Bayesian Persuasion with Detectable Lies*

Florian Ederer† Weicheng Min‡

October 3, 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 that Receiver's equilibrium payoff 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 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. Sender observes the binary state of nature and sends a message to a 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 positive probability q. The Receiver then takes an action. Finally, payoffs 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 Sender. These facts do not necessarily reveal the payoff-relevant state, but only prove that Sender has lied. In job interviews or trial testimonies, Sender may be required to provide details and arguments supporting her statements, and if she is lying, she may be at risk of producing an internally or externally inconsistent account, thereby revealing her 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 but does not have an impact on utilities. 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. In the limit of perfect lie detection the Sender lies all the time in either state.

[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.

Fourth, when the lie detection technology is sufficiently reliable, the Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability.

Three recent papers, including Balbuzanov (2019), Dziuda and Salas (2018), and Jehiel (2019), also investigate the role of lie detectability in communication. The largest difference lies in the commitment assumption of Sender. In all those three papers, the communication game is in the form of cheap talk. The detailed comparison is deferred to Section 4.

Kartik et al. (2006) and Kartik (2009) introduce an exogenous cost of lying tied to the size of the lie. They find that most types inflate their messages, but only up to a point. Full information revelation follows for some (if the type and message space are bounded) or all (if the type and message space are bounded) types. Holm (2010) analyzes a binary state game in which Sender wants to deceive Receiver, and both lies and truths are detectable with some probability. Seidmann (2005) looks at a model with two payoff-relevant types, and allows for stochastic lie detectability. In the equilibrium in which the "good type" reports truthfully, lie detection equals state detection. Seidmann (2005) obtains information revelation by offering an option to remain "silent" and committing Receiver to particular actions after silence and confession. Without reading Holm (2010) and Seidmann (2005), it sounds like our setup is nested by theirs. But it is not, because again there is no commitment power on Sender's side.

Maybe we could also connect our paper to the literature on evidence games.

Finally, a large experimental literature (Gneezy, 2005; Hurkens and Kartik, 2009; Sánchez-Pagés and Vorsatz, 2009; Ederer and Fehr, 2017; Gneezy et al., 2018) provides support for the existence of an exogenous cost of lying. However, Vrij (2008) argues that many people experience no guilt or fear associated with lying. Arguably, this should be true especially for highly sophisticated agents such as politicians or sales people. Is the above argument a defense for why we do not introduce an exogenous cost of lying? Fréchette et al. (2019) investigate models of cheap talk, information disclosure, and Bayesian persuasion, in a unified experimental framework. Their experiments provide general support for the strategic rationale behind the role of commitment and, more specifically, for the Bayesian persuasion model of Kamenica and Gentzkow (2011).

2 Model

2.1 Setup

Let $w \in \{0,1\}$ denote the state of the world and $\Pr(w=1)=\mu \in (0,1)$. Sender (S) observes w and commits to send a message $m \in \{0,1\}$ to Receiver (R). If S lies, *i.e.* $m \neq w$, R is informed with probability $q \in [0,1]$ and learns w perfectly. 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.

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

$$u^S(a,w) = \mathbb{1}_{\{a=1\}} \tag{1}$$

$$u^R(a,w) = (1-t)\cdot \mathbb{1}_{\{a=w=1\}} + t \cdot \mathbb{1}_{\{a=w=0\}}, \quad 0<t<1 \tag{2}$$

That is, S wants R to always take the action $a=1$ regardless of the state, while R 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, R takes action $a=1$ if and only if $\Pr(w=1 \mid m;d) \geq t$, so we could also interpret t as the threshold of R's posterior probability above which she takes $a=1$. Note that if $t \leq \mu$, there is no need to persuade because R will choose S's preferred action $a=1$ even without a message. Therefore, assume $t \in (\mu,1)$ so that the problem is interesting.

2.2 Optimal Messages and Responses

As is common in the Bayesian persuasion literature we assume that the Sender's commitment to an information structure is always binding.[2] The strategy of Sender is a mapping $m : \{0,1\} \longrightarrow \Delta(\{0,1\})$, and the strategy of Receiver is a mapping $a : \{0,1\} \times \{lie, \neg lie\} \longrightarrow \Delta(\{0,1\})$. Formally, S is choosing $m(\cdot)$ to maximize

$$\mathbb{E}[u^S(a(m(w),d(m(w),w)),w)]$$

[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).

where $a(m,d)$ maximizes

$$\mathbb{E}[u^R(a,w)|m;d].$$

Due to the simple structure of the model, it is equivalent that S chooses only two parameters $p_0 = \Pr(m = 0 \mid w = 0)$ and $p_1 = \Pr(m = 1 \mid w = 1)$ 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 S by p_0^* and p_1^*, and the ex-ante payoffs under this reporting probabilities as U^S and U^R. Given S's reporting strategy, R could potentially see four types of events to which he needs to react when choosing action a.

First, R could observe the event $(m=0,d=lie)$ which occurs with probability $\mu(1-p_1)q$. Given the lie detection technology R 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

$$\Pr(w=1\,|\,m=0,d=lie)=1. \tag{3}$$

As a result, R 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, R is uncertain about w because he does not know whether S lied or not. His posterior probability is given by

$$\Pr(w=1\,|\,m=0,d=\neg lie)=\frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p_0}\equiv B. \tag{4}$$

Hence, R takes action $a=1$ if and only if $B\geq t$.

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

$$\Pr(w=1\,|\,m=1,d=lie)=0, \tag{5}$$

which immediately implies the action $a=0$.

Fourth, $(m=1,d=\neg lie)$ occurs with probability $\mu p_1+(1-\mu)(1-p_0)(1-q)$. R is again uncertain

about w. His posterior is given by

$$\Pr(w=1\,|\,m=1,d=\neg lie)=\frac{\mu p_1}{\mu p_1+(1-\mu)(1-p_0)(1-q)}\equiv A \tag{6}$$

and R takes action $a=1$ if and only if $A\geq t$.

Given these optimal responses by R, the relationships between the posteriors A and B and the posterior threshold t divide up the strategy space into four distinct regions which we denote by I, II, III, and IV respectively. Within each region, the response of R as a function of (m,d) is the same, making it easy to find the region-optimal strategy. We are then left to pick the best strategy out of the four candidates for different values of q. These regions are defined as follows:

I. $A<t$, $B<t$: In this region, R only chooses $a=1$ if $(m=0,d=lie)$ and $a=0$ otherwise because the posteriors A and B are insufficiently high to persuade him to choose S's preferred action. Only if S 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 probability that R chooses $a=1$ is given by

$$\Pr_{\mathrm{I}}(a=1)=\max_{p_0,p_1}\quad \mu(1-p_1)q\quad s.t.\,A<t,B<t \tag{7}$$

II. $A<t$, $B\geq t$: In this region, R chooses $a=1$ if $(m=0,d=lie)$ or $(m=0,d=\neg lie)$ and $a=0$ otherwise. The probability that R chooses $a=1$ is given by

$$\Pr_{\mathrm{II}}(a=1)=\max_{p_0,p_1}\quad \mu(1-p_1)+(1-\mu)p_0\quad s.t.A<t,B\geq t \tag{8}$$

III. $A\geq t$, $B<t$: In this region, R chooses $a=1$ if $(m=0,d=lie)$ or $(m=1,d=\neg lie)$ and $a=0$ otherwise. The probability that R chooses $a=1$ is given by

$$\Pr_{\mathrm{III}}(a=1)=\max_{p_0,p_1}\quad \mu p_1+\mu(1-p_1)q+(1-\mu)(1-q)(1-p_0)\quad s.t.\,A\geq t,B<t \tag{9}$$

IV. $A\geq t$, $B\geq t$: In this region, R chooses $a=1$ if $(m=0,d=lie)$, $(m=0,d=\neg lie)$ or $(m=1,d=\neg lie)$.

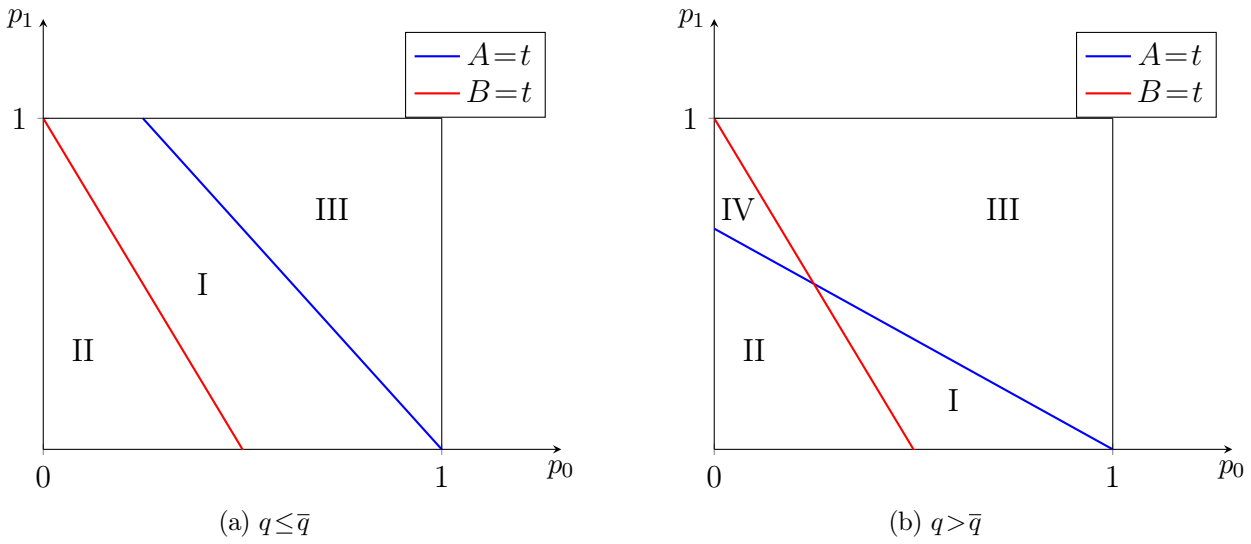
The probability that R chooses $a=1$ is given by

$$\Pr_{\text{IV}}(a=1)=\max_{p_0,p_1} \quad 1-(1-\mu)(1-p_0)q \quad s.t. A\geq t, B\geq t \tag{10}$$

We are now ready to state the main proposition of our model.

Proposition 1. *Let* $\bar{q}=1-\frac{\mu(1-t)}{t(1-\mu)}\in(0,1)$. *If* $q\leq\bar{q}$, *Sender's optimal strategy is in region* III, *in which Sender always tells the truth under* $w=1$, *but lies with positive probability under* $w=0$. *If* $q>\bar{q}$, *Sender's optimal strategy is in region* IV, *in which Sender lies with positive probability under both states.*

In Figure 1 we graphically illustrate how these four regions are divided.



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

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

The proof involves sequential comparisons between the four region-optimal strategies. First, the optimal strategy in region II is better than the optimal strategy in region I. To see this, consider a particular strategy $p_0=p_1=0$ in region II, namely S totally misreports the state. Following this strategy, R takes action $a=1$ if and only $w=1$, which occurs with probability μ. This strategy may not be optimal within region II, but it secures a lower bound μ for the value of the optimal strategy. This lower bound is sufficient to beat all strategies in region I since there $a=1$ only if $w=1$ and $(m=0, d=lie)$, which occurs with a probability less than μ.

Second, the optimal strategy in region III is better than the optimal strategy in region II. Notably, they should be equivalent 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 so that the two maximization problems are identical in II and III. Yet, with the introduction of a lie detection technology, there is an intrinsic meaning of the message sent. Regardless of the committed strategy by S, an on-path message that was not detected as a lie always carries some credibility for the state it is corresponding to. Now the two regions differ in the presence of this additional credibility. In region II, S wants an undetected lie $m=0$ to be indicative of $w=1$. This is more difficult because in that region the additional credibility is towards $w = 0$. On the contrary, in region III, S wants an undetected lie $m = 1$ to be indicative of $w=1$, but this is easier because the additional credibility is towards the same direction. (Still one step missing, why does "harder" suggest lower value?) Therefore, both region I and II are suboptimal, and we just need to focus on the comparison between region III and IV.

Interestingly, as suggested by the two panels of Figure 1, region IV does not exist when q is small. It is the easiest to understand this result when $q=0$, which yields the standard Bayesian Persuasion benchmark. In that case, we know it is impossible to induce $a = 1$ under all messages because by the martingale property, the posteriors after two messages 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). Now, we just need the four posteriors average to the prior. Besides, the posterior following $(1,lie)$ is 0. So, provided q sufficiently large, it is possible to support the two posteriors following $(1,\neg lie)$ and $(0,\neg lie)$ both higher than the prior and also higher than the threshold t. Combining this observation with the previous two arguments, we immediately obtain the first half of the proposition. In particular, if $q \leq \bar{q}$, the optimal strategy takes the following form:

$$p_0^* = \frac{\bar{q}-q}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{11}$$

This is reminiscent of Kamenica and Gentzkow (2011), where R is indifferent between two actions when he takes the preferred action $a=1$, and certain of the state when he takes the least favorite action $a=0$.

If the detection probability $q > \bar{q}$, the optimal strategy in region III involves a corner solution since p_0 cannot drop below 0. In particular, S always claims $m=1$ regardless of the state, so that a message

$m=0$ becomes off-path. Yet this is no more globally optimal as the region IV is now non-empty, and the optimal strategy in region IV is better than the one in region III described above. Specifically, the optimal strategy involves partial lying under both states ($0<p_0,p_1<1$) and is given by

$$p_0^*=\frac{(1-q)}{(2-q)q}(q-\bar{q}) \quad \text{and} \quad p_1^*=\frac{(1-q)}{(2-q)q}\left[\frac{1}{1-\bar{q}}-(1-q)\right] \tag{12}$$

From the above equations it can easily be seen that $p_0^*<p_1^*$.

The reason for this perhaps puzzling behavior of lying in both states is as follows. There are two benefits for sending $m=0$ in the favorable state $w=1$. First, if the lie is detected which happens with probability q, it is a free way to convince R that the state is indeed $w=1$. Second, even if the lie is not detected, it renders $m=0$ more indicative of $w=1$. Thus, the Sender may have a chance to persuade the Receiver in this case. However, these two benefits come at the cost of reducing the credibility of $m=1$. When fewer truthful messages $m=1$ are sent when $w=1$, the Receiver may prefer to take action $a=0$ after seeing ($m=1$,$d=\neg lie$). But with a sufficiently high detection probability q, this cost is equal to 0 because the constraint $A\geq t$ is now slack at the traditional Bayesian persuasion solution ($p_0=0$,$p_1=1$). Therefore, there is room to change p_0 and p_1 locally without affecting R's response after ($m=1$,$d=\neg lie$).

3 Comparative Statics

We now consider the comparative statics of our model with respect to the central parameter q. The results are summarized in Proposition 2 and Proposition 3. and Proposition 3 describes how the utilities of both parties under the optimal signal change as a function of the detection probability.

3.1 Optimal Signals

Proposition 2 describes how the structure of optimal signal (p_0^*,p_1^*) changes as the detection probability varies, which are plotted in Figure 2.

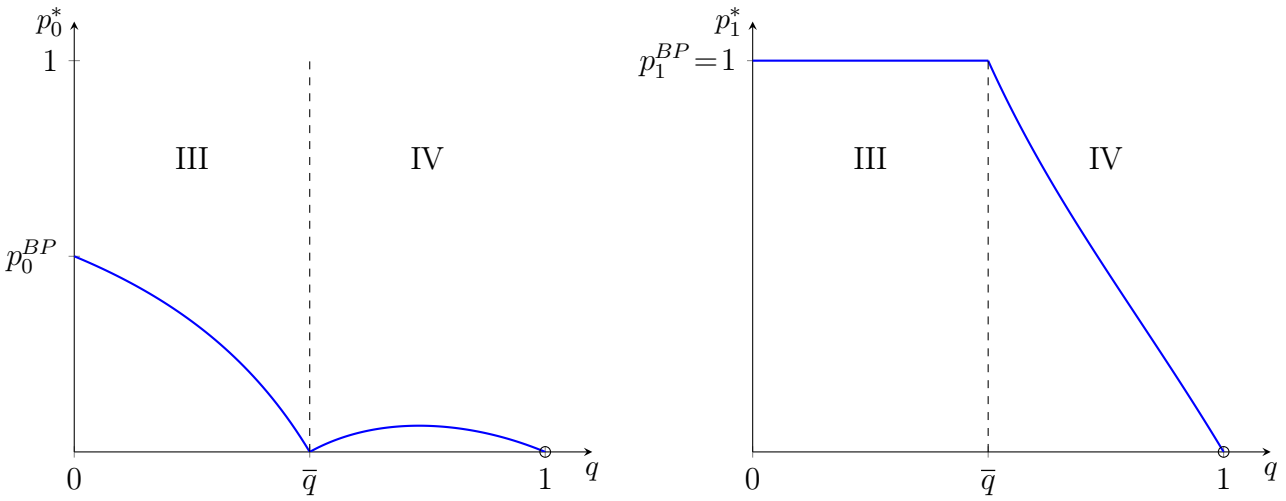
Proposition 2. *As the lie detection probability q increases,*

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

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

If $q \leq \bar{q} = 1 - \frac{\mu(1-t)}{t(1-\mu)}$, p_0^* is decreasing in q and p_1^* is constant at 1. In this range of q, 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 $q>\bar{q}$, p_0^* initially increases and then decreases. In contrast, p_1^* decreases over the entire range of $[\bar{q},1]$. In this range, Sender's optimal strategy lies in IV which involves misreporting both states of the world.



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

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

As can be seen from Figure 2 Panel (b) when q increases, the probability p_1^* that the Sender chooses to report truthfully when the state is $w=1$ decreases.[3] However, the probability p_0^* that the Sender chooses to report truthfully when $w=0$ is non-monotonic in q. This is because under the optimal strategy in region IV the Sender increasingly lies with a message $m=0$ when the state is $w=1$. This, in turn, allows him to send a truthful message $m=0$ when the state is $w=0$ which sufficiently persuades Receiver to choose $a=1$ because $A \geq t$ in region IV. I find this argument incomplete, we need to add another argument explaining why it has to go to 0 as q is close to 1.

[3]Figure 2 Panel (b) would seem to suggest that p_1^* is linear in q for $q>\bar{q}$. However, we can show that the curve is neither convex nor concave. For a large set of parameters, it is close to linear.

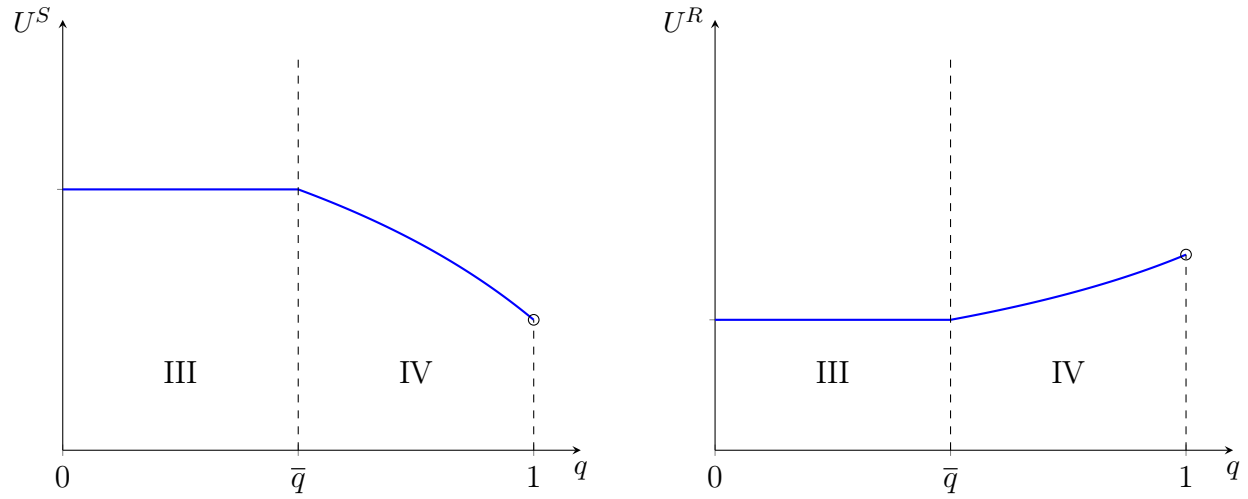
3.2 Utilities

Denote the equilibrium payoffs of Sender and Receiver 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 plotted in Figure 3.

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

1. *U^S is constant over $[0,\bar{q}]$, and decreases over $(\bar{q},1]$.*

2. *U^R is constant over $[0,\bar{q}]$, and increases over $(\bar{q},1]$.*



(a) Sender's Equilibrium Payoff

(b) Receiver's Equilibrium Payoff

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

Sender's equilibrium payoff does not change for $q\leq\overline{q}$ and decreases with q for $q>\overline{q}$. As long as $q\leq\overline{q}$ 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 Sender's utility. In the limit case where $q=1$ 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 Sender's utility, 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 Receiver is just as well off as he would be under perfect information.

4 Conclusion

We analyze the role of probabilistic lie detection in a model of Bayesian persuasion. We show that Sender lies more when the lie detection probability increases. As long as the lie detection probability is sufficiently small Sender's and Receiver's equilibrium payoff are unaffected by the lie detection technology because Sender simply compensates by lying more. Once the lie detection probability is sufficiently high, Sender's (Receiver's) equilibrium payoff decreases (increases) with the lie detection probability.

Our model therefore rationalizes that Senders choose to lie more frequently when it is more likely that their false statements will be flagged as lies.

Balbuzanov (2019) and Dziuda and Salas (2018) also study strategic communication endowed with a lie detection technology, yet in a cheap talk game. Our paper differs from Balbuzanov (2019) in the payoff functions. In Balbuzanov (2019), Sender and Receiver have some degree of common interest, while in our binary model, there is no common interest. Due to this difference, Sender's type-dependent preferences in Balbuzanov (2019) permits 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 given Sender's bias is small. However, Sender in our model never reveals the state perfectly due to the conflict in payoffs.

Dziuda and Salas (2018) does not allow common interest, so that fully revealing equilibria is impossible in their paper either. They have a continuous model so there are many off-path beliefs to be specified. To discipline the off-path beliefs, they impose two refinements, and show that in all such equilibria, the lowest types lie, while *some* higher types tell the truth. Our model violates the second refinement in Dziuda and Salas (2018). So our paper is not nested.

The largest difference between the two papers and ours lies in the commitment power of Sender, (need justification here) In real life, neither full commitment nor no commitment are plausible, but we believe our model is an important step towards studying the communication games with lie detection under partial commitment.

Interestingly, all three papers have some type of non-monotonicity results, though with very different substances. In Balbuzanov (2019), the set of detection probabilities that permits fully revealing equilibrium is not in the form of $[\underline{p},1]$. In Dziuda and Salas (2018), the baseline model features that lower

detection probability leads to less truthtelling. But they show in an extension that if Sender is given an opportunity to costly invest in decreasing the lie detectability, then for intermediate region of cost, the mass of liars could increase in the cost. Since detectability becomes endogenous here, so strictly speaking, this is not a non-monotonicity result in detectability.

Should we keep this paragraph? In addition to the two papers, Jehiel (2019) considers two rounds of communication and a Sender who in the second round cannot remember what lies she sent in the first round. Since Sender remembers only the unconditional distribution of first-round lies in equilibrium, lie detection probability is endogenous to equilibrium and the structure of the equilibria is similar to Dziuda and Salas (2018). As the state space becomes arbitrarily fine, the probability of lie detection goes to 1 (as it is hard to guess exactly the same lie), so only fully revealing equilibria arise.

References

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

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.

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.

Holm, Håkan J, "Truth and lie detection in bluffing," *Journal of Economic Behavior & Organization*, 2010, *76* (2), 318–324.

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," *Working Paper*, 2019.

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, forthcoming*, 2006.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin, "Persuasion via weak institutions," *Available at SSRN 3168103*, 2019.

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.

Seidmann, Daniel J, "The effects of a right to silence," *The Review of Economic Studies*, 2005, *72* (2), 593–614.

Vrij, Aldert, *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2 ed., Wiley Interscience, 2008.

A Proofs

A.1 Proof of Proposition 1

We now show that strategies I and II 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

$$\Pr_{\mathrm{I}}(a=1) \leq \mu \leq \Pr_{\mathrm{II}}(a=1) \tag{13}$$

The second inequality holds because $p_0=p_1=0$ is feasible in II and gives value μ. In fact, Within region II, it is optimal to set $p_1=0$ because this loosens both constraints, but it does not have a direct effect on the objective of maximizing the probability. Given this, we have $A=0<t$ so the optimum requires $B=t$. Therefore, p_0^* solves the following equation

$$\frac{\mu(1-q)}{\mu(1-q)+(1-\mu)p} = t \tag{14}$$

and hence

$$\Pr_{\mathrm{II}}(a=1) = \mu + \left(\frac{\mu}{t} - \mu\right)(1-q) \tag{15}$$

For $\Pr_{\mathrm{III}}(a=1)$ it is optimal to set $p_1=1$ for similar reasons. This yields $B=0<t$ which at the optimum requires p_0 to be small enough, but still ensures that $A \geq t$. Define $\bar{q} \equiv 1 - \frac{\mu(1-t)}{t(1-\mu)} \in (0,1)$, then there are two cases to consider.

- $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$ or $q \leq \bar{q}$. In this case, there exists p_0^* s.t. $A=t$, 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 $q > \bar{q}$. In this case, $A \geq t$ can never bind. So the best option is to set $p=0$ which implies $\Pr_{\mathrm{III}}(a=1) = \mu + (1-\mu)(1-q)$.

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

(1) If $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$, Region IV does not exist, *i.e.*, there is no way to choose p_0,p_1 such that $A \geq t$ and $B \geq t$. 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 $\frac{\mu(1-p_1)(1-q)}{\mu(1-p_1)(1-q)+(1-\mu)p} \geq t$ which would imply

$$\frac{\mu p_1+\mu(1-p_1)}{\mu p_1+\mu(1-p_1)+(1-\mu)(1-p_0)(1-q)+(1-\mu)\frac{p}{1-q}} \geq t \tag{16}$$

and therefore

$$t \leq \frac{\mu}{\mu+(1-\mu)(1-p_0)(1-q)+(1-\mu)\frac{p}{1-q}} \leq \frac{\mu}{\mu+(1-\mu)(1-q)} \tag{17}$$

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 $\frac{\mu}{\mu+(1-\mu)(1-q)} \leq t$, $\Pr_{III}(a=1)$ is optimal with

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

Alternatively,

$$p_0^* = \frac{\bar{q}-q}{1-q} \quad \text{and} \quad p_1^* = 1 \tag{19}$$

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

$$A \geq t \Leftrightarrow (1-t)\mu p_1 \geq t(1-\mu)(1-p_0)(1-q) \tag{20}$$

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

$$B \geq t \Leftrightarrow \mu(1-t)(1-p_1) \geq t(1-\mu)\frac{p}{1-q} \tag{21}$$

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 region IV with

the largest value of p_0. Clearly, this point is O at the intersection of the two lines in Figure 1. At the optimum, this point is given by

$$p_0^*=1-\frac{1-(1-q)\frac{\mu(1-t)}{t(1-\mu)}}{(2-q)q} \quad \text{and} \quad p_1^*=1-\frac{1-(1-q)\frac{t(1-\mu)}{\mu(1-t)}}{(2-q)q} \tag{22}$$

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

$$p_0^*=\frac{(1-q)}{(2-q)q}(q-\bar{q}) \quad \text{and} \quad p_1^*=\frac{(1-q)}{(2-q)q}\left[\frac{1}{1-\bar{q}}-(1-q)\right] \tag{23}$$

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

$$\Pr_{\text{III}}(a=1)=\mu+(1-\mu)(1-q)=1-(1-\mu)q<1-(1-\mu)q(1-p_0^*)=\Pr_{\text{IV}}(a=1). \tag{24}$$

Therefore, the optimal strategy for S involves lying in both states, that is $p_0^*\in(0,1)$ and $p_1^*\in(0,1)$. Even though S wants the action $a=1$ to be taken, he may want to say $m=0$ when $w=1$.

A.2 Proof of Proposition 2

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

$$p_0^*=\frac{\bar{q}-q}{1-q} \quad \text{and} \quad p_1^*=1 \tag{25}$$

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

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

$$p_0^*=\frac{(1-q)}{(2-q)q}(q-\bar{q}) \quad \text{and} \quad p_1^*=\frac{(1-q)}{(2-q)q}\left[\frac{1}{1-\bar{q}}-(1-q)\right] \tag{26}$$

This implies

$$\frac{\partial p_0^*}{\partial q}=\frac{(-2q+1+\bar{q})\cdot(2-q)q-(2-2q)(1-q)(q-\bar{q})}{(2-q)^2q^2} \tag{27}$$

$$=\frac{-q^2+(q^2-2q+2)\bar{q}}{(2-q)^2q^2} \tag{28}$$

Therefore,

$$\frac{\partial p_0^*}{\partial q} \geq 0 \Longleftrightarrow \frac{1}{\bar{q}} \leq \frac{q^2-2q+2}{q^2} = 1+\frac{2-2q}{q^2} \tag{29}$$

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

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

A.3 Proof of Proposition 3

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

- If $q\leq\bar{q}$, then 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,

$$U^S = \mu+(1-\mu)(1-p_0)^*(1-q) = \frac{\mu}{t} \tag{30}$$

  which is constant in q.

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

$$U^S = 1-(1-\mu)(1-p_0^*)q = 1-\frac{t(1-\mu)-\mu(1-t)(1-q)}{t(2-q)} \tag{31}$$

  which is decreasing in q as

$$\frac{\partial U^S}{\partial q} = \frac{-\mu(1-t)t(2-q)-t[t(1-\mu)-\mu(1-t)(1-q)]}{t^2(2-q)^2} \tag{32}$$

$$= \frac{-\mu(1-t)-t(1-\mu)}{t(2-q)^2} \tag{33}$$

$$< 0 \tag{34}$$

The expected payoff of Receiver is $t\cdot\Pr(a=w=0)+(1-t)\cdot\Pr(a=w=1)$. Again, there are two cases.

- If $q \leq \bar{q}$, then 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,

$$U^R = (1-\mu)t \cdot [p_0^* + (1-p_0^*)q] + \mu(1-t) \tag{35}$$

$$= (1-\mu)t \cdot [1-(1-p_0^*)(1-q)] + \mu(1-t) \tag{36}$$

$$= (1-\mu)t \cdot \left[1 - \frac{\mu(1-t)}{t(1-\mu)}\right] + \mu(1-t) \tag{37}$$

$$= (1-\mu)t \tag{38}$$

which is constant in q.

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

$$U^R = (1-\mu)t \cdot (1-p_0^*)q + \mu(1-t) \tag{39}$$

$$= (1-\mu)t \cdot \frac{1-(1-q)\frac{\mu(1-t)}{t(1-\mu)}}{2-q} + \mu(1-t) \tag{40}$$

$$= \frac{(1-\mu)t + t(1-\mu)}{2-q} \tag{41}$$

which is increasing in q.