Bayesian Persuasion with Lie Detection*

Florian Ederer[†] Weicheng Min[‡]

November 12, 2020

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. 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 Philippe Jehiel, Navin Kartik, Elliot Lipnowski, Barry Nalebuff, and Ian Turner for helpful comments.

[†]Yale School of Management and Cowles Foundation, Yale University, 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 US President Donald Trump has made over 20,000 false or misleading claims.[1] But such lies are also detectable. Monitoring and fact-checking should constrain how much licence 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 has increased rather than decreased—a development that runs counter to this intuition.

In this paper we incorporate 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 (or utilities which we use interchangeably) are realized for both parties.

Our main assumption that lies—but not the underlying truth—are detectable is arguably a natural one in many applications. Facts may come to light that contradict the initial claim of the Sender. These facts do not necessarily reveal the payoff-relevant state, but only prove that the Sender has lied. For example, in job interviews or trial testimonies, the Sender may be required to provide details and arguments supporting his statements, and if he is lying, he may be at risk of producing an internally or externally inconsistent account, thereby revealing his lie. Liars may also exhibit physical reactions such as blushing, which reveal the fact of 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 claiming that the state is favorable. That is to say, the lie detection technology changes the Sender's message strategy

[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.

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 also lie more frequently in the favorable state of nature and the Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability.

Two recent papers (Balbuzanov, 2019; Dziuda and Salas, 2018) also investigate the role of lie detectability in communication. The largest difference with respect to our paper lies in the commitment assumption of the Sender. In all those 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 Jehiel (2020) lie detection arises endogenously in a setting with two rounds of communication and a Sender who in the second round cannot remember what lies she sent in the first round. As the state space becomes arbitrarily fine, the probability of lie detection goes to 1 because it is hard to guess exactly the same lie, and therefore only fully revealing equilibria arise.

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 follows 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 own work is Fréchette et al. (2019) 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 and Vorobyev (2020) 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 manipulating the electoral process ex ante and by 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 so that the Receiver is not informed either. 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.

2.2 Utilities

Given both m and d, R 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 $a=1$ if and only if

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

and therefore one could also interpret t as the threshold of the Receiver's posterior belief above which she takes $a=1$. Note that if $t \leq \mu$, there is no need to persuade because the Receiver will choose the Sender's preferred action $a=1$ even without a message. Therefore, we assume $t \in (\mu,1)$.

2.3 Strategies

We assume that the Sender has full commitment power as is common in the Bayesian persuasion framework.[2] 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

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

[2]For a detailed discussion and relaxation of this assumption see Min (2017), Fréchette et al. (2019), Lipnowski et al. (2019), and Nguyen and Tan (2019). Titova (2020) shows that with binary actions and a sufficiently rich enough state space verifiable disclosure enables the sender's commitment solution as an equilibrium.

where $a(m,d)$ maximizes

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

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 expectation sign in the Receiver's utility is only taken with respect to both w. Due to the simple structure of the model, it is without loss of generality to assume that the Sender chooses only two parameters $p_0 = \Pr(m=0\,|\,w=0)$ and $p_1 = \Pr(m=1\,|\,w=1)$ to 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 Receiver is certain that the message $m=0$ is a lie and therefore the state of the world w must be equal to 1, that is

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

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

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

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

Hence, the Receiver takes action $a=1$ if and only if $\mu_0\geq t$. For brevity of notation we denote the posterior following this event by μ_0 (and thus omitting the lie detection outcome $d=\neg lie$). 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, he does not care about how 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

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

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

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

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 μ_1 (and thus omitting the lie detection outcome $d=\neg lie$). Similarly, if $p_0=1,p_1=0$, then this event occurs with 0 probability and the belief μ_1 is not well-defined, but again this does not matter for Sender. For simplicity, define $\mu_1=0$ when $p_0=1,p_1=0$.

Given these optimal responses by R, the relationships between the posteriors μ_0,μ_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 μ_0 and μ_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 R sufficiently convinced that $a=1$ is the right action. The maximal probability that the Receiver chooses $a=1$[3] is given by

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

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

$$\Pr_{\mathrm{II}}(a=1) = \sup_{p_0, p_1 \in [0,1]} \mu(1-p_1) + (1-\mu)p_0 \quad \text{s.t.} \quad \mu_0 \geq t, \mu_1 < t \tag{2}$$

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

$$\Pr_{\mathrm{III}}(a=1) = \sup_{p_0, p_1 \in [0,1]} \mu p_1 + \mu(1-p_1)q + (1-\mu)(1-q)(1-p_0) \quad \text{s.t.} \quad \mu_0 < t, \mu_1 \geq t \tag{3}$$

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

$$\Pr_{\mathrm{IV}}(a=1) = \sup_{p_0, p_1 \in [0,1]} 1-(1-\mu)(1-p_0)q \quad \text{s.t.} \quad \mu_0 \geq t, \mu_1 \geq t \tag{4}$$

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.

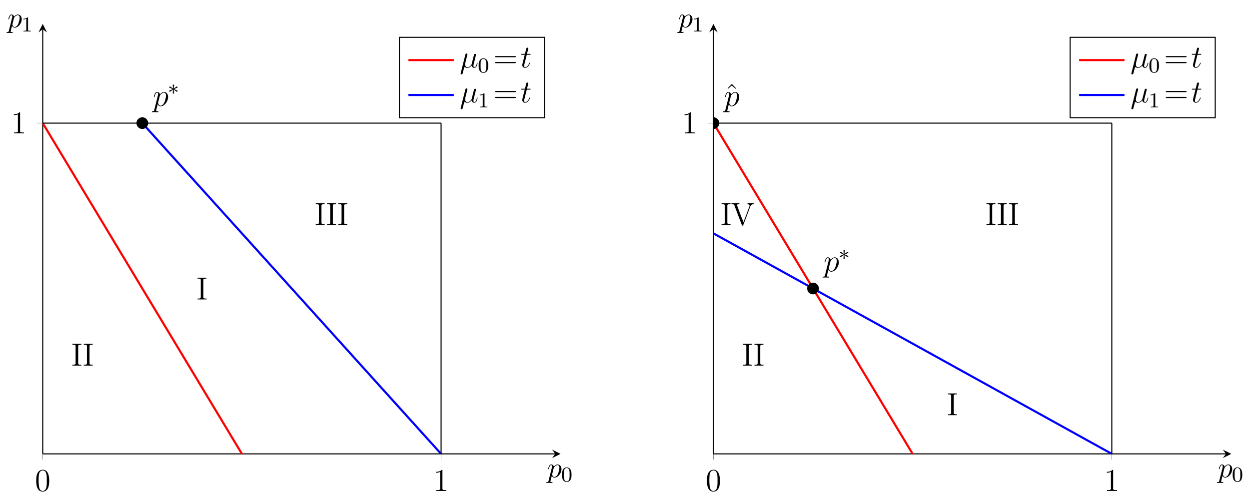
[3] The choice set of the maximization problem is not closed, so the maximum may not be achieved a priori.

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

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

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 $w=1$, which occurs with probability μ. 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 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 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 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 meanings 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 $(m=1,d=\neg lie)$ suggest $w=0$ 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.

In addition, 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:

$$p_0^* = \frac{\overline{q} - q}{1 - q} \quad \text{and} \quad p_1^* = 1 \tag{5}$$

This 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. So, 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 frequent 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 could show a much stronger statement that $\hat{p}$ is worse than any type IV strategy p whenever the latter is feasible. To this end, we decompose the 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 $m=1$ and that is credible. 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 and does so also overall. Mathematically,

$$U_S(\hat{p}) = \underbrace{\mu}_{\Pr(w=1)} \times \overbrace{1}^{\Pr(a=1)} + \underbrace{(1-\mu)}_{\Pr(w=0)} \times \overbrace{1}^{\Pr(m=1)} \times \underbrace{(1-q)}_{\Pr(d=\neg lie)}$$

and

$$U_S(p) = \mu \times 1 + (1-\mu) \times [p_0 \times 1 + (1-p_0) \times (1-q)]$$

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 argued, the main benefit of p relative to $\hat{p}$ is that the "safer" message $m=0$ is sent more frequently in p, so 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 previous arguments (provided that $q > \overline{q}$). The expressions are given by

$$p_0^* = \frac{1-q}{(2-q)q}(q-\overline{q}) \quad \text{and} \quad p_1^* = \frac{1-q}{(2-q)q}\left[\frac{1}{1-\overline{q}} - (1-q)\right] \tag{6}$$

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 in Sender's payoff between 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 with 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 μ 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 μ 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 μ), 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. $p_0^* = Pr(m=0|w=0)$ *is decreasing over* $[0,\overline{q}]$*, and has an inverse U shape over* $(\overline{q},1]$.

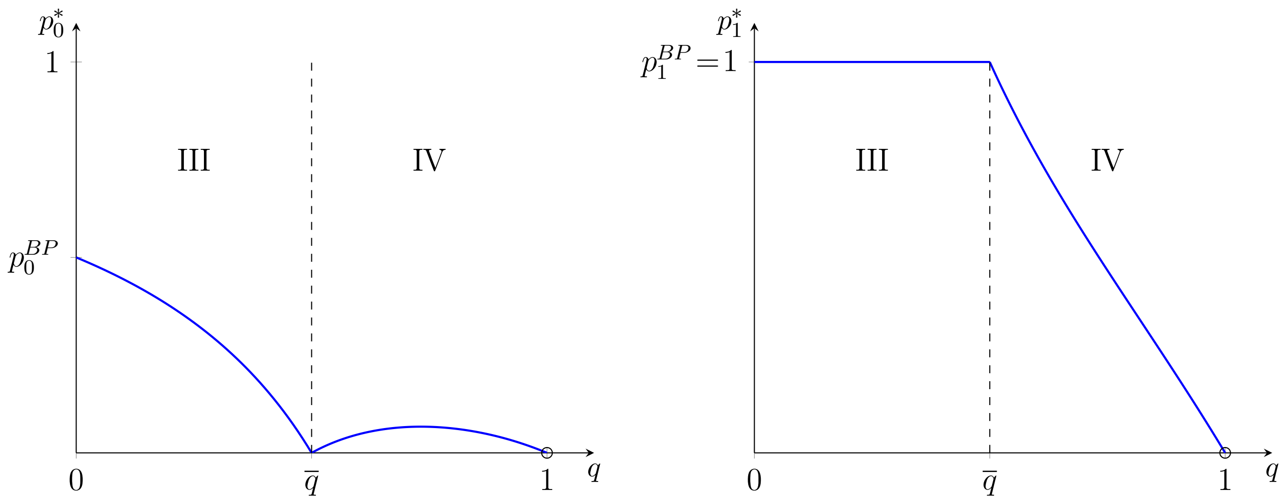
2. $p_1^* = Pr(m=1|w=1)$ *is constant over* $[0,\overline{q}]$*, and decreases over* $(\overline{q},1]$.

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.

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



(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 $t = \frac{1}{2}$

$m=1$ 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 just held to her outside utility of getting no information whatsoever. When lie detection q goes up the Receiver is now more certain that $(m=1, d=\neg lie)$ means $w=1$ and would therefore 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 ever 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 good 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 $a=1$. At $q=\overline{q}$ the Sender therefore 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.

3.2.2 Utilities

Recall that the equilibrium payoffs of the Sender and the Receiver are denoted by U_S and U_R. 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.

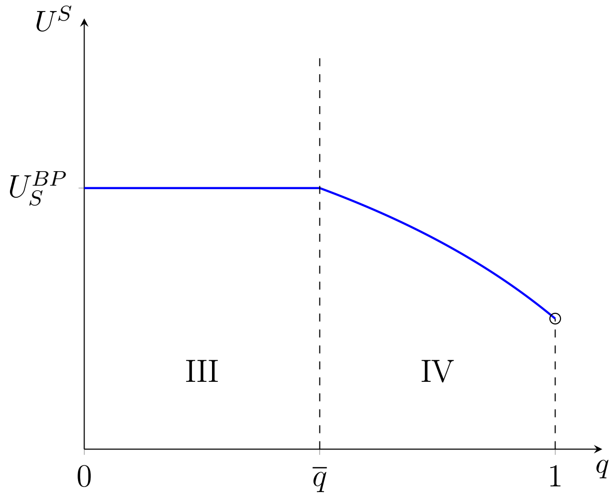
Proposition 3. *As the lie detection probability q increases,*

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

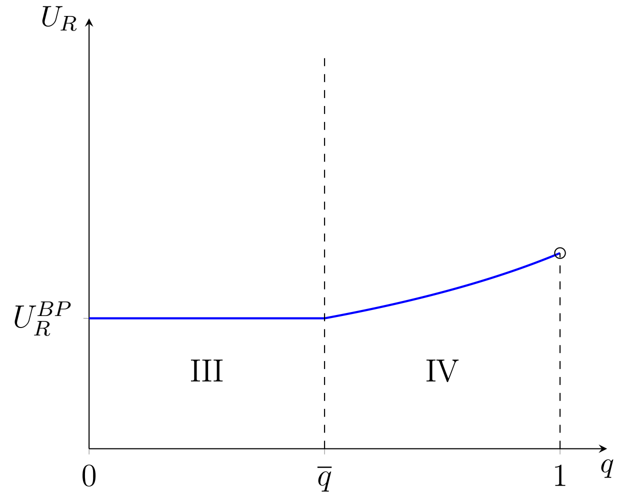
2. *U_R is constant over $[0,\overline{q}]$, and increases over $(\overline{q},1]$.*

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 actually $w=1$ which occurs with probability μ.

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. In the limit, the Receiver is just as well off as she would be under perfect information.



(a) Sender's Equilibrium Payoff



(b) Receiver's Equilibrium Payoff

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

4 Discussion

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 in our model there is no common interest. 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 and show that in all remaining equilibria, the lowest types lie, while *some* higher types tell the truth. Our model violates the second refinement in Dziuda and Salas (2018) and thus irrespective of the commitment power of the Sender our paper is not nested by theirs.

Both papers feature a type of non-monotonicity result with respect to the relationship between lie detection and lying that is similar to that in our paper. In Balbuzanov (2019), the set of detection probabilities that permits fully revealing equilibrium obtains for low lie detection probabilities, but not for higher ones. Although in the baseline model of Dziuda and Salas (2018) a lower lie detection probability leads to less truth-telling, in an extension of their model they show that if the Sender is given an opportunity to make costly investments in decreasing the lie detectability, then for an intermediate region of the cost, the mass of liars can increase with the investment cost.

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 simply compensates by lying more. Once the lie detection probability is sufficiently high, the Sender is no longer able to 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.

Our paper explores the impact of lie detection on communication in a setting with complete commitment to a communication strategy by the sender and thereby establishes a useful benchmark relative to the diametrically opposed assumption of no commitment in existing cheap talk models with lie detection. However, how does lie detection influence communication behavior in intermediate settings with partial commitment? And what role does lie detection play under Bayesian persuasion with a richer state and message space? We leave these and other interesting questions to future research.

References

Balbuzanov, Ivan, "Lies and consequences," *International Journal of Game Theory*, 2019, *48* (4), 1203–1240.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica*, 1982, pp. 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.

Fréchette, Guillaume R, Alessandro Lizzeri, and Jacopo Perego, "Rules and commitment in communication," *CEPR Discussion Paper No. DP14085*, 2019.

Gehlbach, Scott and Dmitriy Vorobyev, "A Model of Censorship and Propaganda," *University of Chicago Working Paper*, 2020.

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.

Jehiel, Philippe, "Communication with Forgetful Liars," *Theoretical Economics*, 2020, *forthcoming.*

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," *Available at SSRN 3168103*, 2019.

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," *Work. Pap., Univ. Ariz., Tucson*, 2017.

Nguyen, Anh and Teck Yong Tan, "Bayesian Persuasion with Costly Messages," *Available at SSRN 3298275*, 2019.

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*, 2020.

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

$$\Pr_{\mathrm{I}}(a=1) \leq \mu \leq \Pr_{\mathrm{II}}(a=1) \tag{7}$$

The second inequality holds because $(p_0, p_1) = (0,0)$ is a type II strategy and gives value μ. 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

$$\mu_0 = \frac{\mu(1-q)}{\mu(1-q)+(1-\mu)p_0} = t \tag{8}$$

and hence

$$\Pr_{\mathrm{II}}(a=1) = \mu + \left(\frac{\mu}{t} - \mu\right)(1-q) \tag{9}$$

Similarly, in the maximization problem within type III strategies, it is optimal to set $p_1 = 1$. Then $\mu_0 = 0 < t$ becomes loos. 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

$$\frac{\mu p_1}{\mu p_1+(1-\mu)(1-p_0)(1-q)} \geq t \tag{10}$$

and

$$\frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p} \geq t \iff \frac{\mu(1-p_1)}{\mu(1-p_1)+(1-\mu)\frac{p}{1-q}} \geq t \tag{11}$$

which would imply

$$\frac{\mu p_1+\mu(1-p_1)}{\mu p_1+\mu(1-p_1)+(1-\mu)(1-p_0)(1-q)+(1-\mu)\frac{p}{1-q}} \geq t \tag{12}$$

and therefore

$$t \leq \frac{\mu}{\mu+(1-\mu)(1-p_0)(1-q)+(1-\mu)\frac{p}{1-q}} \leq \frac{\mu}{\mu+(1-\mu)(1-q)} \tag{13}$$

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

$$p_0^* = 1 - \frac{\frac{\mu(1-t)}{t(1-\mu)}}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{14}$$

Alternatively,

$$p_0^* = \frac{\overline{q}-q}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{15}$$

(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

$$\mu_1 \geq t \Leftrightarrow (1-t)\mu p_1 \geq t(1-\mu)(1-p_0)(1-q) \tag{16}$$

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

$$\mu_0 \geq t \Leftrightarrow \mu(1-t)(1-p_1) \geq t(1-\mu)\frac{p}{1-q} \tag{17}$$

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

$$p_0^* = 1-\frac{1-(1-q)\frac{\mu(1-t)}{t(1-\mu)}}{(2-q)q} \quad \text{and} \quad p_1^* = 1-\frac{1-(1-q)\frac{t(1-\mu)}{\mu(1-t)}}{(2-q)q} \tag{18}$$

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

$$p_0^* = \frac{1-q}{(2-q)q}(q-\overline{q}) \quad \text{and} \quad p_1^* = \frac{1-q}{(2-q)q}\left[\frac{1}{1-\overline{q}}-(1-q)\right] \tag{19}$$

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

$$\Pr_{\text{III}}(a=1)=\mu+(1-\mu)(1-q)=1-(1-\mu)q<1-(1-\mu)q(1-p_0^*)=\Pr_{\text{IV}}(a=1). \tag{20}$$

A.2 Proof of Proposition 2

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

$$p_0^* = \frac{\overline{q}-q}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{21}$$

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

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

$$p_0^* = \frac{1-q}{(2-q)q}(q-\overline{q}) \quad \text{and} \quad p_1^* = \frac{1-q}{(2-q)q}\left[\frac{1}{1-\overline{q}}-(1-q)\right] \tag{22}$$

This implies

$$\frac{\partial p_0^*}{\partial q} = \frac{(-2q+1+\overline{q})\cdot(2-q)q-(2-2q)(1-q)(q-\overline{q})}{(2-q)^2q^2} \tag{23}$$

$$= \frac{-q^2+(q^2-2q+2)\overline{q}}{(2-q)^2q^2} \tag{24}$$

Therefore,

$$\frac{\partial p_0^*}{\partial q} \geq 0 \Longleftrightarrow \frac{1}{\overline{q}} \leq \frac{q^2-2q+2}{q^2} = 1+\frac{2-2q}{q^2} \tag{25}$$

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,

$$U_S = \mu+(1-\mu)(1-p_0)^*(1-q) = \frac{\mu}{t} \tag{26}$$

  which is constant in q.

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

$$U_S = 1-(1-\mu)(1-p_0^*)q = 1-\frac{t(1-\mu)-\mu(1-t)(1-q)}{t(2-q)} \tag{27}$$

which is decreasing in q as

$$\frac{\partial U_S}{\partial q} = \frac{-\mu(1-t)t(2-q)-t[t(1-\mu)-\mu(1-t)(1-q)]}{t^2(2-q)^2} \tag{28}$$

$$= \frac{-\mu(1-t)-t(1-\mu)}{t(2-q)^2} \tag{29}$$

$$< 0 \tag{30}$$

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,

$$U_R = (1-\mu)t\cdot[p_0^* + (1-p_0^*)q] + \mu(1-t) \tag{31}$$

$$= (1-\mu)t\cdot[1-(1-p_0^*)(1-q)] + \mu(1-t) \tag{32}$$

$$= (1-\mu)t\cdot\left[1-\frac{\mu(1-t)}{t(1-\mu)}\right] + \mu(1-t) \tag{33}$$

$$= (1-\mu)t \tag{34}$$

  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,

$$U_R = (1-\mu)t\cdot(1-p_0^*)q + \mu(1-t) \tag{35}$$

$$= (1-\mu)t\cdot\frac{1-(1-q)\frac{\mu(1-t)}{t(1-\mu)}}{2-q} + \mu(1-t) \tag{36}$$

$$= \frac{(1-\mu)t + t(1-\mu)}{2-q} \tag{37}$$

  which is increasing in q.