.

Lie Detection Technology: Messages in our model are defined to have literal meanings. That is to say, a message is classified as a lie if it does not match the true state of nature. We make this assumption for several reasons. First, this assumption is realistic because lie detection and fact-checking in practice involve checking the literal text of statements, not what is implied by them for the receivers (Nyhan et al., 2020). Second, assuming that messages have a literal meaning corresponding to the underlying state (or type) is in line with the work of Dziuda and Salas (2018) and Balbuzanov (2019) and thus allows us to compare our results to theirs. Third, this assumption ensures that the problem remains tractable. If the Sender lies (i.e., $m \neq \omega$), the Receiver is informed with probability $q \in [0,1]$ that the message is a lie. With the remaining probability $1-q$, he is not informed. If the Sender does not lie (i.e., $m = \omega$), the message is never flagged as a lie, and the Receiver is not informed. Formally, the detection technology can be described by the following relation:

$$d(m,\omega) = \begin{cases} lie, & \text{with probability } q \text{ if } m \neq \omega \\ \neg lie, & \text{with probability } 1-q \text{ if } m \neq \omega \\ \neg lie, & \text{with probability } 1 \text{ if } m = \omega \end{cases}$$

With a slight abuse of notation, we denote $d = \{lie, \neg lie\}$ as the outcome of the detection result. The detection technology is common knowledge. In a standard Bayesian persuasion setup, this detection probability q is equal to 0, giving us an easily comparable benchmark.

The lie detection technology in our baseline model does not incorrectly flag truthful messages

et al. (2022), Nguyen and Tan (2021), Perez-Richet and Skreta (2022), and Koessler and Skreta (2023). Titova (2021) shows that with binary actions and a sufficiently rich state space, verifiable disclosure enables the Sender's commitment solution to be an equilibrium. Lin and Liu (2022) reviews the different approaches used by these papers. In Section 4, we show that our results continue to hold under partial commitment.

as lies. However, as we show in Section 5.2.1, even with such false alarms, our main insights continue to hold. Note further that lie detection is different from state detection. While the former informs the Receiver of the true state conditional on a lie, the latter informs him of the true state independently of the message. Section 5.2.2 discusses the differences between the two detection technologies in more detail.

Payoffs: Given an action a under the state ω, the players' payoffs are realized as follows:

$$u_S(a,\omega) = \mathbb{1}_{\{a=1\}}$$
$$u_R(a,\omega) = (1-t)\cdot \mathbb{1}_{\{a=\omega=1\}} + t\cdot \mathbb{1}_{\{a=\omega=0\}}, \quad 0<t<1$$

The Sender wants the Receiver to take the action $a = 1$ regardless of the state, while the Receiver wants to match the state.[6] The payoff from matching state 0 may differ from the payoff from matching state 1. Given the payoff function, the Receiver takes action $a = 1$ if and only if

$$\Pr(\omega = 1 \mid m, d) \geq t$$

and therefore one could also interpret t as the threshold of the Receiver's posterior belief above which he takes action $a = 1$. In the main body, we assume $t \in (\mu, 1)$ to capture the more interesting case in which the Receiver's default action differs from the Sender's preferred action. However, unlike in standard persuasion models, even the case in which $t \in (0, \mu]$ is nontrivial because the couple $(m,\, d)$ necessarily reveals some information to the Receiver. We defer the detailed discussion of this case to Section 5.3.

At this point, it is worth emphasizing that our setup makes a few critical omissions and assumptions. First, the Sender does not intrinsically care if she is detected as having lied. Lie detection simply changes the Receiver's belief about the state.[7] Second, our lie detection technology makes fully uninformative messages impossible. For example, if the Sender commits to the same message

[6] Balbuzanov (2019) allows some degree of common interest between the two players and focuses on the existence of a fully revealing equilibrium. In contrast, a fully revealing equilibrium never exists in our setting because the two players have no common interest, just as in Dziuda and Salas (2018).

[7] Relaxing this assumption would lead to a further trade-off which largely runs counter to our own and which is discussed at length in the literature on lying aversion (Kartik et al., 2007; Kartik, 2009; Ederer and Fehr, 2017).

no matter the state, the lie detection technology necessarily reveals some information about the state, and how much is revealed depends on which constant message the Sender chooses to send. Third, as discussed above, messages are defined to have literal meanings for reasons of realism, simplicity, and comparability with the existing literature.

3 Analysis

The Sender solves the following maximization problem:

$$\begin{aligned} \max_{\sigma\in\Sigma,\, a\in A} \quad & \mathbb{E}_{\omega,m,d}[u_S(a\,(m,d)\,,\omega)] \\ s.t. \quad & a(m,d) \in \operatorname*{arg\,max}_{a\in\{0,1\}} \mathbb{E}_{\omega}[u_R(a,\omega) \mid m,d],\ \forall (m,d) \in \{0,1\} \times \{lie, \neg lie\} \end{aligned}$$

Due to the simple structure of the model, it is without loss of generality to assume that the Sender chooses only two reporting probabilities, $p_0 = \Pr(m = 0 \mid \omega = 0)$ and $p_1 = \Pr(m = 1 \mid \omega = 1)$.[8] We denote the optimal reporting probabilities of the Sender by p_0^* and p_1^* and the ex-ante payoffs under these reporting probabilities by U_S and U_R.[9] Next, we characterize the optimal messaging strategy (p_0^*, p_1^*) and discuss its comparative statics in Section 3.1. Section 3.2 then examines how lie detection affects the equilibrium payoffs U_S and U_R.

3.1 Optimal Messaging Strategy

Given the Sender's strategy (p_0, p_1), the Receiver could potentially observe four types of events. Denote his posterior belief after observing the event (m, d) by $\mu_{m,d}$. By Bayes' rule,

$$\begin{aligned} \mu_{0,lie} &= 1, \qquad & \mu_{0,\neg lie} &= \frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q) + (1-\mu)p_0} \\ \mu_{1,lie} &= 0, \qquad & \mu_{1,\neg lie} &= \frac{\mu p_1}{\mu p_1 + (1-\mu)(1-p_0)(1-q)} \end{aligned}$$

[8] We use the terms messaging strategy and reporting probability interchangeably throughout the paper.

[9] In principle, there may exist multiple Sender-optimal strategies under which the Receiver's payoffs differ. However, as we will show later, such multiplicity does not arise in our setting, ensuring that U_R is well-defined.

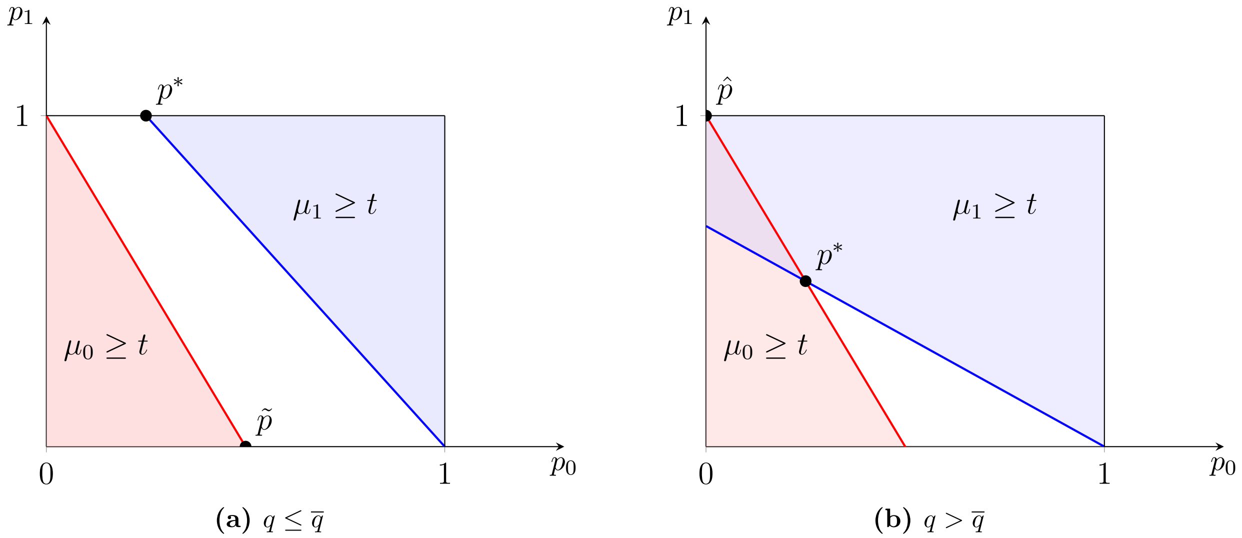


Figure 1: Partition of strategy space and equilibrium strategies for different detection probabilities q. Strategies in the red region induce $\mu_0 \geq t$, while those in the blue region induces $\mu_1 \geq t$.

Specifically, if the Receiver is informed of a lie, his posterior beliefs are degenerate due to the binary state space. Consequently, he chooses $a = 1$ after $(m = 0, d = lie)$ and $a = 0$ after $(m = 1, d = lie)$. With a slight abuse of notation, let $\mu_m \equiv \mu_{m,\neg lie}$. When $p_0 = 0,\ p_1 = 1$, the belief μ_0 is off-path and not restricted by Bayes' rule. However, the off-path belief does not matter for the Sender. For expositional convenience, define $\mu_0 = 0$ in this case. Analogously, if $p_0 = 1,\ p_1 = 0$, define $\mu_1 = 0$.

Figure 1 illustrates how the Sender's strategy space is partitioned into different lie detection probabilities. Strategies in the red region induce $\mu_0 \geq t$, while those in the blue region induces $\mu_1 \geq t$. Proposition 1 characterizes the optimal messaging strategy. There are two cases which depend on the strength of the lie detection technology.

Proposition 1. *Let* $\overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)} \in (0, 1)$.

(a) *If* $q \in (0, \overline{q}]$, *the Sender's optimal strategy is* $p_1^* = 1,\ p_0^* = \frac{\overline{q}-q}{1-q}$, *where the Sender always tells the truth under* $\omega = 1$ *but lies with positive probability under* $\omega = 0$.

(*b*) *If* $q \in (\overline{q}, 1]$*, the Sender's optimal strategy is*[10]

$$p_0^* = \frac{1-q}{(2-q)q}(q - \overline{q}) \quad \text{and} \quad p_1^* = \frac{1-q}{(2-q)q}\left[\frac{1}{1-\overline{q}} - (1-q)\right],$$

where the Sender lies with positive probability under both states.

In case (a), the lie detection probability q is small, so it is impossible to induce both $\mu_0 \geq t$ and $\mu_1 \geq t$. This follows from the martingale property, which requires that the four posteriors average to the prior. When q is small, lie detection is rare, and most of the weights fall on the posteriors μ_0 and μ_1. Thus, μ_0 and μ_1 must approximately average back to the prior μ, suggesting that they cannot simultaneously lie far above the threshold $t > \mu$.

The Sender's optimal messaging strategy in this case, denoted by p^* in Figure 1(a), resembles the optimal experiment in Bayesian persuasion without lie detection: the Sender is fully honest in the favorable state $\omega = 1$, and lies in the unfavorable state $\omega = 0$ just enough to leave the Receiver indifferent between the two actions. Note that when the lie detection technology is unavailable ($q = 0$), messages have no intrinsic meaning and can be freely relabeled. It is therefore equally optimal for the Sender to be totally dishonest in the favorable state and lie just enough in the unfavorable state to satisfy the Receiver's indifference condition. As shown in the figure, this means that the strategy $\tilde{p}$ achieves the same payoff as p^*. However, the introduction of a lie detection ($q > 0$) generates an intrinsic meaning for the message. Any message not flagged as a lie carries credibility for the state it denotes. This additional source of credibility breaks the symmetry, making p^* the unique optimal strategy for $q \in (0, \overline{q}]$.

In case (b), the detection probability q exceeds $\overline{q}$, making it feasible to sustain both $\mu_0 \geq t$ and $\mu_1 \geq t$. Any strategy in the intersection of the red and blue regions achieves this. Interestingly, the optimal strategy, denoted p^*, entails lying even in the favorable state $\omega = 1$. The intuition unfolds in two steps.

First, any strategy $p = (p_0, p_1)$ in the interior of this intersection strictly dominates $\hat{p}$, the best strategy that entails no lying in state $\omega = 1$. Under $\hat{p}$, the Sender always lies by sending $m = 1$ in

[10] To be precise, when $q = 1$ the Receiver learns the true state regardless of the Sender's strategy. Thus, every messaging strategy is optimal for the Sender. To ensure continuity in q, we set $p_0^* = p_1^* = 0$ in this case.

state $\omega = 0$, which induces $a = 0$ when such a lie is detected, an event that occurs with probability $(1-\mu)q$. Under p, the Sender sends $m = 0$ in the favorable state as well, making $m = 0$ a risk-free message to induce $a = 1$, regardless of lie detection. Similarly, this strategy induces $a = 0$ only if the Sender tells a lie in the unfavorable state and it is detected, which occurs with a probability $(1-\mu)q \cdot (1-p_0)$. Since $p_0 < 1$, any such p yields the Sender a strictly higher payoff than $\hat{p}$.

In comparison, the advantage of p relative to $\hat{p}$ is that the "safer" message $m = 0$ is sent more frequently. Thus, the optimal strategy must involve the highest p_0 within the intersection of two colored regions, yielding p^* in Figure 1 (b). Under p^*, we have $\mu_0 = \mu_1 = t$, so the Receiver is indifferent between actions whenever no lie is detected. Although p^* features lying in both states, the Sender lies more frequently in the unfavorable state: $p_0^* < p_1^*$.

Finally, the threshold $\overline{q}$ at which the optimal strategy switches is decreasing in μ and increasing in t. When the Receiver is easily persuaded (i.e., the prior μ is already close to the threshold t), the Sender optimally lies in both states. By contrast, when the Receiver is difficult to persuade (i.e., t is much higher than μ), it is optimal for the Sender to only lie in the unfavorable state.

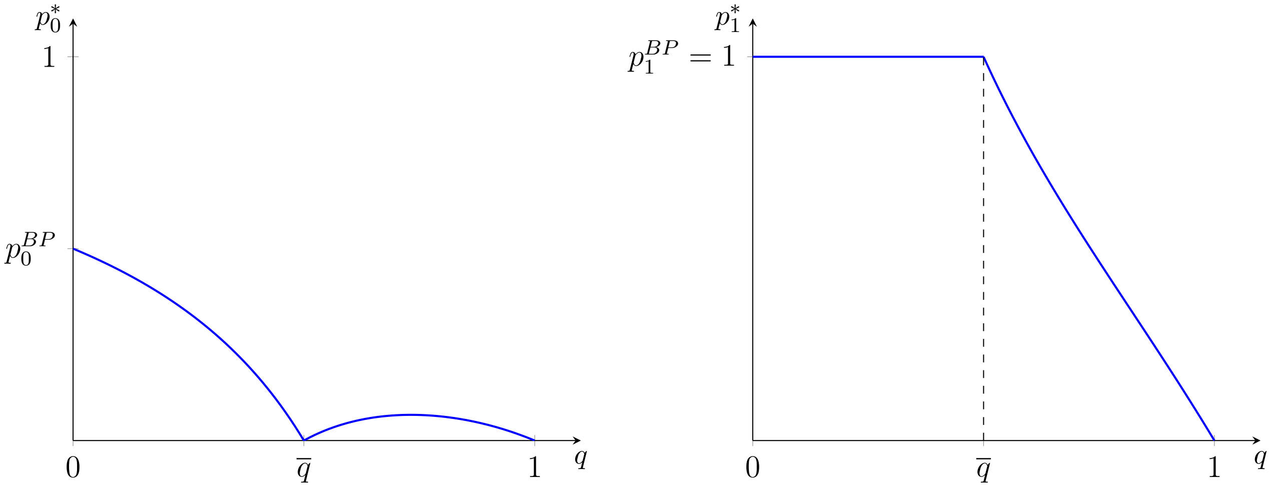
Proposition 2 then describes how the optimal messaging strategy $(p_0^*,\ p_1^*)$ varies with the detection probability. Figure 2 plots these reporting probabilities as functions of q. For comparison, the probabilities p_0^{BP} and p_1^{BP} denote the equilibrium reporting probabilities in a standard Bayesian persuasion model without lie detection.

Proposition 2. *The optimal messaging strategy satisfies the properties with respect to q:*

(a) $p_0^* = Pr(m = 0 \mid \omega = 0)$ *decreases over* $q \in [0, \overline{q}]$ *and has an inverse U shape over* $q \in (\overline{q}, 1]$,

(b) $p_1^* = Pr(m = 1 \mid \omega = 1)$ *is constant over* $q \in [0, \overline{q}]$ *and decreases over* $q \in (\overline{q}, 1]$.

If $q \leq \overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)}$, the Sender's optimal strategy involves truthfully reporting the state $\omega = 1$ but progressively misreporting the state $\omega = 0$ as the lie detection technology improves. This result contrasts with the findings of Dziuda and Salas (2018) who show that lie detection is effective in reducing lying in a cheap talk environment. If $q > \overline{q}$, the Sender lies more often in the favorable state as detection improves, while the probability of lying in the unfavorable state becomes non-monotonic in q: p_0^* first rises and then falls. As $q \to 1$, the Sender lies maximally under both states. Next, we provide the underlying intuition for these comparative statics.



(a) $p_0^* \equiv \Pr(m = 0 \mid \omega = 0)$

(b) $p_1^* \equiv \Pr(m = 1 \mid \omega = 1)$

Figure 2: Equilibrium reporting probabilities p_0^* and p_1^* as a function of q for $\mu = \frac{1}{3}$ and $t = \frac{1}{2}$.

For $q = 0$, recall from Kamenica and Gentzkow (2011) that if an optimal signal induces a belief leading to the worst action for the Sender ($a = 0$), the Receiver is certain of his action at this belief. Conversely, if the optimal signal induces a belief leading to the best action for the Sender ($a = 1$), the Receiver must be indifferent between the two actions at this belief.

Now consider the addition of a lie detection technology. As q increases, the event $(m = 1, d = \neg lie)$ becomes more indicative of the favorable state $\omega = 1$. Holding the Sender's strategy fixed, the Receiver would strictly prefer $a = 1$ after this event. In response, the Sender would like to send the message $m = 1$ more often. Because the Sender already sends $m = 1$ with probability one in state $\omega = 1$, the only way to increase the frequency of $m = 1$ is to send it more often in state $\omega = 0$. Overall, the Sender increases lying about the unfavorable state just enough to keep the Receiver indifferent.

Once the detection probability q exceeds $\overline{q}$, the Sender can no longer rely solely on increasing lies in state $\omega = 0$, because at $\overline{q}$ she already lies maximally in that state. In this case, a lie of the form $m = 0$ in state $\omega = 1$ is very likely to be detected, which the Sender can exploit to ensure that $(m = 0, d = \neg\text{lie})$ still leads the Receiver to choose $a = 1$. Hence, for $q \approx \overline{q}$, the Sender increases the use of message $m = 0$ by both raising p_0 and reducing p_1. Finally, as q approaches one, the Sender cannot continue increasing p_0. When detection is nearly perfect, an unflagged message $m = 0$ becomes too indicative of $\omega = 0$ and can no longer induce $a = 1$. Consequently, the probability of lying in the unfavorable state must follow an inverse U-shaped pattern as q rises from $\overline{q}$ to 1.

The counterintuitive lying phenomenon (i.e., sending a message $m = 0$ when the state is $\omega = 1$) that would seem to hurt the Sender is somewhat reminiscent of Morris (2001) where the Sender also sometimes lies seemingly against her own interest. The underlying sources for such counterintuitive lying, however, differ. Whereas in Morris (2001) the Sender does so in order to enhance her reputation to report truthfully in future periods, in our setting such counterintuitive lying arises not for reputational reasons but because the Sender thereby neutralizes the lie detection technology.

As a complementary result, Proposition 3 analyzes the impact of lie detection on the informativeness of the Sender's strategy. Formally, each Sender's messaging strategy (p_0, p_1) corresponds to an experiment

$$\mathcal{E}(p_0, p_1) = \begin{bmatrix} p_0 & 1 - p_1 \\ 1 - p_0 & p_1 \end{bmatrix}$$

When $q \leq \overline{q}$, $\mathcal{E}(p_0^*, p_1^*)$ becomes Blackwell less informative as q increases, which echoes with our intuition that the Sender lies more to offset the additional information conveyed by lie detection. Interestingly, when $q > \overline{q}$, $\mathcal{E}(p_0^*, p_1^*)$ becomes Blackwell more informative as q increases. Thus, the Sender strategically provides more information when the lie detection is sufficiently strong, suggesting again that lie detection causes a move in the right direction only when the detection technology is good enough.

We can also treat the Sender's message and the lie detection outcome as a joint experiment with four signal realizations and two states, defined as

$$\Gamma(p_0, p_1) = \begin{bmatrix} p_0 & (1-p_1)(1-q) \\ 0 & (1-p_1)q \\ (1-p_0)(1-q) & p_1 \\ (1-p_0)q & 0 \end{bmatrix}.$$

In equilibrium, this joint experiment always becomes Blackwell more informative as q increases.

Proposition 3. *For any q, q' such that $0 \leq q' < q \leq 1$,*

(a) If $q \leq \overline{q}$, then $\mathcal{E}(p_0^*(q'), p_1^*(q'))$ Blackwell dominates $\mathcal{E}(p_0^*(q), p_1^*(q))$.

(b) If $q' \geq \overline{q}$, then $\mathcal{E}(p_0^*(q), p_1^*(q))$ Blackwell dominates $\mathcal{E}(p_0^*(q'), p_1^*(q'))$.

(c) $\Gamma(p_0^*(q), p_1^*(q))$ Blackwell dominates $\Gamma(p_0^*(q'), p_1^*(q'))$.

3.2 Payoffs

We now investigate how the equilibrium payoffs are affected by improvements in the lie detection technology. The results are summarized in Proposition 4 and graphically depicted in Figure 3. For comparison, U_S^{BP} and U_R^{BP} are the equilibrium payoffs that would result in the absence of lie detection, while U_S^F and U_R^F are the payoffs when the Receiver is fully informed about the underlying state.

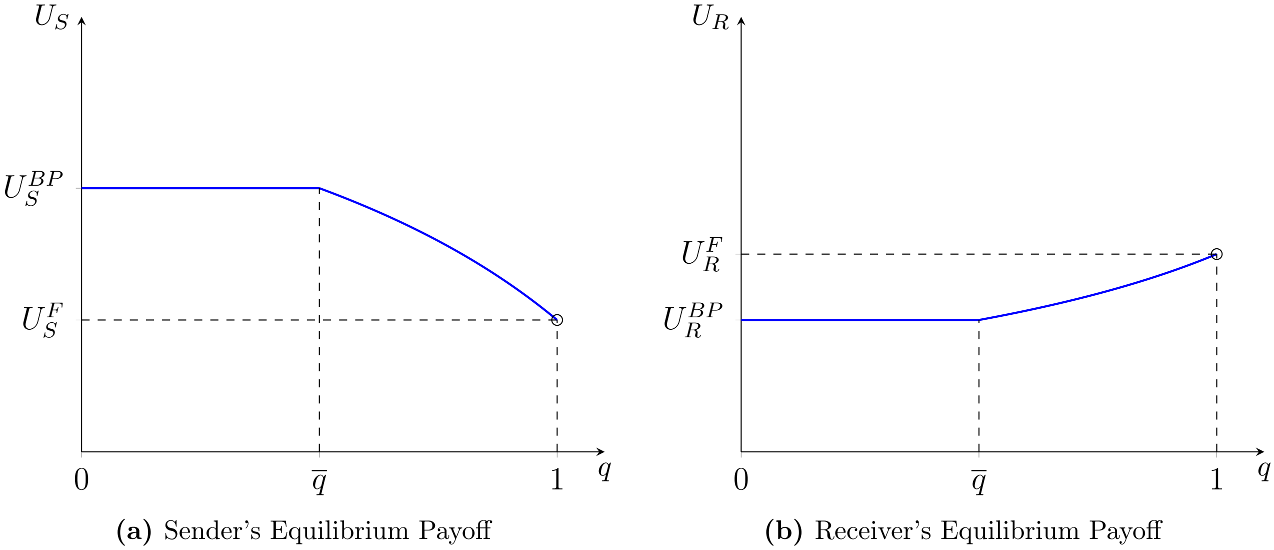
Proposition 4. *As the lie detection probability q increases,*

(a) U_S is constant over $[0, \overline{q}]$ and decreases over $(\overline{q}, 1]$.

(b) U_R is constant over $[0, \overline{q}]$ and increases over $(\overline{q}, 1]$.

The Sender's equilibrium payoff does not change for $q \leq \overline{q}$ and decreases with q for $q > \overline{q}$. As long as $q \leq \overline{q}$, the Sender receives exactly the same payoff that she would receive under the Bayesian persuasion benchmark. Any marginal improvement to the lie detection technology (i.e., an increase in q) is completely offset by less truthful reporting when $\omega = 0$ (i.e., a decrease in p_0^*).

However, for $q > \overline{q}$, any further improvements reduce the Sender's payoff as the strategic effect of less truthful reporting is dominated by the direct effect of improving q. In the limit case where $q = 1$, the Sender has no influence anymore, and the action $a = 1$ is implemented if and only if the state is $\omega = 1$.



(a) Sender's Equilibrium Payoff

(b) Receiver's Equilibrium Payoff

Figure 3: Equilibrium payoffs as a function of q for $\mu = \frac{1}{3}$, $t = \frac{1}{2}$.

Analogously, the Receiver's payoff is constant at the Bayesian persuasion benchmark for $q \leq \overline{q}$ and then increases with q for $q > \overline{q}$.[11] As q approaches 1, the Receiver's payoff approaches his payoff under full information U_R^F. In the canonical Bayesian persuasion benchmark, the Receiver is held to his outside option of obtaining no information whatsoever. Thus, when the lie detection probability q increases, the Receiver is more certain that $(m = 1,\ d = \neg lie)$ means $\omega = 1$ and would obtain a larger surplus from the improvement in the lie detection technology. However, as long as p_0^* is greater than 0, the Sender can simply undo this payoff improvement for the Receiver by lying more about $\omega = 0$ (i.e., reduce p_0^* even further), thereby "signal-jamming" the information obtained by the Receiver. This result is also in line with Nyhan et al. (2020) who find that the impact of lie detection and fact-checking on evaluations of candidates or voting decisions is minimal.[12]

[11] This means that if having access to the lie detection technology required any costly investment, the Receiver would only ever want to invest in improving lie detection if it raised q above the threshold $\overline{q}$.

[12] Strictly speaking, our model is closer to flagging misinformation than traditional fact-checking. In traditional fact-checking, each claim is fact-checked with some probability, and if it is fact-checked, it is usually judged either true or false. In contrast to our setup, this yields three signals (true, false, and unchecked).

4 Partial Commitment

Some of the principal insights of our model continue to hold even under partial commitment. Following Lipnowski et al. (2022) and Min (2021), we assume that the Sender's commitment binds probabilistically. The generalized game with partial commitment proceeds as follows.

The Sender first declares a commitment strategy $(p_0, p_1) \in [0,1]^2$. She then privately learns the true state $\omega \in \{0,1\}$ and whether her commitment is binding. With probability $\alpha \in [0,1]$, her commitment binds, and she must send a message following the prespecified commitment strategy. Otherwise, her commitment is not binding, and she can send any message $m \in \{0,1\}$ at her discretion. Let $(\tilde{p}_0, \tilde{p}_1) \in [0,1]^2$ denote her strategy following a nonbinding commitment, where $\tilde{p}_i$ is the probability that she sends a message $i \in \{0,1\}$ when the true state is i. Finally, define $E[p_i^*] := \alpha p_i^* + (1-\alpha)\tilde{p}_i^*$ as the Sender's expected probability of telling the truth in state $\omega = i$.

The rest of the model is similar to our baseline model. Any message that is inconsistent with the true state is identified as a lie with probability q regardless of the status of the commitment. Last, the Receiver takes an action $a \in \{0,1\}$ after observing both the message and the lie detection outcome. He is aware that the Sender may not abide by her commitment strategy, and the commitment probability α is common knowledge. For simplicity, let the status of commitment be independent of both the true state and the lie detection technology. The payoff functions are identical to those in the baseline model. The baseline model corresponds to the special case $\alpha = 1$, whereas $\alpha = 0$ instead leads to a model of cheap talk with lie detection. Because cheap talk suffers from the issue of equilibrium multiplicity, we focus on the Sender-optimal equilibrium when $\alpha = 0$.

We fully characterize the optimal messaging strategy and the associated equilibrium payoffs for all parameter pairs $(\alpha, q) \in [0,1]^2$. Proposition 5 summarizes the comparative statics of the expected probabilities of lying and equilibrium payoffs in the lie detection probability. The interaction between the Sender's commitment level and the strength of the lie detection gives rise to four distinct cases, corresponding to the four regions depicted in Figure 4.

Proposition 5. *Denote $\overline{q} = \frac{t-\mu}{t(1-\mu)}$, $\underline{\alpha}(q) = \frac{\overline{q}-q}{1-q}$, and $\overline{\alpha}(q) = \frac{1-(1-q)(1-\overline{q})}{2q-q^2}$. Consider four cases:*

(a) If $q \le \overline{q}$ and $\alpha < \underline{\alpha}(q)$, then $\mathbb{E}[p_1^*]$ is constant in q, while $\mathbb{E}[p_0^*]$ is not unique. Moreover, both U_S and U_R strictly increase in q.

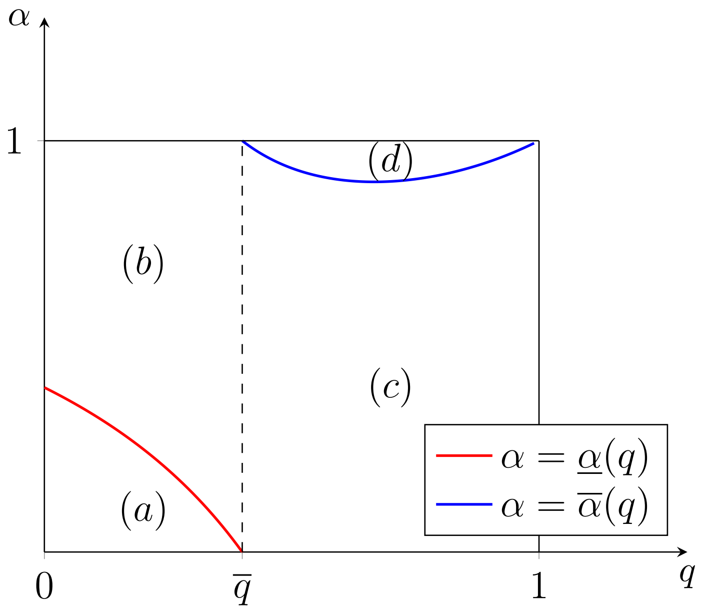


Figure 4: Interplay between commitment and lie detection ($\overline{q} = 0.4$). The regions (a), (b), (c), and (d) correspond to the respective parts of Proposition 5.

(b) *If $q \leq \overline{q}$ and $\alpha \geq \underline{\alpha}(q)$, then $E[p_0^*]$ strictly decreases in q, while $E[p_1^*]$ is constant in q. Moreover, both U_S and U_R are constant in q.*

(c) *If $q > \overline{q}$ and $\alpha < \overline{\alpha}(q)$, then Both $E[p_0^*]$ and $E[p_1^*]$ are constant in q. Moreover, U_S strictly decreases in q while U_R strictly increases in q.*

(d) *If $q > \overline{q}$ and $\alpha \geq \overline{\alpha}(q)$, then $E[p_0^*]$ first strincreases and then decreases in q, while $E[p_1^*]$ strictly decreases in q. Moreover, U_S strictly decreases in q while U_R strictly increases in q.*

Regions (b) and (d) essentially replicate the results of Section 3. In region (b), the Sender always tells the truth in the favorable state $\omega = 1$, and lies just enough in the unfavorable state to maintain the Receiver's indifference condition upon observing $(m = 1, d = \neg lie)$. As q increases, the Sender increases lying such that both U_S and U_R remain constant. In region (d), the Sender engages in counterintuitive lying to persuade the Receiver to take $a = 1$ whenever a message is not flagged as a lie. As in the baseline model, increases in q hurt the Sender and benefit the Receiver. Thus, our main results do not hinge on the full commitment assumption commonly used in Bayesian persuasion models.

In region (c), although lie detection is strong, the commitment is so weak (i.e., $\alpha < \overline{\alpha}(q)$) for the

Sender to induce the Receiver to take $a = 1$ after both $(m = 1, d = \neg lie)$ and $(m = 0, d = \neg lie)$. As a second-best strategy, the Sender always claims $m = 1$, regardless of whether her commitment binds. Since her strategy is independent of q, improvements in lie detection strictly benefit the Receiver and strictly hurt the Sender.

Recall that the Receiver does not know whether the Sender's commitment binds, and simply best responds to the expected informativeness of the prespecified strategy and the revised strategy, weighted by α and $1 - \alpha$, respectively. In region (a), the Sender's commitment is so weak (i.e., $\alpha < \underline{\alpha}(q)$) that even if the prespecified strategy is fully informative, the expected informativeness is still insufficient to induce $a = 1$. In this case, the Sender's best response is to always lie in the favorable state $\omega = 1$, hoping that the lie will be detected. Thus, she prefers stronger lie detection. Moreover, because the Sender's strategic incentive to lie more disappears, the Receiver also prefers stronger lie detection.

Overall, the strategic force of increasing lies to neutralize lie detection arises in regions (b) and (d). This means that for this mechanism to be effective the Sender requires sufficiently strong commitment power for a given lie detection probability.

While payoffs vary monotonically and continuously with q within each region, they need not be monotone or continuous across regions. For example, when α is very small (i.e., $\alpha < \underline{\alpha}(q)$), increasing q from 0 to 1 moves the environment through regions $(a)-(b)-(c)$. Figure 5 shows that the Receiver's payoff first rises linearly, then drops downward and remains constant, and eventually increases up to the full-information payoff. The Receiver is therefore worst off when lie detection is unavailable or only moderate. Indeed, improving lie detection can strictly reduce the Receiver's payoff as the environment transitions from region (a) to (b).

Finally, we can analyze how changes in commitment power α affect equilibrium payoffs. Proposition 6 shows that for a fixed q, the Sender's payoff is always (weakly) increasing in α, while the Receiver's payoff is (weakly) decreasing. This is in line with insights from other communication models, where greater commitment power always benefits the Sender. However, in our model, these benefits of commitment for the Sender (and, analogously, the corresponding disadvantages for the Receiver) are discontinuous and occur when α increases to cross from region (a) to (b) and

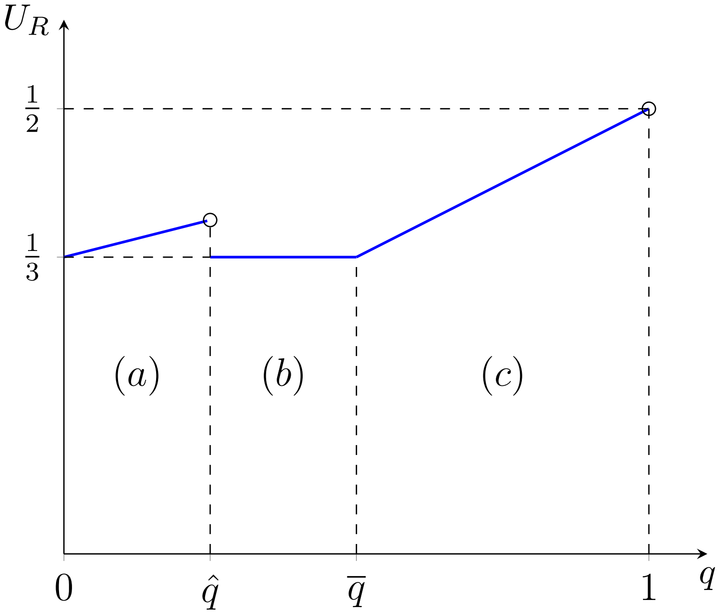


Figure 5: Receiver's equilibrium payoff as a function of q for $\mu = \frac{1}{3}, t = \frac{1}{2}$, and $\alpha = \frac{1}{3}$.

from (c) to (d).

Proposition 6. *Fix any $q \in [0, 1]$, the Sender's payoff is weakly increasing in α and the Receiver's payoff is weakly decreasing in α.*

5 Extensions

Our baseline model considers the role of lie detection in a simple setting with binary states, binary messages, binary actions, and a particular lie detection technology. We now investigate how alternative assumptions about the state space, the message space, the action space, and the detection technology modify our analysis and show that the principal insights of the model remain unchanged. To ensure tractability, we focus on environments in which the Sender has full commitment.

5.1 General Persuasion Environments

Because equilibrium uniqueness is not necessarily guaranteed in general persuasion environments, we do not analyze the comparative statics of optimal messages but instead focus on the comparative statics of equilibrium payoffs.

5.1.1 General State Space

With a non-binary state space, the Receiver's posterior belief is not necessarily degenerate after learning that the Sender has lied. However, this does not mean that the Receiver is better off. On the contrary, we show that the effectiveness of lie detection completely disappears. Both players' payoffs are entirely unaffected by lie detection, no matter how strong it is.

Let $\omega \in \Omega = \{\omega_1, ..., \omega_N\}$ be the state of the world, where $0 = \omega_1 < \omega_2 ... < \omega_N = 1$. Let $(\lambda_1, ..., \lambda_N)$ be the full-support common prior over Ω and μ be the prior mean. As in the baseline model, the message space is identical to the state space, and a lie ($m \neq \omega$) is detected with probability $q \in [0, 1]$. After observing the message and the lie detection outcome, the Receiver takes a binary action $a \in \{0, 1\}$. Both players' ex post payoff functions are given by

$$\begin{aligned}
&u_S(a, \omega) = a \\
&u_R(a, \omega) = \sum_{\omega_i \geq t} (\omega_i - t) \mathbb{1}_{\{a=1,\, \omega=\omega_i\}} + \sum_{\omega_i < t} (t - \omega_i) \mathbb{1}_{\{a=0,\, \omega=\omega_i\}}
\end{aligned}$$

The Sender always prefers $a = 1$ over $a = 0$ regardless of the true state. The Receiver's right action under a state ω_i is $a = 1$ if $\omega_i \geq t$, and is $a = 0$ otherwise. The weights for taking the right action under different states again ensure that the Receiver takes action $a = 1$ if and only if his posterior mean is weakly higher than t. Moreover, assume that $t \in (\mu, 1)$ which guarantees that the Receiver's default action is $a = 0$.

The Sender's strategy is a mapping $\sigma : \Omega \to \triangle(\Omega)$. Since her strategy space is richer than in the baseline model, there may exist multiple Sender-optimal strategies. Moreover, the Receiver's payoff may differ across these Sender-optimal strategies and it is possibly not well-defined. We solve this issue by focusing on the Receiver's highest payoff among these strategies. Formally, denote the set of Sender-optimal strategies by Σ^*. Let $U_S(q)$ be the Sender's (ex-ante) optimal payoff when the lie detection probability is q. Let $U_R(\sigma; q)$ be the Receiver's (ex-ante) payoff under strategy σ when the lie detection probability is q. Finally, let $U_R(q) = \sup_{\sigma \in \Sigma^*} U_R(\sigma; q)$ be the Receiver's highest payoff among all Sender-optimal strategies. Proposition 7 shows that both players' payoffs are independent of lie detection, thereby strengthening Proposition 4.

Proposition 7. *If $N \geq 3$, then $U_S(q) = U_S(0)$ and $U_R(q) = U_R(0)$, for any $q \in [0, 1]$.*

Our proof in Appendix A.4 proceeds as follows. We first construct a Sender-optimal strategy that induces the benchmark payoff pair $(U_S(0), U_R(0))$ when $q = 0$. Then we construct another strategy that induces the same payoff pair for arbitrary q. Since lie detection constrains the set of induced distribution of posteriors, any payoff vector that can be generated when $q > 0$ can also be generated when $q = 0$. It follows that $(U_S(q), U_R(q)) = (U_S(0), U_R(0))$ for $q > 0$.

5.1.2 General Message Space

The baseline model restricts the message space to coincide with the state space, thereby ruling out fuzzy statements such as "the state is either zero or one" (in our binary setting) or "the state belongs to a subset of the state space" (when the state space is more general). As in the case of a general state space in the previous section, we show that allowing for richer messages renders lie detection entirely useless, regardless of its strength.

Formally, given a state space $\Omega = \{\omega_1, \ldots, \omega_N\}$ with $0 = \omega_1 < \omega_2 < \cdots < \omega_N = 1$, let the message space be $\mathcal{M} := 2^\Omega \setminus \{\emptyset\}$, where 2^Ω is the set of all subsets of Ω. The lie detection technology now works in such a way that a message is flagged as a lie if it does not contain the true state.

Consider first a binary state space. The message space is $M = \{0, 1, \{0, 1\}\}$, where the additional message $\{0, 1\}$ represents a fuzzy statement that is never flagged as a lie, regardless of the true state. The Sender can commit to sending $m = \{0, 1\}$ for sure when the state is favorable, and to mixing between $m = \{0, 1\}$ and $m = 0$ when the state is unfavorable, with the mixing probabilities chosen to make the Receiver indifferent upon observing $m = \{0, 1\}$. Under this strategy, the Sender's message is never flagged as a lie. The induced information structure exactly replicates the optimal experiment in Kamenica and Gentzkow (2011), thereby yielding the maximal payoff for the Sender and the minimal one for the Receiver. Consequently, lie detection has no effect in this environment.

In a more general state space, Proposition 7 already shows that lie detection is ineffective even with a smaller message space, as the Sender can still attain the maximal payoff. It therefore follows that enlarging the message space has no effect. The Sender can already achieve the same payoff without resorting to the additional fuzzy messages.

5.1.3 General Action Space

We now return to a binary state and message space but allow for a general action space $A = \{a_1, ..., a_N\}$, where $0 = a_1 < a_2 < ... < a_N = 1$. The Sender's payoff function is still $u_S(a, \omega) = a$. For the Receiver, we assume a payoff function that induces a ladder-shaped best response function, which is qualitatively similar to the one adopted in the baseline model:

$$a^*(\mu) = \sum_{i=1}^{N} a_i \mathbb{1}_{\{\mu \in T_i\}}$$

where $T_i = [t_{i-1}, t_i)$ for $i \in \{1, ..., N-1\}$, $T_N = [t_{N-1}, t_N]$, and $0 = t_0 < t_1 < ... < t_{N-1} < t_N = 1$. Thus, when the state mean satisfies $\mu \in T_n$, the Receiver's optimal action is a_n.

By construction, $a^*(\mu)$ also represents the Sender's payoff function at posterior μ. Following the standard concavification method, the Sender's maximal payoff in the absence of lie detection is the concave closure of $a^*(\mu)$, denoted by $f(\mu)$. Since $a^*(\mu)$ is a ladder function, $f(\mu)$ must be a piece-wise linear and concave function, as illustrated by Figure 6.

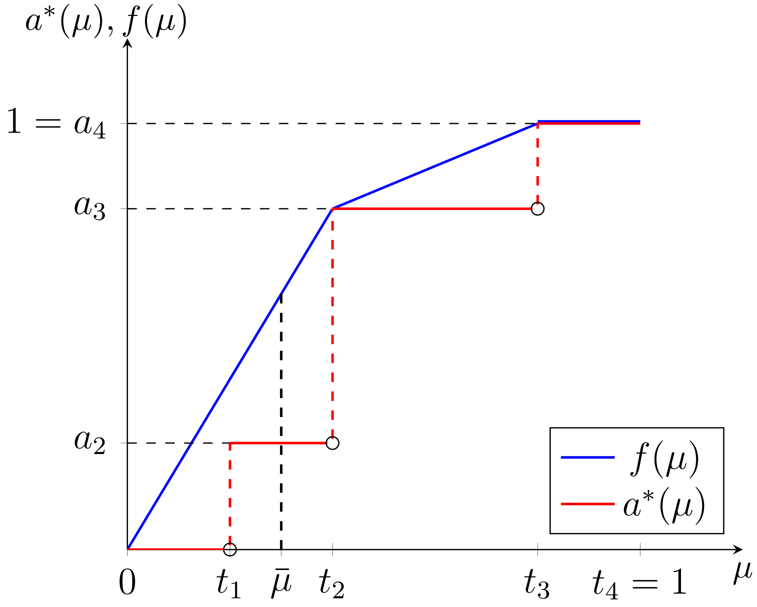


Figure 6: Illustration of $a^*(\mu)$ and $f(\mu)$ when $N = 4$.

In this general setup, Proposition 8 shows that the Sender's payoff is invariant to a small probability of lie detection if and only if the prior mean is sufficiently low or high.

Proposition 8. *Define* $s = \max\{i \in \{1, .., N-1\} \mid \frac{a_{i+1}}{t_i} \geq \frac{a_{j+1}}{t_j}, \forall j = 1, ..., N-1\}$.

(*a*) *If* $\bar{\mu} \in (0, t_s) \cup (t_{N-1}, 1)$*, then* $U_S(q) = U_S(0)$ *for sufficiently small* $q > 0$*.*

(*b*) *If* $\bar{\mu} \in [t_s, t_{N-1}]$*, then* $U_S(q) < U_S(0)$ *for any* $q > 0$*.*

Geometrically, t_s is the largest belief cutoff such that the line connecting the origin and the point (t_s, a_{s+1}) lies above $a^*(\mu)$. In the example illustrated in Figure 6, we have $t_s = t_2$. If there is no lie detection and $\bar{\mu} < t_s$, then one possible Sender-optimal strategy splits the prior into 0 and t_s. Otherwise, any Sender-optimal strategy must split the prior into strictly positive posteriors.

The first half of Proposition 8 generalizes the baseline result. If $q > 0$ and $\bar{\mu} \in (0, t_s)$, then the Sender would strategically lie more under the unfavorable state to offset the effect of lie detection, leaving her payoff unchanged. If $\bar{\mu} \in (t_{N-1}, 1)$, then the Receiver's default action coincides with the Sender's most preferred action. When $q = 0$, the Sender finds it optimal to employ a completely uninformative strategy. Such a strategy is no longer feasible when $q > 0$, yet when q is sufficiently small, Sender has access to a sufficiently uninformative strategy which leads to the same outcome. Section 5.3 further explores this issue in detail.

The second half of Proposition 8 is less straightforward. If $q > 0$ and $\bar{\mu} \in [t_s, t_{N-1}]$, it is not trivial to compute the Sender's maximal payoff because the set of induced distributions of posteriors is not easily characterized. Instead, we bypass the difficulty by showing that $U_S(p_0, p_1; q) < U_S(0)$ for any strategy $(p_0, p_1) \in [0, 1]^2$.

As a corollary of Proposition 8, if the Receiver's default action is either the lowest action a_1 or the highest action a_N, a sufficiently weak lie detection technology is always ineffective. Intuitively, when the Receiver's opinion is extreme and difficult to sway, a little additional information from lie detection is not helpful.

Corollary 1. *If* $a^*(\bar{\mu}) = a_1$ *or* $a^*(\bar{\mu}) = a_N$*, then* $U_S(q) = U_S(0)$ *for sufficiently small* $q > 0$*.*

5.2 Detection Technologies

Our results also continue to hold under different detection technologies that the Receiver can use to inform his choice of action.

5.2.1 Lie Detection with False Alarms

The baseline model considers an extreme form of lie detection technology in which a message that is identified as a lie, is surely a lie. We introduce false alarms by considering the following general lie detection technology:

$$d = \begin{cases} lie, & \text{with probability } q \in [0,1] \text{ if } m \neq \omega \\ lie, & \text{with probability } r \in [0,q] \text{ if } m = \omega. \end{cases}$$

A message is flagged as a lie with probability r even if it is actually not a lie. In particular, $r = 0$ indicates no false alarm and corresponds to the baseline model, while $r = q$ indicates an uninformative lie detection technology and corresponds to a standard persuasion problem without lie detection.

The potential of a false alarm has a non-monotonic effect on the Sender's equilibrium payoff. Consider $q \leq \overline{q}$, in which case we have shown that the Sender's equilibrium payoffs are identical when $r = 0$ or $r = q$. However, as Proposition 9 demonstrates, her equilibrium payoff is strictly lower for any $r \in (0, q)$, and is thus non-monotonic over r. As Kamenica and Gentzkow (2011) explain in the canonical persuasion problem without lie detection, the Sender obtains the payoff $U_S(0,0)$ by either inducing the Receiver to be indifferent between two actions ($\mu = t$) or inducing the worst belief ($\mu = 0$). When $r \in (0,q)$, it is impossible to induce such a distribution of posteriors. Specifically, whenever $\mu_{m,d} = t$ for some event (m, d), it is necessary that $\mu_{m,d'} \neq 0$ and $\mu_{m,d'} \neq t$ for $d' \neq d$.

Proposition 9. *$U_S(q,r) = U_S(0,0)$ if and only if $r = 0$, $q \leq \overline{q}$ or $q = r$.*

This result also suggests that the Sender cannot obtain the benchmark payoff $U_S(0,0)$ and is thus generically hurt by a lie detection technology. However, this does not invalidate one of our baseline results that lie detection does not impact the Sender's payoff. Proposition 10 shows that the Sender is hurt purely by the possibility of a false alarm rather than by the detection of true lies. Formally, a weak lie detection technology (i.e., low q) has no impact on either player's payoff as long as there is a sufficiently low probability r of a false alarm, thereby reaffirming an insight of the baseline model.

Proposition 10. *Fix any $q < \overline{q}$, then there exists $\overline{r} \in (0, q)$ such that for any $r \in [0, \overline{r}]$, $U_S(q, r) = \frac{\mu(1-r)}{t}$ and $U_R(q, r) = t(1 - \mu)$.*

5.2.2 Truth and State Detection

Consider a different detection technology that informs the Receiver with probability r that a message is truthful. Rather than being able to (probabilistically) detect a lie, the Receiver can (probabilistically) detect that a message is truthful.[13]

Truth detection turns out to be payoff-equivalent to lie detection. Therefore, all of our insights about the equilibrium payoffs as a function of the lie detection probability q in Figure 3 also hold for the truth detection probability r. However, under truth detection, the Sender's optimal messaging strategy is completely flipped and has some unnatural features. When the truth detection probability r is low but positive, it is optimal for the Sender to always lie in the favorable state (i.e., $p_1 = 0$) and to choose p_0 such that the Receiver is indifferent between $a = 0$ and $a = 1$ upon a message $m = 0$ that is not marked as truth.

Combining lie detection and truth detection such that they are perfectly positively correlated is equivalent to state detection. Assume that with probability $q = r$, the Receiver learns the state ω regardless of the message sent by the Sender. With such a state detection technology, the analysis becomes much simpler, as we simply return to the Bayesian persuasion benchmark. This is because the Sender's message does not influence at all whether the Receiver learns the state and any message is only relevant whenever the Receiver does not learn the state.

These observations also highlight our interpretation of Bayesian persuasion under lie detection in that the Sender's messages have a literal meaning of truth and lies. Even though the Sender

[13] Truth detection is perhaps a less realistic assumption, as it is arguably easier to detect whether the Sender has lied than whether she has sent a truthful message (Vrij et al., 2011).

is committing to the strategy—or, alternatively speaking, choosing an experiment—the strategies employed by the Sender are not equivalent to just an arbitrary garbling of information about the state of the world.

5.3 Default Action Coincides with Sender's Preferred Action

In standard Bayesian persuasion models without lie detection, the Sender can always remain silent and leave the Receiver totally uninformed by committing to a purely uninformative signal. Therefore, a trivial case obtains if the Receiver's default action coincides with the Sender's preferred action. However, the messages in our model have literal meanings and are subject to lie detection, which forces information transmission from the Sender to the Receiver. Therefore, the Sender cannot leave the Receiver totally uninformed, rendering her optimization problem nontrivial even when the Receiver's default action coincides with her preferred action.

This feature of our model is especially relevant in settings where the Sender is expected or even required to communicate, regardless of whether it benefits them. For instance, in the prosecutor-judge example mentioned earlier, it would be implausible for the prosecutor to simply remain silent. In such scenarios, lie detection can compel the Sender to reveal information even when doing so is against their interest. The possibility that a lie may be exposed deters the Sender from remaining vague or silent, especially when silence itself may be interpreted unfavorably or is not a viable option. This stands in contrast to standard Bayesian persuasion models, where the Sender can always choose to transmit no information at all. Our framework therefore captures an important institutional realism: lie detection technologies can induce informative communication from the Sender even in environments where silence would otherwise be optimal.

We revisit the baseline model under the assumption that the prior mean μ exceeds the threshold t. The results are analogous to those in the baseline model: the effect of lie detection depends on its strength. When lie detection is weak ($q \leq \tilde{q} \equiv \frac{\mu - t}{\mu(1-t)}$), the Sender can still attain the benchmark payoff without lie detection by sending message $m = 0$ with probability one in both states. This uninformative strategy induces the Receiver to always take the Sender's favorable action $a = 1$. However, when lie detection is strong ($q > \tilde{q}$), this outcome is no longer attainable.

Any uninformative strategy requires substantial lying in at least one state, and such lies are frequently detected. Consequently, even an uninformative strategy ends up revealing significant information to the Receiver. In the extreme case $q = 1$, the Receiver effectively learns the true state, and thus chooses $a = 1$ if and only if $\omega = 1$. Consistent with Proposition 4, the Sender's (Receiver's) payoff is constant in q for $q \leq \tilde{q}$ and then strictly decreases (increases) in q for $q > \tilde{q}$.

Admittedly, the fact that the Sender cannot induce the Receiver to always take the favorable action even when $\mu \geq t$ suggests some tension between our model and the standard persuasion paradigm. However, it is easy to reconcile this tension by introducing an additional stage prior to the persuasion game in which the Sender decides whether or not to enter the game. If she enters, the Sender and the Receiver play the persuasion game with lie detection specified in our main analysis. Otherwise, the Sender cannot send any message, and the Receiver takes an action based on his prior. It is straightforward to show that the Sender enters the game if the Receiver's default action does not coincide with her preferred action. Otherwise, the Sender does not enter the game, but the Receiver always takes action $a = 1$, consistent with the standard persuasion paradigm.

6 Conclusion

In this paper, we analyze the role of probabilistic lie detection in a model of Bayesian persuasion between a Sender and a Receiver. We show that the Sender lies more when the lie detection probability increases. As long as the lie detection probability is sufficiently small, the Sender's and the Receiver's equilibrium payoffs are unaffected by lie detection technology because the Sender compensates by lying more. Once the lie detection probability is sufficiently high, the Sender can no longer maximally lie about the unfavorable state, and the Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability. Our model rationalizes that a sender of communication chooses to lie more frequently when it is more likely that her false statements will be flagged as lies.

These insights extend more generally and continue to hold under partial commitment for the Sender as well as in richer state, message, and action spaces, and under different detection technologies that the Receiver can use to inform his decision. Nonetheless, our analysis raises further questions about the role of lie detection under Bayesian persuasion and communication more broadly. For example, messages in our model are defined to have literal meanings, and thus they are classified as lies if they do not match the true state of nature. In other words, the definition of lies is exogenous. But what happens if messages do *not* have a literal meaning and are classified as lies if they induce an action that does not match the true state of nature? In that case, lies are necessarily endogenous and determined only in equilibrium which leaves further discretion as to what truly constitutes a lie. We also assumed that the probability of lie detection is exogenous, but what if this probability is instead a strategic choice of the Receiver or a third party? We leave these and other interesting questions to future research.

References

Allcott, Hunt and Matthew Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 2017, *31* (2), 211–236.

Aral, Sinan, *The hype machine: how social media disrupts our elections, our economy, and our health–and how we must adapt*, Currency, 2021.

Balbuzanov, Ivan, "Lies and Consequences: The Effect of Lie Detection on Communication Outcomes," *International Journal of Game Theory*, 2019, *48* (4), 1203–1240.

Ball, Ian and José Antonio Espín-Sánchez, "Experimental Persuasion," *Cowles Foundation Research Paper 2298*, 2022.

Crawford, Vincent P and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), 1431–1451.

Doval, Laura and Vasiliki Skreta, "Constrained Information Design," *arXiv preprint arXiv:1811.03588*, 2018.

Dziuda, Wioletta and Christian Salas, "Communication with Detectable Deceit," *SSRN Working Paper 3234695*, 2018.

Ederer, Florian and Ernst Fehr, "Deception and Incentives: How Dishonesty Undermines Effort Provision," *Yale SOM Working Paper*, 2017.

Fréchette, Guillaume R., Alessandro Lizzeri, and Jacopo Perego, "Rules and Commitment in Communication: An Experimental Analysis," *Econometrica*, 2022, *90* (5), 2283–2318.

Gehlbach, Scott, Zhaotian Luo, Anton Shirikov, and Dmitriy Vorobyev, "A Model of Censorship and Propaganda," *Working Paper*, 2022.

Gneezy, Uri, "Deception: The Role of Consequences," *American Economic Review*, 2005, *95* (1), 384–394.

Gneezy, Uri, Agne Kajackaite, and Joel Sobel, "Lying Aversion and the Size of the Lie," *American Economic Review*, 2018, *108* (2), 419–53.

Guo, Yingni and Eran Shmaya, "Costly Miscalibration," *Theoretical Economics*, 2021, *16* (2), 477–506.

Hurkens, Sjaak and Navin Kartik, "Would I Lie to You? On Social Preferences and Lying Aversion," *Experimental Economics*, 2009, *12* (2), 180–192.

Ivanov, Maxim, "Optimal Monotone Signals in Bayesian Persuasion Mechanisms," *Economic Theory*, 2021, *72* (3), 955–1000.

Jehiel, Philippe, "Communication with Forgetful Liars," *Theoretical Economics*, 2021, *16* (2), 605–638.

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

Kamenica, Emir, Kyungmin Kim, and Andriy Zapechelnyuk, "Bayesian Persuasion and Information Design: Perspectives and Open Issues," *Economic Theory*, 2021, *72*, 701–704.

Kao, Jeff and Priyanjana Bengani, "How Verified Accounts on X Thrive While Spreading Misinformation About the Israel-Hamas Conflict," *ProPublica*, 2023, *December* (20). Available at https://www.propublica.org/article/x-verified-accounts-misinformation-israel-hamas-conflict.

Kartik, Navin, "Strategic Communication with Lying Costs," *The Review of Economic Studies*, 2009, *76* (4), 1359–1395.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani, "Credulity, Lies, and Costly Talk," *Journal of Economic Theory*, 2007, *134* (1), 93–116.

Koessler, Frédéric and Vasiliki Skreta, "Informed information design," *Journal of Political Economy*, 2023, *131* (11), 3186–3232.

Le Treust, Maël and Tristan Tomala, "Persuasion with Limited Communication Capacity," *Journal of Economic Theory*, 2019, *184*, 104940.

Levkun, Aleksandr, "Communication with strategic fact-checking," *University of Vienna Working Paper*, 2022.

Liedke, Jacob and Luxuan Wang, "News Platform Fact Sheet," *Pew Research Center*, 2023, *11.*

Lin, Xiao and Ce Liu, "Credible Persuasion," *Working Paper*, 2022.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin, "Persuasion via Weak Institutions," *Journal of Political Economy*, 2022, *130* (10), 2705–2730.

Luo, Zhaotian and Arturas Rozenas, "Strategies of Election Rigging: Trade-offs, Determinants, and Consequences," *Quarterly Journal of Political Science*, 2018, *13* (1), 1–28.

Luo, Zhaotian and Arturas Rozenas, "Lying in Persuasion," *SSRN Working Paper*, 2021.

Matyskova, Ludmila and Alfonso Montes, "Bayesian Persuasion with Costly Information Acquisition," *Journal of Economic Theory*, 2023, *211*, 105678.

Min, Daehong, "Bayesian Persuasion under Partial Commitment," *Economic Theory*, 2021, *72*, 743–764.

Morris, Stephen, "Political Correctness," *Journal of Political Economy*, 2001, *109* (2), 231–265.

Naeem, Salman Bin, Rubina Bhatti, and Aqsa Khan, "An exploration of how fake news is taking over social media and putting public health at risk," *Health Information & Libraries Journal*, 2021, *38* (2), 143–149.

Nguyen, Anh and Teck Yong Tan, "Bayesian Persuasion with Costly Messages," *Journal of Economic Theory*, 2021, *193*, 105212.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J Wood, "Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability," *Political Behavior*, 2020, *42*, 939–960.

Perez-Richet, Eduardo and Vasiliki Skreta, "Test Design under Falsification," *Econometrica*, 2022, *90* (3), 1109–1142.

Rayo, Luis and Ilya Segal, "Optimal Information Disclosure," *Journal of Political Economy*, 2010, *118* (5), 949–987.

Sánchez-Pagés, Santiago and Marc Vorsatz, "Enjoy the Silence: An Experiment on Truth-telling," *Experimental Economics*, 2009, *12* (2), 220–241.

Simon-Kerr, Julia, "Unmasking Demeanor," *George Washington Law Review Arguendo*, 2020, *88*, 158.

Sobel, Joel, "Lying and Deception in Games," *Journal of Political Economy*, 2020, *128* (3), 907–947.

Titova, Maria, "Persuasion with Verifiable Information," *UCSD Working Paper*, 2021.

Tsakas, Elias and Nikolas Tsakas, "Noisy Persuasion," *Games and Economic Behavior*, 2021, *130*, 44–61.

Vrij, Aldert, Pär Anders Granhag, Samantha Mann, and Sharon Leal, "Outsmarting the liars: Toward a Cognitive Lie Detection Approach," *Current Directions in Psychological Science*, 2011, *20* (1), 28–32.

A Proofs

Since full commitment is a special case of partial commitment, Propositions 1, 2, and 4 follow immediately from Proposition 5 by setting $\alpha = 1$. So, it suffices to prove Proposition 5 only.

A.1 Proof of Proposition 5

The strategy space splits into four types, determined by whether $(1, \neg lie)$ and $(0, \neg lie)$ induce the Receiver to choose $a = 1$ or $a = 0$. Equivalently, the Receiver takes action $a = 1$ if and only if the posterior likelihood ratio is larger than a threshold $X := \frac{t(1-\mu)}{\mu(1-t)} (= \frac{1}{1-\bar{q}} > 1)$, where the two posterior likelihood ratios are given by

$$l_{1,\neg lie} = \frac{\alpha p_1 + (1-\alpha)\tilde{p}_1}{[\alpha(1-p_0) + (1-\alpha)(1-\tilde{p}_0)](1-q)},$$
$$l_{0,\neg lie} = \frac{[\alpha(1-p_1) + (1-\alpha)(1-\tilde{p}_1)](1-q)}{\alpha p_0 + (1-\alpha)\tilde{p}_0}.$$

For each type, we first identify the conditions under which it is feasible and then characterize the Sender's optimal strategy within that type, conditional on its existence. Finally, we compare across types and derive the globally optimal strategy for all parameters.

(1) Type I: $l_{1,\neg lie} < X$ and $l_{0,\neg lie} < X$. When the Sender's commitment does not bind, it is immediate that $\tilde{p}_1^* = 0$ and $\tilde{p}_0^* \in [0, 1]$. Since $(p_1, p_0, \tilde{p}_1, \tilde{p}_0) = (0, 1-q, 0, 1-q)$ satisfies the constraints, the type I strategy is always feasible. Within this type, the Sender's payoff is equal to $\mu q[\alpha(1-p_1) + 1 - \alpha]$, which is maximized at $p_1^* = 0$. In summary, the optimal strategy is not unique, and the players' equilibrium payoffs are

$$U_{\text{I}}^S = \mu q, \quad U_{\text{I}}^R = \mu q(1-t) + (1-\mu)t.$$

(2) Type II: $l_{1,\neg lie} < X$ and $l_{0,\neg lie} \geq X$. When the Sender's commitment does not bind, it is immediate that $\tilde{p}_1^* = 0$ and $\tilde{p}_0^* = 1$. Within this type, the Sender solves

$$\max_{p_0, p_1 \in [0,1]} \mu[\alpha(1-p_1) + 1 - \alpha] + (1-\mu)(\alpha p_0 + 1 - \alpha) \quad s.t. \quad l_{1,\neg lie} < X, \; l_{0,\neg lie} \geq X.$$

The problem is feasible if and only if $\alpha \geq 1 - \frac{1-q}{X} = 1 - (1-q)(1-\bar{q})$, since otherwise we would have $l_{0,\neg lie} < X$ for any $p_0, p_1 \in [0, 1]$. Given feasibility, the solution is given by $p_1^* = 0$

and $p_0^* = 1 - \frac{1-(1-q)(1-\overline{q})}{\alpha}$, yielding equilibrium payoffs

$$U^S_{\text{II}} = \mu + (1-\mu)(1-q)(1-\overline{q}), \quad U^R_{\text{II}} = \mu(1-t) + (1-\mu)t\left[1-(1-q)(1-\overline{q})\right].$$

(3) Type III: $l_{1,\neg lie} \geq X$ and $l_{0,\neg lie} < X$. When the Sender's commitment does not bind, it is immediate that $\tilde{p}_1^* = 1$ and $\tilde{p}_0^* = 0$. Within this type, the Sender solves

$$\max_{p_0,p_1 \in [0,1]} \mu[\alpha[p_1+(1-p_1)q]+1-\alpha]+(1-\mu)[\alpha(1-p_0)+1-\alpha](1-q) \quad s.t. \quad l_{1,\neg lie} \geq X, \; l_{0,\neg lie} < X.$$

The problem is feasible if and only if $\alpha \geq \underline{\alpha}(q) := \frac{\overline{q}-q}{1-q}$, since otherwise we would have $l_{1,\neg lie} < X$ for any $p_0, p_1 \in [0,1]$. Given feasibility, the solution is given by $p_1^* = 1$ and $p_0^* = \max\left\{0, \frac{\overline{q}-q}{\alpha(1-q)}\right\}$, yielding equilibrium payoffs

$$U^S_{\text{III}} = \begin{cases} \mu + (1-\mu)(1-q), & \text{if } q > \overline{q}, \\ \frac{\mu}{t}, & \text{if } q \leq \overline{q}, \end{cases} \quad \text{and} \quad U^R_{\text{III}} = \begin{cases} \mu(1-t) + (1-\mu)tq, & \text{if } q > \overline{q}, \\ (1-\mu)t, & \text{if } q \leq \overline{q}. \end{cases}$$

(4) Type IV: $l_{1,\neg lie} \geq X$ and $l_{0,\neg lie} \geq X$. When the Sender's commitment does not bind, it is immediate that $\tilde{p}_1^* \in [0,1]$ and $\tilde{p}_0^* = 1$. Within this type, the Sender solves

$$\max_{p_0,p_1,\tilde{p}_1 \in [0,1]} \mu + (1-\mu)[\alpha(p_0 + (1-p_0)(1-q)) + 1 - \alpha] \quad s.t. \quad l_{1,\neg lie} \geq X, \; l_{0,\neg lie} \geq X.$$

The problem is feasible if the two constraints hold simultaneously, which implies that

$$\begin{aligned} \alpha p_1 + (1-\alpha)\tilde{p}_1 &\geq \alpha(1-p_0)(1-q)X, \\ \alpha(1-p_1) + (1-\alpha)(1-\tilde{p}_1) &\geq \frac{1}{1-q}(\alpha p_0 + 1 - \alpha)X. \end{aligned}$$

Summing over the two inequalities yields

$$1 \geq \left[\alpha(1-p_0)(1-q) + \frac{\alpha p_0 + 1 - \alpha}{1-q}\right] \cdot X \geq \left[\alpha(1-q) + \frac{1-\alpha}{1-q}\right] \cdot X,$$

which implies $\alpha \geq \overline{\alpha}(q) := \frac{1-(1-q)(1-\overline{q})}{2q-q^2}$. Given feasibility, the solution solves $\alpha p_1^* + (1-\alpha)\tilde{p}_1^* = \overline{\alpha}(q)\frac{1-q}{1-\overline{q}}$ and $p_0^* = 1 - \frac{\overline{\alpha}(q)}{\alpha}$, which renders both constraints binding. This strategy yields equilibrium payoffs

$$U^S_{\mathrm{IV}} = \mu + (1-\mu)(1-q)\frac{2-\overline{q}}{2-q}, \quad U^R_{\mathrm{IV}} = \frac{\mu(1-t)+(1-\mu)t}{2-q}.$$

By comparing the Sender's payoffs U^S_i, $i \in \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}\}$ together with the feasibility conditions for each strategy type, we obtain the following characterization.

(a) If $q \le \overline{q}$ and $\alpha < \underline{\alpha}(q)$, then only type I strategy is feasible. The optimal strategy satisfies

$$p^*_1 = 0, \ \tilde{p}^*_1 = 0, \ \text{and } \mathbb{E}[p^*_0] > (1-q)(1-\overline{q}).$$

Obviously, $\mathbb{E}[p^*_1] = 0$ is constant in q. Moreover, both $U_S = \mu q$ and $U_R = \mu q(1-t) + (1-\mu)t$ strictly increase in q.

(b) If $q \le \overline{q}$ and $\alpha \ge \underline{\alpha}(q)$, then type IV strategy is infeasible. The optimal strategy is a type III strategy that satisfies

$$p^*_1 = 1, \ \tilde{p}^*_1 = 1, \ \tilde{p}^*_0 = 0, \ \text{and } p^*_0 = \frac{\overline{q}-q}{\alpha(1-q)}.$$

As a result, $E[p^*_0] = \alpha p^*_0$ strictly decreases in q, while $E[p^*_1] = 1$ is constant in q. Moreover, both $U_S = \frac{\mu}{t}$ and $U_R = (1-\mu)t$ are constant in q.

(c) If $q > \overline{q}$ and $\alpha < \overline{\alpha}(q)$, then type IV strategy is still infeasible. The optimal strategy is a type III strategy that satisfies

$$p^*_1 = 1, \ \tilde{p}^*_1 = 1, \ \tilde{p}^*_0 = 0, \ \text{and } p^*_0 = 0.$$

Clearly, both $E[p^*_0] = 1$ and $E[p^*_1] = 0$ are constant in q. Moreover, $U_S = \mu + (1-\mu)(1-q)$ strictly decreases in q while $U_R = \mu(1-t) + (1-\mu)tq$ strictly increases in q.

(d) If $q > \overline{q}$ and $\alpha \ge \overline{\alpha}(q)$, then type IV strategy is feasible. The optimal strategy is a type IV strategy that satisfies

$$\tilde{p}^*_0 = 1, \ p^*_0 = 1 - \frac{\overline{\alpha}(q)}{\alpha}, \ \text{and } \mathbb{E}[p^*_1] = \overline{\alpha}(q)\frac{1-q}{1-\overline{q}}.$$

Clearly, $U_S = \mu + (1-\mu)(1-q)\frac{2-\overline{q}}{2-q}$ strictly decreases in q while $U_R = \frac{\mu(1-t)+t(1-\mu)}{2-q}$ strictly increases in q. We show below that $E[p_0^*] = 1-\overline{\alpha}(q)$ first strictly increases and then decreases in q, while $E[p_1^*]$ strictly decreases in q:

$$\begin{aligned}
\frac{\partial E[p_1^*]}{\partial q} &= \frac{-q^2 - \frac{\overline{q}}{1-\overline{q}}(2-2q+q^2)}{(2-q)^2q^2} < 0, \forall q \in (\overline{q}, 1], \\
\frac{\partial E[p_0^*]}{\partial q} &= \frac{-q^2 + \overline{q}(2-2q+q^2)}{(2-q)^2q^2} \geq 0 \Longleftrightarrow \frac{1}{\overline{q}} \leq 1 + \frac{2-2q}{q^2}.
\end{aligned}$$

Because $\frac{2-2q}{q^2}$ decreases in q, the sign of the derivative $\frac{\partial E[p_0^*]}{\partial q}$ changes at most one time. Since the derivative is positive at $q = \overline{q}$ and negative at $q = 1$, $E[p_0^*]$ must be initially increasing and then decreasing in q over $(\overline{q}, 1]$.

A.2 Proof of Proposition 6

For $q \leq \overline{q}$, the Sender's payoff is an increasing step function in α while the Receiver's payoff is a decreasing step function in α:

$$U^S(q,\alpha) = \begin{cases} \mu q, & \text{if } \alpha < \underline{\alpha}(q), \\ \frac{\mu}{t}, & \text{if } \alpha \geq \underline{\alpha}(q), \end{cases} \qquad U^R(q,\alpha) = \begin{cases} \mu q(1-t) + (1-\mu)t, & \text{if } \alpha < \underline{\alpha}(q), \\ (1-\mu)t, & \text{if } \alpha \geq \underline{\alpha}(q). \end{cases}$$

For $q > \overline{q}$, the Sender's payoff is again an increasing step function in α while the Receiver's payoff is a decreasing step function in α:

$$U^S(q,\alpha) = \begin{cases} \mu + (1-\mu)(1-q), & \text{if } \alpha < \overline{\alpha}(q), \\ \mu + (1-\mu)(1-q)\frac{2-\overline{q}}{2-q}, & \text{if } \alpha \geq \overline{\alpha}(q), \end{cases} \qquad U^R(q,\alpha) = \begin{cases} \mu(1-t) + (1-\mu)tq, & \text{if } \alpha < \overline{\alpha}(q), \\ \frac{\mu(1-t)+t(1-\mu)}{2-q}, & \text{if } \alpha \geq \overline{\alpha}(q). \end{cases}$$

A.3 Proof of Proposition 3

To simplify notations, let $(p_0, p_1) = (p_0^*(q), p_1^*(q))$ and $(p_0', p_1') = (p_0^*(q'), p_1^*(q'))$.

(*a*) Substituting the expressions for the optimal messaging strategy, we obtain

$$\mathcal{E}(p_0, p_1) = \begin{bmatrix} \frac{\overline{q}-q}{1-q} & 0 \\ \frac{1-\overline{q}}{1-q} & 1 \end{bmatrix} \quad \text{and} \quad \mathcal{E}(p_0', p_1') = \begin{bmatrix} \frac{\overline{q}-q'}{1-q'} & 0 \\ \frac{1-\overline{q}'}{1-q'} & 1 \end{bmatrix}$$

So, $\mathcal{E}(p_0', p_1')$ Blackwell dominates $\mathcal{E}(p_0, p_1)$ as there exists $x = \frac{p_0}{p_0'} \in [0,1], y = 1$ such that

$$\mathcal{E}(p_0, p_1) = \begin{bmatrix} x & 1-y \\ 1-x & y \end{bmatrix} \mathcal{E}(p_0', p_1')$$

(b) Define

$$x = \frac{p_0' p_1 + (1-p_1')(p_0 - 1)}{p_0 + p_1 - 1} \quad \text{and} \quad y = \frac{(1-p_0')(p_1 - 1) + p_1' p_0}{p_0 + p_1 - 1}$$

It can be verified that

$$\begin{bmatrix} p_0' & 1-p_1' \\ 1-p_0' & p_1' \end{bmatrix} = \begin{bmatrix} x & 1-y \\ 1-x & y \end{bmatrix} \begin{bmatrix} p_0 & 1-p_1 \\ 1-p_0 & p_1 \end{bmatrix}$$

So, $\mathcal{E}(p_0, p_1)$ Blackwell dominates $\mathcal{E}(p_0', p_1')$ if $x, y \in [0,1]$. To this end, first notice that for $q > \overline{q}$,

$$p_1 + p_0 = \frac{1-q}{(2-q)q}\left(2q + \frac{\overline{q}^2}{1-\overline{q}}\right) < \frac{1-q}{(2-q)q}\left(2q + \frac{q^2}{1-q}\right) = 1$$

Moreover,

$$\frac{p_0}{1-p_1} = (1-q)(1-\bar{q}), \quad \frac{p_0'}{1-p_1'} = (1-q')(1-\bar{q}), \quad \frac{1-p_0}{p_1} = \frac{1-\bar{q}}{1-q}, \quad \frac{1-p_0'}{p_1'} = \frac{1-\bar{q}}{1-q'}$$

Thus,

$$\begin{aligned}
x \geq 0 &\iff \frac{p_0'}{1-p_1'} \leq \frac{1-p_0}{p_1} \iff (1-\bar{q})(1-q') \leq (1-\bar{q})\frac{1}{1-q} \\
x \leq 1 &\iff \frac{1-p_0'}{p_1'} \leq \frac{1-p_0}{p_1} \iff \frac{1-\bar{q}}{1-q'} \leq \frac{1-\bar{q}}{1-q} \\
y \geq 0 &\iff \frac{1-p_0'}{p_1'} \geq \frac{p_0}{1-p_1} \iff \frac{1-\bar{q}}{1-q'} \geq (1-q)(1-\bar{q}) \\
y \leq 1 &\iff \frac{p_0}{1-p_1} \leq \frac{p_0'}{1-p_1'} \iff (1-q)(1-\bar{q}) \leq (1-q')(1-\bar{q})
\end{aligned}$$

Obviously, all four inequalities hold because $\overline{q} \leq q' < q \leq 1$.

(c) It suffices to prove the result in two subcases: $0 \leq q' < q \leq \overline{q}$, and $\overline{q} \leq q' < q \leq 1$. The remaining case, where $0 \leq q' < \overline{q} < q \leq 1$, follows from these two subcases by the transitivity of Blackwell dominance.

(i) Suppose $0 \leq q' < q \leq \overline{q}$. Substituting the equilibrium messaging strategy yields

$$\Gamma(p_0, p_1) = \begin{bmatrix} \frac{\overline{q}-q}{1-q} & 0 \\ 0 & 0 \\ 1-\overline{q} & 1 \\ \frac{(1-\overline{q})q}{1-q} & 0 \end{bmatrix}.$$

By construction,

$$\Gamma(p_0', p_1') = \begin{bmatrix} x & 0 & 0 & x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1-x & 0 & 0 & 1-x \end{bmatrix} \cdot \Gamma(p_0, p_1), \text{ where } x = \frac{\overline{q}-q'}{\overline{q}(1-q')} \in [0,1].$$

Hence, 4×4 matrix is a Markovian matrix, and $\Gamma(p_0, p_1)$ Blackwell dominates $\Gamma(p_0', p_1')$.

(ii) Suppose $\overline{q} \leq q' < q \leq 1$. Substituting the equilibrium messaging strategy yields

$$\Gamma(p_0, p_1) = \begin{bmatrix} \frac{(1-q)(q-\overline{q})}{q(2-q)} & \frac{(1-q)(q-\overline{q})}{q(2-q)(1-\overline{q})} \\ 0 & \frac{q-\overline{q}}{(2-q)(1-\overline{q})} \\ \frac{(1-q)[q+(1-q)\overline{q}]}{q(2-q)} & \frac{(1-q)[q+(1-q)\overline{q}]}{q(2-q)(1-\overline{q})} \\ \frac{q+(1-q)\overline{q}}{2-q} & 0 \end{bmatrix}.$$

By construction,

$$\Gamma(p_0', p_1') = \begin{bmatrix} 1-y & (1-y)(1-z) & 1-y & (1-y)(1-x) \\ 0 & z & 0 & 0 \\ y & y(1-z) & y & y(1-x) \\ 0 & 0 & 0 & x \end{bmatrix} \cdot \Gamma(p_0, p_1),$$

where

$$x = \frac{[q' + (1-q')\overline{q}](2-q)}{[q + (1-q)\overline{q}](2-q')}, \quad y = \frac{q' + (1-q')\overline{q}}{(2-\overline{q})q'}, \quad z = \frac{(q'-\overline{q})(2-q)}{(q-\overline{q})(2-q')}.$$

By assumption, $x, y, z \in [0,1]$. Thus, the 4×4 matrix is a Markovian matrix, implying that $\Gamma(p_0, p_1)$ Blackwell dominates $\Gamma(p'_0, p'_1)$.

A.4 Proof of Proposition 7

The proof is decomposed into three steps. First, we construct a Sender-optimal strategy under which the Receiver obtains the highest payoff among all Sender-optimal strategies. Then, we construct an alternative Sender's strategy that achieves the same payoff pair for any $q \in [0,1]$. Finally, we argue that for any $q \in [0,1]$ and $i \in \{S, R\}$, $U_i(q) = U_i(0)$, concluding the proof.

Step 1: For $k \in \{1, ..., N\}$ and $j \in \{1, ..., N-1\}$, define

$$t_k = \frac{\sum_{i=k}^{N} \lambda_i \omega_i}{\sum_{i=k}^{N} \lambda_i} \quad \text{and} \quad \tilde{t}_j = \frac{\sum_{i=2}^{j+1} \lambda_i \omega_i}{\sum_{i=2}^{j+1} \lambda_i}$$

These thresholds are ranked in the following way.

$$\omega_2 = \tilde{t}_1 < ... < \tilde{t}_{N-1} = t_2 < ... < t_N = \omega_N$$

When $q = 0$ and $t \in [t_k, t_{k+1})$, $\forall k \in \{1, ..., N-1\}$, Ivanov (2021) implies that the following partitional strategy σ^* is optimal for the Sender.

$$\sigma^*(\omega_i) = 1, \quad \text{if } i > k; \quad \sigma^*(\omega_k) = \begin{cases} 1, & w.p. \quad s_k \\ 0, & w.p. \quad 1 - s_k \end{cases} ; \quad \sigma^*(\omega_i) = 0, \quad \text{if } i < k$$

where the mixing probability s_k solves

$$\frac{\sum_{j=k+1}^{N} \lambda_j \omega_j + s_k \lambda_k \omega_k}{\sum_{j=k+1}^{N} \lambda_j + s_k \lambda_k} = t. \tag{1}$$

The Sender's optimal payoff is given by

$$U_S(0) = \frac{\sum_{i=k+1}^{N} \lambda_i (\omega_i - \omega_k)}{t - \omega_k}$$

We then show that, within all Sender-optimal strategies, σ^* also maximizes the Receiver's payoff. The Receiver's payoff in any Sender-optimal strategy $\sigma \in \Sigma^*$ satisfies

$$\begin{aligned}
U_R(\sigma; 0) =& \sum_{i=1}^{N} (\omega_i - t) \text{Pr}^{\sigma}(a = 1, \omega = \omega_i) + \sum_{\omega_i < t} (t - \omega_i) \lambda_i \\
=& \sum_{i=1}^{k-1} (\omega_i - \omega_k) \text{Pr}^{\sigma}(a = 1, \omega = \omega_i) + \sum_{i=k+1}^{N} (\omega_i - \omega_k) \text{Pr}^{\sigma}(a = 1, \omega = \omega_i) \\
&+ (\omega_k - t) \sum_{i=1}^{N} \text{Pr}^{\sigma}(a = 1, \omega = \omega_i) + \sum_{\omega_i < t} (t - \omega_i) \lambda_i \\
\leq& \sum_{i=k+1}^{N} (\omega_i - \omega_k) \text{Pr}^{\sigma}(\omega = \omega_i) + (\omega_k - t) \cdot U_S(0) + \sum_{\omega_i < t} (t - \omega_i) \lambda_i \\
=& \sum_{\omega_i < t} (t - \omega_i) \lambda_i
\end{aligned}$$

The inequality binds if and only if the Receiver always takes action $a = 1$ for $\omega \geq \omega_{k+1}$ and always takes action $a = 0$ for $\omega \leq \omega_{k-1}$, which is exactly achieved by the strategy σ^*.

Step 2: We show that there always exists a strategy σ such that $U_i(\sigma; q) = U_i(0)$, $i \in \{S, R\}$, $\forall q \in [0, 1]$. Consider three possible scenarios.

(a) If $t \in (\mu, \tilde{t}_1)$, the following strategy σ_1 is qualified because under both σ^* and σ_1, the Receiver takes $a = 1$ with probability one if $\omega \geq \omega_2$ and with probability s_1 if $\omega = \omega_1$. Moreover, σ_1

does not depend on q:

$$\sigma_1(\omega_i) = \omega_2, \quad \text{if } i > 2; \quad \sigma_1(\omega_2) = 1; \quad \sigma_1(\omega_1) = \begin{cases} \omega_2, & w.p. \quad u \\ 1, & w.p. \quad s_1 - u \\ 0, & w.p. \quad 1 - s_1 \end{cases}$$

where s_1 solves Equation (1) when $k = 1$ and $u \in [0, s_1]$ solves

$$\frac{\lambda_2 \omega_2}{\lambda_2 + (s_1 - u)\lambda_1} = t$$

(b) If $t \in [\tilde{t}_{k-1}, \tilde{t}_k)$ for some $k \in \{2, ..., N-1\}$, the following strategy σ_2 is qualified because under both σ^* and σ_2, the Receiver takes $a = 1$ with probability one if $\omega \geq \omega_2$ and with probability s_1 if $\omega = \omega_1$. Moreover, σ_2 does not depend on q:

$$\sigma_2(\omega_i) = \omega_k, \quad \text{if } i = k+1, ..., N; \qquad \sigma_2(\omega_i) = 0, \quad \text{if } i = 2, ..., k;$$

$$\sigma_2(\omega_k) = \begin{cases} \omega_k, & w.p. \quad u \\ 0, & w.p. \quad 1 - u \end{cases}; \qquad \sigma_2(\omega_1) = \begin{cases} \omega_k, & w.p. \quad s_1 \\ 1, & w.p. \quad 1 - s_1 \end{cases}$$

where s_1 solves Equation (1) when $k = 1$ and $u \in [0, 1]$ solves

$$\frac{\sum_{j=2}^{k} \lambda_j \omega_j + (1-u)\lambda_{k+1}\omega_{k+1}}{\sum_{j=2}^{k} \lambda_j + (1-u)\lambda_{k+1}} = t.$$

(c) If $t \in [t_k, t_{k+1})$ for some $k \in \{2, ..., N-1\}$, the following strategy σ_3 is qualified because under both σ^* and σ_3, the Receiver takes $a = 1$ with probability one if $\omega > \omega_k$, with probability s_k if $\omega = \omega_k$, and with probability zero if $\omega < \omega_k$. Moreover, σ_3 does not depend on q:

$$\sigma_3(\omega_i) = 0, \quad \text{if } i > k; \quad \sigma_3(\omega_k) = \begin{cases} 0, & w.p. \quad s_k \\ 1, & w.p. \quad 1 - s_k \end{cases}; \quad \sigma_3(\omega_i) = 1, \quad \text{if } i = 2, ..., k$$

where s_k solves Equation (1).

Step 3: Since lie detection restricts the Sender's strategy space and thus the set of induced distribution of posteriors, it follows that $U_S(q) \leq U_S(0)$ for any $q \in [0,1]$. Combined with Step 2, this implies that $U_S(q) = U_S(0)$ for any $q \in [0,1]$. Moreover, it cannot be the case that $U_R(q) > U_R(0)$. Otherwise, by incorporating the additional information from the lie detection into the strategy, the Sender could achieve a payoff pair $(U_S(0), U_R(q))$ when $q = 0$, contradicting with Step 1. Consequently, $U_R(q) \leq U_R(0)$ for any $q \in [0,1]$. Combined with Step 2, this implies that $U_R(q) = U_R(0)$ for any $q \in [0,1]$.

A.5 Proof of Proposition 8

(a) Suppose $q = 0$, then by the concavification method, there exists a Sender-optimal strategy (p_0, p_1) that splits the prior into 0 and t_i for some $t_i > \bar{\mu}$. In particular, $p_1 = 1$ and p_0 solves

$$\frac{\bar{\mu}}{\bar{\mu} + (1-\bar{\mu})(1-p_0)} = t_i$$

In this case, a small $q > 0$ does not affect the Sender's payoff since she induces the same distribution of posteriors and thus obtains the same payoff from the strategy $(\hat{p}_0, \hat{p}_1)$ such that $\hat{p}_1 = 1$ and $1 - \hat{p}_0 = (1-q)(1-p_0)$.

(b) Suppose $q = 0$, then by the concavification method, in any Sender-optimal strategy, neither of the posteriors is 0, which implies that

$$U_S(p_0, 1; 0) < U_S(0), \ \forall p_0 \in [0,1] \tag{2}$$

For any $q > 0$, we show that $U_S(p_0, p_1; q) < U_S(0)$ for any $(p_0, p_1) \in [0,1]^2$. The arguments are different depending on whether $p_1 = 1$ or $p_1 < 1$.

(1) Consider an arbitrary strategy (p_0, p_1) such that $p_1 = 1$. Then by Bayes' rule,

$$\mu_{1,lie} = \mu_{0,\neg lie} = 0, \quad \mu_{0,lie} = 1, \quad \mu_{1,\neg lie} = \frac{\bar{\mu}}{\bar{\mu} + (1-\bar{\mu})(1-p_0)(1-q)}$$

Consequently, the Sender's payoff is

$$U_S(p_0, 1; q) = [\bar{\mu} + (1-\bar{\mu})(1-p_0)(1-q)] \cdot a^*(\mu_{1,\neg lie})$$

Consider (p'_0, p'_1) such that $p'_1 = 1$ and $1 - p'_0 = (1 - p_0)(1 - q)$. When $q = 0$, this strategy induces a pair of posteriors (μ_0, μ_1). By construction, $\mu_0 = \mu_{0,\neg lie} = 0$ and $\mu_1 = \mu_{1,\neg lie}$. It follows that

$$U_S(p'_0, 1; 0) = [\bar{\mu} + (1 - \bar{\mu})(1 - p_0)(1 - q)] \cdot a^*(\mu_{1,\neg lie}) = U_S(p_0, 1; q)$$

Finally, by Inequality (2), for any $(p_0, p_1) \in [0,1]^2$ such that $p_1 = 1$,

$$U_S(p_0, p_1; q) < U_S(0)$$

(2) Consider an arbitrary strategy (p_0, p_1) such that $p_1 < 1$. Then by Bayes' rule, $\mu_{0,lie} = 1$ and it is generated with positive probability. Notice that an unconstrained persuasion problem with a binary message space and lie detection can be alternatively viewed as a constrained persuasion problem with a larger message space and no lie detection. Thus, we introduce an auxiliary persuasion problem where there is no lie detection but the message space is enriched to $\tilde{M} = \{0,1\}^2$. The Sender's strategy space is denoted by $\Sigma = \{\sigma : \{0,1\} \to \triangle(\tilde{M})\}$, where each strategy σ induces a distribution of posteriors as follows. For $m \in \tilde{M}$,

$$\Pr(\mu = \mu_m^\sigma) = \lambda_m^\sigma$$

We further restrict attention to $\hat{\Sigma} = \{\sigma \in \Sigma \,|\, \mu^\sigma_{(0,1)} = 1,\ \lambda^\sigma_{(0,1)} > 0\}$. Then by definition,

$$U_S(p_0, p_1; q) \leq \max_{\sigma \in \hat{\Sigma}} U_S(\sigma; 0) \tag{3}$$

Moreover, it must be that

$$\max_{\sigma \in \hat{\Sigma}} U_S(\sigma; 0) < \max_{\sigma \in \Sigma} U_S(\sigma; 0) \tag{4}$$

To this end, suppose a strategy σ^* maximizes $U_S(\sigma; 0)$ within $\hat{\Sigma}$. Then consider another strategy $\tilde{\sigma} \in \Sigma$ such that

$$\mu_m^{\tilde{\sigma}} = \mu_m^{\sigma^*} \quad \text{and} \quad \lambda_m^{\tilde{\sigma}} = \lambda_m^{\sigma^*} \quad \text{for } m \neq (0,1)$$
$$\mu_m^{\tilde{\sigma}} = t_{N-1} \quad \text{and} \quad \lambda_m^{\tilde{\sigma}} = \frac{\lambda_m^{\sigma^*}}{t_{N-1}} \quad \text{for } m = (0,1)$$

Such a strategy $\tilde{\sigma}$ always exists by Bayes plausibility and $\bar{\mu} < t_{N-1}$. Since $a^*(\mu_{(0,1)}^{\tilde{\sigma}}) = a^*(\mu_{(0,1)}^{\sigma^*}) = A_n$ and $\lambda_m^{\tilde{\sigma}} > \lambda_m^{\sigma^*}$, it follows that the Sender obtains a strictly higher payoff under $\tilde{\sigma}$ than under σ^*.

$$U_S(\sigma^*; 0) = \sum_{m \in \tilde{M}} \lambda_m^{\sigma^*} a^*(\mu_m^{\sigma^*}) < \sum_{m \neq (0,1)} \lambda_m^{\sigma^*} a^*(\mu_m^{\sigma^*}) + \lambda_{(0,1)}^{\tilde{\sigma}} a^*(\mu_{(0,1)}^{\tilde{\sigma}}) = U_S(\tilde{\sigma}; 0)$$

Finally, observe that in the absence of lie detection, there always exists a Sender-optimal strategy that splits the prior into two posteriors. So, the Sender does not benefit from a larger message space, i.e.,

$$\max_{\sigma \in \Sigma} U_S(\sigma; 0) = \max_{(p_0, p_1) \in [0,1]^2} U_S(p_0, p_1; 0) \tag{5}$$

Hence, by Inequalities (3)-(5), for any $(p_0, p_1) \in [0,1]^2$ such that $p_1 < 1$,

$$U_S(p_0, p_1; q) < \max_{(p_0, p_1) \in [0,1]^2} U_S(p_0, p_1; 0) \equiv U_S(0)$$

concluding the proof.

A.6 Proof of Proposition 9

The sufficiency is trivial. So, we focus on the necessity part and show that $U_S(q, r) < U_S(0, 0)$ whenever $0 < r < q \leq 1$. By Bayes' rule, the posterior after observing an event (m, d) is given by

$$\mu_{m,d} = \frac{\mu \cdot \Pr(m, d | \omega = 1)}{\mu \cdot \Pr(m, d | \omega = 1) + (1 - \mu) \cdot \Pr(m, d | \omega = 0)}.$$

Since $q > r$, it follows that $\frac{q}{r} > \frac{1-q}{1-r}$ and

$$\mu_{0,lie} = \frac{\mu(1 - p_1)q}{\mu(1 - p_1)q + (1 - \mu)p_0 r} \geq \frac{\mu(1 - p_1)(1 - q)}{\mu(1 - p_1)(1 - q) + (1 - \mu)p_0(1 - r)} = \mu_{0,\neg lie}. \tag{6}$$

Moreover, the inequality is strict if $p_1 \neq 1$ and $p_0 \neq 0$. Similarly,

$$\mu_{1,lie} = \frac{\mu p_1 r}{\mu p_1 r + (1-\mu)(1-p_0)q} \leq \frac{\mu p_1(1-r)}{\mu p_1(1-r) + (1-\mu)(1-p_0)(1-q)} = \mu_{1,\neg lie}. \tag{7}$$

where the inequality is strict if $p_1 \neq 0$ and $p_0 \neq 1$. Recall that the Sender's payoff $U_S(q, r)$ is upper bounded by the benchmark payoff $U_S(0,0) = \frac{\mu}{t}$. Moreover, the upper bound is attained if and only if for $\forall(m,d) \in E$, either $\Pr(m,d,\omega = 1) = 0$ or $\mu_{m,d} = t$. However, we show below that this is impossible. To this end, consider three types of Sender's strategies.

(1) If $p_1 = 1$, then $\Pr(m = 0, d = \neg lie, \omega = 1) = \Pr(m = 0, d = lie, \omega = 1) = 0$ and $\Pr(m = 1, d = \neg lie, \omega = 1)$, $\Pr(m = 1, d = lie, \omega = 1) > 0$. But then by Equation (7), it is impossible that $\mu_{1,lie} = \mu_{1,\neg lie} = t$.

(2) If $p_1 = 0$, then $\Pr(m = 1, d = \neg lie, \omega = 1) = \Pr(m = 1, d = lie, \omega = 1) = 0$ and $\Pr(m = 0, d = \neg lie, \omega = 1)$, $\Pr(m = 0, d = lie, \omega = 1) > 0$. But then by Equation (6), it is impossible that $\mu_{0,lie} = \mu_{0,\neg lie} = t$.

(3) If $p_1 \in (0,1)$, then $\Pr(m,d,\omega = 1) > 0$ for any $(m,d) \in E$. Again, by Equation (6) and (7), it is impossible that $\mu_{m,d} = t$ for any $(m,d) \in E$.

In summary, the benchmark payoff is never attainable if $0 < r < q \leq 1$.

A.7 Proof of Proposition 10

Let $E_1 \subset E$ be the set of events (m,d) such that the Receiver responds by taking action $a = 1$, or alternatively, $\mu_{m,d} \geq t$. As in the baseline model, we analogously partition the Sender's strategy space according to E_1. By Inequality (6), (7) and $q < \overline{q}$, there are seven cases. We solve the Sender's optimal payoff in each case and then pick the highest one when r is sufficiently small.

(I) $E_1 = \emptyset$. In this case, $U_S^{\mathrm{I}} = 0$, which is clearly not globally optimal.

(II) $E_1 = \{(1, \neg lie)\}$. In this case, by usual arguments, it is optimal to set $p_1 = 1$ and $\mu_{1,\neg lie} = t$, which gives rise to the payoff $U_S^{\mathrm{II}} = \frac{\mu(1-r)}{t}$.

(III) $E_1 = \{(0, lie)\}$. In this case, by usual arguments, it is optimal to set $p_1 = 0$ and $\mu_{0,lie} = t$, which gives rise to the payoff $U_S^{\mathrm{III}} = \frac{\mu q}{t} < U_S^{\mathrm{II}}$ for sufficiently small r.

(IV) $E_1 = \{(0, lie), (1, \neg lie)\}$. In this case, $\mu_{0,lie} \geq t$ and $\mu_{1,\neg lie} \geq t$. Thus, $U_S^{\text{IV}} = \Pr(0, lie) + \Pr(1, \neg lie) \leq \frac{\Pr(0,lie,\omega=1)+\Pr(1,\neg lie,\omega=1)}{t} = \frac{\mu[(1-p_1)q+p_1(1-r)]}{t} < U_S^{\text{II}}$ for sufficiently small r.

(V) $E_1 = \{(0, lie), (0, \neg lie)\}$. In this case, by usual arguments, it is optimal to set $p_1 = 0$ and $\mu_{0,\neg lie} = t$, which gives rise to the payoff $U_S^{\text{V}} = \mu + \frac{\mu(1-q)(1-t)}{(1-r)t} < U_S^{\text{II}}$ for sufficiently small r.

(VI) $E_1 = \{(1, lie), (1, \neg lie)\}$. In this case, by usual arguments, it is optimal to set $p_1 = 1$ and $\mu_{1,lie} = t$, which gives rise to the payoff $U_S^{\text{VI}} = \mu + \frac{\mu r(1-t)}{qt} < U_S^{\text{II}}$ for sufficiently small r.

(VII) $E_1 = \{(0, lie), (1, \neg lie), (1, lie)\}$. In this case, by usual arguments, it is optimal to set $\mu_{1,lie} = \mu_{0,lie} = t$, which yields the strategy

$$p_0^{\text{VII}} = \frac{q^2 - qr(1-\overline{q})}{q^2 - r^2} \quad \text{and} \quad p_1^{\text{VII}} = \frac{q^2 - \frac{qr}{1-\overline{q}}}{q^2 - r^2}.$$

As r goes to zero, p_0^{VII}, $p_1^{\text{VII}} \to 1$, and $U_S^{\text{VII}} = \mu[p_1^{\text{VII}} + (1-p_1^{\text{VII}})q] + (1-\mu)[1-p_0^{\text{VII}} + p_0^{\text{VII}} r] \to \mu$. Thus, $U_S^{\text{VII}} < U_S^{\text{II}}$ for sufficiently small r.

In summary, when r is sufficiently small, it is optimal to choose a type II strategy such that $p_1^* = 1$ and $p_0^* = 1 - \frac{(1-r)(1-\overline{q})}{1-q}$. Consequently,

$$U_S(q,r) = \frac{\mu(1-r)}{t},\ U_R(q,r) = \mu(1-t)p_1^*(1-r) + t(1-\mu)[1-(1-p_0^*)(1-q)] = t(1-\mu).$$