# Bayesian Persuasion with Lie Detection*

Florian Ederer† Weicheng Min‡

May 4, 2022

**Abstract**

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 the lie detection probability is sufficiently small, the Sender's and the Receiver's equilibrium payoffs are unaffected by the lie detection technology because the Sender simply compensates by lying more. However, when the lie detection probability is sufficiently high, the Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability.

JEL Codes: D83, D82, K40, D72

*We are particularly grateful to Andrew Little for inspiring our initial analysis. We also thank Joshua Gans, Scott Gehlbach, Matt Gentzkow, Philippe Jehiel, Navin Kartik, Kohei Kawamura, Elliot Lipnowski, Zhaotian Luo, Barry Nalebuff, Larry Samuelson, Joel Sobel, and Ian Turner for helpful comments.

†Yale School of Management and NBER, florian.ederer@yale.edu

‡Department of Economics, Yale University, weicheng.min@yale.edu

# 1 Introduction

Lies are a pervasive feature of communication, even when 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. But, interestingly, in the face of increased fact-checking and media focus, the rate of Trump's lying increased rather than decreased—a development that runs counter to this intuition.

This paper incorporates probabilistic lie detection in an otherwise standard model of Bayesian persuasion (Kamenica and Gentzkow, 2011; Kamenica, 2019). Two players, a Sender and a Receiver, engage in one round of communication. The Sender observes the binary state of nature and sends a message to the Receiver. To clearly define whether a message is a lie or not, we assume that the message space and the state space are the same. The Receiver observes the message, and if the message is a lie, it is flagged as such with some probability. The Receiver 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, payoffs are realized for both parties.

Our main assumption that lies are detectable is a natural one in many applications. Ekman and Frank (1993) argue that there are two basic reasons why lies fail due to detection: facts and emotions. First, facts may surface that contradict the message of the Sender. These facts do not necessarily tell all the information about the state of the world, but they reveal that the Sender's message was a lie. Second, lies can be detected because emotions and physical reactions such as blushing or sweating provide strong clues that the Sender is lying.

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

[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.

claiming that the state is favorable. That is to say, the lie detection technology changes the Sender's message strategy but does not have an impact on the utilities of both players. 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 the lie detection probability.

Our framework is sufficiently tractable to analyze a number of extensions. First, our main results also continue to hold under partial commitment for the Sender. Second, our results do not rely on the fully revealing nature of a lie in our model. They continue to hold even when the state is not binary and thus the detection of a lie does not fully reveal the state of nature. Third, we consider alternative detection technologies such as truth or state detection and show that the central insights of our model continue to hold. Fourth, we analyze the (non-trivial) 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.

Two recent papers (Balbuzanov, 2019; Dziuda and Salas, 2018) also investigate the role of lie detectability in communication. The most significant difference with respect to our paper lies in the commitment assumption of the Sender. In both papers, the communication game takes the form of cheap talk (Crawford and Sobel, 1982) rather than Bayesian persuasion as in our paper. We defer a detailed comparison between these papers and our work to Section 4. In addition, 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 what exact lies 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.

Related theoretical work on lying in communication games also 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.

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. (forthcoming) 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).

Finally, our paper is related to recent work on communication in political science. Whereas we focus on an improvement of the Receiver's communication technology (i.e., lie detection), Gehlbach et al. (2022) analyze how improvements that benefit the Sender (e.g., censorship and propaganda) impact communication under Bayesian persuasion. In a related framework which can be recast as Bayesian persuasion, Luo and Rozenas (2018) study how the electoral mechanism performs when the government (the Sender) can rig elections by manipulating the electoral process ex ante and falsifying election returns ex post.

# 2 Model

Consider the following simple model of Bayesian persuasion in the presence of lie detection. Let $w \in \{0,1\}$ denote the state of the world and $\Pr(w=1) = \mu \in (0,1)$. The Sender ($S$, he) observes $w$ and sends a message $m \in \{0,1\}$ to the Receiver ($R$, she). In Section 2.3 we specify the exact nature of the Sender's communication strategy.

## 2.1 Lie Detection Technology

If the Sender lies (i.e., $m \neq w$), the Receiver is informed with probability $q \in [0,1]$ that it is a lie and thus learns $w$ perfectly. With remaining probability $1-q$, she is not informed. If the Sender does not lie (i.e., $m = w$), the message is never flagged as a lie and the Receiver is not informed.

Formally, the detection technology can be described by the following relation

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

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 immediately comparable benchmark.

Note that lie detection here is different from state detection. While the former would inform the Receiver the true state conditional on a lie, the latter would inform her the true state independently of the message. Section 4 discusses their differences in more detail.[2]

## 2.2 Utilities

Given both $m$ and $d$, the Receiver takes an action $a \in \{0,1\}$, and the payoffs are realized. The payoffs are defined as follows.

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

That is, the Sender wants the Receiver to always take the action $a = 1$ regardless of the state, while the Receiver wants to match the state. The payoff from matching the state 0 may differ from the payoff from matching the state 1. Given the payoff function, the Receiver takes action $a = 1$

[2]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. An alternative definition of messages and lies views a message as a lie if, in equilibrium, this message induces an action that is inconsistent with the true state of nature. This alternative definition is more complicated and involves calculating a fixed point. Essentially, one starts with an arbitrary lying set $L \subset \{(m=1,\omega=1),(m=1,\omega=0)\} \times \{(m=0,\omega=1),(m=0,\omega=0)\}$ and then solves for the Sender's optimal solution when any pair in $L$ triggers a lie detection with probability $q$. This provisional solution generates a new lying set $L'$. A consistency condition $L = L'$ is thus required to close the model. We do not adopt this alternative definition because it leads to a multiplicity of equilibria which makes the comparative statics difficult.

if and only if

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

and therefore one could also interpret $t$ as the threshold of the Receiver's posterior belief above 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, different from standard persuasion models, it is non-trivial even if $t \leq \mu$ because there exists no message that is purely uninformative. We defer a detailed discussion of this case to Section 4.4.

## 2.3 Strategies

We assume that the Sender has full commitment power as is common in the Bayesian persuasion framework.[3] Specifically, the strategy of the Sender is a mapping $m : \{0, 1\} \longrightarrow \Delta(\{0, 1\})$, and the strategy of the Receiver is a mapping $a : \{0, 1\} \times \{lie, \neg lie\} \longrightarrow \Delta(\{0, 1\})$. Formally, the Sender is choosing $m(\cdot)$ to maximize

$$\mathbb{E}_{w,d,m}[u_S(a\left(m(w), d(m(w), w)\right), w)]$$

where $a(m, d)$ maximizes

$$\mathbb{E}_w[u_R(a, w) \mid m, d].$$

The two expectation signs are taken with respect to different variables. The expectation sign in the Sender's utility is taken with respect to both $w$, $d$, and perhaps $m$ if the strategy is mixed, whereas the (conditional) expectation sign in the Receiver's utility is only taken with respect to $w$. Due to the simple structure of the model, it is without loss of generality to assume that the Sender chooses only two parameters $p_0 = \Pr(m = 0 \mid w = 0)$ and $p_1 = \Pr(m = 1 \mid w = 1)$ to

[3]For a detailed discussion and relaxation of this assumption see Min (2021), Fréchette et al. (forthcoming), Lipnowski et al. (forthcoming), and Nguyen and Tan (2021). Titova (2021) shows that with binary actions and a sufficiently rich enough state space verifiable disclosure enables the Sender's commitment solution as an equilibrium. In Section 4.1 we show that our results continue to hold under partial commitment.

maximize $\Pr(a(m,d)=1)$ which we write as $\Pr(a=1)$ henceforth for brevity of notation. We denote the optimal reporting probabilities of the Sender by $p_0^*$ and $p_1^*$, and the ex-ante payoffs under this reporting probabilities as $U_S$ and $U_R$.

# 3 Analysis

## 3.1 Optimal Messages

Given the Sender's reporting strategy, the Receiver could potentially see four types of events to which she needs to react when choosing action $a$.

First, the Receiver could observe the event $(m=0,\ d=lie)$ which occurs with probability $\mu(1-p_1)q$. Given the lie detection technology, the Receiver is certain that the message $m=0$ is a lie. Therefore, the state of the world $w$ must be equal to 1, that is

$$\Pr(w=1 \mid m=0, d=lie)=1.$$

As a result, the Receiver optimally chooses $a=1$. [4]

Second, the event $(m=0, d=\neg lie)$ could occur with probability $\mu(1-p_1)(1-q)+(1-\mu)p_0$. In that case, the Receiver is uncertain about $w$ because she does not know whether the Sender lied or not. Her posterior probability is given by

$$\Pr(w=1 \mid m=0, d=\neg lie)=\frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p_0}\equiv\mu_0.$$

Hence, the Receiver takes action $a=1$ if and only if $\mu_0 \geq t$. We denote the posterior following this event by $\mu_0$ (and thus omitting the lie detection outcome $d=\neg lie$) for brevity of notation. When $p_0=0,\ p_1=1$, this event occurs with 0 probability, so the belief is off-path and not restricted by Bayesian updating. However, the off-path belief does not matter for the Sender, because if the Sender chooses the strategy that renders $(m=0, d=\neg lie)$ a zero probability event,

[4] The posterior belief following a lie detection is always degenerate because the state space is binary. However, the binary structure is not driving the main results on players' equilibrium payoffs. See Section 4.2 for a detailed discussion.

he does not care about how the Receiver responds to that event. For expositional convenience, define $\mu_0 = 0$ when $p_0 = 0,\ p_1 = 1$.

Third, $(m = 1,\ d = lie)$ occurs with probability $(1-\mu)(1-p_0)q$. Because a lie was detected, the Receiver is again certain about $w$ and therefore her posterior probability is given by

$$\Pr(w = 1 \mid m = 1, d = lie) = 0,$$

which immediately implies the action $a = 0$.

Fourth, $(m = 1, d = \neg lie)$ occurs with probability $\mu p_1 + (1-\mu)(1-p_0)(1-q)$. The Receiver is again uncertain about $w$. Her posterior is given by

$$\Pr(w = 1 \mid m = 1, d = \neg lie) = \frac{\mu p_1}{\mu p_1 + (1-\mu)(1-p_0)(1-q)} \equiv \mu_1$$

and the Receiver takes action $a = 1$ if and only if $\mu_1 \geq t$. Analogously, for brevity of notation, we denote the posterior following this event by $\mu_1$ (and thus omitting the lie detection outcome $d = \neg lie$). Similarly, if $p_0 = 1,\ p_1 = 0$, this event occurs with 0 probability, and the belief $\mu_1$ is not well-defined, but again this does not matter for the Sender. For simplicity, define $\mu_1 = 0$ when $p_0 = 1,\ p_1 = 0$.

Given these optimal responses by the Receiver, the relationships between the posteriors $\mu_0, \mu_1$ and the posterior threshold $t$ divide up the strategy space into four different types of strategies which we denote by I, II, III, and IV respectively. For each strategy type, the Receiver's response as a function of $(m, d)$ is the same, making it then easy to find the specific optimal strategy. We are then left to pick the best strategy out of the four candidates. These types of strategies are defined as follows:

I. $\mu_0 < t,\ \mu_1 < t$: For this type of strategy, the Receiver only chooses $a = 1$ if $(m = 0, d = lie)$ and $a = 0$ otherwise because the posteriors $\mu_0$ and $\mu_1$ are insufficiently high to persuade her to choose $S$'s preferred action. Only if the Sender lies in state $w = 1$ and his message is detected as a lie, is the Receiver sufficiently convinced that $a = 1$ is the right action. The

maximal probability that the Receiver chooses $a = 1$[5] is given by

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

II. $\mu_0 \geq t,\ \mu_1 < t$: The Receiver chooses $a = 1$ if $(m = 0, d = lie)$ or $(m = 0, d = \neg lie)$ and $a = 0$ otherwise. The maximal probability that the Receiver chooses $a = 1$ is given by

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

III. $\mu_0 < t,\ \mu_1 \geq t$: The Receiver chooses $a = 1$ if $(m = 0, d = lie)$ or $(m = 1, d = \neg lie)$ and $a = 0$ otherwise. The maximal probability that the Receiver chooses $a = 1$ is given by

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

IV. $\mu_0 \geq t,\ \mu_1 \geq t$: The Receiver chooses $a = 1$ if $(m = 0, d = lie)$, $(m = 0, d = \neg lie)$ or $(m = 1, d = \neg lie)$ and $a = 0$ otherwise. The maximal probability that the Receiver chooses $a = 1$ is given by

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

Table 1 summarizes when the Receiver chooses $a = 1$ under the different types of strategies. Notably, given the definition of off-path beliefs, $(0, 1)$ is a type III strategy and $(1, 0)$ is a type I strategy. We are now ready to state the main proposition of our model.

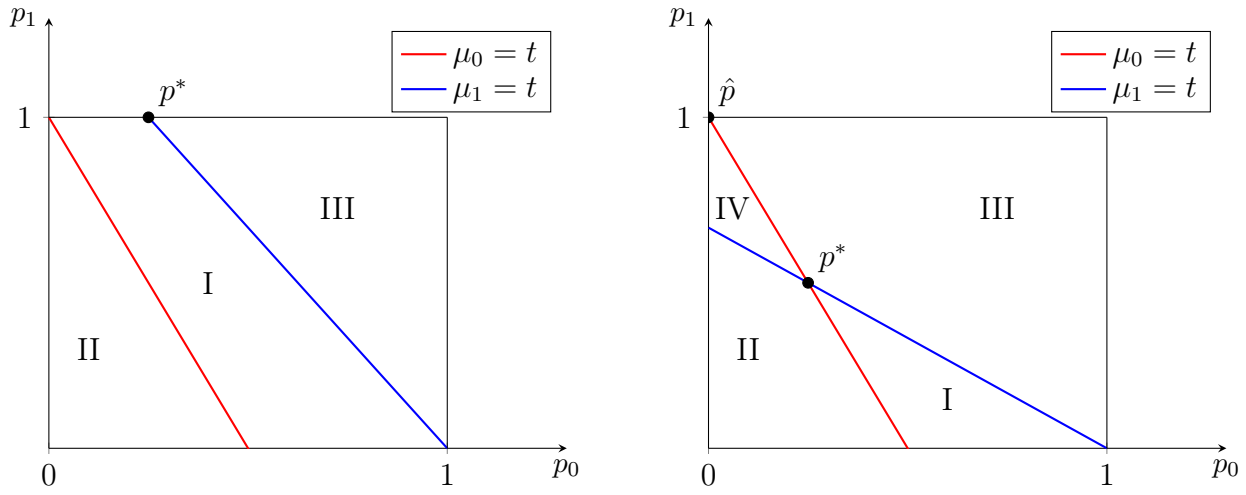
| | $d = lie$ | $d = \neg lie$ |
|---|---|---|
| $m = 0$ | I, II, III, IV | II, IV |
| $m = 1$ | | III, IV |

Table 1: Cases where the Receiver chooses $a = 1$ under I, II, III, and IV.

[5]The choice set of the maximization problem is not closed, so the maximum may not be achieved a priori.

**Proposition 1.** *Let* $\overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)} \in (0,1)$. *If* $q \leq \overline{q}$, *the Sender's optimal strategy is a type* III *strategy, in which the Sender always tells the truth under* $w = 1$, *but lies with positive probability under* $w = 0$. *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 how these four strategy types are divided. The proof involves sequential comparisons between the four type-optimal strategies. First, there exists *some* type II strategy that is better than *all* type I strategies. Consider a particular strategy $p_0 = p_1 = 0$ of type II (i.e., the Sender totally misreports the state). Following this strategy, the Receiver takes action $a = 1$ if and only if $w = 1$, which occurs with probability $\mu$. This strategy may not be optimal among all type II strategies, but it is sufficient to beat all strategies of type I since for those strategies the Receiver takes action $a = 1$ only if $w = 1$ and $(m = 0,\ d = lie)$, which occurs with a probability less than $\mu$.




(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 just need to focus on the ones with $p_1 = 0$ because lying more under state $w = 1$ relaxes both constraints and is beneficial for the Sender. Now, for any type II strategy of the form $(p_0, 0)$, consider a strategy $(\tilde{p}_0, 1)$ such that $p_0 = (1 - \tilde{p}_0)(1 - q)$. It can then be verified that this is a type III strategy. Moreover, this new strategy is equally good as $(p_0, 0)$ for the Sender

by construction.

To see the intuition for this result, note that the type II and III strategies are totally symmetric if the lie detection technology is not available ($q = 0$) since in that 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 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 $w = 1$ with a sufficiently high probability, while $(m = 1,\ d = \neg lie)$ suggests $w = 0$ 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 type III strategies, and we only need to focus on the comparison between type III and type IV strategies. Interestingly, as suggested by Figure 1 $(a)$, type IV strategies do not exist when $q$ is small. The proof is given in Appendix A. Intuitively, when $q = 0$ our setup yields the standard Bayesian Persuasion benchmark, which essentially only involves two events $(m = 0, d = \neg lie)$ and $(m = 1, d = \neg lie)$. In that case, we know it is impossible to induce $a = 1$ under both events because by the martingale property, the posteriors following two events must average to the prior, suggesting some posterior is lower than the prior and must induce $a = 0$. However, the presence of lie detection extends the information from $m$ to a couple $(m,\ d)$, and the martingale property only requires the four posteriors' average over the prior. Furthermore, the posterior following $(m = 1,\ d = lie)$ is 0. Therefore if $q$ is sufficiently large, it is possible to support the two posteriors following $(m = 1,\ d = \neg lie)$ and $(m = 0,\ d = \neg lie)$ to be both higher than the prior and even higher than the threshold $t$.

As shown by Figure 1 $(a)$, the constraint $\mu_0 < t$ is implied by the constraint $\mu_1 \geq t$. Hence, the set of type III strategies is compact, and the associated maximization problem admits a solution. Combining this observation with the previous arguments, we immediately obtain the first half of Proposition 1, *i.e.*, the Sender's optimal strategy is a type III strategy if $q \leq \overline{q}$. In particular, the

optimal strategy takes the following form:

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

This result is reminiscent of Kamenica and Gentzkow (2011) where the Receiver is indifferent between two actions when she takes the preferred action $a = 1$ and certain of the state when she takes the less preferred action $a = 0$.

If the detection probability $q$ is larger than $\overline{q}$, the two lines that characterize the constraints in the right panel of Figure 1 intersects, implying the set of type III strategies is not closed anymore. However, the associated maximization problem still admits a solution: $(p_0, p_1) = (0, 1)$, which is a type III strategy according to off path beliefs specified earlier. This strategy can be shown to be optimal within type III strategies in two steps. First, increasing $p_1$ relaxes both constraints and improves the Sender's expected payoff at the same time (i.e., being more sincere in the favorable state benefits the Sender unambiguously). Thus, the optimal type III strategy, if it exists, must be of the form $(p_0, 1)$. Second, the whole segment from $(0, 1)$ to $(1, 1)$ are type III strategies when $q > \overline{q}$. Hence, the optimal strategy on this segment is the leftmost point $(0, 1)$ as it involves sending the persuasive message $m = 1$ as frequently as possible.

Yet, this optimal type III strategy, denoted as $\hat{p}$ in Figure 1 $(b)$, is no longer globally optimal because the set of type IV strategies is non-empty, and the optimal type IV strategy is better than $\hat{p}$. In fact, we can prove a stronger statement that $\hat{p}$ is worse than any type IV strategy $p$ whenever the latter is feasible. To this end, we decompose the value of a strategy for the Sender into two parts: the expected payoff in the favorable state $w = 1$ and the expected payoff in the unfavorable state $w = 0$. The strategy $\hat{p}$ induces $a = 1$ for sure when $w = 1$ 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 definition of type IV strategies. The last event directly informs the Receiver that $w = 1$, so it also induces $a = 1$. Hence, the strategy $\hat{p}$ and $p$ agree in the expected payoff in the favorable state $w = 1$. However, they differ in the expected payoff in

the unfavorable state $w = 0$. Given $\hat{p}$, the Sender always lies and sends the message $m = 1$ when $w = 0$, 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 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 $w = 0$. Hence, the strategy $p$ results in a higher expected payoff for Sender in the unfavorable state as well as overall. Mathematically,

$$U_S(\hat{p}) = \underbrace{\mu}_{\Pr(w=1)} \times \overbrace{1 \times 1}^{\Pr(a=1|w=1;\hat{p}_1)} + \underbrace{(1-\mu)}_{\Pr(w=0)} \times \overbrace{1 \times (1-q)}^{\Pr(a=1|w=0;\hat{p}_0)}$$

and

$$U_S(p) = \underbrace{\mu}_{\Pr(w=1)} \times \overbrace{[p_1 \times 1 + (1-p_1) \times (1-q) + (1-p_1) \times q]}^{\Pr(a=1|w=1;p_1)} + \underbrace{(1-\mu)}_{\Pr(w=0)} \times \overbrace{[p_0 \times 1 + (1-p_0) \times (1-q)]}^{\Pr(a=1|w=0;p_0)}$$

where the first term ($\mu \times 1$) is the same for the two expressions, but the second term is larger for $U_S(p)$ since $p_0$ is not multiplied by $1 - q$ but instead by 1. 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 ($p_0^* < p_1^*$).

Interestingly, the difference between the Sender's payoffs of the strategy $\hat{p}$ and $p^*$ is non-monotone in the detection probability $q$. When $q = \overline{q}$, $\hat{p}$ coincides with $p^*$, so they are equally

good. When $q = 1$, it is as if the Receiver is informed about the state with probability 1, so any strategy results in the same payoff for the Sender. Only when $q \in (\overline{q}, 1)$, $p^*$ yields a strictly higher payoff than $\hat{p}$.

Finally, the threshold $\overline{q}$ where the optimal strategy switches from a type III to a type IV strategy, is decreasing in $\mu$ and increasing in $t$. To see the intuition for this result, fix the lie detection probability $q \in (0, 1)$. If a weak signal is sufficient to persuade the Receiver (i.e., the prior $\mu$ is already close to the threshold $t$), a type IV strategy is optimal for the Sender. On the other hand, if the signal has to be very convincing to persuade the Receiver (i.e., the threshold $t$ is much larger than the prior $\mu$), a type III strategy is optimal for the Sender.

## 3.2 Comparative Statics

We now consider the comparative statics of our model with respect to the central parameter of the lie detection probability $q$ to show how the optimal communication and the utilities of the communicating parties changes as the lie detection technology improves.

### 3.2.1 Optimal Messages

Proposition 2 describes how the structure of the optimal message 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.

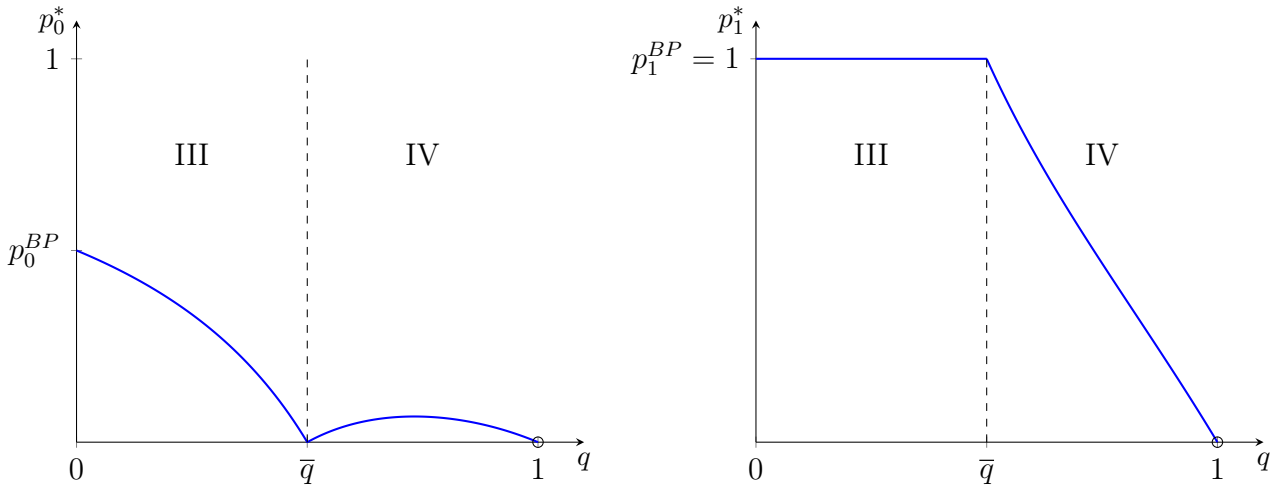
**Proposition 2.** *As the lie detection probability $q$ increases,*

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

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

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 lies in III, which involves truthfully reporting the state $w = 1$ (i.e., $p_1 = 1$), but progressively misreporting the state $w = 0$ as the lie detection technology improves (i.e., $p_0 < 1$ and decreasing with $q$).

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 lies in IV which involves misreporting both states of the world.




(a) $p_0^* \equiv \Pr(m = 0 \mid w = 0)$ (b) $p_1^* \equiv \Pr(m = 1 \mid w = 1)$

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 $q = 0$ we have the Bayesian benchmark. 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 $w = 1$, 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 the action $a = 1$. Because the Sender already sends the message $m = 1$ with probability 1 under $w = 1$, the only way to increase the frequency of $m = 1$ is to send such a message more often in the unfavorable state $w = 0$ (i.e., lie more frequently if $w = 0$). In other words, the Sender increases the frequency of lying just enough about the unfavorable state ($w = 0$) to make the Receiver indifferent when choosing the favorable action

$a = 1$.

Recall that in the canonical Bayesian persuasion setup, the Receiver is held to her outside utility of getting no information whatsoever. Thus, when the lie detection probability $q$ increases, the Receiver is more certain that $(m = 1,\ d = \neg lie)$ means $w = 1$ 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 improvement by lying more about $w = 0$ (i.e., reduce $p_0^*$ even further), thereby "signal-jamming" the information obtained by the Receiver.

However, once the detection probability $q$ rises above $\overline{q}$ it is no longer possible for the Sender to just 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 when $q > \overline{q}$. Under a type IV strategy, the Receiver only takes the unfavorable action $a = 0$ if he receives a message $m = 1$ that is flagged as a lie. This is because with a type IV strategy the Receiver has access to such a reliable lie detection technology that a lie involving the message $m = 1$ is sufficiently likely to be detected as a lie and will then induce the unfavorable action $a = 0$. At the same time, the Receiver is also very likely to be notified of a lie involving the message $m = 0$ which the Sender can use to his advantage to ensure that the Receiver chooses the favorable action $a = 1$. Therefore at $q = \overline{q}$, the Sender wants to increase the frequency of the message $m = 0$ which he achieves 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 $m = 1$ which 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 $p_1$.

These perhaps surprising comparative statics, especially those of the type IV strategy, are partly due to the asymmetric nature of the signal structure (as in Engers et al. (1999)) which in our case only detects lies rather than detecting both lies and truths, and partly 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.

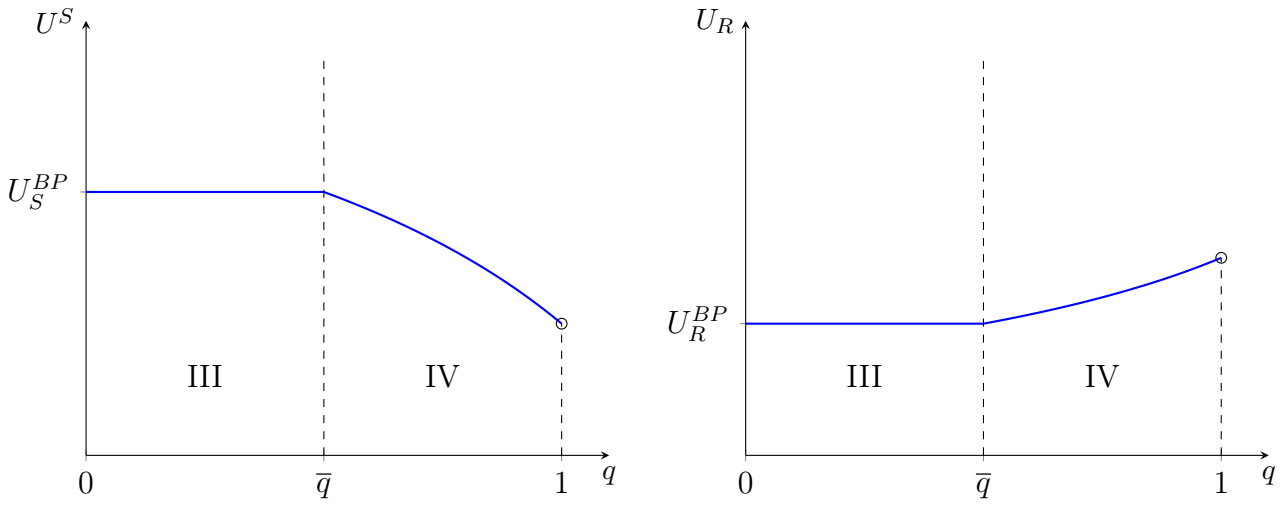
### 3.2.2 Utilities

Recall that $U_S$ and $U_R$ denote the equilibrium payoffs of the Sender and the Receiver. We now investigate how $U_S$ and $U_R$ are affected by improvements in the lie detection technology. The results are summarized in Proposition 3 and graphically depicted in Figure 3. For comparison, the utilities $U_S^{BP}$ and $U_R^{BP}$ are the equilibrium utilities that would result in a standard Bayesian persuasion setup without lie detection.

**Proposition 3.** *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) The Sender's Equilibrium Payoff

(b) The 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 utility that he would receive under the Bayesian Persuasion benchmark. Any marginal improvement in the lie detection technology (i.e., increase in $q$) is completely offset by less truthful reporting when $w = 0$ (i.e., decrease in $p_0^*$). However, for $q > \overline{q}$ any further improvements reduce the Sender's utility. In the limit case where $q = 1$

the Sender has no influence anymore and the action $a = 1$ in only implemented when the state is $w = 1$ which occurs with probability $\mu$.

Analogously for the case of the Sender's utility, the Receiver's utility 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. 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 $\overline{q}$. In the limit, the Receiver is just as well off as she would be under perfect information.

# 4 Extensions and Discussion

Our baseline model considers the role of lie detection in a simple setting with full commitment and binary states. We now investigate how alternative assumptions about the Sender's commitment, the state 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. This is also true in our model if the Sender cannot commit at all. However, under partial commitment the main insights of our baseline model continue to hold.

If the Sender cannot commit at all to a communication strategy, then for any $q \in [0, 1]$, the derived optimal messaging strategy $(p_0^*, p_1^*)$ in the baseline model fails to be part of an equilibrium. For the purpose of illustration, let us restrict attention to the case in which the lie detection technology is not too strong (i.e., $q \leq \overline{q}$). In this case, the optimal messaging strategy $(p_0^*, p_1^*)$ is equal to $\left(\frac{\overline{q}-q}{1-q}, 1\right)$. Given this strategy, $(m = 1, \neg lie)$ induces $a = 1$ while $(m = 0, \neg lie)$ induces $a = 0$. Thus, the Sender would like to send message $m = 1$ with probability one in both states if his commitment is not binding. The Receiver would then find it optimal to take action $a = 0$ even after observing $(m = 1, \neg lie)$. It can be shown that when the Sender lacks all commitment, the Receiver is always strictly better off with a stronger lie detection technology (i.e., higher detection

probability $q > 0$), which is in stark contrast to Proposition 3.

We now demonstrate the robustness of our main results to partial commitment. Following Min (2021) we assume that the Sender's commitment only binds probabilistically. The generalized game with partial commitment therefore 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 $\alpha$ his commitment binds and he has to send a message following the pre-specified 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 non-binding commitment where $\tilde{p}_i$ is the probability that he sends a message $i \in \{0,1\}$ when the true state is $i$ and the commitment does not bind. 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 lie detection outcome. She is aware that the Sender may not abide by his commitment strategy and the probability $\alpha$ 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 the one given in Section 2.2. 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.

As usual, we focus on the Sender's preferred equilibrium. Although a complete loss of commitment drastically changes the equilibrium and the corresponding payoffs, the following proposition suggests that a small loss of commitment has no impact on the key features of the equilibrium strategy 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.

**Proposition 4.** *Assume* $q \leq \overline{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)}$ *and* $\alpha \geq \overline{\alpha} = \frac{\overline{q}-q}{1-q}$. *Then in the Sender's preferred equilibrium,* $p_1^* = 1$, $\tilde{p}_0^* = 0$, $\tilde{p}_1^* = 1$, *and* $p_0^*$ *is such that the Receiver is indifferent between* $a = 0$ *and* $a = 1$ *after observing* $(m = 1, d = \neg lie)$. *Moreover,* $U_S(q)$ *and* $U_R(q)$ *are both constant in* $q$.

By analogous arguments, if the lie detection technology is weak, it is impossible to induce $a = 1$ for both $(m = 1, d = \neg lie)$ and $(m = 0, d = \neg lie)$. The Sender can at most induce the favorable

action for one of the two events and he prefers to induce it for the first event. Thus, 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 strategy following a non-binding commitment ($\tilde{p}_0$, $\tilde{p}_1$) must be $(0, 1)$ (i.e., he prefers to send $m = 1$ regardless of the state). Hence, the message generated from ($\tilde{p}_0$, $\tilde{p}_1$) is totally uninformative per se, but the combination of $m$ and $d$ still provides some information.

Intuitively, it is as if the Receiver receives a signal of unknown informativeness. If we ignore the lie detection technology, then with probability $\alpha$, 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 must still satisfy the Receiver's indifference condition. Thus, the Sender must commit to a more informative messaging strategy than in the baseline model. In fact, in the Sender's preferred equilibrium, the probability of sending $m = 0$ under the favorable state in the pre-specified strategy satisfies

$$\alpha[1 - p_0^*(\alpha)] + 1 - \alpha = \frac{1 - \overline{q}}{1 - q}, \ \forall \alpha \geq \overline{\alpha}.$$

It follows then that $p_0^*$ decreases in $\alpha$. The lower bound on the probability that the commitment binds also has a natural interpretation. If the average informativeness leaves the Receiver indifferent, then the Receiver must prefer to take action $a = 1$ when the Sender commits to the most informative signal (i.e., the truth-telling strategy ($p_0$, $p_1$) = (1, 1)). This condition implies the lower bound on $\alpha$.

On the other hand, $p_0^*$ again 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 payoff 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.

## 4.2 Non-revealing Lie Detection

In a binary-state environment, the lie detection technology considered in this paper is quite special in the sense that whenever the Receiver learns that the Sender has lied, she immediately learns the true state. However, Proposition 3 is not driven by this special feature.[6] Intuitively, in a binary-state environment, the lie detection technology forces the release of too much information to the Receiver. Still, we can show that the Receiver does not benefit from a weak lie detection technology, even in this case. Following this reasoning, when the Receiver's posterior beliefs after a lie detection are non-degenerate, the Receiver obtains less information. Therefore, we would expect Proposition 3 to be strengthened instead of weakened. In fact, we show that in a three-state environment, lie detection technology is completely useless in the sense that both the Sender's and the Receiver's payoffs are unaffected by the strength of lie detection. This suggests that fully revealing lie detection is not the driving force of our main results.[7]

Formally, let $\omega \in \{0, \lambda, 1\}$ be the state of the world and $(P_0, P_\lambda, P_1)$ be the full-support prior belief where $\lambda \in (0, 1)$. The message space is again restricted to be identical to the state space and a lie is detected with probability $q \in [0, 1]$ whenever the message is inconsistent with the true state. For simplicity, keep the player's utility functions unchanged. In particular, the Sender always prefers $a = 1$ over $a = 0$ regardless of the true state, whereas the Receiver takes an action $a = 1$ if and only if her posterior mean is higher than an action threshold $t$. Assume $t \in (\mu, \lambda)$, where $\mu = P_1 + \lambda P_\lambda$ is the prior mean.[8] The implication of this restriction is twofold. First, the Receiver's default action is $a = 0$. Second, if the Receiver knows the state is $\lambda$, she prefers to take an action $a = 1$. In other words, both $\omega = 1$ and $\omega = \lambda$ are favorable states for the Sender. To

[6] In the three-state environment, the uniqueness of our equilibrium is not necessarily guaranteed. Thus, it is hard to generalize the comparative statics of Proposition 2. However, in all of these equilibria the Sender still has a strategic incentive to change his messaging strategy in response to changes in the lie detection probability.

[7] An alternative approach to model non-revealing lie detection is to modify the lie detection technology by introducing false alarms. In other words, a lie may be detected even if the message is consistent with the true state. However, this approach is more complicated. So, we adopt the approach of expanding the state space.

[8] The choice of $t$ is not important for the extension. However, it affects two things and complicates the exposition. First, it affects the Receiver's utility function. Second, it affects the Sender's equilibrium payoff in the benchmark.

quantify the Receiver's utility explicitly, let her payoff function and expected payoff be

$$u_R(a,\omega) = (1-t)\cdot \mathbb{1}_{\{a=1,\omega=1\}} + (\lambda - t)\cdot \mathbb{1}_{\{a=1,\omega=\lambda\}} + t\cdot \mathbb{1}_{\{a=0,\omega=0\}}.$$

and

$$U^R = (1-t)\cdot \Pr(a=1,\omega=1) + (\lambda - t)\cdot \Pr(a=1,\omega=\lambda) + t\cdot \Pr(a=0,\omega=0).$$

This utility function is analogous to the one in the main body. The Receiver would like to take the right action for each state but assigns different weights for different states. The particular choice of weights induces a decision rule that it is optimal to take $a = 1$ if and only if the posterior is higher than $t$.

The goal here is to show that both players' payoffs are constant in $q$. To this end, we first compute the Sender's payoff in the benchmark scenario ($q = 0$) and then construct a strategy that leads to the same payoff for the Sender with any detection probability. Last, we show that in any Sender's preferred equilibrium, the Receiver's payoff is constant. Moreover, this constant is independent of $q$.

The optimal signal/messaging strategy in the classical Bayesian Persuasion framework with a binary action has been analyzed in the literature. Ivanov (2021) shows that in a binary-action and continuous-state environment, there exists an optimal strategy with a partitional structure where the Sender sends a message if the state is above some threshold and sends another message otherwise. By applying the insight to our discrete-state model, there exists an optimal strategy with the following properties. The Sender sends one (another) message if the state is strictly higher (lower) than some threshold state. Moreover, he mixes between two messages at the threshold state.

In particular, the following strategy achieves the optimum.

$$\textbf{Strategy 1:} \qquad \begin{aligned} \omega = 1 &\longrightarrow m = 1 \\ \omega = \lambda &\longrightarrow m = 1 \\ \omega = 0 &\longrightarrow m = \begin{cases} 1, & w.p. \quad r \\ 0, & w.p. \quad 1 - r, \end{cases} \end{aligned}$$

where $r$ solves

$$\frac{P_1 + \lambda P_\lambda}{P_1 + P_\lambda + rP_0} = t.$$

Essentially, the mixing probability $r$ ensures the Receiver to be indifferent after observing $m = 0$. Given Strategy 1, the Receiver takes the favorable action if and only if she receives $m = 1$. Thus,

$$U^S(0) = P_1 + P_\lambda + rP_0 = \frac{\mu}{t}.$$

Now, suppose there is a lie detection probability $q > 0$. In principle, this limits the Sender's scope to manipulate the Receiver's posterior beliefs, and thus potentially lowers the Sender's payoff. However, the following strategy yields the Sender the same payoff as in the benchmark. Moreover, this strategy is independent of $q$, suggesting that lie detection has not impact on the Sender's payoff at all.

$$\textbf{Strategy 2:} \qquad \begin{aligned} \omega = 1 &\longrightarrow m = \lambda \\ \omega = \lambda &\longrightarrow m = 1 \\ \omega = 0 &\longrightarrow m = \begin{cases} 1, & w.p. \quad r \\ \lambda, & w.p. \quad s \\ 0, & w.p. \quad 1 - r - s \end{cases} \end{aligned}$$

where $r$ and $s$ respectively solve

$$\begin{cases} \frac{\lambda P_\lambda}{P_\lambda + P_0 r} = t, \\ \frac{P_1}{P_1 + P_0 s} = t. \end{cases}$$

The assumption $\mu < t < \lambda$ ensures that $s, r, s + r \in (0, 1)$. Given Strategy 2, the Receiver is indifferent after observing $(m = 1, d = \neg lie)$, $(m = 1, d = lie)$, $(m = \lambda, d = \neg lie)$, and $(m = \lambda, d = lie)$. So, she takes the favorable action if and only if she receives $m = 1$ or $m = \lambda$, regardless of the lie detection outcome. It follows that

$$U^S(q) = P_1 + P_0 r + P_\lambda + P_0 s = \frac{\lambda P_\lambda}{t} + \frac{P_1}{t} = \frac{\mu}{t} = U^S(0), \quad \forall q \in (0, 1]. \tag{1}$$

Lie detection is not useful here because conditional on this particular strategy, the message $\lambda$ and the message 1 are always lies, whereas a message 0 is never a lie. Moreover, the probability of lie detection is constant as long as the Sender is lying. Thus, lie detection does not provide any additional information for the Receiver, no matter how strong it is.

Since we focus on the Sender's preferred equilibrium, the Receivers' payoff is potentially non-unique. Fortunately, that is not the case here. Lemma 1 guarantees that the Receiver's payoff is unique. In addition, it is always linear in the Sender's equilibrium payoff with a negative slope.

**Lemma 1.** *Fix a lie detection probability $q \in [0, 1]$. If $\mu < t < \lambda$, then in any Sender's preferred equilibrium,*

$$U^R(q) = (1 - t)P_1 + (\lambda - t)P_\lambda + t[1 - U^S(q)].$$

Here is the key intuition of this result. Note that both $\omega = 1$ and $\omega = \lambda$ are favorable states for the Sender. Thus, it is optimal to induce $a = 1$ under those two states, which suggests $\Pr(a = 1, \omega = 1) = \Pr(a = 1, \omega = \lambda) = 1$ in any Sender's preferred equilibrium. Then, the Receiver's expected payoff only depends on $\Pr(a = 0, \omega = 0)$. But this probability is completely determined by the Sender's optimality and is therefore linked to the Sender's equilibrium payoff.

Roughly speaking, the Sender wishes to minimize this probability while conditional on $\Pr(a = 1, \omega = 1) = \Pr(a = 1, \omega = \lambda) = 1$.

Combining equation (1) and Lemma 1, it is immediate that the Receiver's equilibrium payoff is also independent of $q$ and given by

$$U^R(q) = (1 - t)P_1 + (\lambda - t)P_\lambda + t - \mu = tP_0.$$

Our analysis thus suggests that fully revealing lie detection does not drive Proposition 3.

### 4.3 Truth and State Detection

First, consider a different detection technology that informs the Receiver with probability $r$ that a message is truthful. That is to say, 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 often easier to detect whether the Sender has lied than whether he has sent a truthful message.

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 message 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.

Second, combining lie detection and truth detection such that they are perfectly positively correlated is equivalent to state detection. With probability $q = r$ the Receiver learns the state $w$ regardless of the message sent by the Sender. With such a state detection technology the analysis becomes much simpler as we just 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 $m$ is only relevant whenever the Receiver does not learn the state. This finding contrasts with the literature on noisy cheap talk in which adding communication error or noise influences

the messaging strategies and can improve welfare (Blume et al., 2007).

These observations 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 the state.

## 4.4 Default Action Coincides with Sender's Preferred Action

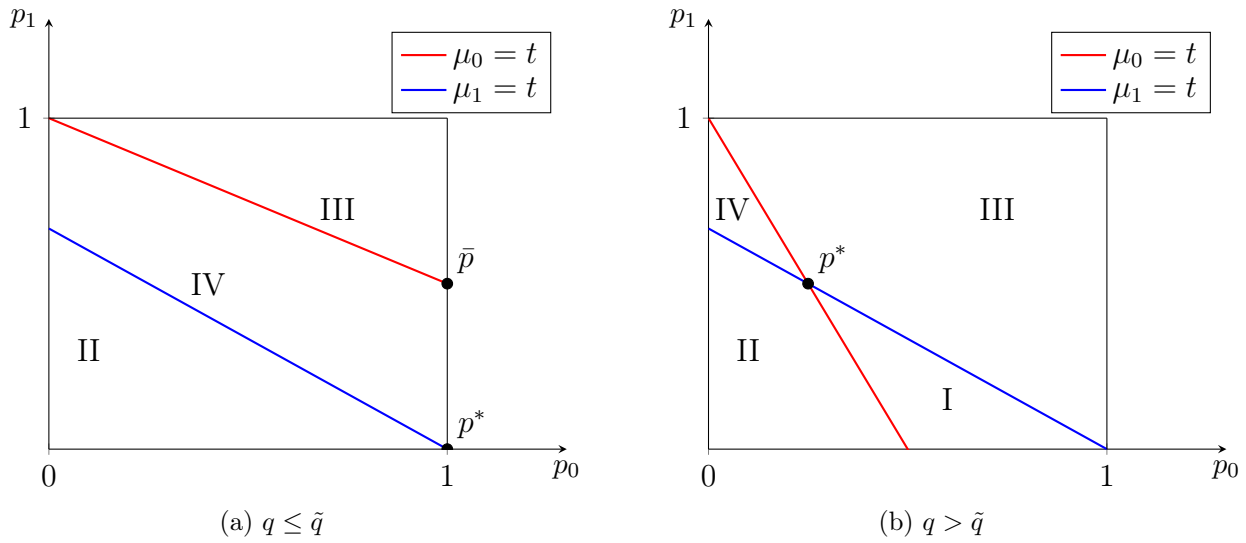
In standard Bayesian persuasion models without lie detection the Sender can always send a purely uninformative signal. Therefore, a trivial case obtains if the Receiver's default action coincides with the Sender's preferred action because the Sender can induce the Receiver to take this action with probability one by committing to an uninformative signal. However, the messages in our model have literal meanings and are subject to lie detection. Therefore, a purely uninformative signal is unavailable to the Sender. Intuitively, lie detection forces information transmission from the Sender to the Receiver which makes the Sender's optimization problem nontrivial even when the Receiver's default action coincides with the Sender's preferred action.

In this extension, we analyze the scenario in which the prior mean $\mu$ is higher than the action 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 sub-problems. The only change relative to the baseline model is that Region IV now exists for any $q \in [0, 1]$ as shown in Figure 4.

The optimal messaging strategy $p^*$ is always in Region IV. When $q \in \tilde{q} \equiv 1 - \frac{t(1-\mu)}{\mu(1-t)}$, the strategy $(p_0, p_1) = (1, 0)$ would induce the Receiver to take $a = 1$ with probability one and is thus optimal.[9] Under this strategy, the Sender reports $m = 0$ with probability one in both states. If it is flagged as a lie, the Receiver immediately learns that the true state is $\omega = 1$. Otherwise, her posterior mean would drop by the martingale property. 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

[9]It is not unique though because the Receiver actually strictly prefers to take $a = 1$ when she observes $(m = 0, \neg lie)$. Other optimal strategies include the segments from $p^*$ to $\bar{p}$.




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

Figure 4: Equilibrium message strategies for different detection probabilities $q$ ($\mu \geq t$).

$a = 1$ with probability one. For example, in the extreme case where $q = 1$, it is as if the Receiver 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 4 (b).

Given the discussion above, the Sender's (Receiver's) payoff is initially constant in $q$ when $q \leq \tilde{q}$, and then decreasing (increasing) in $q$ when $q > \tilde{q}$. This is consistent with Proposition 3.

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. In the standard paradigm without lie detection this case is trivial whereas in ours it is not. However, it is easy to reconcile this tension by introducing an additional stage prior to the persuasion game, in which the Sender decides whether or not to enter the game. If 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.

## 4.5 Related Literature

Balbuzanov (2019) and Dziuda and Salas (2018) also study strategic communication in the presence of a lie detection technology but in a cheap talk setting. The largest difference between these two papers and ours therefore lies in the commitment power of the Sender. 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 setting, we believe our model is an important step towards studying the communication games with lie detection under (partial) commitment.

Our paper also differs from Balbuzanov (2019) in the payoff functions. In Balbuzanov (2019) the Sender and the Receiver have some degree of common interest whereas there is no common interest in our model. Due to this difference the Sender's type-dependent preferences in Balbuzanov (2019) permit fully revealing equilibria in some cases as it allows the Receiver to tailor message-specific punishment actions. In particular, fully revealing equilibria exist for some intermediate degree of lie detectability if the Sender's bias is small. However, the Sender in our model never reveals the state perfectly due to the conflict in payoffs.

Dziuda and Salas (2018) do not allow for common interest and therefore, like in our paper, fully revealing equilibria are impossible in their paper. In their continuous state model, there are many off-path beliefs to be specified. To discipline these off-path beliefs, they impose two refinements. They show that in all remaining equilibria, the lowest types lie but *some* higher types tell the truth. However, the assumptions of our model allow the second refinement required by Dziuda and Salas (2018) to be violated. Therefore, irrespective of the commitment power of the Sender, our model is not nested by theirs. Furthermore, in the baseline model of Dziuda and Salas (2018), a higher lie detection probability leads to more truth-telling, which is the exact opposite of our finding.

# 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 payoff are unaffected by the 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 their 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 spaces, and under different detection technologies that inform the Receiver's action. Nonetheless, our analysis raises further questions about the role of lie detection under Bayesian persuasion and communication more generally. For example, how does a richer action space modify our insights? 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? We leave these and other interesting questions to future research.

# References


**Balbuzanov, Ivan**, "Lies and Consequences: The effect of lie detection on communication outcomes," *International Journal of Game Theory*, 2019, *48* (4), 1203–1240.

**Blume, Andreas, Oliver J. Board, and Kohei Kawamura**, "Noisy Talk," *Theoretical Economics*, 2007, *2* (4), 395–440.

**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.

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

**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.

**Ekman, Paul and Mark G. Frank**, "Lies that Fail," in Lewis M. and C. Saarni, eds., *Lying and Deception in Everyday Life*, The Guilford Press, 1993, pp. 184–201.

**Engers, Maxim, Joshua S. Gans, Simon Grant, and Stephen P. King**, "First-author Conditions," *Journal of Political Economy*, 1999, *107* (4), 859–883.

**Fréchette, Guillaume R., Alessandro Lizzeri, and Jacopo Perego**, "Rules and Commitment in Communication: An Experimental Analysis," *Econometrica*, forthcoming.

**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.

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

**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**, "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 2019, *11*, 249–272.

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

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

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

**Lipnowski, Elliot, Doron Ravid, and Denis Shishkin**, "Persuasion via Weak Institutions," *Journal of Political Economy*, forthcoming.

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

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

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

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

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

# A Proofs

## A.1 Proof of Proposition 1

We now show that type I and II strategies are suboptimal because the resulting implementation probabilities $\Pr_{\mathrm{I}}(a=1)$ and $\Pr_{\mathrm{II}}(a=1)$ are dominated by the probability $\Pr_{\mathrm{III}}(a=1)$ resulting from III. To see this, note first that

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

The second inequality holds because $(p_0, p_1) = (0,0)$ is a type II strategy and gives value $\mu$. In fact, for a type II strategy, it is optimal to set $p_1 = 0$ because this loosens both constraints, and improves the objective. Given this, $\mu_1 = 0 < t$ is loose. Hence the optimum requires

$$\mu_0 = \frac{\mu(1-q)}{\mu(1-q) + (1-\mu)p_0} = t$$

and hence

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

Similarly, in the maximization problem within type III strategies, it is optimal to set $p_1 = 1$. Then $\mu_0 = 0 < t$ becomes loose. The optimum requires $p_0$ to be as small as possible while ensuring that $\mu_1 \geq t$. Define $\overline{q} \equiv 1 - \frac{\mu(1-t)}{t(1-\mu)} \in (0,1)$, then there are two cases to consider.

- $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$ or $q \leq \overline{q}$. In this case, there exists $p_0^*$ s.t. $\mu_1 = t$, that is $\frac{\mu}{\mu+(1-\mu)(1-p_0^*)(1-q)} = t$. Therefore, $\Pr_{\mathrm{III}}(a=1) = \frac{\mu}{t}$.

- $\frac{\mu}{\mu+(1-\mu)(1-q)} > t$ or $q > \overline{q}$. In this case, $\mu_1 \geq t$ can never bind. Thus, the best option is to set $p = 0$ which implies $\Pr_{\mathrm{III}}(a=1) = \mu + (1-\mu)(1-q)$.

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

(1) If $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$, the 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 that were the case we would have

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

and

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

which would imply

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

and therefore

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

where the last inequality is binding if $q = 0$ or $p = 0$. This in turn yields $t < \frac{\mu}{\mu+(1-\mu)(1-q)}$ which is a contradiction. Hence, if $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$, the optimal strategy is

$$p_0^* = 1 - \frac{\frac{\mu(1-t)}{t(1-\mu)}}{1-q} \quad \text{and} \quad p_1^* = 1.$$

Alternatively,

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

(2) If $\frac{\mu}{\mu+(1-\mu)(1-q)} > t$, it is now possible to induce $\mu_1 \geq t, \mu_0 \geq t$. In particular, the constraints can be rewritten as two lines where the coordinates are $p_0$ and $p_1$. In particular, we have

$$\mu_1 \geq t \Leftrightarrow (1-t)\mu p_1 \geq t(1-\mu)(1-p_0)(1-q)$$

which passes through $(1,0)$ and $\left(0, \frac{t(1-\mu)(1-q)}{(1-t)\mu}\right)$ where $\frac{t(1-\mu)(1-q)}{(1-t)\mu} < 1$ by assumption. We also

have

$$\mu_0 \geq t \Leftrightarrow \mu(1-t)(1-p_1) \geq t(1-\mu)\frac{p}{1-q}$$

which passes through $(0,1)$ and $\left(\frac{\mu(1-t)(1-q)}{t(1-\mu)}, 0\right)$ where $\frac{\mu(1-t)(1-q)}{t(1-\mu)} < 1$ because $t > \mu$.

Since the objective is to maximize $1-(1-\mu)(1-p_0)q$, we want to find the point in type IV strategies with the largest value of $p_0$. Clearly, this point is at the intersection of the two lines in Figure 1$(b)$, given by

$$p_0^* = 1 - \frac{1-(1-q)\frac{\mu(1-t)}{t(1-\mu)}}{(2-q)q} \quad \text{and} \quad p_1^* = 1 - \frac{1-(1-q)\frac{t(1-\mu)}{\mu(1-t)}}{(2-q)q}$$

where $\frac{\mu(1-t)}{t(1-\mu)} \in (1-q, 1)$ by assumption. Alternatively,

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

As a result, we have $\Pr_{\text{III}}(a=1) < \Pr_{\text{IV}}(a=1)$ because the following inequality holds

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

## A.2 Proof of Proposition 2

- If $q \leq \overline{q}$,

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

Clearly, $p_0^* = 1 - \frac{1-\overline{q}}{1-q}$ decreases in $q$ and $p_1^*$ is constant in $q$.

- If $q > \overline{q}$,

$$p_0^* = \frac{1-q}{(2-q)q}(q-\overline{q}) \quad \text{and} \quad p_1^* = \frac{1-q}{(2-q)q}\left[\frac{1}{1-\overline{q}} - (1-q)\right] \tag{3}$$

This implies

$$\begin{aligned}\frac{\partial p_0^*}{\partial q} &= \frac{(-2q+1+\overline{q})\cdot(2-q)q-(2-2q)(1-q)(q-\overline{q})}{(2-q)^2q^2} \\ &= \frac{-q^2+(q^2-2q+2)\overline{q}}{(2-q)^2q^2}\end{aligned}$$

Therefore,

$$\frac{\partial p_0^*}{\partial q} \geq 0 \iff \frac{1}{\overline{q}} \leq \frac{q^2-2q+2}{q^2} = 1+\frac{2-2q}{q^2} \tag{4}$$

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

On the other hand, $p_1^*$ can be written as a product of $\frac{(1-q)}{(2-q)}$ and $\frac{\frac{1}{1-\overline{q}}-(1-q)}{q}$. Each term decreases in $q$, the it follows that $p_1^*$ decreases in $q$ over $(\overline{q},1]$.

## A.3 Proof of Proposition 3

The expected payoff of the Sender is $\Pr(a=1)$. There are two cases depending on whether $q>\overline{q}$.

- If $q \leq \overline{q}$, then the Receiver chooses $a=1$ whenever $(m=1,\ d=\neg lie)$ or $(m=0,\ d=lie)$. But the latter occurs with probability 0 in the equilibrium. So,

$$U_S = \mu+(1-\mu)(1-p_0^*)(1-q) = \frac{\mu}{t} \tag{5}$$

  which is constant in $q$. Essentially, any marginal improvement in the lie detection technology (i.e., increase in $q$) is completely offset by less truthful reporting when $w=0$ (i.e., decrease in $p_0^*$).

- If $q>\overline{q}$, then the Receiver chooses $a=1$ always unless $(m=1,\ d=lie)$. So,

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

which is decreasing in $q$ as

$$\begin{aligned}\frac{\partial U_S}{\partial q} &= \frac{-\mu(1-t)t(2-q) - t[t(1-\mu) - \mu(1-t)(1-q)]}{t^2(2-q)^2} \\ &= \frac{-\mu(1-t) - t(1-\mu)}{t(2-q)^2} \\ &< 0\end{aligned}$$

The Receiver's expected payoff is $t \cdot \Pr(a = w = 0) + (1-t) \cdot \Pr(a = w = 1)$. Again, there are two cases.

- If $q \leq \overline{q}$, then the Receiver matches the state $w = 0$ correctly if $(w = 0, m = 0)$ or if $(w = 0, m = 1,\ d = lie)$, and matches the state $w = 1$ correctly if $w = 1$. In sum,

$$\begin{aligned}U_R &= (1-\mu)t \cdot [p_0^* + (1-p_0^*)q] + \mu(1-t) \\ &= (1-\mu)t \cdot [1 - (1-p_0^*)(1-q)] + \mu(1-t) \\ &= (1-\mu)t \cdot \left[1 - \frac{\mu(1-t)}{t(1-\mu)}\right] + \mu(1-t) \\ &= (1-\mu)t\end{aligned}$$

  which is constant in $q$.

- If $q > \overline{q}$, then the Receiver matches the state $w = 0$ correctly if $(w = 0, m = 1,\ d = lie)$, and matches the state $w = 1$ correctly if $w = 1$. In sum,

$$\begin{aligned}U_R &= (1-\mu)t \cdot (1-p_0^*)q + \mu(1-t) \\ &= (1-\mu)t \cdot \frac{1 - (1-q)^{\frac{\mu(1-t)}{t(1-\mu)}}}{2-q} + \mu(1-t) \\ &= \frac{(1-\mu)t + t(1-\mu)}{2-q}\end{aligned}$$

  which is increasing in $q$.

## A.4 Proof of Lemma 1

In Step 1, we characterize properties of the Sender's preferred equilibria and show that in any Sender's preferred equilibrium, the Receiver always takes $a = 1$ under state 1 and $\lambda$. In Step 2,

we decompose the payoff functions $U^S(q)$ and link it to $U^R(q)$.

**Step 1**:

Let the Sender's strategy be represented by $\boldsymbol{a} = \{a_{ij}\}_{i,j\in\{0,\,\lambda,\,1\}}$, where $a_{ij}$ is the probability of sending message $j$ under state $i$. Let $X$ be the set of pairs $(m,\, d)$ where $(m, d) \in \{0,\, \lambda,\, 1\} \times \{lie,\, \neg lie\}$. Let $\mu_{m,d}$ denote the posterior mean after observing $(m, d) \in \{0,\, \lambda,\, 1\} \times \{lie,\, \neg lie\}$. The formulas of the posterior means are given by

$$
\begin{aligned}
\mu_{1,\,lie} &= \frac{q \cdot \lambda P_\lambda a_{\lambda 1}}{q \cdot (P_\lambda a_{\lambda 1} + P_0 a_{01})} \\
\mu_{1,\,\neg lie} &= \frac{P_1 a_{11} + \lambda P_\lambda a_{\lambda 1}(1-q)}{P_1 a_{11} + P_\lambda a_{\lambda 1}(1-q) + P_0 a_{01}(1-q)} \\
\mu_{\lambda,\,lie} &= \frac{q \cdot P_1 a_{1\lambda}}{q \cdot (P_1 a_{1\lambda} + P_0 a_{0\lambda})} \\
\mu_{1,\,\neg lie} &= \frac{\lambda P_\lambda a_{\lambda\lambda} + P_1 a_{1\lambda}(1-q)}{P_\lambda a_{\lambda\lambda} + P_1 a_{1\lambda}(1-q) + P_0 a_{0\lambda}(1-q)} \\
\mu_{0,\,lie} &= \frac{q \cdot (P_1 a_{10} + \lambda P_\lambda a_{\lambda 0})}{q \cdot (P_1 a_{10} + P_\lambda a_{\lambda 0})} \\
\mu_{0,\,\neg lie} &= \frac{(P_1 a_{10} + \lambda P_\lambda a_{\lambda 0})(1-q)}{P_0 a_{00} + (P_1 a_{10} + P_\lambda a_{\lambda 0})(1-q)}
\end{aligned}
$$

where the off-path beliefs are equal to zero. Moreover, let $num(x) = \mu_x \cdot \Pr(x)$ to be the numerator of $\mu_x$. Denote $X_1 = \{(m,\, d) \in X \,|\, \mu_{m,d} \geq t\}$ as the set of message-detection pairs under which the Receiver takes action $a = 1$. An observation is that the sum of six numerators equal $\mu$ and the sum of six denominators equal 1. Thus, for any strategy $\boldsymbol{a}$, the Sender's payoff is equal to

$$
\sum_{x\in X_1} \Pr(x) \leq \frac{\displaystyle\sum_{x\in X_1} \mu_x \cdot \Pr(x)}{t} \leq \frac{\mu}{t} \tag{7}
$$

We know from equation (1) that $\frac{\mu}{t}$ is exactly the Sender's optimal payoff in this case. Thus, it suffices to find conditions on $\boldsymbol{a}$ such that both equalities in equation (7) are attained. The first equality requires that $\forall x \in X_1$, the Receiver is indifferent after observing $x$, i.e., $\mu_x = t$. This immediately implies $a_{10} = a_{\lambda 0} = 0$ because otherwise $\mu_{0,\,lie} > t$ by assumption $t < \lambda$. Next, we consider two cases.

If $a_{1\lambda} > 0$, then the second equality requires $(\lambda, lie) \in X_1$. Otherwise,

$$\sum_{x \in X_1} num(x) \leq \mu - q \cdot P_1 a_{1\lambda} < \mu.$$

Analogously, $(\lambda, \neg lie) \in X_1$. However, if $(\lambda, lie)$, $(\lambda, \neg lie) \in X_1$, then by the implication of the first equality, it must be that $a_{\lambda\lambda} = 0$ and thus $a_{\lambda 1} = 1$. Repeat the arguments, the second equality requires $(1, lie)$, $(1, \neg lie) \in X_1$, and the first inequality requires $a_{11} = 0$ and thus $a_{1\lambda} = 1$. In summary, the Sender always sends message 1 under state $\lambda$ and sends message $\lambda$ under state 1. Moreover, $(\lambda, \neg lie)$, $(\lambda, lie)$ $(1, \neg lie)$, $(1, lie)$ all induce action $a = 1$. Thus, $\Pr(a = 1, \omega = 1) = \Pr(a = 1, \omega = \lambda) = 1$.

In the second case, suppose $a_{1\lambda} = 0$, which implies $a_{11} = 1$. By the second equality, $(1, \neg lie) \in X_1$. Now, $a_{\lambda 1}$ must be 0. Otherwise, the second equality also requires $(1, lie) \in X_1$. However, then the first equality is violated as $t \leq \mu_{1,\neg lie} < \mu_{1,lie}$. In summary, the Sender is totally truthful under state 1 and $\lambda$. Moreover, $(\lambda, \neg lie)$, $(1, \neg lie)$ both induce action $a = 1$. Again, $\Pr(a = 1, \omega = 1) = \Pr(a = 1, \omega = \lambda) = 1$.

**Step 2**:

Note that the Sender's equilibrium payoff can be decomposed in the following way.

$$U^S(q) = \Pr(a = 1) = P_1 \cdot \Pr(a = 1, \omega = 1) + P_\lambda \cdot \Pr(a = 1, \omega = \lambda) + P_0 \cdot \Pr(a = 1, \omega = 0).$$

Step 1 implies $\Pr(a = 1, \omega = 1) = \Pr(a = 1, \omega = \lambda) = 1$ so that

$$U^S(q) = P_1 + P_\lambda + P_0 \cdot [1 - \Pr(a = 0, \omega = 0)].$$

At the same time, the Receiver's expected payoff is reduced to

$$\begin{aligned} U^R(q) &= (1 - t)P_1 + (\lambda - t)P_\lambda + tP_0 \cdot \Pr(a = 0, \omega = 0) \\ &= (1 - t)P_1 + (\lambda - t)P_\lambda + t[1 - U^S(q)]. \end{aligned}$$

which concludes the proof.

## A.5 Proof of Proposition 4

Due to the binary structure, if $d = lie$, the true state is perfectly revealed. We only need to focus on the Receiver's responses when $d = \neg lie$. The formulas of the relevant posterior likelihood ratios are given by

$$l_0 \equiv l|_{m=0,d=\neg lie} = \frac{\alpha\mu(1-p_1)(1-q) + (1-\alpha)\mu(1-\tilde{p}_1)(1-q)}{\alpha(1-\mu)p_0 + (1-\alpha)(1-\mu)\tilde{p}_0},$$
$$l_1 \equiv l|_{m=1,d=\neg lie} = \frac{\alpha\mu p_1 + (1-\alpha)\mu\tilde{p}_1}{\alpha(1-\mu)(1-p_0)(1-q) + (1-\alpha)(1-\mu)(1-\tilde{p}_0)(1-q)}.$$

With a slight abuse of notation, we can partition the set of Sender's strategies $(p_0, p_1, \tilde{p}_0, \tilde{p}_1)$ into four types. Type I strategy implies $l_0, l_1 < \frac{t}{1-t}$. Type II strategy implies $l_0 \geq \frac{t}{1-t}$, $l_1 < \frac{t}{1-t}$. Type III strategy implies $l_0 < \frac{t}{1-t}$, $l_1 \geq \frac{t}{1-t}$. Type IV strategy implies $l_0, l_1 \geq \frac{t}{1-t}$. We first demonstrate that type IV strategy does not exist if $q \leq \overline{q}$. We then compute the optimal type III strategy and argue that it dominates any type I and II strategies.

First, for the purpose of contradiction, suppose a type IV strategy exists, then

$$\frac{\alpha\mu(1-p_1)(1-q) + (1-\alpha)\mu(1-\tilde{p}_1)(1-q)}{\alpha(1-\mu)p_0 + (1-\alpha)(1-\mu)\tilde{p}_0} \geq \frac{t}{1-t},$$

which implies

$$\frac{\alpha\mu(1-p_1) + (1-\alpha)\mu(1-\tilde{p}_1)}{\alpha(1-\mu)p_0(1-q) + (1-\alpha)(1-\mu)\tilde{p}_0(1-q)} > \frac{t}{1-t}. \tag{8}$$

At the same time, $l_1 \geq \frac{t}{1-t}$ suggests that

$$\frac{\alpha\mu p_1 + (1-\alpha)\mu\tilde{p}_1}{\alpha(1-\mu)(1-p_0)(1-q) + (1-\alpha)(1-\mu)(1-\tilde{p}_0)(1-q)} \geq \frac{t}{1-t}. \tag{9}$$

Combining the two inequalities (8) and (9) yields

$$\frac{\mu}{(1-\mu)(1-q)} > \frac{t}{1-t}.$$

This is in contradiction with the assumption that $q \leq \overline{q}$. Therefore, a type IV strategy is not

available. Next, we derive the optimal type III strategy where the objective function is given by

$$U_S^{\text{III}} = \mu\{\alpha[p_1 + (1-p_1)q] + (1-\alpha)[\tilde{p}_1 + (1-\tilde{p}_1)q]\} \\ + (1-\mu)[\alpha(1-p_0)(1-q) + (1-\alpha)(1-\tilde{p}_0)(1-q)].$$

Note that in any type III strategy, it must be that $\tilde{p}_0 = 0$ and $\tilde{p}_1 = 1$ because the message $m = 1$ induces $a = 1$ with a higher probability than the message $m = 0$. Furthermore, at the optimum, it must be that $p_1 = 1$ because increasing $p_1$ relaxes the constraint $l_1 \geq \frac{t}{1-t}$ while improving the objective. Finally, the constraint on $l_1$ must bind at the optimum, which, after plugging in $p_1$, $\tilde{p}_0$, $\tilde{p}_1$, implies

$$\alpha(1-p_0) + 1 - \alpha = \frac{\mu(1-t)}{t(1-\mu)(1-q)}. \tag{10}$$

The assumptions $q \leq \overline{q}$ and $\alpha \geq \underline{\alpha}$ ensure that the solution to equation (10) belongs to $[0, 1]$. Given this strategy, the Sender's payoff is equal to

$$\begin{aligned} U_S^{\text{III}} &= \mu[\alpha + (1-\alpha)] + (1-\mu)[\alpha(1-p_0)(1-q) + (1-\alpha)(1-q)] \\ &= \mu + \frac{\mu(1-t)}{t} = \frac{\mu}{t}. \end{aligned}$$

This strategy dominates any type I strategy because the Sender there obtains a positive payoff only if $\omega = 1$, $m = 0$, $d = lie$, which occurs with a probability bounded above by $\mu q$. Finally, we need to derive the optimal type II strategy, where the objective function is given by

$$U_S^{\text{II}} = \mu[\alpha(1-p_1) + (1-\alpha)(1-\tilde{p}_1)] + (1-\mu)[\alpha p_0 + (1-\alpha)\tilde{p}_0].$$

Analogously, in any type II strategy, it must be that $\tilde{p}_0 = 1$ and $\tilde{p}_1 = 0$ because the message $m = 0$ induces $a = 1$ with a higher probability than the message $m = 1$. Furthermore, at the optimum, it must be that $p_1 = 0$ because decreasing $p_1$ relaxes the constraint $l_0 \geq \frac{t}{1-t}$ while improving the objective. Finally, the constraint on $l_0$ must bind at the optimum, which, after plugging in $p_1$, $\tilde{p}_0$, $\tilde{p}_1$, implies

$$\alpha p_0 + 1 - \alpha = \frac{\mu(1-t)(1-q)}{t(1-\mu)}.$$

Thus, the Sender's payoff is reduced to

$$U_S^{\mathrm{II}} = \mu + \frac{\mu(1-t)(1-q)}{t},$$

which, for any $q > 0$, is smaller than $U_S^{\mathrm{III}}$. Therefore, the globally optimal strategy is given by $(p_0^*,\, 0,\, 1,\, 0)$, where $p_0^*$ solves equation (10). It has been shown that the Sender's equilibrium payoff equals $\frac{\mu}{t}$. Last, the Receiver's equilibrium payoff equals

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

Both payoffs are constant in $q$.