Strategic Disinformation Generation and Detection

Wenxiao Yang[1], Yunfei (Jesse) Yao[2], and Pengxiang Zhou[*3]

[1]University of California, Berkeley, ywenxiao@berkeley.edu
[2]The Chinese University of Hong Kong, jesseyao@cuhk.edu.hk
[3]Hong Kong University of Science & Technology, pxzhou@ust.hk

February 27, 2025

Latest version of the paper

Abstract

Disinformation detection is becoming increasingly important and relevant because it is easier than ever to create and disseminate disinformation. How does detection ability affect the incentive to generate disinformation? Given the practical constraints of classification technology, how should a detector be designed? To answer these questions, this paper studies the problem where a sender strategically communicates his type (high or low) to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver then infers the sender's type both through the message from the sender and through the signal from the detector. We find a non-monotonic relationship between the probability that the low-type sender is lying and the accuracy of detection. More accurate detection (a higher true-positive rate and a lower false-positive rate) increases the probability of lying when the true-positive rate is low, because of a persuasive effect. By contrast, more accurate detection decreases the probability of lying when the true-positive rate is high, due to an dissuasive effect. We also characterize the optimal detector design. The designer always chooses the lowest feasible false-positive rate for any true-positive rate. The possibility of false-positive alarms implies that the designer chooses an intermediate true-positive rate rather than the highest true-positive rate. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

[*]Corresponding author. Authors are ordered alphabetically.

"It is better that ten guilty persons escape than that one innocent suffer."

— William Blackstone, *Commentaries on the Laws of England*

1 Introduction

There is widespread disinformation nowadays, including fake reviews, ad fraud, manipulated transactions, fraudulent resumes, and misleading posts (Anderson and Simester, 2014; Mayzlin, Dover, and Chevalier, 2014; Luca and Zervas, 2016; Gordon et al., 2021; He et al., 2022; He, Hollenbeck, and Proserpio, 2022).[1] The emerging generative AI technologies further exacerbate such deceptive practices. These issues have attracted much attention in places such as online platforms, political realms, and justice systems, where trust and integrity are paramount. Yet, detecting disinformation remains challenging (Callander and Wilkie, 2007; Dziuda and Salas, 2018; Mattes, Popova, and Evans, 2023). In response to the ubiquitous deceptive activities, platforms and regulators have devised various ways of detecting and raising an alarm about disinformation, usually with the help of sophisticated algorithms. For example, Yelp utilizes automated systems to identify compensated or incentivized reviews and flags businesses with suspicious activities.[2] Using an internal system, Twitter labels false or misleading content to help people "find credible and authentic information" and "make informed decisions."[3] To fight against fake accounts and fraudulent activities, LinkedIn has built "automated detection systems at scale."[4] Such detection and alarm attempts help individuals decide whether they will go to a particular restaurant, re-post a social media post, or connect with a Linkedin account.

When developing mechanisms to detect deceptive information, the designer typically faces a dilemma between increasing the likelihood of correctly recognizing deceptive content (true positives) and reducing the probability of falsely identifying genuine content as deceptive (false positives). Such a trade-off between Type I error (false-positive) and Type II error (false-negative) is a well-known statistical challenge faced by many fields (Goodin, 1985; Buckland and Gey, 1994; Lieberman and Cunningham, 2009; Cappelen,

[1] There are two related terms commonly used in the media and literature - misinformation and disinformation. According to Dictionary.com, misinformation refers to "false information that is spread, regardless of whether there is intent to mislead", whereas disinformation means "deliberately misleading or biased information." By focusing on the strategic incentive of the information provider (sender), this paper studies disinformation. We thank one of the anonymous reviewers for raising this important distinction between misinformation and disinformation.

[2] https://trust.yelp.com/trust-and-safety-report/2023-report/ and https://trust.yelp.com/consumer-alerts/quarterly-alerts/.

[3] https://blog.x.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

[4] https://www.linkedin.com/blog/engineering/trust-and-safety/automated-fake-account-detection-at-linkedin.

Cappelen, and Tungodden, 2023).

How does the detection ability affect the incentive to generate disinformation? Given the practical constraints of classification technology, how should the detectors be designed? To answer these questions, this paper studies the problem where a sender strategically communicates his type (high or low) to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message.[5] The receiver then infers the sender's type both through messages from the sender and through signals from the detector. Previous work has considered the possibility that a detector may fail to send an alarm when there is disinformation (false negative). Observing that the detector may make another type of mistake by sending a false alarm in the absence of disinformation (false positive), a key contribution of our paper is to allow for both types of mistakes in disinformation detection. In addition to being more realistic, it also leads to qualitatively different insights about the relationship between the probability that the low-type sender is lying and the accuracy of detection. The other main contribution of this paper is to endogenize the design of the detector rather than treating the detection technology as exogenously given. The optimal detector design is also qualitatively different with and without consideration of false-positive alarms.

Specifically, this paper considers a model where a receiver makes a binary decision between actions r_H and r_L. The sender may be either the H type (the high type) or the L type (the low type); this is his private information. The sender always wants the receiver to take action r_H, whereas the receiver prefers to take action r_H if the sender's type is H and to take action r_L if the sender's type is L. The sender can send a strategic message about his type (m_H for high type and m_L for low type) to the receiver, while a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver infers the sender's type both through the message from the sender and through the signal from the detector, and then makes a decision.

Due to practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender lies (false negative). It may also send a false alarm when the sender is truthful (false positive). (CMA, 2015; Lappas, Sabnis, and Valkanas, 2016). Previous work has focused on the first type of mistake by implicitly assuming that the false-positive rate is zero (Becker and Stigler, 1974; Dziuda and Salas, 2018; Balbuzanov, 2019). Given the practical constraints of classification technology, however, the sender cannot avoid making the second type of mistake (false positive) unless he never sends an alarm.[6]

[5] We refer to the sender as "he" and the receiver as "she" throughout the paper

[6] Similarly, the sender cannot avoid making the first type of mistake (false negative) unless he always sends an alarm.

Moreover, false positives are not only ubiquitous but also economically significant. J.P. Morgan views false positives as a multi-billion dollar problem.[7] Global business loses more than $100 billion every year due to false positives, which is even more than the actual fraud costs.[8] In order to understand the strategic impact of disinformation detection in a realistic setting, this paper explicitly considers the possibility of false-positive alarms.

We first study how the detection technology affects the equilibrium outcomes. We find a non-monotonic relationship between the probability that the low-type sender is lying and the accuracy of detection: more accurate detection (a higher true-positive rate and a lower false-positive rate) increases the probability of lying when the true-positive rate is low and decreases it when the true-positive rate is high. Two effects drive the non-monotonicity. Because the detector is more likely to send no alarm when the sender is high-type than when he is low-type, the receiver becomes more certain that the sender is high-type if she receives no alarm. The presence of a detector persuades the receiver to trust the sender's m_H message more in this case. We call this posterior belief-enhancing effect a *persuasive effect*. Because the detector is more likely to send an alarm when the sender is low-type than when he is high-type, the receiver becomes more certain that the sender is low-type if she receives an alarm. The presence of an alarm causes the receiver to have less trust about the sender's m_H message. We call this posterior belief-reducing effect an *dissuasive effect*. When the true-positive rate is low, the receiver adopts a mixed strategy between actions r_H and r_L after observing message m_H and no alarm.[9] As the detector becomes more accurate, for a fixed sender's strategy, the receiver's posterior belief after observing no alarm will be higher due to the larger persuasive effect. So, the low-type sender can afford to lie more frequently in equilibrium. When the true-positive rate is high, the detector will catch a high proportion of low-type senders who are lying. This creates a low incentive for lying. Consequently, the receiver always takes the sender's desired action if there is no alarm and adopts a mixed strategy after seeing an alarm. As the detector becomes more accurate, for a given sender's strategy, the receiver's posterior belief after observing an alarm will be lower due to a larger dissuasive effect. In equilibrium, the low-type sender needs to lie less frequently in order for there to be a positive probability that the receiver will take the sender's desired action even after observing an alarm.

We then characterize the optimal detector design. The receiver and both types of sender all benefit from

[7] https://www.jpmorgan.com/insights/payments/analytics-and-insights/cnp-fraud-prevention-combat-chargebacks.

[8] https://www.vesta.io/blog/false-positives-and-how-to-prevent-them.

[9] In other words, the receiver takes action r_H with some probability and takes action r_L with some probability after observing message m_H and no alarm.

a lower false-positive rate, whereas the low-type sender is hurt by a higher true-positive rate. Therefore, the designer always chooses the lowest feasible false-positive rate for any given true-positive rate. The possibility of false-positive alarms implies that the designer will not choose the largest true-positive rate. Instead, the designer chooses different intermediate true-positive rates for different objectives. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

Related Literature

Our research is most closely related to the strategic communication literature. One stream of the literature on verifiable disclosure, initiated by Grossman (1981) and Milgrom (1981), assumes that information is verifiable and thus agents can withhold it but cannot lie. Another stream of the literature on cheap talk, developed by Crawford and Sobel (1982), considers a model where information is unverifiable and thus agents can freely send deceptive messages. Later work studies strategic communication in markets where firms and consumers interact (Villas-Boas, 2004; Shin, 2005; Guo, 2009; Guo and Zhao, 2009; Kuksov, 2009; Kuksov and Lin, 2010; Mayzlin and Shin, 2011; Sun, 2011; Zhang, 2013; Branco, Sun, and Villas-Boas, 2016; Iyer and Singh, 2018; Sun and Tyagi, 2020; Iyer and Singh, 2022; Lauga, Ofek, and Katona, 2022; Zheng and Singh, 2023; Lee, Shin, and Yu, 2024). In the verifiable disclosure literature, senders can be viewed as having an infinite lying cost and therefore never lie, whereas they have zero lying cost in the cheap talk literature. The more recent literature on the theory of costly lying (Kartik, Ottaviani, and Squintani, 2007; Kartik, 2009) assumes that the sender has a finite lying cost and can be viewed as the middle ground of two extreme cases. The presence of lying cost in the above papers allows messages to have a signaling role. However, the information is still completely unverifiable. Recent work (Dziuda and Salas, 2018; Balbuzanov, 2019) starts considering the possibility of an imperfect detector that may detect the lie with some probability. In such cases, the sender's message becomes partially verifiable.

Due to practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender lies (false negative). It also may send a false alarm when the sender is telling the truth (false positive). By assuming that the detector detects the lie with some probability, previous work implicitly assumes that there is no false positive (the false-positive rate is zero). A major contribution of our paper is to allow for both types of mistakes by studying a detector with general true-positive and false-positive rates. The other key contribution is to endogenize the design of the detector rather than treating the detection

technology as exogenously given. One reason that previous literature has focused on exogenous detectors is that, in the absence of false positives, the receiver's payoff, the high-type sender's payoff, and social welfare are all (weakly) increasing in the true-positive rate of the detector; we will discuss this as a benchmark situation in section 3.2. So, there is no trade-off, and the designer always wants to maximize the true-positive rate of the detector. In contrast, we will show that, in the presence of false-negative alarms, the designer prefers an intermediate true-positive rate to the highest true-positive rate. In this case, the optimal detector design becomes both non-trivial and managerially important.

We use an information design framework to study the general design of the detector. Since Rayo and Segal (2010) and Kamenica and Gentzkow (2011) initiated the study of the optimal design of flexible information provision with commitment, researchers have found its applications in various areas, including advertising, recommendation algorithms, influencer marketing, search, and online platforms (Jerath and Ren, 2021; Pei and Mayzlin, 2022; Ke, Lin, and Lu, 2022; Iyer and Zhong, 2022; Berman, Zhao, and Zhu, 2022; Shin and Wang, 2024; Shulman and Gu, 2024; Yao, 2024). Unlike papers in this literature, the information design problem in our model is just a subgame of a costly signaling game. Due to the presence of the sender's private information, the sender sends strategic signaling messages to influence the receiver's decision, on top of the information design of the detection technology. In addition, we are studying the design of detection technology rather than the information itself.

Our paper is also related to the growing literature on strategic interactions between humans and algorithms. Liang (2019); Miklós-Thal and Tucker (2019); Calvano et al. (2020); Salant and Cherry (2020); O'Connor and Wilson (2021) and Montiel Olea et al. (2022) study competitive dynamics among multiple algorithms. Berman and Katona (2013, 2020) study the impact of online algorithms on advertisers' and social media users' behavior. Eliaz and Spiegler (2019) and Björkegren, Blumenstock, and Knight (2020) look at algorithm design with information manipulation by strategic agents. Qian and Jain (2024) investigate the impact of recommendation systems on digital content creation. Iyer and Ke (2024) study on strategic model selection in competitive environments. Iyer, Yao, and Zhong (2024) examine the precision-recall trade-off in the deployment of machine learning algorithms for targeting. We focus instead on the design of a disinformation detector in a strategic communication game.

Lastly, our paper is related to the literature on information misrepresentation. Anderson and Simester (2014); Mayzlin, Dover, and Chevalier (2014); Lappas, Sabnis, and Valkanas (2016) and Luca and Zervas (2016) provide empirical evidence about the prevalence of strategic review manipulation. Mayzlin (2006)

and Dellarocas (2006) theoretically study firms' costly misrepresentation of product quality. A growing literature investigates deceptive advertising practices that promote false claims about product quality (Piccolo, Tedeschi, and Ursino, 2015; Zinman and Zitzewitz, 2016; Rao and Wang, 2017; Piccolo, Tedeschi, and Ursino, 2018; Rhodes and Wilson, 2018) Jin, Yang, and Hosanagar (2023) examines a widespread phenomenon on e-commerce platforms where sellers place fake orders to boost the search ranking of their products. Some papers also examine the regulation and policy implications of deceptive activities (Piccolo, Tedeschi, and Ursino, 2015; Rhodes and Wilson, 2018; Papanastasiou, 2020; Chen and Papanastasiou, 2021). We contribute to this literature by considering the trade-offs between the false-positive and true-positive rates of the detection algorithm and studying the interaction between the sender's strategic communication strategy and the detection technology.

The rest of this paper is organized as follows. Section 2 introduces the main model. Section 3 presents several benchmarks. Section 4 solves the equilibrium and compares the results with the benchmarks. Section 5 concludes.

2 Model

2.1 States, Actions, and Payoffs

There are two players: a sender (S) and a receiver (R). The receiver makes a binary decision between actions r_H and r_L. Depending on the specific applications, the receiver's action can be purchasing a product from an e-commerce seller, visiting a restaurant, re-posting social media content, sending a business contact request, etc. The sender is the H type with probability ρ and the L type with probability $1-\rho$ (we will use type H/L and high-type/low-type interchangeably throughout the paper). The sender's type is his private information and can be interpreted as the quality of an online marketplace seller's product, the quality of a restaurant, the trustworthiness of a social media content creator, the authenticity of an online business account, etc. The sender always wants the receiver to take action r_H, whereas the receiver prefers to take action r_H if the sender's type is H and to take action r_L if the sender's type is L. Table 1 summarizes players' payoffs.

(sender payoff, receiver payoff)	action r_H	action r_L
type H sender	$(\Delta_H^S > 0, \Delta_H^R > 0)$	$(0, 0)$
type L sender	$(\Delta_L^S > 0, \Delta_L^R < 0)$	$(0, 0)$

Table 1: Players' Payoffs

The sender always earns a positive payoff if the receiver takes action r_H, $\Delta_H^S, \Delta_L^S > 0$. The receiver's payoff is positive if she takes action r_H when the sender's type is H, $\Delta_H^R > 0$, and negative if she takes action r_H when the state is L, $\Delta_L^R < 0$. Both players' payoffs are normalized to zero if the receiver takes action r_L. Denote by $\hat{\rho} \in (\rho, 1)$ the *critical belief* such that the receiver is indifferent between taking either action, $(1-\hat{\rho})\Delta_L^R + \hat{\rho}\Delta_H^R = 0$. We study the non-trivial case where the receiver will take action r_L without any information by assuming that $(1-\rho)\Delta_L^R + \rho\Delta_H^R < 0 \Leftrightarrow \rho < \hat{\rho}$. We also focus on the case where $\Delta_L^S + \Delta_L^R \leq 0$, which means that successful lying by a low-type sender does not increase social welfare.[10]

The sender can send a non-verifiable but detectable message $m \in \{m_H, m_L\}$ about his type to the receiver. The sender is lying if his message is not aligned with his type (i.e., sending message m_H if his type is L or sending message m_L if his type is H). Consistent with previous literature, the sender needs to incur a positive cost of C if he lies. The lying cost may come from the sender's intrinsic aversion to lying (Gneezy, 2005), the potential ex-post penalty for lying, or the effort of manipulating the information. We rule out the uninteresting case where the sender never lies due to a high lying cost by assuming that $C < \min\{\Delta_H^S, \Delta_L^S\}$.

A lie detector generates a noisy signal $l \in \{a, na\}$ on the truthfulness of the sender's message if the sender sends m_H. The detector will send an alarm, $l = a$, to the receiver if it thinks the message m_H is sent by a type L sender. It will send a no-alarm signal, $l = na$, to the receiver if it thinks the message m_H is sent by a type H sender or the message is m_L.

Eventually, the receiver infers the sender's type through messages from the sender and the detector and then makes a decision.

The timing of the game is as follows:

1. The designer designs the lie detector (the details are in the following paragraphs).

2. Nature draws the sender's type $\theta \in \{H, L\}$.

[10] Our analyses can be easily extended to the case where $\Delta_L^S + \Delta_L^R > 0$.

3. The sender sends a message $m \in \{m_H, m_L\}$ to the receiver.

4. The detector sends a signal $l \in \{a, na\}$ to the receiver.

5. The receiver takes an action $r \in \{r_H, r_L\}$.

Detector Design

A designer designs the detector. The designer's goal depends on the specific contexts, including maximizing the receiver's expected payoff, maximizing the high-type sender's expected payoff, and maximizing social welfare. We assume that the designer has access to an exogenously given classifier that generates a prediction for the message's trustworthiness when the sender sends message m_H.[11] The designer then decides whether to send an alarm based on the prediction. More specifically, the classifier generates a binary outcome $s \in \{s_L, s_H\}$. We assume without loss of generality that the message is more likely to be truthful if the outcome is s_H. Formally, denote the probability of outcome s conditional on the sender's true type θ by $\phi(s|\theta)$. Then, $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$.[12] We refer to ϕ as the classifier's capacity, as it reflects the quality of the classification. For example, the classifier perfectly reveals the truthfulness of the message if $\phi(s_H|\theta = H) = \phi(s_L|\theta = L) = 1$, whereas it is not very informative if $\phi(s_H|\theta = H)$ is close to $\phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L)$ is close to $\phi(s_L|\theta = H)$. In reality, the classifier will not be perfect at classification due to practical limitations. So, we assume that $0 < \phi(s|\theta) < 1$ for $s \in \{s_H, s_L\}$ and $\theta \in \{H, L\}$.

Given the classifier's prediction, the designer decides whether to send an alarm. The designer's decision can be characterized by the probability of sending an alarm given classification outcome s_L, $\lambda_L = \Pr(l = a|s_L)$, and the probability of sending an alarm given classification outcome s_H, $\lambda_H = \Pr(l = a|s_H)$. We will refer to $\{\lambda_L, \lambda_H\}$ as the alarm rule. In a perfect world, a detector sends an alarm if and only if a low-type sender sends a deceptive message m_H. Therefore, the detector's **true-positive rate, denoted by** β, is the probability of sending an alarm when a type L sender sends message m_H, $\Pr(l = a \mid m = m_H, \theta = L)$. The detector will send a false alarm if a high-type sender sends a message m_H. So, the detector's **false-positive rate, denoted by** α, is $\Pr(l = a \mid m = m_H, \theta = H)$. For a given alarm rule $\{\lambda_L, \lambda_H\}$, the

[11] We will show that, in equilibrium, the sender must be the low type if he sends message m_L; there is no uncertainty about the sender's type when the receiver sees message m_L.

[12] Conditions $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$ are equivalent to $Pr(\theta = H|s_H) > Pr(\theta = H|s_L)$ by Bayes' rule.

true-positive rate is $\beta = \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H$ and the false-positive rate is $\alpha = \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H$, according to Bayes' rule. We refer to α and β as the detector's capacity because they reflect the quality of the detection. We can represent a detector by its capacity, $\{\beta, \alpha\}$, or by its alarm rule and the capacity of the classifier, $\{\lambda_L, \lambda_H, \phi\}$. Intuitively, a stronger detector correctly alarms a lie more frequently and mistakenly alarms a truth-telling message less frequently. This leads to the following definition, which is useful in the subsequent analyses.

Definition 1. A detector $\{\beta', \alpha'\}$ is stronger than a detector $\{\beta, \alpha\}$ if and only if the following conditions hold: $\beta' \geq \beta$, $\alpha' \leq \alpha$, and at least one of the inequalities is strict.

Receiver's Belief Updating

Without any information, the receiver's *prior belief* that the sender is high-type is ρ. The receiver updates her belief after the sender chooses a message based on the sender's communication strategy and Bayes' rule. The updated belief is the *intermediate belief* of the receiver. The receiver updates the belief again after observing the detector's signal. Because this belief takes into account all the available information the receiver can obtain, it is the *posterior belief* of the receiver. Figure 1 illustrates the receiver's belief updating processes.

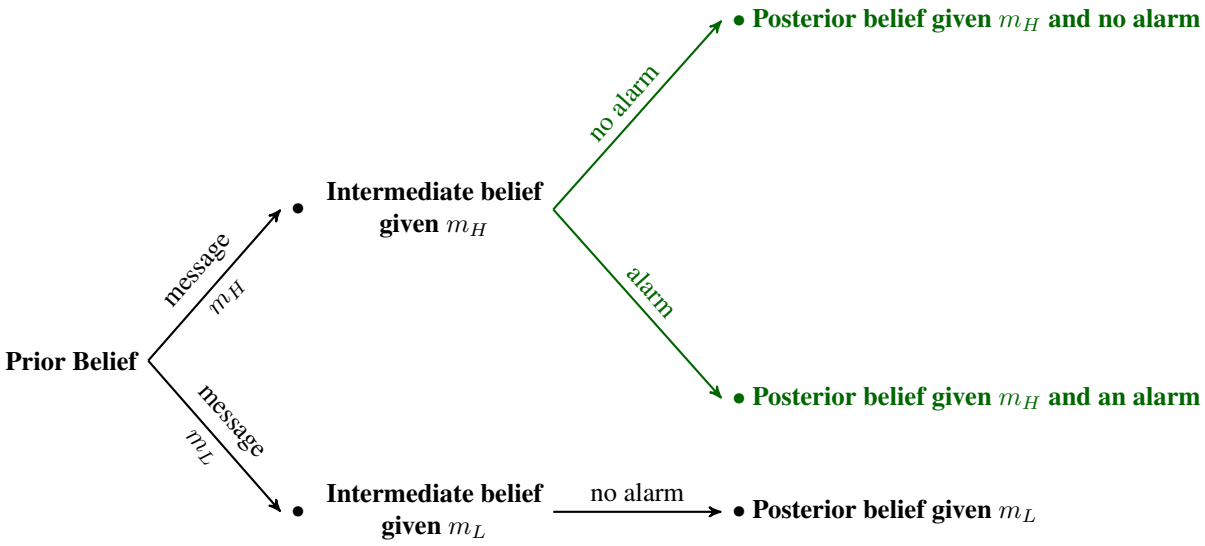

Figure 1: Receiver's Belief Updating Processes

In reality, the receiver may not observe the sender's message and the alarm signal sequentially. For example, they may see both the reviews and alarm information about a restaurant on Yelp. In such cases,

the receiver will directly reach the posterior belief given the information. This does not change our analysis because we introduce the intermediate belief in order to present the intuition and underlying mechanisms, rather than to be interpreted literally.

2.2 Strategies and Equilibrium Concept

Because we are studying a multi-stage game with incomplete information, we consider the Perfect Bayesian Equilibrium (PBE hereafter). We denote the sender's strategy by $\sigma^S(m \mid \theta)$, the probability of sending message $m \in \{m_L, m_H\}$ when the sender's type is $\theta \in \{L, H\}$. We denote the receiver's posterior belief about the sender's type by $b(t \mid m, l)$, the probability that the sender's type is $t \in \{L, H\}$ given the sender's message $m \in \{m_L, m_H\}$ and the detector's signal $l \in \{a, na\}$. We denote the receiver's strategy by $\sigma^R(r \mid m, l)$, the probability that the receiver takes action $r \in \{r_L, r_H\}$ given the sender's message $m \in \{m_L, m_H\}$ and the detector's signal $l \in \{a, na\}$.

One can see that the receiver will take action r_H if her posterior belief about the sender's probability of being type H is higher than $\hat{\rho}$ and will take action r_L if her posterior belief is lower than $\hat{\rho}$. Intuitively, the sender has no incentive to pretend to be type L by sending costly disinformation when he is type H. The following lemma formalizes the intuition that the high-type sender is always truth-telling in equilibrium.

Lemma 1. *In any PBE, type H sender always sends the message $m = m_H$. The receiver always takes action $r = r_L$ after receiving message $m = m_L$, if the sender sends message m_L with a positive probability in equilibrium.*

Proof. See A.1. □

To simplify notation, we denote the low-type sender's strategy by $\sigma^S \equiv \sigma^S(m_H \mid L)$ and denote the receiver's strategy by $\sigma^R_{na} \equiv \sigma^R(r_H \mid m_H, na), \sigma^R_a \equiv \sigma^R(r_H \mid m_H, a)$, and $\sigma^R_{L,na} \equiv \sigma^R(r_H \mid m_L, na)$. Because Lemma 1 has pinned down the strategy of type H sender and the strategy of the receiver upon receiving message m_L, in the subsequent analyses, we will use $\{\sigma^{S*}, \sigma^{R\,*}_{na}, \sigma^{R*}_a, \alpha^*, \beta^*\}$ to denote the entire equilibrium, and will use $\{\sigma^{S*}, \sigma^{R\,*}_{na}, \sigma^{R*}_a\}$ to denote the equilibrium with an exogenous lie detector.

3 Some Benchmarks

3.1 No alarm

Lie detection plays an important role in our model. To better understand its strategic role, we consider a benchmark where the detector never sends an alarm (both the true positive and false positive rates equal zero).

Lemma 2. *Suppose the detector always sends a no-alarm signal, na. In the unique PBE, the low-type sender sends message m_H with probability $-\rho\Delta_H^R/[(1-\rho)\Delta_L^R] \in (0,1)$; the receiver has a posterior belief of $\hat{\rho}$ and takes action r_H with probability $C/\Delta_L^S \in (0,1)$ upon observing message m_H. Both the low-type sender and the receiver obtain zero expected payoff. The high-type sender obtains an expected payoff of $\Delta_H^S C/\Delta_L^S$.*

Proof. See A.2. □

In this benchmark, the unique PBE is a semi-pooling equilibrium. A high-type sender always sends a truthful message, whereas a low-type sender uses a mixed strategy, with some probability of sending a truthful message m_L and some probability of pretending to be the high-type by sending message m_H. The probability of lying is such that the receiver has a posterior belief of $\hat{\rho}$ upon seeing m_H, and is indifferent between taking either action. When the prior belief that the sender is type H, ρ, is higher, the receiver is more inclined to believe that the sender is type H upon receiving message m_H for a given sender's strategy. Therefore, the type L sender is more likely to send a deceptive message m_H. Upon receiving message m_H, the receiver does not know whether the sender is a truth-telling high-type sender or a deceptive low-type sender and uses a mixed strategy between actions r_H and r_L. The probability of taking action r_H is such that a low-type sender is indifferent between lying and truth-telling. The sender's cost of lying increases in C. For him to be indifferent between lying and not lying, the benefit of lying must also increase in C. So, in equilibrium, upon observing message m_H, the receiver takes the sender's desired action r_H more frequently when the lying cost C increases.

3.2 No false-positive alarm

As discussed in the introduction, previous work on lie detection under strategic communication implicitly assumes that there is no false-positive alarm. Those studies only consider one type of mistake, where

a detector may fail to send an alarm when there is disinformation (false negative). This section considers a benchmark consistent with previous literature by assuming that there is a detector that never sends a false-positive alarm ($\alpha = 0$).

Lemma 3 (Exogenous detector)**.** *Suppose the detector's false-positive rate is zero. The low-type sender is lying with a higher likelihood as the detector's true-positive rate β increases but stops lying when β exceeds $\hat{\beta} := 1 - C/\Delta_L^S$. The PBEs are the following:*

1. *High lying cost $C \geq -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$*

$$\begin{cases} \sigma^{S*} = -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^{R\,*} = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < \hat{\beta} \\ \sigma^{S*} \in \left[0, -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}\right], \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

2. *Low lying cost $C < -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$*

$$\begin{cases} \sigma^{S*} = -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^{R\,*} = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, \sigma_{na}^{R\,*} \in \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right], \sigma_a^{R*} = 0, & \beta = 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} < \beta < \hat{\beta} \\ \sigma^{S*} \in [0,1], \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, \sigma_{na}^{R\,*} = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

Proof. See A.3. □

Figure 2 illustrates the low-type sender's equilibrium probability of lying as a function of the detector's true-positive rate β. In the absence of false-positive alarms, the receiver may see an alarm only if the sender is low-type. In that case, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. The receiver never takes the sender's desired action when there is an alarm. If a low-type sender is caught lying by the detector, he obtains no benefit from lying but instead incurs the cost of lying. When the true-positive rate is high, the expected payoff from lying is negative because the low-type sender has a high chance of being detected. So, the sender never lies and there is no disinformation when the true-positive rate β exceeds a threshold $\hat{\beta}$. We will

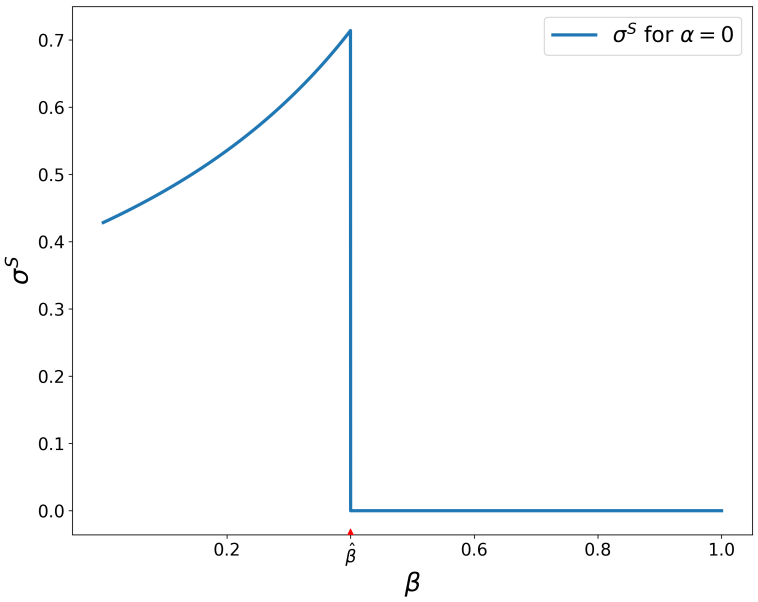

Figure 2: Probability of lying under different detectors without false-positive alarms for $\Delta_H^R = 0.5, \Delta_L^R = -0.5, \Delta_L^S = 0.5, C = 0.3, \rho = 0.3$, and any $\Delta_H^S > 0$.

show in the main model that this is not the case when we consider false-positive alarms.

We now study the endogenous detector design. To have a well-defined equilibrium payoff, we select the Pareto-optimal equilibrium for those cases with multiple equilibria. For a given detector $(\beta, 0)$, denote the receiver's expected equilibrium payoff by $\mathbb{E}U_0^R(\beta)$, the high-type sender's expected equilibrium payoff by $\mathbb{E}U_{0,H}^S(\beta)$, and the low-type sender's expected equilibrium payoff by $\mathbb{E}U_{0,L}^S(\beta)$. The social welfare is $\mathbb{E}W_0(\beta) = \mathbb{E}U_0^R(\beta) + \rho\mathbb{E}U_{0,H}^S(\beta) + (1-\rho)\mathbb{E}U_{0,L}^S(\beta)$. We consider three types of objectives by the designer, including choosing β to maximize the receiver's expected payoff $\mathbb{E}U_0^R(\beta)$, the high-type sender's expected payoff $\mathbb{E}U_{0,H}^S(\beta)$, or the social welfare $\mathbb{E}W_0(\beta)$.

Lemma 4 (Endogenous detector)**.** *The receiver's expected payoff, the high-type sender's expected payoff, and the social welfare all (weakly) increase in the true-positive rate β. The optimal true positive rate for the receiver is any $\beta \geq \hat{\beta}$. The optimal true positive rate for the high-type sender is any $\beta \geq \min\left\{1 + \rho\Delta_H^R/[(1-\rho)\Delta_L^R], \hat{\beta}\right\}$. The optimal true positive rate for social welfare is any $\beta \geq \hat{\beta}$.*

Proof. See A.5. □

The receiver benefits from better distinguishing two types of senders and making a more informed decision. As long as the true-positive rate exceeds $\hat{\beta}$, the low-type sender stops lying due to the high likelihood of being caught. So, the receiver can perfectly infer the sender's type from the sender's message

m, and can achieve the highest payoff for any β that is high enough. There is never an alarm when the sender is high type. So, a high-type sender only cares about the receiver's action upon seeing no alarm. Because the receiver always takes the sender's desired action when $\beta \geq \min\left\{1+\rho\Delta_H^R/[(1-\rho)\Delta_L^R], \hat{\beta}\right\}$, the high-type sender achieves the highest payoff for any such β. Because society benefits from a lower level of disinformation, social welfare achieves its maximum when β is high enough such that the sender stops lying.

The general message from the lemma is simple and intuitive: when there is no false positive, the more accurate the detector is, the better. So, there is no trade-off, and the detector designer always prefers a higher true-positive rate. In reality, the detector may make another type of mistake by sending a false alarm in the absence of disinformation (false positive). It is generally impossible to eliminate either type of mistake unless the detector always or never sends alarms. We will show in the main model that the designer strictly prefers an intermediate true-positive rate to the highest true-positive rate in the presence of false-negative alarms. A higher true-positive rate may reduce the receiver's expected payoff, the high-type sender's expected payoff, and the social welfare. In this case, the optimal detector design becomes both non-trivial and managerially important.

4 Equilibrium

4.1 Equilibrium with an exogenous detector

We first consider the equilibrium with an exogenous detector $\{\beta, \alpha\}$ for two reasons. First, the entire equilibrium is complicated with three strategic players: the sender, the receiver, and the designer/detector. The equilibrium with only the sender and the receiver is simpler to solve and serves as a building block to solve the entire equilibrium. Second, by abstracting away the strategic role of the designer/detector, we can more clearly understand the driving force of different results.

The detector is uninformative if $\alpha = \beta$. In such cases, the equilibrium outcome is essentially the same as the equilibrium of the no-alarm benchmark in section 3.1.[13] We will present the formal characterization of the equilibrium in the appendix, and will show that an uninformative detector is never optimal in equilibrium. For any detector such that $\alpha > \beta$, we can obtain essentially the same equilibrium outcome with an alternative

[13]Technically, there is a subtle difference between the two cases because we need to specify the receiver's strategies both upon receiving an alarm and upon receiving no alarm in the $\alpha = \beta > 0$ case, though either strategy is the same as the receiver's strategy in the no alarm benchmark because the detector is not informative.

detector whose $\alpha < \beta$. Therefore, we focus on the interesting case where the detector may send both types of alarms and the false-positive rate is lower than the true-positive rate, $0 < \alpha < \beta$.

Proposition 1 (Equilibrium with an Exogenous Detector)**.** *Suppose the detector $\{\beta, \alpha\}$, $0 < \alpha < \beta$, is exogenously given. The receiver's posterior belief about the sender being type H upon observing message m_H and a noisy signal $l \in \{n, na\}$ is the following.*

$$b(t \mid m_H, l) = \begin{cases} \frac{\alpha\rho}{\alpha\rho+\beta\sigma^S(1-\rho)}, & l = a \\ \frac{(1-\alpha)\rho}{(1-\alpha)\rho+(1-\beta)\sigma^S(1-\rho)}, & l = na \end{cases}$$

The low-type sender's probability of lying first increases and then decreases in the detector's true-positive rate β. The PBEs are the following.

$$\sigma^{S*} \begin{cases} = \min\left\{-\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}, & \beta < \hat{\beta} \\ \in \left[-\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{-\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right], & \beta = \hat{\beta} \\ = -\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, & \beta > \hat{\beta} \end{cases}$$

$$\sigma_{na}^{R\,*} \begin{cases} = 1, & \beta > 1 + \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \in \left[\min\left\{\frac{C}{(1-\beta)\Delta_L^S}, 1\right\}, 1\right], & \beta = 1 + \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ = \min\left\{\frac{C}{(1-\beta)\Delta_L^S}, 1\right\}, & \beta \in (\alpha, 1 + \frac{(1-\alpha)\rho\Delta_H^R}{(1-\rho)\Delta_L^R}) \end{cases}$$

$$\sigma_a^{R*} = \max\left\{\frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, 0\right\}$$

Table 2 summarizes the PBEs when $\alpha > 0$.

β Range \ α Range	$\alpha < 1 + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha = 1 + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha \in \left(1 + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta), \beta\right)$	$\alpha = \beta$
$\beta \in \left(\hat{\beta}, 1\right]$	$\sigma^S = -\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}$			$\sigma^S = -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}$,
$\beta = \hat{\beta}$	$\sigma^S \in \left[-\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{-\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right], \sigma_{na}^R = 1, \sigma_a^R = 0$			$\sigma_{na}^R, \sigma_a^R$
$\beta \in \left(0, \hat{\beta}\right)$	$\sigma^S = 1$, $\sigma_{na}^R = 1$, $\sigma_a^R = 0$	$\sigma^S = 1$, $\sigma_{na}^R \in \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right]$, $\sigma_a^R = 0$	$\sigma^S = -\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}$, $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}$, $\sigma_a^R = 0$	such that $\beta\sigma_a^R + (1-\beta)\sigma_{na}^R = \frac{C}{\Delta_L^S}$

Table 2: Equilibria with an exogenous detector

Proof. See A.3. □

4.1.1 Effect of Lie Detection on Receiver's Posterior Belief

After observing message m_H but before observing the detector's signal, the receiver's intermediate belief about the sender's type is $\Pr(\theta = H|m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$. We now disentangle two effects of lie detection on the receiver's posterior belief, illustrated by Figure 3. We also examine how a stronger lie detector changes the belief-updating process by comparing the receiver's posterior beliefs under detector $\{\beta, \alpha\}$ and a stronger detector $\{\beta', \alpha'\}$.

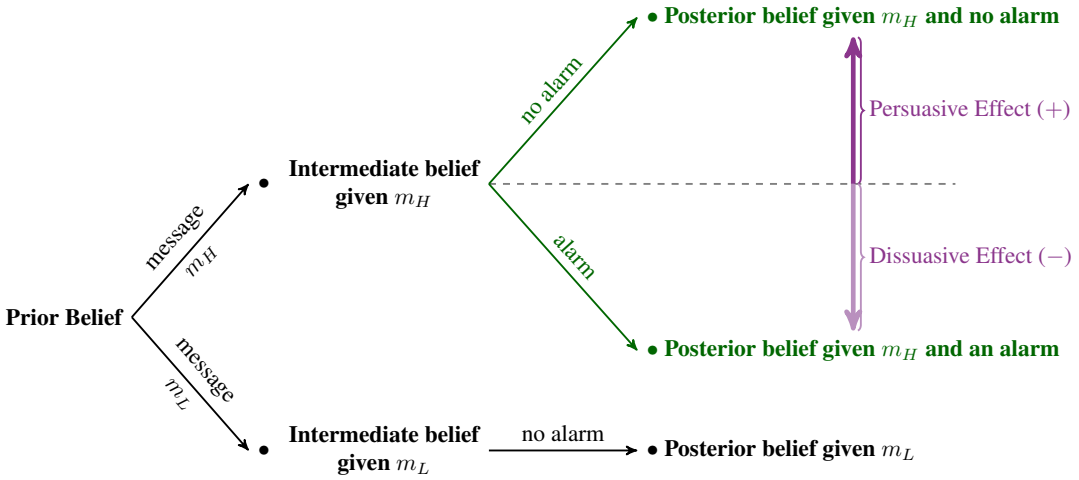

Figure 3: The Effect of Lie Detection on the Receiver's Belief.

1. Persuasive effect: Because the detector is more likely to send no alarm when the sender is high-type than when he is low-type, the receiver becomes more certain that the sender is high-type if she receives no alarm. The presence of a detector persuades the receiver to trust the sender's m_H message more in this case. We call this posterior belief-enhancing effect a persuasive effect.

   Formally, the persuasive effect raises the receiver's belief from the intermediate belief, $\Pr(\theta = H|m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$, to the posterior belief, $\Pr(\theta = H|m = m_H, l = na) = (1-\alpha)\rho/[(1-\alpha)\rho + (1-\beta)\sigma^S(1-\rho)]$. The posterior belief increases in the true-positive rate and decreases in the false-positive rate. Therefore, *the persuasive effect is larger under a stronger detector*, as illustrated by Figure 4.

   Upon receiving message m_H and no alarm, the receiver takes the sender's desired action $r = r_H$ if the posterior belief exceeds $\hat{\rho}$, $\Pr(\theta = H|m = m_H, l = na) \geq \hat{\rho}$. This condition is more likely to be satisfied under a stronger detector because a stronger detector generates a larger persuasive effect.

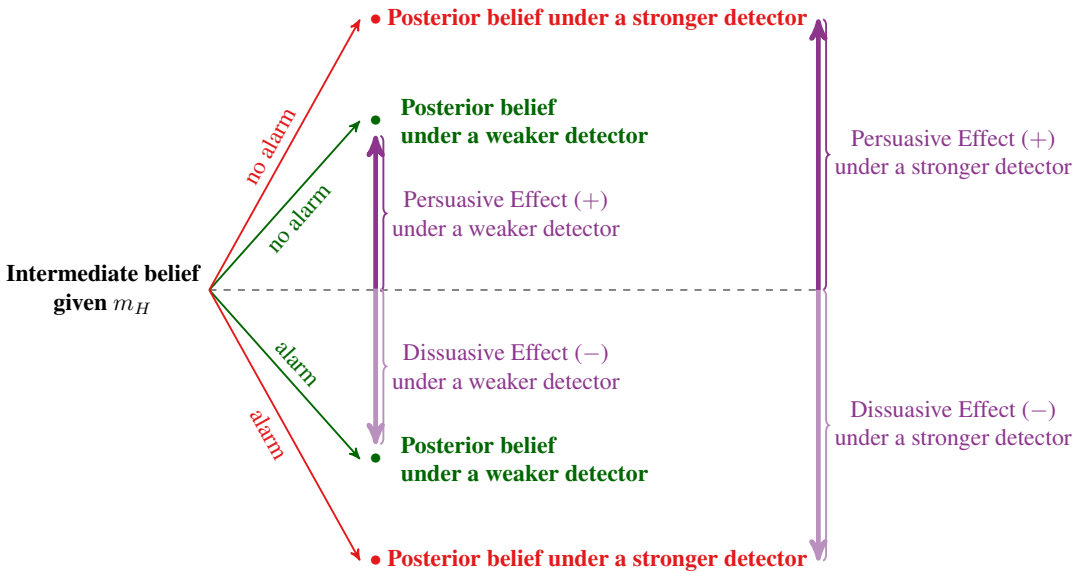

Figure 4: The Effect of a Stronger Lie Detector on the Receiver's Belief.

2. Dissuasive effect: Because the detector is more likely to send an alarm when the sender is low-type than when he is high-type, the receiver becomes more certain that the sender is low-type if she receives an alarm. The presence of an alarm makes the receiver less trustful about the sender's m_H message. We call this posterior belief-reducing effect a dissuasive effect.

   Formally, the dissuasive effect reduces the receiver's belief from the intermediate belief, $\Pr(\theta = H|m = m_H) = \rho/[\sigma^S(1-\rho)+\rho]$, to the posterior belief, $\Pr(\theta = H|m = m_H, l = a) = \alpha\rho/[\alpha\rho + \beta\sigma^S(1-\rho)]$. The posterior belief decreases in the true-positive rate and increases in the false-positive rate. Therefore, *the (absolute value of the) dissuasive effect is larger under a stronger detector*, as illustrated by Figure 4.

   A key difference between our setting and the no false-positive alarm benchmark is related to the dissuasive effect. In the absence of false-positive alarms ($\alpha = 0$), the receiver may see an alarm only if the sender is low-type. Therefore, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. Thus, the dissuasive effect does not depend on the true-positive rate of the detector; two detectors with very different β generate the same effect on the posterior belief if they send an alarm. In contrast, in the presence of false-positive alarms, two detectors with the same false-positive rate but different true-positive rates generate different dissuasive effects. As we will show in the next subsection, variations in the

dissuasive effects lead to qualitatively different equilibrium outcomes.

Upon receiving message m_H and an alarm, the receiver takes the sender's desired action $r = r_H$ if the posterior belief exceeds $\hat{\rho}$, $\Pr(\theta = H|m = m_H, l = a) \geq \hat{\rho}$. This condition is less likely to be satisfied under a stronger detector because a stronger detector generates a larger (negative) dissuasive effect.

4.1.2 Non-monotonic Relationship Between the Detector's Capacity and the Sender's Probability of Lying

According to Proposition 1, there is a non-monotonic relationship between the detector's capacity α and β and a low-type sender's probability of lying σ^S. Figure 5 illustrates such non-monotonicity by plotting a low-type sender's probability of lying as a function of the detector's true-positive rate for three fixed false-positive rates. As we can see from the figure, a stronger detector increases the probability of lying when the true-positive rate is low and decreases the probability of lying when the true-positive rate is high. Below, we discuss the underlying mechanism and intuition in detail.

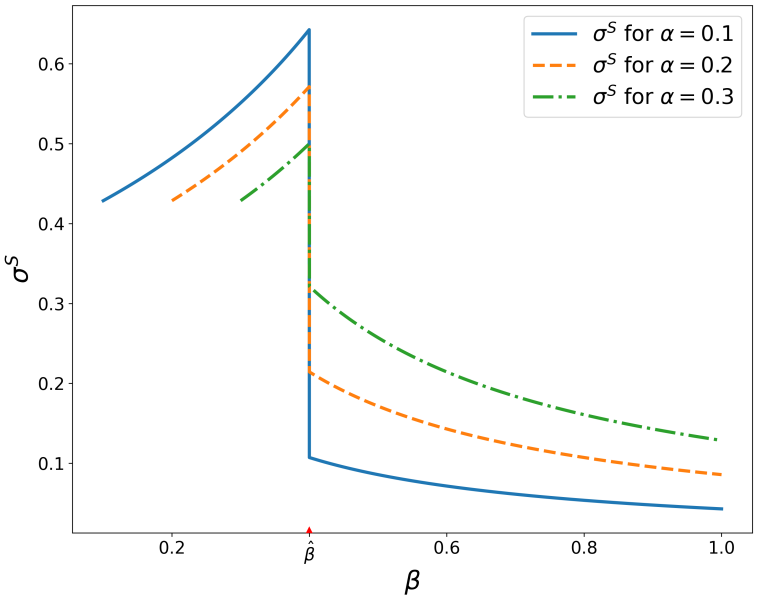

Figure 5: Probability of lying under different detectors for $\Delta_H^R = 0.5, \Delta_L^R = -0.5, \Delta_L^S = 0.5, C = 0.3, \rho = 0.3$, and any $\Delta_H^S > 0$.

Low True-positive Rate

When the detector's true-positive rate β is low, the detector will fail to catch many low-type senders who are lying. This creates a strong incentive for a low-type sender to pretend to be a high-type. So, the probability of lying is at a relatively high level. This leads to a lower posterior belief. Therefore, the receiver will never take the sender's desired action upon observing an alarm. A low-type sender pretending to be a high type will not be detected with probability $1 - \beta$.

If the receiver always takes the sender's desired action r_H after observing message m_H and no alarm, the expected benefit of lying is $(1 - \beta)\Delta_L^S$, which is larger than the lying cost C when β is low. However, this implies that a low-type sender will always lie, and that the receiver should not take action r_H upon observing no alarm. This is a contradiction. If the receiver never takes the sender's desired action r_H after observing message m_H and no alarm, no low-type sender will lie. But this implies that only the high type will send message m_H and that the receiver's posterior belief upon observing m_H will be one. Hence, the receiver should always take action r_H upon observing message m_H. This is also a contradiction. In sum, the only equilibrium for the receiver is to follow a mixed strategy after observing message m_H and no alarm. This implies that her posterior belief after observing message m_H and no alarm is exactly $\hat{\rho}$.

For a fixed sender's strategy, as the detector becomes stronger, the receiver's posterior belief after observing no alarm will be higher due to the larger persuasive effect. Given that the posterior belief in equilibrium remains $\hat{\rho}$, one can see that the low-type sender lies more frequently, which lowers the base rate of a high-type sender conditional on sending message m_H. Therefore, the equilibrium probability of lying increases as the detector becomes stronger. Figure 6 illustrates the mechanism.

High True-positive Rate

When the detector's true-positive rate β is high, the detector will catch a high proportion of low-type senders who are lying. This creates a low incentive for a low-type sender to pretend to be a high type. So, the probability of lying is at a relatively low level. This implies that the base rate of a low-type sender conditional on sending message m_H is low. So, after observing message m_H and before observing the alarm, the receiver has a high intermediate belief about the sender being high-type. Consequently, the receiver always takes the sender's desired action if there is no alarm. Additionally, even after the receiver observes an alarm that reduces her belief, the posterior belief is still high enough such that the receiver may

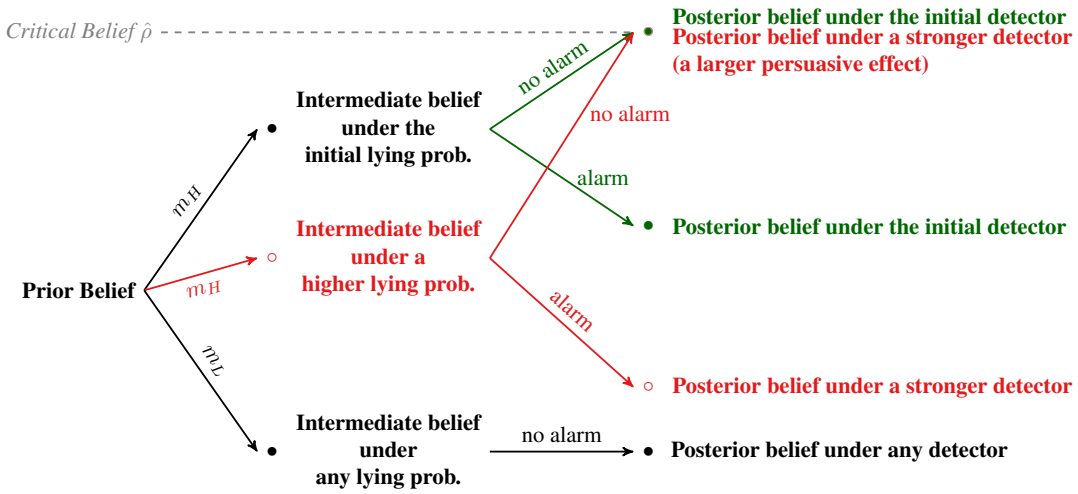

Figure 6: How a stronger lie detector increases the probability of lying when β is low.

take the sender's desired action with a positive probability.

Suppose the receiver's posterior belief after observing an alarm is higher than $\hat{\rho}$. The receiver always takes the sender's desired action regardless of the alarm. Then, the low-type sender will always lie because the benefit of lying Δ_L^S is larger than the lying cost C. In this scenario, the prior belief equals the intermediate belief because all senders send the same message. But then the receiver's posterior belief after observing an alarm, which is always lower than the intermediate belief, will be lower than the prior belief $\rho < \hat{\rho}$. This is a contradiction. Suppose, instead, that the posterior belief after observing an alarm is lower than $\hat{\rho}$. The receiver will never take action r_H after observing an alarm. A low-type sender pretending to be a high type will not be detected with probability $1 - \beta$. Even if the receiver always takes the sender's desired action upon receiving no alarm, the expected benefit of lying is $(1 - \beta)\Delta_L^S$, which is smaller than the lying cost C when β is high. So, no low-type sender will lie. However, in that case, an alarm can only be a false-positive alarm, and the receiver's posterior belief after seeing message m_H will be one regardless of the alarm signal. This is also a contradiction. In sum, the receiver's posterior belief after observing an alarm is exactly $\hat{\rho}$. This implies that the receiver adopts a mixed strategy after observing message m_H and an alarm.

For a given sender's strategy, as the detector becomes stronger, the receiver's posterior belief after observing an alarm will be lower due to a larger dissuasive effect. In order to maintain a posterior belief of $\hat{\rho}$, the low-type sender needs to lie less frequently, which raises the base rate of a high-type sender conditional on sending message m_H and increases the posterior belief. Therefore, the equilibrium probability of lying decreases as the detector becomes stronger. Figure 7 illustrates the mechanism. This decreasing pattern of

the probability of lying is absent in the benchmark (with no false-positive alarms) because there is no variation in the dissuasive effect when $\alpha = 0$: a stronger detector with a higher β does not affect the receiver's inference about the sender's type conditional on seeing an alarm.

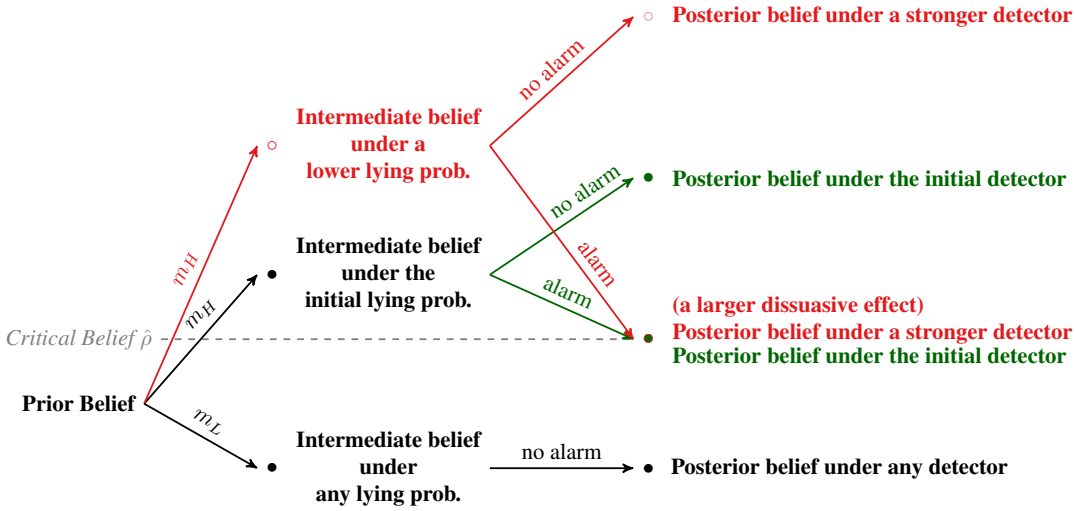

Figure 7: How a stronger lie detector decreases the probability of lying when β is high.

In sum, the persuasive and dissuasive effects jointly drive the non-monotonic relationship between the detector's capacity and the sender's probability of lying. The persuasive effect leads to an increasing pattern when the detector is weak, whereas the dissuasive effect generates a decreasing pattern when the detector is strong.

4.1.3 Effect of Lie Detection on the Payoffs

We need a well-defined equilibrium payoff to study the effect of lie detection on welfare. According to Proposition 1, the equilibrium is unique as long as $\alpha \neq 1 + [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta), \beta \neq \hat{\beta}$, and $\alpha \neq \beta$. For those cases with multiple equilibria, we select the Pareto-optimal equilibrium.[14] Section A.4 in the appendix contains details of the refinement. The next proposition summarizes the effect of lie detection on the expected payoff of the receiver, low-type sender, and high-type sender.

Proposition 2. *1. For a given true-positive rate β, the expected payoff of the receiver, the expected payoff of the low-type sender, and the expected payoff of the high-type sender are all weakly decreasing in the false-positive rate α.*

[14]The refinement does not drive the results on payoffs or welfare because the area in the detector's capacity space $\{(\alpha,\beta)|0 < \alpha \leq \beta \leq 1\}$ with multiple equilibria, $\{(\alpha,\beta)|\alpha = 1 + [(1-\rho)\Delta_L^R/(\rho\Delta_H^R)](1-\beta)$ *or* $\beta = \hat{\beta}$ *or* $\alpha = \beta\}$, has measure zero.

2. *For a given false-positive rate α, the expected payoff of the receiver and the expected payoff of the high-type sender are weakly increasing in the true-positive rate β, and the expected payoff of the low-type sender is weakly decreasing in the true-positive rate β.*

The receiver benefits from a more informed decision. She wants to better match the action with the sender's true type. A stronger detector (higher true-positive rate and lower false-positive rate) helps separate high-type and low-type senders. So, the receiver payoff increases in the true-positive rate and decreases in the false-positive rate. Similarly, a high-type sender wants to distinguish himself from a low-type sender and, therefore, benefits from a higher true-positive rate and a lower false-positive rate.

The effect of lie detection on a low-type sender's payoff is more complicated. A low-type sender benefits from successfully pretending to be a high-type sender. A higher true-positive rate raises the likelihood that the low-type sender will be caught by the detector. So, a low-type sender's payoff decreases in the true-positive rate. Interestingly, a low-type sender's payoff also decreases in the false-positive rate, though a higher false-positive rate makes it harder to distinguish between the two types. The reason is as follows. A low-type sender obtains a positive payoff only if there is no alarm and the receiver takes action r_H in the absence of an alarm. A higher false-positive rate does not reduce the low-type sender's likelihood of being detected. However, it leads to a smaller persuasive effect when there is no alarm (please refer to Figure 4 and the discussion in Section 4.1.1). The receiver's posterior belief after observing no alarm decreases in α, and the receiver uses a mixed strategy rather than always taking action r_H when the belief hits $\hat{\rho}$; this hurts the low-type sender's payoff.

As one can see from the proposition, a lower false-positive rate makes all players better off. This property plays an important role when we study the optimal design of the detector.

4.2 Entire equilibrium

We now study the entire equilibrium where the detector is endogenously determined. Proposition 1 in section 4.1 has characterized the sender's and the receiver's strategies for any given detector. So, we only need to determine the designer's strategy.

4.2.1 Optimal False-positive Rate and Alarm Rule Given True-positive Rate

Due to the constraint of the classifier's capacity, the designer cannot obtain all detectors $(\beta, \alpha) \in \{(\beta, \alpha)|0 \leq \alpha < \beta \leq 1\}$ by choosing an alarm rule $\{\lambda_H, \lambda_L\}$. In particular, the space of the feasible detectors given a classifier ϕ is $\{(\beta, \alpha)|\beta = \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H, \alpha = \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H, \lambda_L \in [0,1], \lambda_H \in [0,1]\}$. According to Proposition 2, a lower false-positive rate increases the receiver's and both senders' expected payoffs for a given true-positive rate. Thus, the designer always chooses the *lowest feasible false-positive rate* for any true-positive rate.

Lemma 5 (Optimal False-positive Rate and Alarm Rule Given True-positive Rate)**.** *For a given true-positive rate β, the detector's optimal false-positive rate, denoted by $\alpha^*(\beta; \phi)$, is*

$$\alpha^*(\beta;\phi) = \begin{cases} \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta, & if\ \beta \leq \phi(s_L|\theta = L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\beta + 1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta = L), \end{cases}$$

which increases in β. The detector $\{\beta, \alpha^*(\beta; \phi)\}$ can be achieved by the alarm rule

$$\lambda_L^*(\beta) = \begin{cases} \frac{\beta}{\phi(s_L|\theta=L)}, & if\ \beta \leq \phi(s_L|\theta = L) \\ 1, & if\ \beta > \phi(s_L|\theta = L) \end{cases}, \ \lambda_H^*(\beta) = \begin{cases} 0, & if\ \beta \leq \phi(s_L|\theta = L) \\ \frac{\beta - \phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, & if\ \beta > \phi(s_L|\theta = L). \end{cases}$$

Proof. See A.6 □

When choosing the alarm rule, the designer wants to achieve a given true-positive rate while minimizing the false-positive rate. Since the sender is more likely to be a low type under signal s_L than under signal s_H, the detector sends fewer false alarms, conditional on sending the same amount of true alarms, by sending alarms after getting prediction s_L rather than s_H from the classifier. As a result, the designer prefers sending an alarm after getting prediction s_L. To achieve a low true-positive rate, the detector does not need to send any alarms after getting prediction s_H. So, $\lambda_H^*(\beta) = 0$ and $\lambda_L^*(\beta)$ increases in β for low β, as illustrated by Figure 8. Since each alarm falsely recognizes a high-type sender as a low-type with some probability, the false-positive rate also increases in β.

The true-positive rate is capped by $\phi(s_L|\theta = L)$ even if the detector always sends an alarm after getting prediction s_L. So, the detector must also sometimes send an alarm after getting prediction s_H to achieve a true-positive rate above $\phi(s_L|\theta = L)$. In such cases, $\lambda_L^*(\beta) = 1$ and $\lambda_H^*(\beta)$ increases in β. In addition,

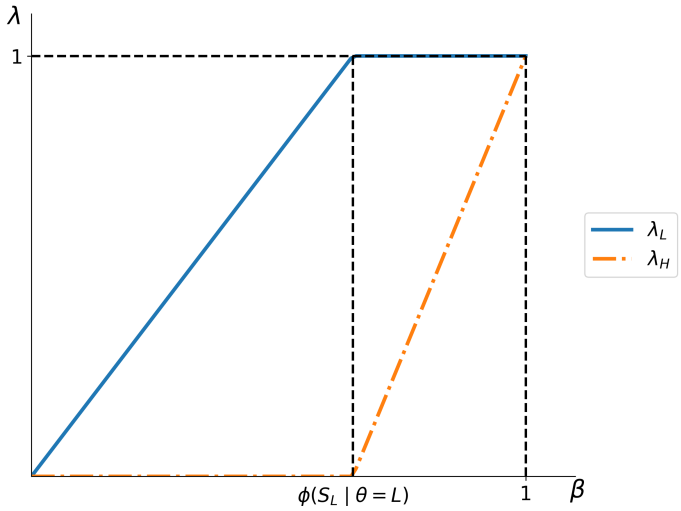

Figure 8: Optimal alarm rule for a given true-positive rate

the likelihood of the sender being a low type is smaller given prediction s_H rather than s_L. Compared to the low β case, the detector needs to raise the alarm probability by a larger value to add a unit to the true-positive rate. Consequently, $\lambda_H^*(\beta)$ and $\alpha^*(\beta;\phi)$ increase in β in this case at a higher rate than do $\lambda_L^*(\beta)$ and $\alpha^*(\beta;\phi)$ in the low β case.

Figure 9 demonstrates the classification performance of a given classifier under different designs of the detector. The receiver operating characteristic curve (ROC curve) represents the Pareto frontier of the classification outcome. The optimal detector must be a point on the ROC curve. As we can see, the false-positive rate increases in the true-positive rate. The designer needs to sacrifice one metric in order to improve the other. Furthermore, the false-positive rate increases in the true-positive rate at a lower rate when β is low and at a higher rate when β is high.

4.2.2 Optimal Design of Lie Detector

We now study the detector designer's equilibrium strategy, which is the optimal choice of a feasible detector, $\{\beta^*, \alpha^*\}$. We have shown in Lemma 5 that the optimal false-positive rate is $\alpha^*(\beta;\phi)$ given any true-positive rate β. Therefore, we only need to pin down the optimal true-positive rate β^*.

The classifier is more likely to generate outcome s_H if the sender's type is H rather than L, $\phi(s_H|\theta = H) > \phi(s_H|\theta = L)$, and more likely to generate outcome s_L if the sender's type is L rather than H, $\phi(s_L|\theta = L) > \phi(s_L|\theta = H)$. It is more informative if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L)$ and $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$ are higher. This motivates the following definition, which is useful in the subsequent

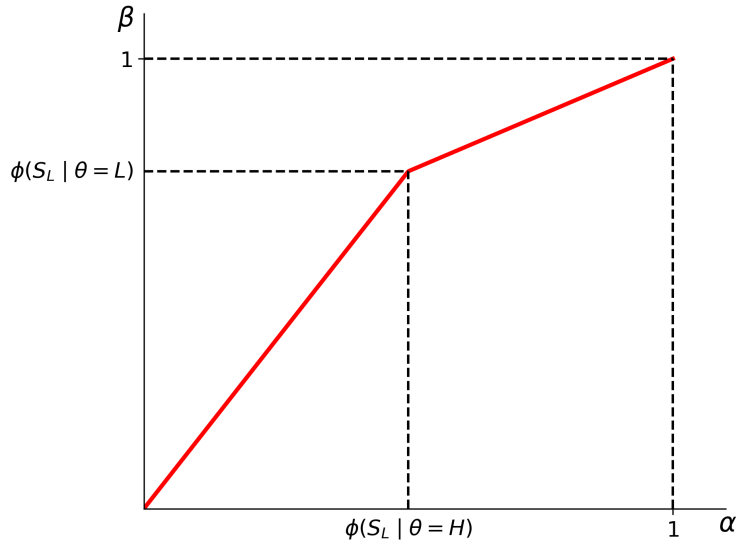

Figure 9: Receiver operating characteristic (ROC) curves

analyses.

Definition 2. A classifier has a high capacity if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq -(1-\rho)\Delta_L^R/(\rho\Delta_H^R)$ and $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq (\Delta_L^S - C)\rho\Delta_H^R/[\Delta_L^S\rho\Delta_H^R + (1-\rho)\Delta_L^R C]$.[15] Otherwise, it has a low capacity.

Maximizing Receiver's Payoff

Proposition 3. *The optimal true-positive rate of the detector that maximizes the receiver's expected payoff is any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$*, which minimizes the low-type sender's equilibrium probability of lying, if the lying cost is high,* $C \geq \hat{C}$*, and is* $\beta = \phi(s_L|\theta = L)$ *if the lying cost is low,* $C < \hat{C}$*.*[16]

Proof. See A.7. □

The receiver benefits from making a more informed decision. She wants to better distinguish between two types of senders. A low-type sender has a strong incentive to lie if the detector's true positive rate is low. The receiver will take action r_L and obtain zero payoff upon observing an alarm because the sender is highly likely to be type L. Even in the absence of an alarm, the combination of a high probability of lying and a low detection rate implies that the receiver still has much uncertainty about the sender's type (a weak

[15] We can set the thresholds to be any constants higher than one. The specific numbers are chosen because they are the cutoffs affecting the optimal detector design.

[16] The threshold $\hat{C} := \dfrac{\left[\phi(s_H|\theta=H) + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H|\theta=L)\right]\phi(s_H|\theta=L)}{\left[\phi(s_H|\theta=H) + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H|\theta=L) - 1\right]\phi(s_H|\theta=L) + \phi(s_H|\theta=H)}\Delta_L^S$

detector generates a small persuasive effect). Therefore, the receiver either follows a mixed strategy and obtains zero payoff or takes action r_H but earns a low payoff due to the high chance of taking the wrong action when the sender's true type is L. So, a low true-positive rate is not optimal for the receiver.

Upon observing an alarm, the receiver will be fairly confident that the sender is a low type if the detector's true positive rate is high due to the large dissuasive effect of a strong detector. In the meantime, the receiver will have a high posterior belief about the sender's type and will always take action r_H after receiving no alarm (a strong detector generates a large persuasive effect). In $\rho(1-\alpha)$ amount of time, the sender is a high type, and the receiver earns a positive payoff of Δ_H^R. In $(1-\rho)\sigma^S(1-\beta)$ amount of time, the sender is a low type, and the receiver earns a negative payoff of Δ_L^R. As the true-positive rate β in increases, the false-positive rate α also increases. So, both the benefit and the cost of action r_H are reduced. The receiver needs to make a trade-off. According to Lemma 5 and Figure 8, the detector's false-positive rate increases faster in its true-positive rate when the true-positive rate is high, $\beta > \phi(s_L \mid \theta = L)$. In such cases, the benefit of action r_H decreases at a high rate as β increases. So, a high true-positive rate is also not optimal for the receiver.

In equilibrium, the optimal true-positive rate of the detector either minimizes the low-type sender's probability of lying or takes full advantage of the region where a unit increase in the true-positive rate corresponds to a small increase in the false-positive rate. The intuition is that the receiver benefits from a low percentage of disinformation and a good detection technology. The low-type sender's equilibrium probability of lying has a discrete downward jump at $\hat{\beta}$ and is minimized at $[\hat{\beta}, \max\{\hat{\beta}, \phi(s_L \mid \theta = L)\}]$. When the lying cost is high, $\hat{\beta}$ is low. A reasonable true-positive rate suffices to reduce the low-type sender's probability of lying by a lot. The designer does not generate lots of false positive alarms by choosing such a true-positive rate. A low equilibrium percentage of disinformation helps the receiver take the right action and maximizes her expected payoff. When the lying cost is low, $\hat{\beta}$ is high. In order to induce a low probability of lying, the designer must choose a high true-positive rate, which comes with a high false-positive rate. Despite a low equilibrium percentage of disinformation, the receiver will suffer from a false-positive alarm and thus take a wrong action with a high likelihood. In such cases, a detector that generates a lower proportion of false-positive alarms relative to true-positive alarms maximizes the receiver's expected payoff.

Maximizing High-type Sender's Payoff

Proposition 4. *The optimal true-positive rate of the detector that maximizes the high-type sender's expected payoff is* $\beta_1 := [(1-\rho)\Delta_L^R + \rho\Delta_H^R]/[(1-\rho)\Delta_L^R + \rho\Delta_H^R\phi(s_L|\theta = H)/\phi(s_L|\theta = L)]$*, which is lower than* $\hat{\beta}$ *and decreases in* $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$*, if the classifier has a high capacity, and is any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ *if the classifier has a low capacity.*

Proof. See A.8. □

A high-type sender wants the receiver to take action r_H as frequently as possible. In equilibrium, the receiver may take action r_H or may adopt a mixed strategy if she observes message m_H and no alarm. Clearly, the sender prefers the receiver to always take action r_H to increase his payoff. The lowest true-positive rate that induces such behavior when the classifier has a high capacity, β_1, is lower than the lowest true-positive rate that induces such behavior when the classifier has a low capacity, $\hat{\beta}$. This is because the detector with the same true-positive rate has a lower false-positive rate and is thus a stronger detector under a higher-capacity classifier. A stronger detector leads to a larger persuasive effect and makes it easier to induce the receiver to take action r_H.

The high-type sender does not want to further increase β once it is high enough such that the receiver will always take action r_H after observing message m_H and no alarm. A higher true-positive rate corresponds to a higher false-positive rate. As a result, the detector is more likely to send a false alarm when β is higher. Because the receiver takes action r_L with a positive probability when there is an alarm, the more frequent false alarm hurts the sender's payoff.

The classifier is better at distinguishing two types of senders if $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$ is higher. When the classifier has a high capacity, we find that, *counter-intuitively*, the optimal true-positive rate is decreasing in the classifier's capacity - the optimal detector alarms a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type. The mechanism is the following. The high-type sender wants to choose the lowest true-positive rate that induces the receiver to always take the sender's desired action r_H upon seeing message m_H and no alarm. Fixing a true-positive rate, the detector has a lower false-positive rate if the classifier has a higher capacity. So, the receiver's posterior belief after observing message m_H and no alarm is higher. The detector can induce the receiver to take action r_H even if its true-positive rate is adjusted downward. This way, the sender can still obtain the highest payoff without a false-positive alarm and is less likely to be the object of a false alarm.

Maximizing Social Welfare

The designer, such as a platform, may care about both the sender and the receiver. For a given detector $(\beta, \alpha^*(\beta;\phi))$, denote the receiver's expected equilibrium payoff by $\mathbb{E}U^R(\beta)$, the high-type sender's expected equilibrium payoff by $\mathbb{E}U_H^S(\beta)$, and the low-type sender's expected equilibrium payoff by $\mathbb{E}U_L^S(\beta)$. The social welfare is $\mathbb{E}W(\beta) = \mathbb{E}U^R(\beta) + \rho\mathbb{E}U_H^S(\beta) + (1-\rho)\mathbb{E}U_L^S(\beta)$. The designer's objective in this case is to choose β to maximize $\mathbb{E}W(\beta)$.

Proposition 5. *If the classifier has a low capacity, the optimal true-positive rate of the detector is any* $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$. *If the classifier has a high capacity, the optimal true-positive rate of the detector must fall in* $\left[\beta_1, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}\right]$ *if the lying cost is high,* $C \geq \hat{C}$*, and must fall in* $[\beta_1, \phi(s_L|\theta = L)]$ *if the lying cost is low,* $C < \hat{C}$.

Proof. See A.9. □

If the designer has access to a classifier with a low capacity, any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ maximizes the receiver's and both senders' expected payoff.[17] So, it also maximizes social welfare. If the designer has access to a classifier with a high capacity, the receiver and the sender's preferences towards the detector are not aligned. We have shown that the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L \mid \theta = L)\}]$ if the lying cost is high and is $\phi(s_L|\theta = L)$ if the lying cost is low. In contrast, the sender prefers a lower true-positive rate: the unique optimal true-positive rate for both types of sender is β_1. When the designer's objective is maximizing social welfare, the optimal detector is a compromise between the receiver's and the sender's preferences.

Comparison With the No False-positive Alarm Benchmark

Compared to the no false-positive benchmark in section 3.2, the consideration of false positives leads to qualitatively different results of detector design regardless of the designer's objective. When there are no false-positive alarms, there is no trade-off in the detector design, and the detector designer always prefers a higher true-positive rate. In contrast, we have shown that the designer strictly prefers an intermediate true-positive rate to the highest true-positive rate in the presence of false-negative alarms. A higher true-positive rate may reduce the receiver's expected payoff, the high-type sender's expected payoff, and social welfare when we take into account the possibility of false-positive alarms and the players' strategic responses.

[17] We characterize the optimal true-positive rate for the low-type sender in A.8

5 Discussion and Concluding Remarks

Disinformation detection is becoming increasingly important and relevant because it is easier than ever to create and disseminate disinformation. To study the strategic interaction between disinformation generation and detection, this paper considers a game-theoretic model where a sender strategically communicates his type to a receiver, and a lie detector generates a noisy signal on the truthfulness of the sender's message. The receiver then infers the sender's type through messages from the sender and the detector.

Due to practical limitations, the detector may make two types of mistakes. It may fail to send an alarm when the sender is lying (false negative). It may also send a false alarm when the sender is truthful (false positive). Previous work has focused on the first type of mistake by implicitly assuming that the false-positive rate is zero. In reality, the sender cannot avoid making the second type of mistake unless he never sends an alarm. A key contribution of this paper is to explicitly consider the practical constraints of classification technology by allowing for both types of mistakes in disinformation detection. The other main contribution of this paper is to endogenize the design of the detector rather than treating the detection technology as exogenously given.

We first study how the detection technology affects the equilibrium outcomes. We find a non-monotonic relationship between the sender's probability of lying and the detection accuracy. A stronger detector increases the sender's probability of lying when the true-positive rate is low, because of a persuasive effect, whereas a stronger detector decreases the sender's probability of lying when the true-positive rate is high, due to a dissuasive effect.

We then characterize the optimal detector design that maximizes the receiver's payoff, the high-type sender's payoff, or social welfare. The receiver and both types of sender all benefit from a lower false-positive rate, whereas the low-type sender is hurt by a higher true-positive rate. Therefore, the designer always chooses the lowest feasible false-positive rate given any true-positive rate. The possibility of false-positive alarms implies that the designer will not choose the largest true-positive rate. Instead, the designer chooses different intermediate true-positive rates for different objectives. Counter-intuitively, the optimal detector may raise an alarm about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type.

Our results have important managerial implications. Regarding the descriptive value, we find qualitatively different insights about the relationship between the sender's probability of lying and the detector's

accuracy when we allow for false-positive alarms. Without false-positive alarms, an alarm eliminates all the uncertainty about the sender's type. The receiver's posterior belief goes all the way to zero after observing an alarm. Thus, two detectors with different true-positive rates generate the same *dissuasive effect*. In contrast, in the presence of false-positive alarms, two detectors with the same false-positive rate but different true-positive rates generate different dissuasive effects. Variations in the dissuasive effects lead to the non-monotonic relationship between the sender's probability of lying and the detector's accuracy.

Regarding the prescriptive value, we characterize the optimal design of the detector in the presence of practical limitations. Importantly, the possibility of false-positive alarms implies that the designer should not choose the largest true-positive rate. Instead, the designer should choose different intermediate true-positive rates given different objectives. The optimal detector may even raise alarms about a smaller percentage of disinformation when its underlying classifier is better at distinguishing the sender's type. The qualitatively different and counter-intuitive findings highlight the importance of considering the interaction between senders' strategic behavior and both types of mistakes by the detection technology in practice.

There are some interesting areas for future research. In this paper, the sender cannot affect the detector's ability. Future research can consider the possibility that a sender can make an effort (and incur a cost) to affect the detector's ability. It also would be interesting to extend the sender's type from binary to a continuous type, which may generate additional insights. Lastly, we study the optimal detector design for a given classifier. This setup reflects the fact that it is much harder to change the capacity of the classifier than to change how to use an endowed classifier to detect disinformation and send alarms because it takes lots of time, money, and data to train the classifier. Nevertheless, it may be interesting to also endogenize the capacity of the classifier when it is feasible to change the classifier in some applications.

Appendix A Proof

A.1 Proof of Lemma 1

We first prove an intuitive result that the receiver is less likely to take the sender's desired action r_H when there is an alarm than when there is no alarm.

Lemma 6. *Given $\beta > \alpha$, $\sigma^R(r_H \mid m_H, a) \leq \sigma^R(r_H \mid m_H, na)$. More specifically, $\sigma^R(r_H \mid m_H, na) \in [0,1) \Rightarrow \sigma^R(r_H \mid m_H, a) = 0$ and $\sigma^R(r_H \mid m_H, a) \in (0,1) \Rightarrow \sigma^R(r_H \mid m_H, na) = 1$.*

Proof. We first show that $m = m_H$ is an on-path message in equilibrium. Suppose that $\sigma^S(m_H \mid H) = \sigma^S(m_H \mid L) = 0$ in a PBE. The corresponding receiver's strategy after receiving $m = m_L$ is $\sigma^R(r_L \mid m_L, na) = 1$ because $(1-\rho)\Delta_L^R + \rho\Delta_H^R < 0$. So, deviating to $\sigma^S(m_H \mid H) > 0$ is always profitable for the sender with type $\theta = H$, which gives a contradiction. Hence, there is no PBE such that $\sigma^S(m_H \mid H) = \sigma^S(m_H \mid L) = 0$, and thereby $m = m_H$ must be an on-path message.

By the definition of PBE, the on-path belief satisfies

$$b(H \mid m_H, a) = \frac{\alpha\sigma^S(m_H \mid H)\rho}{\alpha\sigma^S(m_H \mid H)\rho + \beta\sigma^S(m_H \mid L)(1-\rho)}$$
$$b(H \mid m_H, na) = \frac{(1-\alpha)\sigma^S(m_H \mid H)\rho}{(1-\alpha)\sigma^S(m_H \mid H)\rho + (1-\beta)\sigma^S(m_H \mid L)(1-\rho)}$$

One can see that $b(H \mid m_H, na) > b(H \mid m_H, a)$ given $\beta > \alpha$.

Based on these beliefs, the best responses of the receiver should follow

$$\sigma^R(r_H \mid m_H, a) \begin{cases} = 1 & b(H \mid m_H, a)\Delta_H^R + (1 - b(H \mid m_H, a))\Delta_L^R > 0 \\ \in [0,1] & b(H \mid m_H, a)\Delta_H^R + (1 - b(H \mid m_H, a))\Delta_L^R = 0 \\ = 0 & b(H \mid m_H, a)\Delta_H^R + (1 - b(H \mid m_H, a))\Delta_L^R < 0 \end{cases}$$
$$\sigma^R(r_H \mid m_H, na) \begin{cases} = 1 & b(H \mid m_H, na)\Delta_H^R + (1 - b(H \mid m_H, na))\Delta_L^R > 0 \\ \in [0,1] & b(H \mid m_H, na)\Delta_H^R + (1 - b(H \mid m_H, na))\Delta_L^R = 0 \\ = 0 & b(H \mid m_H, na)\Delta_H^R + (1 - b(H \mid m_H, na))\Delta_L^R < 0 \end{cases}$$

Since $b(H \mid m_H, na) > b(H \mid m_H, a)$, $b(H \mid m_H, a)\Delta_H^R + (1 - b(H \mid m_H, a))\Delta_L^R < b(H \mid m_H, na)\Delta_H^R + (1 - b(H \mid m_H, na))\Delta_L^R$. Hence, $\sigma^R(r_H \mid m_H, a) \leq \sigma^R(r_H \mid m_H, na)$. □

We now prove that $\sigma^S(m_H \mid H) = 1$. Suppose that there is an equilibrium where $\sigma^S(m_L \mid H) > 0$.

Then,

$$\sigma^R(r_H \mid m_L, na)\Delta_H^S - C \geq \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_H^S$$
$$\Leftrightarrow -C \geq \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na) - \sigma^R(r_H \mid m_L, na)\right)\Delta_H^S$$
$$\Rightarrow \sigma^R(r_H \mid m_L, na) > \alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na) \tag{1}$$

We now show that a type L sender obtains a lower expected payoff from sending message $m = m_H$ rather than $m = m_L$:

1. If $\alpha = \beta$,
$$\left(\beta\sigma^R(r_H \mid m_H, a) + (1-\beta)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C$$
$$= \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C$$
$$\overset{(1)}{<} \sigma^R(r_H \mid m_L, na)\Delta_L^S - C < \sigma^R(r_H \mid m_L, na)\Delta_L^S$$

2. If $\alpha < \beta$,
$$\left(\beta\sigma^R(r_H \mid m_H, a) + (1-\beta)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C$$
$$\overset{Lemma\ 6}{\leq} \left(\alpha\sigma^R(r_H \mid m_H, a) + (1-\alpha)\sigma^R(r_H \mid m_H, na)\right)\Delta_L^S - C$$
$$\overset{(1)}{<} \sigma^R(r_H \mid m_L, na)\Delta_L^S - C < \sigma^R(r_H \mid m_L, na)\Delta_L^S$$

Hence, the sender with $\theta = L$ always sends the message $m = m_L$, $\sigma^S(m_L \mid L) = 1$.

Since $\sigma^S(m_L \mid L) = 1$, the receiver's expected payoff given $(m = m_L, l = na)$ from taking $r = r_L$ is always higher than it from taking $r = r_H$, $(1-\rho)\Delta_L^R + \sigma^S(m_L \mid H)\rho\Delta_H^R \leq (1-\rho)\Delta_L^R + \rho\Delta_H^R < 0$. Hence, $\sigma^R(r_H \mid m_L, na) = 0$, which contradicts to (1). Hence, $\sigma^S(m_H \mid H) = 1$.

An immediate implication is that the receiver always takes action r_L after receiving message m_L, if $m = m_L$ is an on-path message.

A.2 Proof of Lemma 2

Since $\alpha = \beta = 0$, the receiver may only observe either $\{m = m_H, l = na\}$ or $\{m = m_L, l = na\}$. So, σ_{na}^R completely characterizes the receiver's strategy. The best responses of the receiver and the sender to the

opponent's strategy are

$$\sigma^R_{na,\text{BR}}(\sigma^S)\begin{cases}=1, & \sigma^S<-\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}\\ \in[0,1], & \sigma^S=-\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}\\ =0, & \sigma^S>-\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}\end{cases};\ \sigma^S_{\text{BR}}(\sigma^R_{na})\begin{cases}=1, & \sigma^R_{na}>\frac{C}{\Delta^S_L}\\ \in[0,1], & \sigma^R_{na}=\frac{C}{\Delta^S_L}\\ =0, & \sigma^R_{na}<\frac{C}{\Delta^S_L}\end{cases}$$

A PBE satisfies $\sigma^{S*}=\sigma^S_{\text{BR}}({\sigma^R_{na}}^*)$ and ${\sigma^R_{na}}^*=\sigma^R_{na,\text{BR}}(\sigma^{S*})$. Hence, there exists a unique equilibrium,

$$\sigma^{S*}=-\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L},\ {\sigma^R_{na}}^*=\frac{C}{\Delta^S_L}$$

In this equilibrium, the utility of the receiver is $\mathbb{E}U^R=0$, the utility of the type L sender is $\mathbb{E}U^S_L=0$, and the utility of the type H sender is $\mathbb{E}U^S_H=\frac{C}{\Delta^S_L}\Delta^S_H$.

A.3 Proof of Lemma 3 and Proposition 1

For completeness, we also include the case of $0<\alpha=\beta$ in the proof. We first consider the equilibria where a low-type sender always lies.

A.3.1 Equilibria where $\sigma^S=1$

If the low-type sender always sends message $m=m_H$ (i.e., $\sigma^S=1$), $m=m_L$ is an off-path message. The receiver's belief and his action after receiving $m=m_L$ can be arbitrary in a PBE as long as the sender does not have a profitable deviation by sending $m=m_L$.

Lemma 7. *There exists a PBE with $\sigma^S=1$ if and only if $\alpha\le 1+\frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$ and $\beta\le\hat{\beta}:=1-\frac{C}{\Delta^S_L}$. Specifically, the set of equilibria is*

$$\left\{(\sigma^R_a=0,\sigma^R_{na}=1,\sigma^R_{L,na},\sigma^S=1):\sigma^R_{L,na}\le 1-\beta-\frac{C}{\Delta^S_L}\right\}$$

if $\alpha<1+\frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$ and $\beta\le\hat{\beta}$, and the set of equilibria is

$$\left\{(\sigma^R_a=0,\sigma^R_{na},\sigma^R_{L,na},\sigma^S=1):\sigma^R_{L,na}\le(1-\beta)\sigma^R_{na}-\frac{C}{\Delta^S_L},\sigma^R_{na}\in\left[\frac{C}{(1-\beta)\Delta^S_L},1\right]\right\}$$

if $\alpha=1+\frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$ and $\beta\le\hat{\beta}$.

Proof. If $\alpha = \beta$ and the low-type sender always lies, then the receiver always takes action $r = r_L$ since she does not get new information about the sender's type other than the prior. But then, the low-type sender can be better off by not lying. There is no such PBE. So, we only need to consider $\beta > \alpha$.

In a PBE where m_L is an off-path message, both types of sender send m_H. Avoiding profitable deviation of both sender types requires $\sigma^R_{L,na} \leq \beta\sigma^R_a + (1-\beta)\sigma^R_{na} - \frac{C}{\Delta^S_L}$ and $\sigma^R_{L,na} - \frac{C}{\Delta^S_H} \leq \alpha\sigma^R_a + (1-\alpha)\sigma^R_{na}$. By Lemma 6, $\beta\sigma^R_a + (1-\beta)\sigma^R_{na} \leq \alpha\sigma^R_a + (1-\alpha)\sigma^R_{na}$, so we only requires $\sigma^R_{L,na} \leq \beta\sigma^R_a + (1-\beta)\sigma^R_{na} - \frac{C}{\Delta^S_L}$.

With $\sigma^S = 1$, the receiver's on-path beliefs are $b(H \mid m_H, a) = \frac{\alpha\rho}{\alpha\rho+\beta(1-\rho)} \leq \rho$ and $b(H \mid m_H, na) = \frac{(1-\alpha)\rho}{(1-\alpha)\rho+(1-\beta)(1-\rho)}$. We have $\sigma^R_a = 0$ since $b(H \mid m_H, a)\Delta^R_H + (1 - b(H \mid m_H, a))\Delta^R_L < 0$. Then, the set of equilibria can be represented as

$$\left\{(\sigma^R_a = 0, \sigma^R_{na}, \sigma^R_{L,na}, \sigma^S = 1) : \sigma^R_{L,na} \leq (1-\beta)\sigma^R_{na} - \frac{C}{\Delta^S_L}, \sigma^R_{na} \in \left[\frac{C}{(1-\beta)\Delta^S_L}, 1\right]\right\}$$

There are two potential cases:

1. $\frac{(1-\alpha)\rho}{(1-\alpha)\rho+(1-\beta)(1-\rho)}\Delta^R_H + \frac{(1-\beta)(1-\rho)}{(1-\alpha)\rho+(1-\beta)(1-\rho)}\Delta^R_L > 0$ (i.e., $\alpha < 1 + \frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$): $\sigma^R_{na} = 1$. In this case, to make the set above nonempty, we need $\beta \leq \hat{\beta}$.

2. $\frac{(1-\alpha)\rho}{(1-\alpha)\rho+(1-\beta)(1-\rho)}\Delta^R_H + \frac{(1-\beta)(1-\rho)}{(1-\alpha)\rho+(1-\beta)(1-\rho)}\Delta^R_L = 0$ (i.e., $\alpha = 1 + \frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$): $\sigma^R_{na} \in [0, 1]$. In this case, to make the set above nonempty, we also need $\beta \leq \hat{\beta}$.

All in all, there exists a PBE with $\sigma^S = 1$ and $\sigma^R_{L,na} \in [0, 1]$ if and only if $\alpha \leq 1 + \frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H}(1-\beta)$ and $\beta \leq \hat{\beta}$. □

We then consider the equilibria where a low-type sender does not always lie.

A.3.2 Equilibria where $\sigma^S < 1$

In a PBE where $m = m_L$ is an on-path message (i.e., $\sigma^S < 1$), the definition of PBE requires the receiver to have a consistent belief, $b(L \mid m_L, na) = 1$, which means that anyone who sends $m = m_L$ must be type $\theta = L$. Hence, the receiver always takes action $r = r_L$ after receiving $m = m_L$ (i.e., $\sigma^R_{L,na} = 0$).

The sender with $\theta = L$ has expected payoff $\mathbb{E}U^S_0(\sigma^S; \{\sigma^R_{na}, \sigma^R_a\}) = \sigma^S[\left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L - C]$

by playing strategy σ^S, whose best response to $\{\sigma^R_{na}, \sigma^R_a\}$ is

$$\sigma^S_{\text{BR}}(\{\sigma^R_{na}, \sigma^R_a\}) \begin{cases} = 1, & \left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L > C \\ \in [0,1], & \left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L = C \\ = 0, & \left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L < C \end{cases} \tag{BR1}$$

The best response of R with consistent belief to σ^S is given by

$$\sigma^R_{a,\text{BR}}(\sigma^S) \begin{cases} = 1, & \sigma^S < -\frac{\alpha\rho\Delta^R_H}{\beta(1-\rho)\Delta^R_L} \\ \in [0,1], & \sigma^S = -\frac{\alpha\rho\Delta^R_H}{\beta(1-\rho)\Delta^R_L} \\ = 0, & \sigma^S > -\frac{\alpha\rho\Delta^R_H}{\beta(1-\rho)\Delta^R_L} \end{cases}; \; \sigma^R_{na,\text{BR}}(\sigma^S) \begin{cases} = 1, & \sigma^S < -\frac{(1-\alpha)\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L} \\ \in [0,1], & \sigma^S = -\frac{(1-\alpha)\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L} \\ = 0, & \sigma^S > -\frac{(1-\alpha)\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L} \end{cases} \tag{BR2}$$

We now summarize the equilibrium for a given detector.

Equilibrium when $\alpha = \beta$ Given $\alpha = \beta \in (0,1]$, the best response of the receiver can be written as

$$\begin{cases} \sigma^R_{a,\text{BR}}(\sigma^S) = \sigma^R_{na,\text{BR}}(\sigma^S) = 1, & \sigma^S < -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L} \\ \sigma^R_{a,\text{BR}}(\sigma^S) \in [0,1], \sigma^R_{na,\text{BR}}(\sigma^S) \in [0,1], & \sigma^S = -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L} \\ \sigma^R_{a,\text{BR}}(\sigma^S) = \sigma^R_{na,\text{BR}}(\sigma^S) = 0, & \sigma^S > -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L} \end{cases}$$

- Firstly, we consider the equilibrium with $\sigma^S < -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}$. Based on the best response of the receiver, we have $\sigma^R_a = \sigma^R_{na} = 1$ in equilibrium. Since $\sigma^S \in [0,1)$, we need $\Delta^S_L = \left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L \leq C$ to make the equilibrium holds (the inequality becomes equality if $\sigma^S > 0$). However, $C < \min\{\Delta^S_H, \Delta^S_L\}$, there is a contradiction. So, there is no such equilibrium.

- Secondly, we consider the equilibrium with $\sigma^S = -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}$. Based on the best response of the receiver and the sender, σ^R_a and σ^R_{na} in equilibrium satisfy $\left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L = C$. That is, there exists an equilibrium $\left\{\sigma^S = -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}, \sigma^R_{na}, \sigma^R_a\right\}$ with $\left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L = C$.

- Thirdly, we consider the equilibrium with $\sigma^S > -\frac{\rho\Delta^R_H}{(1-\rho)\Delta^R_L}$. Based on the best response of the receiver, we have $\sigma^R_a = \sigma^R_{na} = 0$ in equilibrium. Since $\sigma^S > 0$, we need $0 = \left(\beta\sigma^R_a + (1-\beta)\sigma^R_{na}\right)\Delta^S_L \geq C$ to make the equilibrium holds, which is impossible. So, there is no such equilibrium.

Equilibrium when $0 < \alpha < \beta$ (Proposition 1)

1. Separating equilibrium with $\sigma^S = 0$ implies $\sigma_{na}^R = \sigma_a^R = 1$, which requires $\Delta_L^S \leq C$. There is a contradiction to the definition of C, so there is no complete separating equilibrium.

2. Semi-pooling equilibrium with $\sigma^S \in (0,1)$ requires $\left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S = C$. By the Lemma 6 and (BR2), we can discuss potential equilibria by following cases:

   (a) $\sigma_{na}^R \in (0,1)$ and $\sigma_a^R = 0$: this case requires

   - $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S} \in (0,1)$ (i.e., $\beta < \hat{\beta}$) and
   - $\sigma^S = -\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R} < 1$ (i.e., $\alpha > 1 + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$).

   (b) $\sigma_{na}^R = 1$ and $\sigma_a^R = 0$: this case requires

   - $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S} = 1$ (i.e., $\beta = \hat{\beta}$) and
   - $\sigma^S \in \left[-\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \min\left\{-\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right]$.

   (c) $\sigma_{na}^R = 1$ and $\sigma_a^R \in (0,1)$: this case requires

   - $\sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta} \in (0,1)$ (i.e., $\beta > \hat{\beta}$) and
   - $\sigma^S = -\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}$

   Table 2 summarizes the PBEs when $\alpha > 0$.

Equilibrium when $0 = \alpha < \beta$ (Lemma 3)

1. Separating equilibrium with $\sigma^S = 0$ implies $\sigma_{na}^R = 1$ and requires $(\beta\sigma_a^R + 1 - \beta)\Delta_L^S \leq C$, i.e., $\sigma_a^R \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}$. $\frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta} \geq 0$ only when $\beta \geq \hat{\beta}$.

2. Semi-pooling equilibrium with $\sigma^S \in (0,1)$ requires $\left(\beta\sigma_a^R + (1-\beta)\sigma_{na}^R\right)\Delta_L^S = C$. By the Lemma 6 and (BR2), we can discuss potential equilibria by following cases:

   (a) $\sigma_{na}^R \in (0,1)$ and $\sigma_a^R = 0$: this case requires

   - $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S} \in (0,1)$ (i.e., $\beta < \hat{\beta}$) and
   - $\sigma^S = -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R} < 1$ (i.e., $\beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}$).

   (b) $\sigma_{na}^R = 1$ and $\sigma_a^R = 0$: this case requires

   - $\sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S} = 1$ (i.e., $\beta = \hat{\beta}$) and

- $\sigma^S \in \left[0, \min\left\{-\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, 1\right\}\right]$.

(c) $\sigma_{na}^R = 1$ and $\sigma_a^R \in (0,1)$: this case requires

- $\sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta} \in (0,1)$ (i.e., $\beta > \hat{\beta}$) and
- $\sigma^S = 0$

By Lemma 7, the PBEs with $\alpha = 0$ are the following:

1. if $C \geq -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$,

$$\begin{cases} \sigma^{S*} = -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, {\sigma_{na}^R}^* = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < \hat{\beta} := \hat{\beta} \\ \sigma^{S*} \in \left[0, -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}\right], {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

2. if $C < -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$,

$$\begin{cases} \sigma^{S*} = -\frac{\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, {\sigma_{na}^R}^* = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^{R*} = 0, & 0 < \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, {\sigma_{na}^R}^* = \left[\frac{C}{(1-\beta)\Delta_L^S}, 1\right], \sigma_a^{R*} = 0, & \beta = 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \sigma^{S*} = 1, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} < \beta < \hat{\beta} \\ \sigma^{S*} \in [0,1], {\sigma_{na}^R}^* = 1, \sigma_a^{R*} = 0, & \beta = \hat{\beta} \\ \sigma^{S*} = 0, {\sigma_{na}^R}^* = 1, \sigma_a^{R*} \leq \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta > \hat{\beta} \end{cases}$$

A.4 Refinement on Multiple Equilibria

When $\beta = \hat{\beta}$, $\{-\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, 1, 0\}$ Pareto dominates other equilibria. When $\beta \in \left(0, \hat{\beta}\right)$ and $\alpha = 1 + \frac{(1-\rho)\Delta_L^R(1-\beta)}{\rho\Delta_H^R}$, $\{1,1,0\}$ Pareto dominates other equilibria. Table 3 summarizes the equilibrium payoff under this refinement.

If $\beta \in \left[\hat{\beta}, 1\right)$, the equilibrium is $\sigma^S = -\frac{\alpha\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}$. Since it is an equilibrium with mixed strategies σ^S and σ_a^R, the low-type sender's expected payoff is 0 and the receiver's expected payoff given an alarm is 0. So, the receiver's expected payoff is $\rho(1-\alpha)\Delta_H^R + (1-\rho)(1-\beta)\sigma^S\Delta_L^R = \left(1 - \frac{\alpha}{\beta}\right)\rho\Delta_H^R$ and the high-type sender's expected payoff is $(1 - \alpha + \alpha\sigma_a^R)\Delta_H^S = \Delta_H^S - \frac{(\Delta_L^S - C)\Delta_H^S}{\Delta_L^S}\frac{\alpha}{\beta}$.

β Range \ α Range	$\alpha \leq 1+\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$	$\alpha \in \left(1+\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta), \beta\right]$
$\left[\hat{\beta}, 1\right)$	$\mathbb{E}U^R = \left(1-\frac{\alpha}{\beta}\right)\rho\Delta_H^R,\ \mathbb{E}U_L^S = 0,\ \mathbb{E}U_H^S = \Delta_H^S - \frac{(\Delta_L^S - C)\Delta_H^S}{\Delta_L^S}\frac{\alpha}{\beta}$	
$\left(0, \hat{\beta}\right)$	$\mathbb{E}U^R = (1-\beta)(1-\rho)\Delta_L^R + (1-\alpha)\rho\Delta_H^R,$ $\mathbb{E}U_L^S = (1-\beta)\Delta_L^S - C,$ $\mathbb{E}U_H^S = (1-\alpha)\Delta_H^S$	$\mathbb{E}U^R = 0,$ $\mathbb{E}U_L^S = 0,$ $\mathbb{E}U_H^S = \frac{C}{\Delta_L^S}\Delta_H^S\frac{1-\alpha}{1-\beta}$

Table 3: Equilibrium payoff under detector $\{\beta, \alpha\}$

If $\beta \in \left(0, \hat{\beta}\right)$ and $\alpha \leq 1+\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$, the equilibrium is $\sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0$. The receiver's expected payoff is $(1-\beta)(1-\rho)\Delta_L^R + (1-\alpha)\rho\Delta_H^R$, the low-type sender's expected payoff is $(1-\beta)\Delta_L^S - C$, and the high-type sender's expected payoff is $(1-\alpha)\Delta_H^S$.

If $\beta \in \left(0, \hat{\beta}\right)$ and $\alpha > 1+\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}(1-\beta)$, the equilibrium is $\sigma^S = -\frac{(1-\alpha)\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0$. Since it is an equilibrium with mixed strategies σ^S and σ_{na}^R, the low-type sender's expected payoff is 0 and the receiver's expected payoff given no alarm is 0. As the receiver also gets zero payoff given an alarm, the receiver's expected payoff is 0. And the high-type sender's expected payoff is $(1-\alpha)\sigma_{na}^R\Delta_H^S = \frac{C}{\Delta_L^S}\Delta_H^S\frac{1-\alpha}{1-\beta}$.

A.5 Proof of Lemma 4

When there are multiple equilibria, we select the Pareto-optimal equilibrium. The refinement does not drive the results because the area in the detector's capacity space $\{(0,\beta)|0 \leq \beta \leq 1\}$ with multiple equilibria, $\{(0,\beta)|\beta = \hat{\beta}\ or\ \beta = 1+\rho\Delta_H^R/[(1-\rho)\Delta_L^R]\ and\ C < -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S\}$, has measure zero.

1. High lying cost $C \geq -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$

$$\mathbb{E}U_L^S(\beta) = 0,\ \mathbb{E}U_H^S(\beta) = \begin{cases} \frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < \hat{\beta} \\ \Delta_H^S, & \beta \geq \hat{\beta} \end{cases},\ \mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta < \hat{\beta} \\ \rho\Delta_H^R, & \beta \geq \hat{\beta} \end{cases}$$

$$\mathbb{E}W(\beta) = \mathbb{E}U^R(\beta) + \rho\mathbb{E}U_H^S(\beta) + (1-\rho)\mathbb{E}U_L^S(\beta) = \begin{cases} \rho\frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < \hat{\beta} \\ \rho(\Delta_H^S + \Delta_H^R), & \beta \geq \hat{\beta} \end{cases}$$

One can see that $\mathbb{E}U^R(\beta),\ \mathbb{E}U_H^S(\beta)$, and $\mathbb{E}W(\beta)$ all (weakly) increase in β. They achieve the maximum value at any $\beta \geq \hat{\beta}$.

2. Low lying cost $C < -\frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}\Delta_L^S$

$$\mathbb{E}U_L^S(\beta) = \begin{cases} 0, & \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ (1-\beta)\Delta_L^S - C, & \beta \in \left[1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ 0, & \beta \geq \hat{\beta} \end{cases},$$

$$\mathbb{E}U_H^S(\beta) = \begin{cases} \frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ \Delta_H^S, & \beta \geq 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \end{cases},$$

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ (1-\beta)(1-\rho)\Delta_L^R + \rho\Delta_H^R, & \beta \in \left[1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ \rho\Delta_H^R, & \beta \geq \hat{\beta} \end{cases}$$

$$\mathbb{E}W(\beta) = \begin{cases} \rho\frac{C}{(1-\beta)\Delta_L^S}\Delta_H^S, & \beta < 1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R} \\ (1-\beta)(1-\rho)(\Delta_L^S + \Delta_L^R) + \rho(\Delta_H^S + \Delta_H^R) - (1-\rho)C, & \beta \in \left[1 + \frac{\rho\Delta_H^R}{(1-\rho)\Delta_L^R}, \hat{\beta}\right) \\ \rho(\Delta_H^S + \Delta_H^R), & \beta \geq \hat{\beta} \end{cases}$$

One can see that $\mathbb{E}U^R(\beta)$, $\mathbb{E}U_H^S(\beta)$, and $\mathbb{E}W(\beta)$ all (weakly) increase in β. $\mathbb{E}U^R(\beta)$ and $\mathbb{E}W(\beta)$ achieve the maximum value at any $\beta \geq \hat{\beta}$. $\mathbb{E}U_H^S(\beta)$ achieves the maximum value at any $\beta \geq 1 + \rho\Delta_H^R/[(1-\rho)\Delta_L^R]$.

A.6 Proof of Lemma 5

$$\alpha^*(\beta) = \min_{\lambda_L, \lambda_H \in [0,1]} \phi(s_L|\theta = H)\lambda_L + \phi(s_H|\theta = H)\lambda_H,$$

$$\text{s.t. } \phi(s_L|\theta = L)\lambda_L + \phi(s_H|\theta = L)\lambda_H = \beta$$

The constraint implies that

$$\lambda_L = \frac{\beta - \phi(s_H|\theta = L)\lambda_H}{\phi(s_L|\theta = L)} \tag{C1}$$

Substituting (C1) into the objective function, one can see that the coefficient of λ_H is positive:

$$\begin{aligned}
&\phi(s_H|\theta=H) - \frac{\phi(s_H|\theta=L)\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \\
=&1-\phi(s_L|\theta=H) - \frac{(1-\phi(s_L|\theta=L))\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \\
=&1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} > 0
\end{aligned}$$

Therefore, the optimal λ_H^* with $\lambda_L = \frac{\beta-\phi(s_H|\theta=L)\lambda_H}{\phi(s_L|\theta=L)}$ should be the minimum feasible λ_H. The λ_H has restrictions: $\frac{\beta-\phi(s_H|\theta=L)\lambda_H}{\phi(s_L|\theta=L)} \in [0,1]$ and $\lambda_H \in [0,1]$. So, the minimum feasible λ_H is $\max\left\{\frac{\beta-\phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, 0\right\}$.

All in all,

$$\lambda_H^* = \max\left\{\frac{\beta-\phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, 0\right\},\ \lambda_L^* = \frac{\beta}{\phi(s_L|\theta=L)} - \frac{\phi(s_H|\theta=L)}{\phi(s_L|\theta=L)}\lambda_1^*$$

If $\beta \le \phi(s_L|\theta=L)$, then we have

$$\lambda_H^* = 0, \lambda_L^* = \frac{\beta}{\phi(s_L|\theta=L)}, \alpha^*(\beta) = \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta$$

If $\beta > \phi(s_L|\theta=L)$, then we have

$$\lambda_H^* = \frac{\beta-\phi(s_L|\theta=L)}{\phi(s_H|\theta=L)}, \lambda_L^* = 1$$

$$\begin{aligned}
\alpha^*(\beta) &= \phi(s_L|\theta=H) + \phi(s_H|\theta=H)\frac{\beta-\phi(s_L|\theta=L)}{\phi(s_H|\theta=L)} \\
&= \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\beta + \left(1-\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right)
\end{aligned}$$

A.7 Proof of Proposition 3

We first characterize the sender's equilibrium strategy for a given detector $\{\beta, \alpha^*(\beta;\phi)\}$.

Lemma 8. *For a classifier with a high capacity, the equilibria are*

$$\begin{cases}
\sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\
\sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \\
\sigma^S = -\frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right)
\end{cases}$$

where $\beta_1 := \frac{(1-\rho)\Delta_L^R+\rho\Delta_H^R}{(1-\rho)\Delta_L^R+\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R} \leq \hat{\beta}$. *For a classifier with a low capacity, the equilibria are*

$$\begin{cases} \sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in \left[0, \hat{\beta}\right) \\ \sigma^S = -\frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Proof. By Proposition 1, the equilibria with $\beta \geq \hat{\beta}$ are directly given. Now, we focus on $\beta \in \left[0, \hat{\beta}\right)$, in which case the form of equilibrium is determined by the value of $\frac{1-\alpha^*(\beta;\phi)}{1-\beta}$. Specifically, the equilibrium is $\{\sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0\}$ if $\frac{1-\alpha^*(\beta;\phi)}{1-\beta} \geq -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ and is $\{\sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0\}$ if $\frac{1-\alpha^*(\beta;\phi)}{1-\beta} < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$.

Based on Lemma 5,

$$\begin{aligned} \frac{1-\alpha^*(\beta;\phi)}{1-\beta} &= \begin{cases} -\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}, & if\ \beta \leq \phi(s_L|\theta=L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} & if\ \beta > \phi(s_L|\theta=L), \end{cases} \\ &= \min\left\{-\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}, \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right\} \in \left[1, \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right] \end{aligned}$$

Let $g(\beta) := -\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}$, which is increasing in β and $g(\hat{\beta}) = -\frac{\Delta_L^S - C}{C}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{\Delta_L^S}{C}$.

- Given $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ or $g(\hat{\beta}) < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ ($\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} > \frac{\Delta_L^S}{\Delta_L^S - C} + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\frac{C}{\Delta_L^S - C}$), the $\frac{1-\alpha^*(\beta;\phi)}{1-\beta} < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ must be satisfied for all $\beta \in \left[0, \hat{\beta}\right)$. For all $\beta \in \left[0, \hat{\beta}\right)$, the equilibrium is $\{\sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0\}$.
- Given $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} \geq -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ and $g(\hat{\beta}) \geq -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ ($\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \leq \frac{\Delta_L^S}{\Delta_L^S - C} + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\frac{C}{\Delta_L^S - C}$), one can see that $\beta_1 \in \left(0, \hat{\beta}\right]$ satisfies $g(\beta_1) = -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$. Then, the equilibrium is

$$\begin{cases} \sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \end{cases}$$

□

According to Lemma 8 and Table 3, $\mathbb{E}U^R(\beta) = \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta_H^R$ is decreasing in β when $\beta \in \left[\hat{\beta}, 1\right)$.

- If the classifier has a low capacity, $\mathbb{E}U^R(\beta) = 0$ for $\beta \in \left[0, \hat{\beta}\right)$. For $\beta \geq \hat{\beta}$, $\mathbb{E}U^R(\beta) = \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta_H^R$, which is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ and is decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$.

- If the classifier has a high capacity, $\mathbb{E}U^R(\beta)$ can be written as

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\beta)(1-\rho)\Delta_L^R + (1-\alpha^*(\beta;\phi))\rho\Delta_H^R, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta_H^R, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Lemma 9. *If* $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} \geq -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$, *then* $\phi(s_L|\theta = L) \geq \beta_1$.

Proof. One can see that $\phi(s_L|\theta = L) = \frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}} \geq \frac{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1}{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}} = \beta_1$ because $\frac{x-1}{x - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}$ increases in x. □

Recall that

$$\alpha^*(\beta;\phi) = \begin{cases} \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta, & \beta \leq \phi(s_L|\theta = L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\beta + \left(1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right), & \beta > \phi(s_L|\theta = L) \end{cases}$$

The high capacity requires $\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} \leq \frac{\Delta_L^S}{\Delta_L^S - C} + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\left(\frac{\Delta_L^S}{\Delta_L^S - C} - 1\right)$, i.e., $\hat{\beta} \geq \frac{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1}{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}$.

Note that $\frac{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1}{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}} = \frac{1}{\frac{1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}{-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1} + 1} \leq \frac{1}{\frac{1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1} + 1} = \phi(s_L|\theta = L)$, the relationship between $\phi(s_L|\theta = L)$ and $\hat{\beta}$ is undetermined. So, we discuss the following two cases:

(a) If $\hat{\beta} > \phi(s_L|\theta = L)$,

$$\frac{\partial \mathbb{E}U^R(\beta)}{\partial \beta} = \begin{cases} -(1-\rho)\Delta_L^R - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R > 0, & \beta \in [\beta_1, \phi(s_L|\theta = L)) \\ -(1-\rho)\Delta_L^R - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\rho\Delta_H^R \leq 0, & \beta \in \left[\phi(s_L|\theta = L), \hat{\beta}\right) \end{cases}$$

So, $\mathbb{E}U^R(\beta)$ is maximized at $\phi(s_L|\theta = L)$ for $\beta \in [\beta_1, \hat{\beta})$. Since $\mathbb{E}U^R(\beta) = 0$ for all $\beta \leq \beta_1$ and $\mathbb{E}U^R(\beta)$ is maximized at $\hat{\beta}$ for $\beta \in [\hat{\beta}, 1)$, the maximizer of $\mathbb{E}U^R(\beta)$ among $\beta \in [0, 1]$ must be $\hat{\beta}$

or $\phi(s_L|\theta=L)$. Thus, we only need to compare $\mathbb{E}U^R\left(\hat{\beta}\right)$ and $\mathbb{E}U^R\left(\phi(s_L|\theta=L)\right)$:

$$\mathbb{E}U^R\left(\hat{\beta}\right)=\left(\frac{1}{\hat{\beta}}-1\right)\left(\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-1\right)\rho\Delta_H^R$$

$$\begin{aligned}\mathbb{E}U^R\left(\phi(s_L|\theta=L)\right)&=\phi(s_H|\theta=L)(1-\rho)\Delta_L^R+\phi(s_H|\theta=H)\rho\Delta_H^R\\&=\phi(s_H|\theta=L)\left((1-\rho)\Delta_L^R+\rho\Delta_H^R\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right)\end{aligned}$$

$\mathbb{E}U^R\left(\hat{\beta}\right)\geq\mathbb{E}U^R\left(\phi(s_L|\theta=L)\right)$ if and only if

$$\begin{aligned}\hat{\beta}\leq&\frac{1}{\frac{\phi(s_H|\theta=H)-\left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H|\theta=L)\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-1}+1}=\frac{1}{\frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-\left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}\frac{1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-1}+1}\\=&\frac{\phi(s_H\mid\theta=H)-\phi(s_H\mid\theta=L)}{\left[\phi(s_H|\theta=H)+\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H\mid\theta=L)\right]\phi(s_H\mid\theta=L)+\phi(s_H\mid\theta=H)-\phi(s_H\mid\theta=L)}\\=:&\hat{\beta}_{\text{critical}},\end{aligned}$$

where $\hat{\beta}_{\text{critical}}$ is increasing in $\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}$.

(b) If $\hat{\beta}\leq\phi(s_L|\theta=L)$,

$$\frac{\partial\mathbb{E}U^R(\beta)}{\partial\beta}=-(1-\rho)\Delta_L^R-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R>0,\beta\in\left[\beta_1,\hat{\beta}\right)$$

Hence, all $\beta\in\left[\beta_1,\hat{\beta}\right)$ is dominated by $\beta=\hat{\beta}$. We can also find $\mathbb{E}U^R(\beta)$ is constant for $\beta\in[\hat{\beta},\phi(s_L|\theta=L)]$ and decreasing for $\beta>\phi(s_L|\theta=L)$.

All in all, if the classifier has a high capacity, the optimal true-positive rate for the receiver is any $\beta\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}]$ if $\hat{\beta}\leq\hat{\beta}_{\text{critical}}$ and is $\beta=\phi(s_L|\theta=L)$ if $\hat{\beta}>\hat{\beta}_{\text{critical}}$.

We can prove that $\hat{\beta}_{\text{critical}}>\hat{\beta}$ always holds if the classifier has a low capacity:

Lemma 10. *If the classifier has a low capacity, then $\hat{\beta}_{critical}>\hat{\beta}$.*

Proof. The threshold $\hat{\beta}_{\text{critical}}$ can be rewritten as

$$\hat{\beta}_{\text{critical}}=\frac{1}{\frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-\left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-1}-\frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-\left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}}+1}$$

A classifier with a low capacity means $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ or $\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} > \frac{\Delta_L^S}{\Delta_L^S-C} + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\frac{C}{\Delta_L^S-C}$. If $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} < -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$, $\hat{\beta}_{\text{critical}} > 1 > \hat{\beta}$. Then, we consider the case that $\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} \geq -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}$ and $\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} > \frac{\Delta_L^S}{\Delta_L^S-C} + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\frac{C}{\Delta_L^S-C} = -\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1\right)\frac{1}{\hat{\beta}}$. In this case,

$$\hat{\beta}_{\text{critical}} > \frac{1}{\frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1} - \frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right) + \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1\right)\frac{1}{\hat{\beta}}} + 1$$

The right-hand side $\geq \hat{\beta}$

$$\Leftrightarrow \frac{1}{\hat{\beta}} + \frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right) + \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R} - 1\right)\frac{1}{\hat{\beta}}} \geq \frac{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - \left(-\frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\right)}{\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1} + 1$$

Consider a function $f(x) := x + \frac{k}{k+tx}$, where k, t are non-negative constant and $x \in [1, \infty)$. Since $f'(x) = 1 - \frac{tk}{(k+tx)^2} \geq 1 - \frac{1}{4x} > 0$ for all $x \in [1, \infty)$, $f(x)$ is minimized at $x = 1$. So, the above inequality always holds. Therefore, $\hat{\beta}_{\text{critical}} > \hat{\beta}$ holds. □

We can conclude that the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ if $\hat{\beta} \leq \hat{\beta}_{\text{critical}}$ and is $\beta = \phi(s_L|\theta = L)$ if $\hat{\beta} > \hat{\beta}_{\text{critical}}$. Lastly, one can see that

$$\hat{\beta} \leq \hat{\beta}_{\text{critical}}$$

$$\Leftrightarrow\ C \geq \hat{C} = \frac{\left[\phi(s_H|\theta = H) + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L)\right]\phi(s_H \mid \theta = L)}{\left[\phi(s_H|\theta = H) + \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L) - 1\right]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H)}\Delta_L^S.$$

A.8 Proof of Proposition 4

1. Classifier with a low capacity

   According to Lemma 8, the equilibria are

$$\begin{cases} \sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in \left[0, \hat{\beta}\right) \\ \sigma^S = -\frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

So, the payoff of the low-type sender is $\mathbb{E}U_L^S(\beta) = 0$ and the payoff of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta}, & \beta \in \left[0, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

For $\beta \in [0, \hat{\beta})$, $\mathbb{E}U_H^S(\beta)$ is weakly increasing and is dominated by $\beta = \hat{\beta}$. $\mathbb{E}U_H^S(\beta)$ is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ and is decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$. So, in this case, $\mathbb{E}U_H^S(\beta)$ is maximized at $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$.

2. Classifier with a high capacity

According to Lemma 8, the equilibria are

$$\begin{cases} \sigma^S = -\frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \sigma^S = -\frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

So, the payoff of the low-type sender is

$$\mathbb{E}U_L^S(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1-\beta)\Delta_L^S - C & \beta \in \left[\beta_1, \hat{\beta}\right) \\ 0, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

and the utility of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0, \beta_1) \\ (1-\alpha^*(\beta;\phi))\Delta_H^S, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Note that we have proved $\phi(s_L|\theta = L) \geq \beta_1$ for a strong classifier in Lemma 9. Hence, for $\beta \in [0, \beta_1)$, $\mathbb{E}U_H^S(\beta)$ is increasing in β. Moreover, $\mathbb{E}U_H^S(\beta)$ is decreasing in $\beta \in \left[\beta_1, \hat{\beta}\right)$, constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$, and decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$.

Since $\mathbb{E}U_H^S(\hat{\beta}) = \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\hat{\beta};\phi)}{\hat{\beta}} = (1-\alpha^*(\hat{\beta};\phi))\Delta_H^S \leq (1-\alpha^*(\beta_1;\phi))\Delta_H^S = \mathbb{E}U_H^S(\beta_1)$, $\mathbb{E}U_H^S(\beta)$ is maximized at β_1. One can see that $\mathbb{E}U_L^S(\beta)$ is also maximized at β_1.

To show that $\beta_1 = \frac{(1-\rho)\Delta_L^R + \rho\Delta_H^R}{(1-\rho)\Delta_L^R + \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R}$ decreases in $\frac{\phi(s_L|\theta=L)}{\phi(s_L|\theta=H)}$, one just need to observe that the numerator of β_1 is negative, $(1-\rho)\Delta_L^R < 0$, and $\rho\Delta_H^R > 0$.

A.9 Proof of Proposition 5

Firstly, we consider the situation where the classifier has a low capacity. According to Proposition 3 and Proposition 4, $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ is the optimal detector for both the receiver and the sender. Hence, $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ maximizes social welfare.

Secondly, we consider the situation where the classifier has a high capacity. The social welfare can be given as follows:

$$\begin{aligned}
\mathbb{E}W(\beta) &= \mathbb{E}U^R(\beta) + (1-\rho)\mathbb{E}U_L^S(\beta) + \rho\mathbb{E}U_H^S(\beta) \\
&= \begin{cases} \rho C\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0, \beta_1) \\ \rho(1-\alpha^*(\beta;\phi))(\Delta_H^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L^S + \Delta_L^R) - C\right] & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \rho\left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)(\Delta_H^S + \Delta_H^R) + \rho C\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}
\end{aligned}$$

According to Proposition 3 and Proposition 4, $\{\beta^* \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}], \alpha^* = \alpha^*(\beta^*;\phi)\}$ is the optimal lie detector for the receiver if $\hat{\beta} \le \hat{\beta}_{\text{critical}}$, $\{\beta^* = \phi(s_L \mid \theta = L), \alpha^* = \alpha^*(\beta^*;\phi)\}$ is the optimal lie detector for the receiver if $\hat{\beta} \ge \hat{\beta}_{\text{critical}}$, and $\{\beta^* = \beta_1, \alpha^* = \alpha^*(\beta^*;\phi)\}$ is the optimal lie detector for the sender.

If $\hat{\beta} \le \hat{\beta}_{\text{critical}}$, we have

$$\mathbb{E}W(\beta) = \underbrace{\mathbb{E}U^R(\beta)}_{\text{maximized at any}\beta\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}]} + \underbrace{(1-\rho)\mathbb{E}U_L^S(\beta) + \rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at }\beta=\beta_1}$$

One can see that $\mathbb{E}W(\beta)$ is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$, decreases in β for $\beta \in [\max\{\hat{\beta}, \phi(s_L|\theta = L)\}, 1)$, and increases in β for $\beta \in [0, \beta_1)$. So, there exists an optimal β that falls in $\left[\beta_1, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}\right]$.

If $\hat{\beta} > \hat{\beta}_{\text{critical}}$, we have

$$\mathbb{E}W(\beta) = \underbrace{\mathbb{E}U^R(\beta)}_{\text{maximized at }\beta=\phi(s_L|\theta=L)} + \underbrace{(1-\rho)\mathbb{E}U_L^S(\beta) + \rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at }\beta=\beta_1}$$

According to Lemma 9, $\phi(s_L \mid \theta = L) \ge \beta_1$. One can see that $\mathbb{E}W(\beta)$ decreases in β for $\beta \in$

$[\phi(s_L \mid \theta = L), 1)$ and increases in β for $\beta \in [0, \beta_1)$. So, the optimal true-positive rate of the detector β must fall in $[\beta_1, \phi(s_L \mid \theta = L)]$.

Lastly, we have shown that $\hat{\beta} \leq \hat{\beta}_{\text{critical}} \Leftrightarrow C \geq \hat{C}$ in A.7.

References

Anderson, E. T. and Simester, D. I. (2014). Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research*, 51(3):249–269.

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

Becker, G. S. and Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3(1):1–18.

Berman, R. and Katona, Z. (2013). The role of search engine optimization in search marketing. *Marketing Science*, 32(4):644–651.

Berman, R. and Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316.

Berman, R., Zhao, H., and Zhu, Y. (2022). Strategic recommendation algorithms: Overselling and demarketing information designs. *Available at SSRN 4301489*.

Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.

Branco, F., Sun, M., and Villas-Boas, J. M. (2016). Too much information? information provision and search costs. *Marketing Science*, 35(4):605–618.

Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

Callander, S. and Wilkie, S. (2007). Lies, damned lies, and political campaigns. *Games and Economic Behavior*, 60(2):262–286.

Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

Cappelen, A. W., Cappelen, C., and Tungodden, B. (2023). Second-best fairness: The trade-off between false positives and false negatives. *American Economic Review*, 113(9):2458–2485.

Chen, L. and Papanastasiou, Y. (2021). Seeding the herd: Pricing and welfare effects of social learning manipulation. *Management Science*, 67(11):6734–6750.

CMA (2015). Online reviews and endorsements. Available at https://goo.gl/GxZ4J7. Published on February 26.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593.

Dziuda, W. and Salas, C. (2018). Communication with detectable deceit. *Available at SSRN 3234695*.

Eliaz, K. and Spiegler, R. (2019). The model selection curse. *American Economic Review: Insights*, 1(2):127–140.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Goodin, R. E. (1985). Erring on the side of kindness in social welfare policy. *Policy Sciences*, 18(2):141–156.

Gordon, B. R., Jerath, K., Katona, Z., Narayanan, S., Shin, J., and Wilbur, K. C. (2021). Inefficiencies in digital advertising markets. *Journal of Marketing*, 85(1):7–25.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483.

Guo, L. (2009). Quality disclosure formats in a distribution channel. *Management Science*, 55(9):1513–1526.

Guo, L. and Zhao, Y. (2009). Voluntary quality disclosure and market interaction. *Marketing Science*, 28(3):488–501.

He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from amazon. *Proceedings of the National Academy of Sciences*, 119(47):e2211932119.

He, S., Hollenbeck, B., and Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, 41(5):896–921.

Iyer, G. and Ke, T. T. (2024). Competitive model selection in algorithmic targeting. *Marketing Science*.

Iyer, G. and Singh, S. (2018). Voluntary product safety certification. *Management Science*, 64(2):695–714.

Iyer, G. and Singh, S. (2022). Persuasion contest: Disclosing own and rival information. *Marketing Science*, 41(4):682–709.

Iyer, G., Yao, Y. J., and Zhong, Z. Z. (2024). Precision-recall tradeoff in competitive targeting. Unpublished Working Paper.

Iyer, G. and Zhong, Z. (2022). Pushing notifications as dynamic information design. *Marketing Science*, 41(1):51–72.

Jerath, K. and Ren, Q. (2021). Consumer rational (in) attention to favorable and unfavorable product information, and firm information design. *Journal of Marketing Research*, 58(2):343–362.

Jin, C., Yang, L., and Hosanagar, K. (2023). To brush or not to brush: Product rankings, consumer search, and fake orders. *Information Systems Research*, 34(2):532–552.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, 76(4):1359–1395.

Kartik, N., Ottaviani, M., and Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic theory*, 134(1):93–116.

Ke, T. T., Lin, S., and Lu, M. Y. (2022). Information design of online platforms. *HKUST Business School Research Paper*, (2022-070).

Kuksov, D. (2009). Communication strategy in partnership selection. *Quantitative Marketing & Economics*, 7(3).

Kuksov, D. and Lin, Y. (2010). Information provision in a vertically differentiated competitive marketplace. *Marketing Science*, 29(1):122–138.

Lappas, T., Sabnis, G., and Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, 27(4):940–961.

Lauga, D. O., Ofek, E., and Katona, Z. (2022). When and how should firms differentiate? quality and advertising decisions in a duopoly. *Journal of Marketing Research*, 59(6):1252–1265.

Lee, J.-Y., Shin, J., and Yu, J. (2024). Communicating attribute importance under competition. Unpublished Working Paper.

Liang, A. (2019). Games of incomplete information played by statisticians. *arXiv preprint arXiv:1910.07018*.

Lieberman, M. D. and Cunningham, W. A. (2009). Type i and type ii error concerns in fmri research: re-balancing the scale. *Social cognitive and affective neuroscience*, 4(4):423–428.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management science*, 62(12):3412–3427.

Mattes, K., Popova, V., and Evans, J. R. (2023). Deception detection in politics: Can voters tell when politicians are lying? *Political Behavior*, 45(1):395–418.

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing science*, 25(2):155–163.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–2455.

Mayzlin, D. and Shin, J. (2011). Uninformative advertising as an invitation to search. *Marketing science*, 30(4):666–685.

Miklós-Thal, J. and Tucker, C. (2019). Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science*, 65(4):1552–1561.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, pages 380–391.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M., and Prat, A. (2022). Competing models. *The Quarterly Journal of Economics*, 137(4):2419–2457.

O'Connor, J. and Wilson, N. E. (2021). Reduced demand uncertainty and the sustainability of collusion: How ai could affect competition. *Information Economics and Policy*, 54:100882.

Papanastasiou, Y. (2020). Fake news propagation and detection: A sequential model. *Management Science*, 66(5):1826–1846.

Pei, A. and Mayzlin, D. (2022). Influencing social media influencers through affiliation. *Marketing Science*, 41(3):593–615.

Piccolo, S., Tedeschi, P., and Ursino, G. (2015). How limiting deceptive practices harms consumers. *The RAND Journal of Economics*, 46(3):611–624.

Piccolo, S., Tedeschi, P., and Ursino, G. (2018). Deceptive advertising with rational buyers. *Management Science*, 64(3):1291–1310.

Qian, K. and Jain, S. (2024). Digital content creation: An analysis of the impact of recommendation systems. *Management Science*.

Rao, A. and Wang, E. (2017). Demand for "healthy" products: False claims and ftc regulation. *Journal of Marketing Research*, 54(6):968–989.

Rayo, L. and Segal, I. (2010). Optimal information disclosure. *Journal of political Economy*, 118(5):949–987.

Rhodes, A. and Wilson, C. M. (2018). False advertising. *The RAND Journal of Economics*, 49(2):348–369.

Salant, Y. and Cherry, J. (2020). Statistical inference in games. *Econometrica*, 88(4):1725–1752.

Shin, J. (2005). The role of selling costs in signaling price image. *Journal of Marketing Research*, 42(3):302–312.

Shin, J. and Wang, C.-Y. (2024). The role of messenger in advertising content: Bayesian persuasion perspective. *Marketing Science*.

Shulman, J. D. and Gu, Z. (2024). Making inclusive product design a reality: How company culture and research bias impact investment. *Marketing Science*, 43(1):73–91.

Sun, M. (2011). Disclosing multiple product attributes. *Journal of Economics & Management Strategy*, 20(1):195–224.

Sun, M. and Tyagi, R. K. (2020). Product fit uncertainty and information provision in a distribution channel. *Production and Operations Management*, 29(10):2381–2402.

Villas-Boas, J. M. (2004). Communication strategies and product line design. *Marketing Science*, 23(3):304–316.

Yao, Y. (2024). Dynamic persuasion and strategic search. *Management Science*, 70(10):6778–6803.

Zhang, J. (2013). Policy and inference: The case of product labeling. Unpublished Working Paper.

Zheng, X. and Singh, S. (2023). Ambiguous expert communication. *Available at SSRN 4393315*.

Zinman, J. and Zitzewitz, E. (2016). Wintertime for deceptive advertising? *American Economic Journal: Applied Economics*, 8(1):177–192.