NBER WORKING PAPER SERIES

BAYESIAN PERSUASION WITH LIE DETECTION

Florian Ederer
Weicheng Min

Working Paper 30065
http://www.nber.org/papers/w30065

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2022, Revised February 2024

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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Bayesian Persuasion with Lie Detection
Florian Ederer and Weicheng Min
NBER Working Paper No. 30065
May 2022, Revised February 2024
JEL No. D72,D82,D83,K40,M31

ABSTRACT

How does lie detection constrain the potential for one person to persuade another to change her action?We consider a model of Bayesian persuasion in which the Receiver can detect lies with positive probability. We show that the Sender lies more when the lie detection probability increases. As long as 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. However, when the lie detection probability is sufficiently high, the Sender's equilibrium payoff decreases and the Receiver's equilibrium payoff increases with the lie detection probability. We explore several extensions including partial commitment, general state and action spaces, and different detection technologies and show that our model's main insights continue to hold.

Florian Ederer
Questrom School of Business
Boston University
595 Commonwealth Avenie
Boston, MA 02215
and NBER
florian.ederer@gmail.com

Weicheng Min
Yale University
Department of Economics
28 Hillhouse Avenue
New Haven, CT 06511
minweicheng1994@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).

By incorporating probabilistic lie detection in a model of Bayesian persuasion (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011) this paper shows that a sender of communication may optimally choose to lie more frequently when it is more likely that his false statements will be flagged as lies and that this behavior renders lie detection ineffective for the receiver for a large set of parameter values. The central innovation of our model—that lies are sometimes detectable—is a natural assumption for many applications of (Bayesian) persuasion including political campaigns,

[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 this behavior.

[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.

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 causes the Sender to lie more frequently in the favorable state of nature, and the Sender's (Receiver's) equilibrium payoff decreases (increases) with this 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

[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).

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 an 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.

Our framework is sufficiently tractable to analyze a number of extensions. First, our main results continue to hold under partial commitment for the Sender. Second, they continue to hold in more general persuasion environments with richer state and action spaces. Specifically, with a larger state 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. Third, 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. Fourth, 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) which 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 both the state and 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.

Two recent papers (Balbuzanov, 2019; Dziuda and Salas, 2018) specifically investigate the role of lie detection in communication. 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

[4]In a somewhat related vein, 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.

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 his commitment payoff. In contrast, if the Sender in our model loses all commitment power, he 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 he 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 ($S$, he) sends a message $m \in \{0,1\}$ to the Receiver (R, she). We assume that the Sender has full commitment power, as is common in the Bayesian persuasion framework.[5] Specifically, the strategy of the Sender is a mapping $\sigma : \{0,1\} \longrightarrow \Delta(\{0,1\})$. Denote the Sender's strategy space by Σ. 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 $a : \{0,1\} \times \{lie, \neg lie\} \longrightarrow \Delta(\{0,1\})$.

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 both for simplicity as well as for realism 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). 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 remaining probability $1-q$, she 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

[5] For a detailed discussion and relaxation of this assumption, see Min (2021), Fréchette et al. (2022), Lipnowski 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 enough 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 4.1, we show that our results continue to hold under partial commitment.

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 as lies. However, as we show in Section 4.3.1, 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 her of the true state independently of the message. Section 4.3.2 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

[6] Balbuzanov (2019) allows some degree of common interest between two players and focuses on the existence of 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).

which she takes $a = 1$. 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 4.4.

3 Analysis

3.1 Optimal Messages

The Sender solves the following maximization problem:

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

Due to the simple structure of the model, it is without loss of generality to assume that the Sender chooses only two two reporting probabilities, $p_0 = \Pr(m = 0 \mid \omega = 0)$ and $p_1 = \Pr(m = 1 \mid \omega = 1)$.[7] We denote the optimal reporting probabilities of the Sender by p_0^* and p_1^* and the ex-ante payoffs under this reporting probabilities by U_S and U_R.[8]

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

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

Specifically, if the Receiver is informed of a lie, her posterior beliefs are degenerate due to the binary state space. With a slight abuse of notation, let $\mu_m \equiv \mu_{m,\neg lie}$. Observe that when $p_0 = 0,\ p_1 = 1$,

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

[8] 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.

the event $(m = 0, d = \neg lie)$ occurs with 0 probability, so the belief μ_0 is not restricted by Bayes' rule. However, the off-path belief does not matter for the Sender. For convenience, define $\mu_0 = 0$ in this case. Analogously, if $p_0 = 1,\ p_1 = 0$, define $\mu_1 = 0$.

The Receiver optimally takes action $a = 1$ if her posterior belief exceeds the threshold t. Consequently, she takes action $a = 1$ after observing $(m = 0, d = lie)$ and takes action $a = 0$ after observing $(m = 1, d = lie)$. It remains to compare μ_0, μ_1 with t, which partitions the Sender's strategy space into four types. These types of strategies, which we denote by I, II, III, and IV, are defined in Table 1. Within each type of Sender's strategy, it is easy to find the optimal strategy since the Receiver's best response function is the same. We are then left to pick the best strategy out of four candidates, which is characterized by Proposition 1.

Type	Posteriors
I	$\mu_0 < t,\ \mu_1 < t$
II	$\mu_0 \geq t,\ \mu_1 < t$
III	$\mu_0 < t,\ \mu_1 \geq t$
IV	$\mu_0 \geq t,\ \mu_1 \geq t$

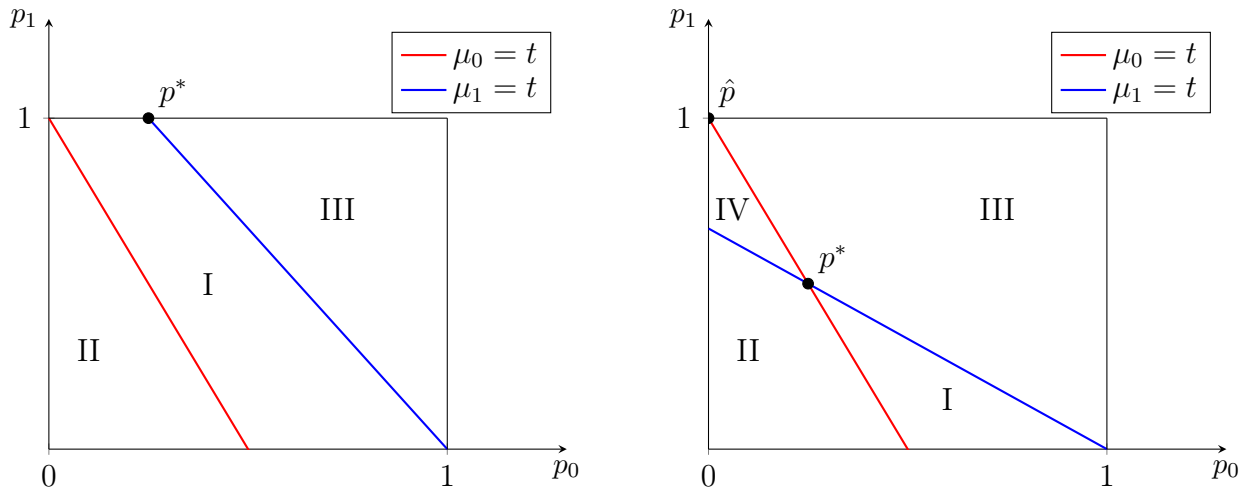
Table 1: The partition of the Sender's strategy space.

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

(*a*) *If* $q \leq \overline{q}$*, the Sender's optimal strategy is a type* III *strategy, in which the Sender always tells the truth under* $\omega = 1$ *but lies with positive probability under* $\omega = 0$.

(*b*) *If* $q > \overline{q}$*, the Sender's optimal strategy is a type* IV *strategy, in which the Sender lies with positive probability under both states.*

In Figure 1, we graphically illustrate the partition of the Sender's strategy space. The proof involves comparisons between the four optimal strategies within each type. First, there exists *some* type II strategy $(0, 0)$ that is better than *all* type I strategies. Following this strategy, the Sender totally misreports the state, and the Receiver takes action $a = 1$ if and only if $\omega = 1$, which occurs with probability μ. In contrast, given any type I strategy, the Receiver takes action $a = 1$ only if $\omega = 1$ and $(m = 0,\ d = lie)$, which occurs with a probability less than μ.



(a) $q \leq \overline{q}$ (b) $q > \overline{q}$

Figure 1: Equilibrium message strategies for different detection probabilities q.

Second, there exists *some* type III strategy that is better than *all* type II strategies. Within type II strategies, we need to focus only on the ones with $p_1 = 0$ because lying more under state $\omega = 1$ relaxes both constraints and is beneficial for the Sender. Now, for any type II strategy of the form $(p_0, 0)$, consider a type III strategy $(\tilde{p}_0, 1)$ such that $p_0 = (1-\tilde{p}_0)(1-q)$. By construction, this new strategy is equally as good as $(p_0, 0)$ for the Sender.

To see the intuition, note that the type II and III strategies are totally symmetric if lie detection technology is not available ($q = 0$) since, in this case, the messages have no intrinsic meaning and we could always rename the messages. However, the introduction of a lie detection technology ($q > 0$) generates an intrinsic meaning for the message that the Sender uses. In particular, an on-path message that was not detected as a lie always carries some credibility for the state to which it corresponds. Now, this additional source of credibility breaks the symmetry. By definition, type II strategies are such that $(m = 0,\ d = \neg lie)$ suggests $\omega = 1$ with a sufficiently high probability while $(m = 1,\ d = \neg lie)$ suggests $\omega = 0$ with a sufficiently high probability. Loosely speaking, it is harder to persuade the Receiver to take $a = 1$ using type II strategies since the Sender needs to counter the intrinsic credibility of messages.

By transitivity, both type I and type II strategies are suboptimal relative to the optimal type III strategy. Moreover, in the optimal type III strategy, the Sender must choose $p_1 = 1$ because

lying less under state $\omega = 1$ relaxes both constraints and is beneficial for the Sender. Finally, it remains to compare the optimal type III and type IV strategy. Interestingly, as suggested by Figure 1 (a), type IV strategies do not exist when the detection probability q is smaller than a threshold $\overline{q}$. Thus, we immediately obtain Proposition 1(a). Intuitively, the martingale property requires that the four posteriors average to the prior. However, when q is sufficiently small, the lie detection occurs rarely, and most weights are assigned to the posteriors μ_0 and μ_1. So, the martingale property requires that μ_0 and μ_1 approximately average back to the prior, suggesting that they cannot be both too much higher than the prior. It follows that a type IV strategy does not exist.

However, if the detection probability q is larger than $\overline{q}$, then it is possible to support the two posteriors μ_0 and μ_1 to be sufficiently higher than the prior and even higher than the threshold t, i.e., the set of type IV strategies is nonempty. In this case, the optimal type III strategy, denoted by $\hat{p}$ in Figure 1 (b), is to always send message $m = 1$ regardless of the state. However, this strategy is no longer globally optimal because $\hat{p}$ is worse than *any* type IV strategy p. To this end, we decompose the value of a strategy for the Sender into two parts: the expected payoff in the favorable state $\omega = 1$ and the expected payoff in the unfavorable state $\omega = 0$.

The strategy $\hat{p}$ induces $a = 1$ for sure when $\omega = 1$ because the Sender always truthfully sends $m = 1$, which is credible and is never flagged as a lie. Meanwhile, any strategy p of type IV also induces $a = 1$ for sure. Such a strategy could induce three different events: $(m = 1,\ d = \neg lie)$, $(m = 0,\ d = \neg lie)$, $(m = 0,\ d = lie)$. The first two events successfully persuade the Receiver to take $a = 1$ by the definition of type IV strategies. The last event directly informs the Receiver that $\omega = 1$, so it also induces $a = 1$. Hence, the strategies $\hat{p}$ and p align in the expected payoff in the favorable state $\omega = 1$. However, they differ in the expected payoff in the unfavorable state $\omega = 0$. Given $\hat{p}$, the Sender always lies and sends the message $m = 1$ when $\omega = 0$, which induces $a = 1$ only if the lie is not detected. Given p, the Sender sometimes tells the truth by sending the message $m = 0$ as well, but by the definition of type IV strategies, $m = 0$ is now a risk-free way to induce $a = 1$ since it will never be flagged as a lie in the unfavorable state $\omega = 0$. Hence, the strategy p results in a higher expected payoff for the Sender in the unfavorable state

as well as overall. Mathematically,

$$U_S(\hat{p}) = \underbrace{\mu}_{\Pr(\omega=1)} \times \overbrace{1}^{\Pr(a=1|\omega=1;\hat{p})} + \underbrace{(1-\mu)}_{\Pr(\omega=0)} \times \overbrace{(1-q)}^{\Pr(a=1|\omega=0;\hat{p})}$$

$$< U_S(p) = \underbrace{\mu}_{\Pr(\omega=1)} \times \overbrace{1}^{\Pr(a=1|\omega=1;p)} + \underbrace{(1-\mu)}_{\Pr(\omega=0)} \times \overbrace{[p_0 \times 1 + (1-p_0) \times (1-q)]}^{\Pr(a=1|\omega=0;p)}$$

As we argued above, the main benefit of p relative to $\hat{p}$ is that the "safer" message $m = 0$ is sent more frequently in p. Thus, the optimal type IV strategy must involve the highest p_0 or the least lying in the unfavorable state. Such a strategy, given by p^* in Figure 1 (b), is also globally optimal by the previous arguments provided that $q > \overline{q}$. The expressions are given by

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

Although the optimal strategy features partial lying under both states, the Sender still lies more in the unfavorable state than in the favorable state ($p_0^* < p_1^*$).

Finally, the threshold $\overline{q}$ where the optimal strategy switches from a type III to a type IV strategy is decreasing in μ and increasing in t. Specifically, fix the lie detection probability q. If the Receiver is easily persuaded (i.e., the prior μ is already close to the threshold t), a type IV strategy is optimal for the Sender. On the other hand, if the Receiver is hard to persuade (i.e., the threshold t is much higher than the prior μ), a type III strategy is optimal for the Sender.

3.2 Comparative Statics

We now show how the optimal message strategy and the equilibrium payoffs of the communicating parties change as the lie detection technology improves.

3.2.1 Optimal Messaging Strategy

Proposition 2 describes how the structure of the optimal messaging strategy $(p_0^*,\ p_1^*)$ changes as the detection probability varies. Figure 2 plots these optimal reporting probabilities as a function

of q. For comparison, the probabilities p_0^{BP} and p_1^{BP} are the equilibrium reporting probabilities that would result in a standard Bayesian persuasion setup without lie detection.

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

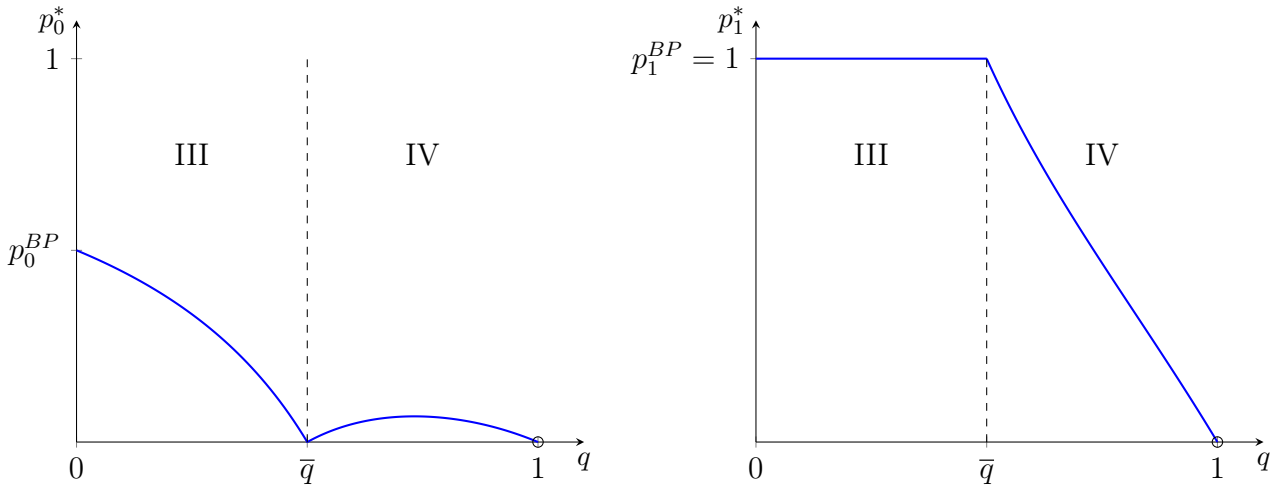
(a) $p_0^* = Pr(m = 0 \mid \omega = 0)$ *is decreasing over* $q \in [0, \overline{q}]$ *and has an inverse U shape over* $q \in (\overline{q}, 1]$,

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

(c) *the aggregate probability of lying,* $\mu(1 - p_1^*) + (1 - \mu)(1 - p_0^*)$, *increases over* $q \in [0, 1]$ *if and only if* $\mu \leq \frac{t^2}{1-2t+2t^2}$.

If $q \leq \overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)}$, p_0^* is decreasing in q, and p_1^* is constant at 1. In this range of q, the Sender's optimal strategy involves truthfully reporting the state $\omega = 1$ (i.e., $p_1 = 1$) but progressively misreporting the state $\omega = 0$ as the lie detection technology improves (i.e., $p_0 < 1$ and p_0 decreases in q). 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 $q > \overline{q}$, p_0^* initially increases and then decreases. In contrast, p_1^* decreases over the entire range of $[\overline{q}, 1]$. In this range, the Sender's optimal strategy involves misreporting both states.



(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 that leads to the worst action for the Sender ($a = 0$ in our case), the Receiver is certain of her action at this belief. In addition, if the optimal signal induces a belief that leads to the best action for the Sender ($a = 1$ in our case), the Receiver is indifferent between the two actions at this belief.

Now consider the addition of a lie detection technology. As the lie detection probability q increases, $(m = 1, d = \neg lie)$ becomes more indicative of the favorable state $\omega = 1$, and therefore the Receiver would strictly prefer to take the favorable action $a = 1$. As a response, the Sender would like to send the message $m = 1$ more often while still maintaining that $(m = 1,\ d = \neg lie)$ sufficiently persuades the Receiver to take action $a = 1$. Because the Sender already sends the message $m = 1$ with probability 1 under $\omega = 1$, the only way to increase the frequency of $m = 1$ is to send such a message more often in the unfavorable state $\omega = 0$ (i.e., lie more frequently if $\omega = 0$). In other words, the Sender increases the frequency of lying just enough about the unfavorable state ($\omega = 0$) to make the Receiver indifferent when choosing the favorable action $a = 1$.

However, once the detection probability q rises above $\overline{q}$, the Sender can no longer simply lie about the unfavorable state because he already maximally lies about it at $\overline{q}$. His optimal messaging strategy is now a type IV strategy, under which the Receiver takes the unfavorable action $a = 0$ only if he receives a message $m = 1$ that is flagged as a lie. In this case, a lie involving the message $m = 0$ is sufficiently likely to be detected, which the Sender can use to his advantage to ensure that the Receiver chooses the favorable action $a = 1$ when she observes $(m = 0, d = \neg lie)$. Therefore, at $q = \overline{q}$, the Sender increases the frequency of the message $m = 0$ by both increasing p_0 and decreasing p_1. However, when the detection probability is close to 1 (i.e., the lie detection technology is almost perfect), p_1 is close to 0, and any message $m = 1$ is very likely to be a lie. To make sure that a message $m = 1$ that is not detected as a lie still sufficiently persuades the Receiver to choose $a = 1$ (i.e., does not violate the constraints $\mu_0 \geq t$ and $\mu_1 \geq t$ required for a type IV strategy), the Sender also has to decrease p_0 while decreasing p_1.[9]

While p_0^* varies nonmonotonically with q when $q \geq \overline{q}$, Proposition 2(c) asserts that the aggregated probability of lying monotonically increases in $q \in [0, 1]$ provided that μ is not too high. In

[9]These perhaps surprising comparative statics of the type IV strategy are due to the persuasion game leading to a mixed strategy equilibrium. Such mixed strategy equilibria often have counterintuitive comparative statics properties, as Crawford and Smallwood (1984) point out.

particular, if $t \geq \frac{1}{2}$, the inequality condition is always satisfied.

Alternatively, we could analyze the impact of lie detection on the informativeness of the Sender's strategy per se. Formally, each Sender's messaging strategy (p_0, p_1) corresponds to an experiment

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

Clearly, 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.

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 $\mathcal{E}(p_0^*(q'), p_1^*(q'))$.*

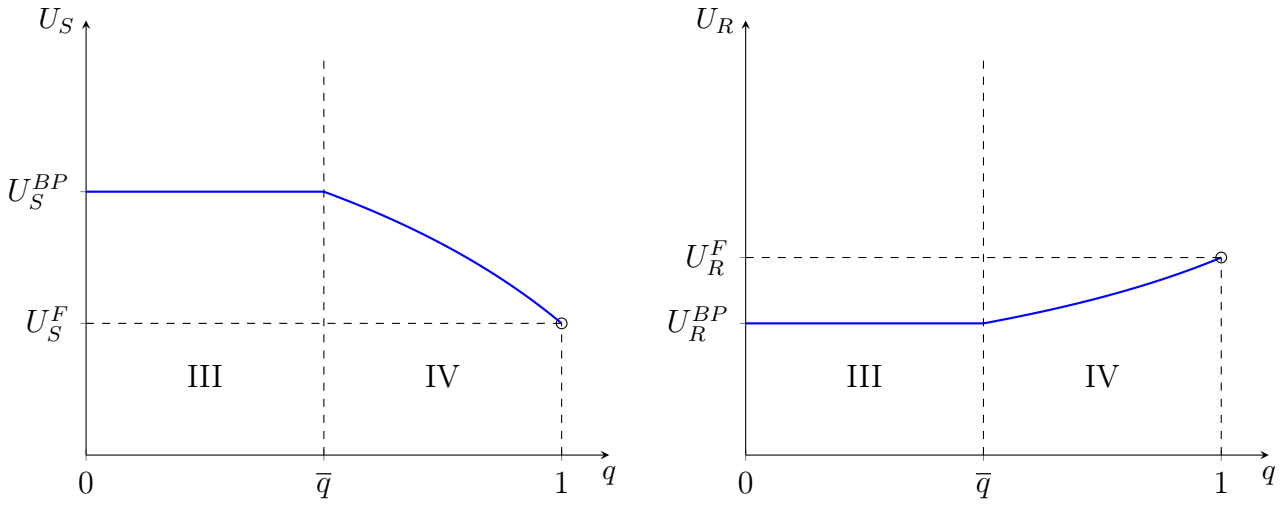
3.2.2 Payoffs

Recall that U_S and U_R denote the equilibrium payoffs of the Sender and the Receiver. We now investigate how they 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,*

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

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



(a) Sender's Equilibrium Payoff

(b) Receiver's Equilibrium Payoff

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

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 he 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 only when the state is $\omega = 1$, which occurs with probability μ.

Analogously to the case of the Sender's payoff, the Receiver's payoff is also constant at the Bayesian persuasion benchmark as long as $q \leq \overline{q}$ and then increases with q for $q > \overline{q}$ as the lie detection technology starts to bite. In the canonical Bayesian persuasion benchmark, the Receiver is held to her 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.

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 $\overline{q}$. As q approaches 1, the Receiver's payoff approaches her payoff under full information U_R^F. Intuitively, whenever a lie is not detected, the Receiver infers that the message is equal to the state with a probability close to 1. Thus, regardless of the Sender's strategy, the Receiver obtains almost full information.

4 Extensions and Discussion

Our baseline model considers the role of lie detection in a simple setting with full commitment, binary states, binary actions, and a particular lie detection technology. We now investigate how alternative assumptions about the Sender's commitment, the state space, the action space, and the detection technology modify our analysis.

4.1 Partial Commitment

In many communication models, the predictions crucially depend on the Sender's ability to commit to a particular messaging strategy. However, the main insights of our baseline model are not confined to the extreme case of full commitment but continue to hold even under partial commitment. Following Lipnowski et al. (2022) and Min (2021), we assume that the Sender's commitment binds only probabilistically. The generalized game with partial commitment proceeds as follows.

The Sender first declares a commitment strategy $(p_1, p_1) \in [0,1]^2$. He then privately learns the true state $\omega \in \{0,1\}$ and whether his commitment is binding. With probability α, his commitment binds, and he has to send a message following the prespecified commitment strategy. Otherwise, his commitment is not binding, and he can send any message $m \in \{0,1\}$ at his discretion. Let $(\tilde{p}_0, \tilde{p}_1) \in [0,1]^2$ denote his strategy following a nonbinding commitment, where $\tilde{p}_i$ is the probability that he sends a message $i \in \{0,1\}$ when the true state is 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. She is aware that the Sender may not abide by his 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 given in the baseline model. Thus, the baseline model corresponds to the special case $\alpha = 1$, whereas $\alpha = 0$ instead leads to a model of cheap talk with lie detection.

Proposition 5 characterizes the role of commitment in the equilibrium and the corresponding payoffs. We focus on the more relevant case $q \leq \overline{q}$ because that is where the equilibrium payoffs are constant in q in the baseline model. Part (a) of the proposition shows that a small loss of commitment has no impact on the key features of the equilibrium and the corresponding payoffs. Part (b) implies that lie detection is Pareto-improving if the Sender's commitment is sufficiently weak.

Proposition 5. *Assume* $q \leq \overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)}$.

(a) *If* $\alpha \geq \underline{\alpha} = \frac{\overline{q}-q}{1-q}$, *then* $p_1^* = 1$, $\tilde{p}_0^* = 0$, *and* $\tilde{p}_1^* = 1$. *Moreover,* p_0^* *decreases in* q, *while* U_S *and* U_R *are both constant in* q.

(b) *If* $\alpha < \underline{\alpha} = \frac{\overline{q}-q}{1-q}$, *then* $p_1^* = 0$ *and* $\tilde{p}_1^* = 0$. *Moreover, both* U_S *and* U_R *strictly increase in* q.

Since the lie detection technology is weak, a type IV strategy does not exist. It is impossible to induce the Receiver to choose $a = 1$ after observing both ($m = 1$, $d = \neg lie$) and ($m = 0$, $d = \neg lie$). We first focus on $\alpha \geq \underline{\alpha}$, in which case the Sender prefers to induce $a = 1$ after the Receiver observes ($m = 1$, $d = \neg lie$) as in the baseline model. In response, the Receiver takes action $a = 1$ if and only if ($m = 1$, $d = \neg lie$) or ($m = 0$, $d = lie$). Given the Receiver's best response, the Sender's optimal strategy following a nonbinding commitment ($\tilde{p}_0^*$, $\tilde{p}_1^*$) must be $(0, 1)$ (i.e., he prefers to send $m = 1$ regardless of the state). Moreover, the optimal prespecified strategy preserves the structure of the optimal messaging strategy in Proposition 1: $p_1^* = 1$ and p_0^* leaves the Receiver indifferent after observing $(m = 1, d = \neg lie)$.

Again, p_0^* decreases in q, highlighting the Sender's strategic incentives to lie more in the presence of a stronger lie detection technology. As in the baseline model, this strategic effect exactly offsets

the positive effect of increasing q because the probability of observing an undetected lie that induces $a = 1$ is equal to

$$(1-\mu)(1-q)\cdot[\alpha(1-p_0^*)+1-\alpha] = (1-\mu)(1-\overline{q})$$

which is constant in q. Consequently, both the Sender's and the Receiver's equilibrium payoffs are also constant in q as long as $q \leq \overline{q}$. Thus, our main results do not hinge on the full commitment assumption commonly used in Bayesian persuasion models.

When the Sender's commitment binds probabilistically, it is as if the Receiver receives a signal of unknown informativeness. If we ignore the lie detection technology, then with probability α, the message is generated according to (p_0^*, p_1^*) and is partially informative. With probability $1-\alpha$, the message is generated according to $(\tilde{p}_0^*, \tilde{p}_1^*)$ and is totally uninformative. However, the average informativeness of these messages needs to leave the Receiver indifferent. Thus, the Sender must commit to a more informative messaging strategy than in the baseline model. In fact, it can be easily shown that as the Sender's commitment power grows, the Sender's prespecified strategy becomes less informative.

The lower bound $\underline{\alpha}$ also has a natural interpretation. If the average informativeness of an uninformative message and a partially informative message leaves the Receiver indifferent, then the Receiver must strictly prefer to take action $a = 1$ when the Sender commits to the fully informative strategy, which yields the lower bound on α. If α is lower than this bound, then even if the prespecified strategy is fully informative, the average informativeness is too low to induce the favorable action $a = 1$. In this case, the best that the Sender could do is to always lie in the favorable state $\omega = 1$, hoping that it is detected. Thus, he prefers a stronger lie detection technology. On the other hand, the Sender's strategic effect to lie more disappears and it follows that the Receiver also prefers a stronger lie detection technology.

4.2 General Persuasion Environments

Our stylized baseline model delivers interesting results, yet it is unclear whether they are driven by the simple structure of the persuasion problem, such as the binary state space and the binary

action space. This section explores more general persuasion environments.

Equilibrium uniqueness is not necessarily guaranteed in general environments. Thus, it is hard to generalize the comparative statics of Proposition 2. Instead, we focus on the comparative statics of equilibrium payoffs with respect to the lie detection probability. In Section 4.2.1, we demonstrate that with a larger state space, both players' payoffs are completely independent of the lie detection technology, thereby strengthening Proposition 4. Subsequently, in Section 4.2.2, we show that with a larger action space, Proposition 4 continues to hold if and only if the prior is sufficiently low or high.

4.2.1 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

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

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 her 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 his 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 6 asserts that both players' payoffs are independent of lie detection, thereby strengthening Proposition 4.

Proposition 6. *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 strategy is 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 $q > 0$. In this section, we illustrate the idea behind this construction for a special case: $N = 3$ and $t \in (\mu, \omega_2)$. The constructions for remaining cases follow a similar approach and are detailed in Appendix A.6.

In the benchmark scenario where $q = 0$, the result of Ivanov (2021) implies that the following (monotone partitional) strategy σ^* is optimal for the Sender:

$$\sigma_1(1) = 1, \quad \sigma_1(\omega_2) = 1, \quad \sigma_1(0) = \begin{cases} 1, & w.p. \quad s \\ 0, & w.p. \quad 1 - s \end{cases}$$

where the mixing probability s leaves the Receiver indifferent when she observes a message $m = 1$. Building on his results, we further show that the Receiver's payoff under σ^* is maximal among all Sender-optimal strategies. Subsequently, we allow for lie detection and construct a strategy σ_1

that is independent of q:

$$\sigma_1(1) = \omega_2, \quad \sigma_1(\omega_2) = 1, \quad \sigma_1(0) = \begin{cases} 1, & w.p. \quad u \\ \omega_2, & w.p. \quad s-u \\ 0, & w.p. \quad 1-s \end{cases}$$

where the mixing probability $u \in [0, s]$ leaves the Receiver indifferent when she observes $(m = 1, d = \neg lie)$, $(m = 1, d = lie)$, $(m = \gamma, d = \neg lie)$, and $(m = \gamma, d = lie)$. Observe that under both σ^* and σ_1, the Receiver chooses $a = 1$ with probability one if $\omega = \omega_3$ or $\omega = \omega_2$, and with probability s if $\omega = \omega_1$. Hence, they induce the same payoff pair as desired.

Intuitively, the strategy σ_1 is immune to lie detection because the information from lie detection is already embedded in the strategy per se. Specifically, even in the absence of a lie detection technology, the Receiver understands that messages $m = \gamma$ and $m = 1$ are surely lies while a message $m = 0$ is never a lie. Consequently, lie detection does not provide any additional information to the Receiver. As the state space (and the message space) further grows, the set of strategies that are immune to lie detection becomes even larger. Therefore, it should be not surprising that lie detection continues to be ineffective.

4.2.2 General Action Space

In this section, we maintain a binary state space $\Omega = \{0, 1\}$ with a prior mean $\bar{\mu} \in (0, 1)$, but we consider 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 consider 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 4.

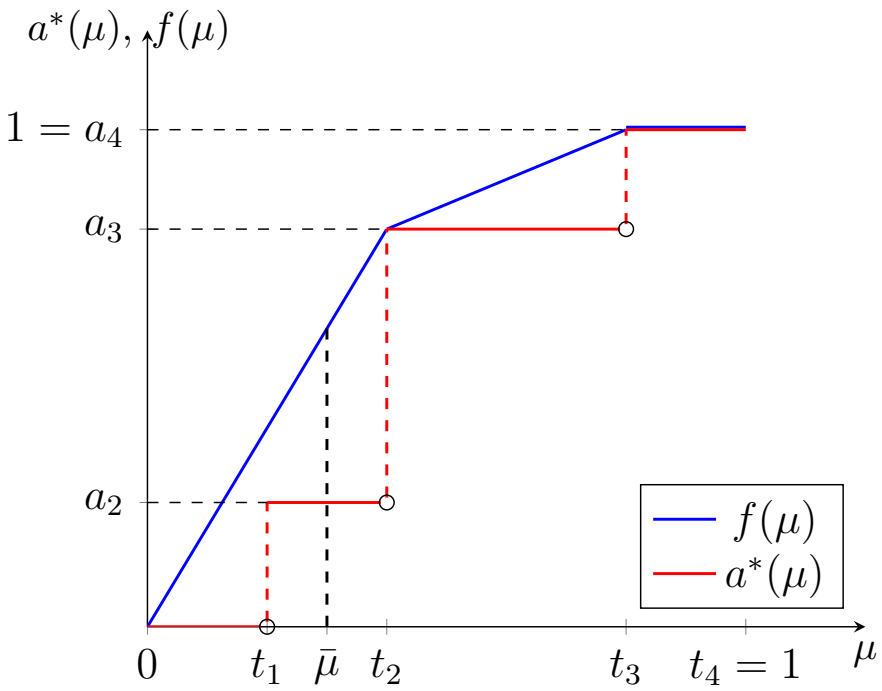


Figure 4: Illustration of $a^*(\mu)$ and $f(\mu)$ when $N = 4$

In this general setup, Proposition 7 indicates 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 7. *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 there exists a sufficiently small* $q > 0$ *such that* $U_S(q) = U_S(0)$.

(*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 4, 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 7 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 his 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 4.4 further explores this issue in detail.

The second half of Proposition 7 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 $(p_0, p_1) \in [0, 1]^2$. Consider two types of strategies as follows.

If $p_1 = 1$, then by reversing the arguments in part (a), we construct an alternative strategy $(\hat{p}_0, 1)$ such that $U_S(p_0, 1; q) = U_S(\hat{p}_0, 1; 0)$. However, $(\hat{p}_0, 1)$ cannot be a Sender-optimal strategy when $\bar{\mu} \in [t_s, t_{N-1}]$ as it induces the posterior $\mu = 0$ with positive probability. It follows by transitivity that $U_S(p_0, 1; q) < U_S(0)$.

If $p_1 < 1$, then the strategy (p_0, p_1) induces the posterior $\mu = 1$ and thus action $a_N = 1$ with positive probability. Nonetheless, this is a waste of credibility because the Receiver is willing to take the same action as long as her posterior exceeds t_{N-1}. Following the intuition, we show that the Sender could be strictly better off in a persuasion problem with no lie detection but a larger message space. Finally, we make the observation that when there is no lie detection, the Sender does not benefit from a larger message space. It follows again by transitivity that $U_S(p_0, 1; q) < U_S(0)$ when $p_1 < 1$.

As a corollary of Proposition 7, 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' 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 $q > 0$.*

Lastly, while we do not specify the Receiver's payoff in this section, for any reasonable payoff

function that generates the ladder-shaped best response function $a^*(\mu)$, we expect Proposition 7 to apply to the Receiver's payoff function as well.

4.3 Detection Technologies

We now show that our results continue to hold under different detection technologies that the receiver can use to inform her choice of action.

4.3.1 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. In this section, we introduce the possibility of a false alarm by considering the following general lie detection technology:

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

This means that 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. To see this, 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 8 demonstrates, his 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 8. *$U_S(q,r) = U_S(0,0)$ 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 necessarily imply that our baseline result is a knife-edge case. We argue in Proposition 9 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 reconfirming the main insight of this paper.

Proposition 9. *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 $U_R(q,r) = t(1-\mu)$.*

4.3.2 Truth and State Detection

Consider a different detection technology that informs the Receiver with probability r that a message is truthful. That is, rather than being able to (probabilistically) detect a lie, the Receiver can (probabilistically) detect that a message is truthful. Truth detection is perhaps a less realistic assumption, as it is arguably easier to detect whether the Sender has lied than whether he has sent a truthful message (Vrij et al., 2011).

In our setting, 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 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.

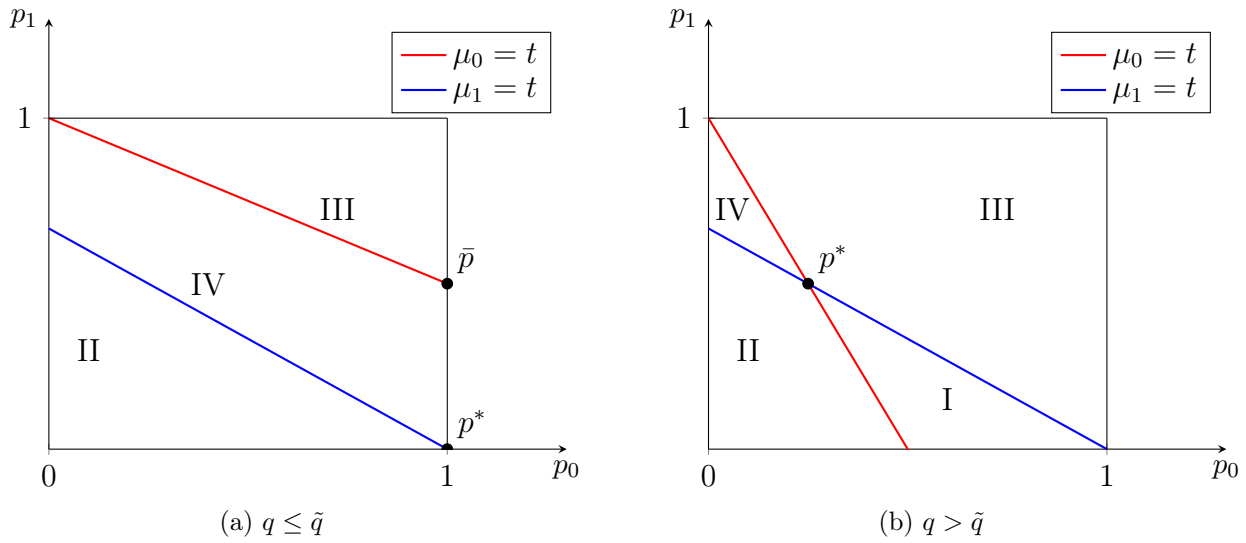
4.4 Default Action Coincides with Sender's Preferred Action

In standard Bayesian persuasion models without lie detection, the Sender can always stay 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 his optimization problem nontrivial even when the Receiver's default action coincides with his preferred action.

In this extension, we analyze the scenario in which the prior mean μ is higher than the threshold t. The results are analogous to those in the baseline model. As before, the Sender's maximization problem is solved by considering the four subproblems. The only change relative to the baseline model is that Region IV now exists for any $q \in [0, 1]$, as shown in Figure 5.

The optimal messaging strategy p^* is always in Region IV. When $q \leq \tilde{q} \equiv 1 - \frac{t(1-\mu)}{\mu(1-t)}$, the strategy $(p_0, p_1) = (1, 0)$ would induce the Receiver to take $a = 1$ with probability one and is thus optimal. Under this strategy, the Sender reports $m = 0$ with probability one in both states. If the message is flagged as a lie, the Receiver immediately learns that the true state is $\omega = 1$. Otherwise, by the martingale property, her posterior mean would drop. Nonetheless, if q is sufficiently small, her posterior mean would be close to the prior mean, which is still higher than the action threshold. Hence, the Receiver is always willing to take the favorable action.

If q is sufficiently large, such a strategy is no longer sustainable and it is impossible to induce $a = 1$ with probability one. For example, in the extreme case where $q = 1$, it is as if the Receiver



(a) $q \leq \tilde{q}$ (b) $q > \tilde{q}$

Figure 5: Equilibrium message strategies for different detection probabilities q ($\mu \geq t$).

learns the true state. Hence, it must be that the Receiver takes action $a = 1$ if and only if $\omega = 1$. In fact, the Sender's optimal messaging strategy is again characterized by the intersection of two indifference conditions: $\mu_0 = t$ and $\mu_1 = t$, as in Figure 5 (b).

Given the discussion above, the Sender's (Receiver's) payoff is initially constant in q when $q \leq \tilde{q}$ and then is decreasing (increasing) in q when $q > \tilde{q}$. This is consistent with Proposition 4.

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 to enter the game. If he 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 her prior. It is straightforward to show that the Sender enters the game if the Receiver's default action does not coincide with his preferred action. Otherwise, the Sender does not enter the game, but the Receiver always takes action $a = 1$, consistent with the standard persuasion paradigm.[10]

[10]On the other hand, endowing the Sender with the option to remain silent significantly changes the results. Suppose the Sender could send a message $m = \emptyset$ that is never flagged as a lie, then his maximal payoff is essentially independent of the strength of the lie detection technology. For example, the Sender could commit to sending $m = \emptyset$ for sure under the favorable state and mixing between $m = \emptyset$ and $m = 0$ under the unfavorable state such that the mixing probabilities leave the Receiver indifferent whenever she observes $m = \emptyset$. By adopting such a strategy, the Sender never lies, his message is never flagged as a lie, and thus lie detection has no bite.

5 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 his false statements will be flagged as lies.

These insights extend more generally and continue to hold under partial commitment for the Sender, in richer state and action spaces, and under different detection technologies that the Sender can use to inform her decision. Nonetheless, our analysis raises further questions about the role of lie detection under Bayesian persuasion and communication more generally. 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 Dennis E Smallwood, "Comparative Statics of Mixed-strategy Equilibria in Noncooperative Two-person Games," *Theory and Decision*, 1984, *16* (3), 225–232.

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.

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

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

A.1 Proof of Proposition 1

For type $i \in \{\text{I}, \text{II}, \text{III}, \text{IV}\}$ strategy, the maximal probability that the Receiver chooses $a = 1$, denoted by $\Pr_i(a = 1)$, is given by

$$\Pr_{\text{I}}(a=1) = \sup_{p_0,p_1\in[0,1]} \quad \mu(1-p_1)q \quad \text{s.t.} \quad \mu_0 < t,\ \mu_1 < t$$

$$\Pr_{\text{II}}(a=1) = \sup_{p_0,p_1\in[0,1]} \quad \mu(1-p_1) + (1-\mu)p_0 \quad \text{s.t.} \quad \mu_0 \geq t,\ \mu_1 < t$$

$$\Pr_{\text{III}}(a=1) = \sup_{p_0,p_1\in[0,1]} \quad \mu p_1 + \mu(1-p_1)q + (1-\mu)(1-q)(1-p_0) \quad \text{s.t.} \quad \mu_0 < t,\ \mu_1 \geq t$$

$$\Pr_{\text{IV}}(a=1) = \sup_{p_0,p_1\in[0,1]} \quad 1-(1-\mu)(1-p_0)q \quad \text{s.t.} \quad \mu_0 \geq t,\ \mu_1 \geq t$$

First, observe that $(p_0, p_1) = (0, 0)$ is a type II strategy and yields a payoff μ to the Sender. So,

$$\Pr_{\text{I}}(a=1) < \mu \leq \Pr_{\text{II}}(a=1)$$

Second, within type II strategies, it is optimal to set $p_1 = 0$ because this loosens both constraints and improves the objective. Given this, $\mu_1 = 0 < t$, and the optimum requires $\mu_0 = t$.

$$p_0 = \frac{\mu(1-q)(1-t)}{t(1-\mu)} \Longrightarrow \Pr_{\text{II}}(a=1) = \mu + \left(\frac{\mu}{t} - \mu\right)(1-q)$$

Similarly, within type III strategies, it is optimal to set $p_1 = 1$. Given this, $\mu_0 = 0 < t$, and the optimum requires p_0 to be as small as possible while preserving $\mu_1 \geq t$. Define $\overline{q} \equiv \frac{t-\mu}{t(1-\mu)} \in (0, 1)$; then, there are two cases.

- If $q \leq \overline{q}$, then $\mu_1 \geq t \iff p_0 \geq \frac{\overline{q}-q}{1-q}$. Thus, it is optimal to set $p_0 = \frac{\overline{q}-q}{1-q}$, which implies $\Pr_{\text{III}}(a = 1) = \frac{\mu}{t}$.

- If $q > \overline{q}$, then $\mu_1 \geq t$ for any $p_0 \in [0, 1]$. Thus, it is optimal to set $p_0 = 0$, which implies $\Pr_{\text{III}}(a = 1) = \mu + (1-\mu)(1-q)$.

Clearly, in either case, we have $\Pr_{\text{III}}(a = 1) > \Pr_{\text{II}}(a = 1)$, and therefore both type I and II strategies are suboptimal. It therefore remains for us to compare $\Pr_{\text{III}}(a = 1)$ and $\Pr_{\text{IV}}(a = 1)$.

(*a*) If $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$, type IV strategies do not exist; i.e., there is no way to choose p_0, p_1 such that $\mu_1 \geq t$ and $\mu_0 \geq t$. If this were the case, we would have

$$\frac{\mu p_1}{\mu p_1 + (1-\mu)(1-p_0)(1-q)} \geq t$$

and

$$\frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p_0} \geq t \iff \frac{\mu(1-p_1)}{\mu(1-p_1)+(1-\mu)\frac{p_0}{1-q}} \geq t$$

which would imply

$$t \leq \frac{\mu p_1 + \mu(1-p_1)}{\mu p_1 + \mu(1-p_1) + (1-\mu)(1-p_0)(1-q) + (1-\mu)\frac{p_0}{1-q}} \leq \frac{\mu}{\mu + (1-\mu)(1-q)}$$

Moreover, it is impossible that all inequalities bind at the same time. Thus, $t < \frac{\mu}{\mu+(1-\mu)(1-q)}$, leading to a contradiction. In summary, the optimal strategy in this case is a type III strategy and takes the following form.

$$p_0^* = \frac{\overline{q}-q}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{1}$$

(*b*) If $\frac{\mu}{\mu+(1-\mu)(1-q)} > t$, type IV strategies exist. Within type IV strategies, both constraints $\mu_0 \geq t$ and $\mu_1 \geq t$ bind at the optimum. If $\mu_0 > t$ at the optimum, then the Sender benefits from slightly increasing p_0. Similarly, if $\mu_1 > t$ at the optimum, then the Sender benefits from slightly decreasing p_1 and increasing p_0 at the same time so that μ_0 is unchanged. Since $\mu_0 = t$ at the optimum, it must be that $p_0 > 0$ at the optimum and consequently,

$$\Pr_{\text{III}}(a=1) = \mu + (1-\mu)(1-q) = 1-(1-\mu)q < 1-(1-\mu)q(1-p_0) = \Pr_{\text{IV}}(a=1)$$

In summary, the optimal strategy in this case is a type IV strategy, which take the following form by the two binding constraints.

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

A.2 Proof of Proposition 2

(a) If $q \leq \overline{q}$, then by Equation (1), p_0^* obviously decreases in q, while p_1^* is constant in q.

(b) If $q > \overline{q}$, then by Equation (2),

$$\frac{\partial p_0^*}{\partial q} = \frac{-q^2 + \overline{q}(2 - 2q + q^2)}{(2-q)^2 q^2} \quad \text{and} \quad \frac{\partial p_1^*}{\partial q} = \frac{-q^2 - \frac{\overline{q}}{1-\overline{q}}(2 - 2q + q^2)}{(2-q)^2 q^2} \tag{3}$$

Therefore, p_1^* decreases over $q \in (\overline{q}, 1]$ and

$$\frac{\partial p_0^*}{\partial q} \geq 0 \Longleftrightarrow \frac{1}{\overline{q}} \leq \frac{q^2 - 2q + 2}{q^2} = 1 + \frac{2 - 2q}{q^2}$$

The RHS decreases in q, meaning that the sign of the derivative $\frac{\partial p_0^*}{\partial q}$ changes at most one time. Since the derivative is positive at $q = \overline{q}$ and negative at $q = 1$, we conclude that p_0^* is initially increasing and then decreasing in q over $(\overline{q}, 1]$.

(c) This is clearly true when $q \leq \overline{q}$, So we focus on $q > \overline{q}$, where

$$\frac{\partial[\mu p_1^* + (1-\mu)p_0^*]}{\partial q} = \frac{-q^2 + \left[(1-\mu)\overline{q} - \mu\frac{\overline{q}}{1-\overline{q}}\right](2 - 2q + q^2)}{(2-q)^2 q^2}$$

Thus,

$$\frac{\partial[\mu p_1^* + (1-\mu)p_0^*]}{\partial q} \leq 0 \Longleftrightarrow \frac{q^2}{2 - 2q + q^2} \geq \frac{(1-2t)(t-\mu)}{t(1-t)}$$

Since $\frac{q^2}{2-2q+q^2}$ increases in q, the above inequality holds for all $q \in (\overline{q}, 1]$ if and only if

$$\frac{\overline{q}^2}{2 - 2\overline{q} + \overline{q}^2} \geq \frac{(1-2t)(t-\mu)}{t(1-t)}$$

which, after tedious algebra, can be reduced to

$$\mu \leq \frac{t^2}{1 - 2t + 2t^2}$$

In particular, if $t \geq \frac{1}{2}$, then this inequality is implied by the assumption that $\mu < t$.

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) By Equation (1),

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

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 verified easily 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 by Equation (2) and $q > \overline{q}$,

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

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 $\bar{q} \leq q' < q \leq 1$.

A.4 Proof of Proposition 4

(a) If $q \leq \bar{q}$, the Receiver chooses $a = 1$ whenever $(m = 1,\ d = \neg lie)$ or $(m = 0,\ d = lie)$. However, the latter occurs with probability 0 in the equilibrium. Hence,

$$U_S(q) = \mu + (1-\mu)(1-p_0^*)(1-q) = \frac{\mu}{t} \tag{4}$$

which is constant in q. Moreover,

$$U_R(q) = (1-\mu)t \cdot [p_0^* + (1-p_0^*)q] + \mu(1-t) = (1-\mu)t \tag{5}$$

which is also constant in q.

(b) If $q > \bar{q}$, the Receiver always chooses $a = 1$ unless $(m = 1,\ d = lie)$. Thus,

$$U_S(q) = 1 - (1-\mu)(1-p_0^*)q = 1 - \frac{t(1-\mu) - \mu(1-t)(1-q)}{t(2-q)} \tag{6}$$

which is decreasing in q. Moreover,

$$U_R(q) = (1-\mu)t \cdot (1-p_0^*)q + \mu(1-t) = \frac{(1-\mu)t + t(1-\mu)}{2-q} \tag{7}$$

which is increasing in q.

A.5 Proof of Proposition 5

(*a*) Let $E \equiv \{0,1\} \times \{lie, \neg lie\}$, then notice that the Sender's payoff always satisfies

$$U_S(q) = \sum_{(m,d)\in E} \Pr(m,d) \cdot \mathbb{1}_{\{\mu_{m,d} \geq t\}} \leq \sum_{(m,d)\in E} \frac{\Pr(m,d,\omega=1)}{t} = \frac{\mu}{t} \tag{8}$$

Moreover, the bound is attained if and only if for any $(m,d) \in E$, either $\Pr(m,d,\omega=1)=0$ or $\mu_{m,d} = t$. Since $\mu_{0,lie} = 1 > t$, it follows that

$$0 = \Pr(0, lie, \omega = 1) = \alpha\mu(1-p_1)(1-q) + (1-\alpha)\mu(1-\tilde{p}_1)(1-q)$$

which implies that $p_1 = \tilde{p}_1 = 1$. Furthermore, $\Pr(1, \neg lie, \omega = 1) > 0$ and $\mu_{0,\neg lie} = 0$. So, if the upper bound is achieved, it must be that

$$\mu_{1,\neg lie} = \frac{\mu}{\mu + (1-\mu)(1-q)[\alpha(1-p_0) + (1-\alpha)(1-\tilde{p}_0]} = t \tag{9}$$

In addition, $\tilde{p}_0 = 0$ because the message $m = 1$ induces $a = 1$ with a higher probability than the message $m = 0$ does. So, Equation (9) yields

$$p_0^* = \frac{1}{\alpha}\left[1 - \frac{\mu(1-t)}{t(1-\mu)(1-q)}\right] \in [0,1] \tag{10}$$

by $q \leq \overline{q}$ and $\alpha \geq \underline{\alpha}$. In summary, the upper bound of the Sender's payoff is uniquely achieved by the strategy $(p_0^*,\ 1,\ 0,\ 1)$. Obviously, p_0^* decreases in q and the Sender's equilibrium payoff is constant in q. Lastly, the Receiver's equilibrium payoff is constant in q because

$$U_R = \mu(1-t) + (1-\mu)t \cdot \{\alpha[p_0^* + (1-p_0^*)q] + (1-\alpha)q\} = (1-\mu)t \tag{11}$$

(*b*) If $\alpha < \underline{\alpha}$, it can be analogously shown that $\mu_{1,\neg lie} \geq t$ and $\mu_{0,\neg lie} \geq t$ cannot hold at the same time. So we are left with three potential cases.

(1) If $\mu_{1,\neg lie} \geq t$ and $\mu_{0,\neg lie} < t$, it is immediate that $\tilde{p}_0^* = 0$ and $\tilde{p}_1^* = 1$, implying that

$$\begin{aligned}\mu_{1,\neg lie} =& \frac{\mu(\alpha p_1 + 1 - \alpha)}{\mu(\alpha p_1 + 1 - \alpha) + (1-\mu)(1-q)[\alpha(1-p_0) + 1 - \alpha]} \\ <& \frac{\mu}{\mu + (1-\mu)(1-q)(1-\alpha)} \\ <& t\end{aligned}$$

by $\alpha < \underline{\alpha}$ and $q \leq \overline{q}$. Contradiction.

(2) If $\mu_{1,\neg lie} < t$ and $\mu_{0,\neg lie} \geq t$, it is immediate that $\tilde{p}_0^* = 1$ and $\tilde{p}_1^* = 0$, implying that

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

by $\alpha < \underline{\alpha}$ and $q \leq \overline{q}$. Contradiction.

(3) If $\mu_{1,\neg lie} < t$ and $\mu_{0,\neg lie} < t$, then $U_S = \mu q[\alpha(1-p_1) + (1-\alpha)(1-\tilde{p}_1)]$. The Sender optimally chooses $p_1^* = \tilde{p}_1^* = 0$, which in turn ensures that $\mu_{1,\neg lie} < t$. Thus, any p_0 and $\tilde{p}_0$ such that $\mu_{0,\neg lie} < t$ is part of an equilibrium. Lastly, $U_S = \mu q$ is strictly increasing in q, while $U_R = \mu q(1-t) + (1-\mu)t$ is strictly increasing in q.

A.6 Proof of Proposition 6

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

$$\frac{\sum_{j=k+1}^{N} \lambda_j \omega_j + s_k \lambda_k \omega_k}{\sum_{j=k+1}^{N} \lambda_j + s_k \lambda_k} = t. \tag{12}$$

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

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

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: Suppose $q \geq 0$. 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 it is 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 is immune to lie detection.

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

where s_1 solves Equation (12) 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 is immune to lie detection.

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

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

where s_1 solves Equation (12) 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 is immune to lie detection.

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

where s_k solves Equation (12).

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.7 Proof of Proposition 7

(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 he 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

$$U_S(p_0, 1; 0) < U_S(0), \ \forall p_0 \in [0, 1] \tag{13}$$

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 (13), 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 $m \in \tilde{M}$,

$$\Pr(\mu = \mu^\sigma_m) = \lambda^\sigma_m$$

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

$$U_S(p_0, p_1; q) \leq \max_{\sigma \in \hat{\Sigma}} U_S(\sigma; 0) \tag{14}$$

Moreover, it must be that

$$\max_{\sigma \in \hat{\Sigma}} U_S(\sigma; 0) < \max_{\sigma \in \Sigma} U_S(\sigma; 0) \tag{15}$$

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

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

Such a strategy $\tilde{\sigma}$ always exists by Bayes plausibility and $\bar{\mu} < t_{N-1}$. Since $a^*(\mu^{\tilde{\sigma}}_{(0,1)}) = a^*(\mu^{\sigma^*}_{(0,1)}) = A_n$ and $\lambda^{\tilde{\sigma}}_m > \lambda^{\sigma^*}_m$, 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.,

$$\max_{\sigma\in\Sigma} U_S(\sigma;0) = \max_{(p_0,p_1)\in[0,1]^2} U_S(p_0,p_1;0) \tag{16}$$

Hence, by Inequalities (14)-(16), 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.8 Proof of Proposition 8

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

$$\mu_{0,lie} = \frac{\mu(1-p_1)q}{\mu(1-p_1)q+(1-\mu)p_0 r} \geq \frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p_0(1-r)} = \mu_{0,\neg lie}. \tag{17}$$

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

$$\mu_{1,lie} = \frac{\mu p_1 r}{\mu p_1 r+(1-\mu)(1-p_0)q} \leq \frac{\mu p_1(1-r)}{\mu p_1(1-r)+(1-\mu)(1-p_0)(1-q)} = \mu_{1,\neg lie}. \tag{18}$$

where the inequality is strict if $p_1\neq 0$ and $p_0\neq 1$. By Equation (8), 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 (18), 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 (17), 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 (17) and (18), 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.9 Proof of Proposition 9

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 (17), (18) 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, $U_S^{\mathrm{IV}} = \Pr(0, lie) + \Pr(1, \neg lie) \leq \frac{\Pr(0,lie,\omega=1)+\Pr(1,\neg lie,\omega=1)}{t} = \frac{\mu[(1-p_1)q+p_1(1-r)]}{t} < U_S^{\mathrm{II}}$ 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 $U_S^{\mathrm{V}} = \mu + \frac{\mu(1-q)(1-t)}{(1-r)t} < U_S^{\mathrm{II}}$ 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 $U_S^{\mathrm{VI}} = \mu + \frac{\mu r(1-t)}{qt} < U_S^{\mathrm{II}}$ 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 $\mu_{1,lie} = 0$ and $\mu_{0,lie} = t$, 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,

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