Agreeing to be fooled:
Optimal ignorance about information sources

Takuma Habu[*]

18th February 2024

Abstract

Should a decision-maker learn whether an information source is reliable? I consider a persuasion game in which the sender is sometimes unreliable—i.e., can covertly manipulate the signal used to persuade the receiver—and the receiver can costlessly investigate the sender's reliability. I show that the receiver benefits from committing to investigations that do not always reveal reliability. Even without the ability to commit to ignorance, I demonstrate that the receiver benefits from delegating investigations to someone partially adversarial to the sender and partially aligned with the receiver. My results shed light on the efficacy of cross-examination, audits, and *ad hominem* arguments.

Suppose a receiver obtains a piece of information from a sender but is worried that the information might not be reliable. Should the receiver investigate and learn whether the information is reliable before deciding on an action?

Once the receiver has obtained the information, learning about reliability is beneficial because it allows him to avoid making decisions based on unreliable information.[1] However, in many economic situations, the sender is likely to change the kinds of information she provides based on what she expects the receiver to

[*] The Department of Economics, Cornell University, takumahabu@cornell.edu. I thank my committee members Emir Kamenica, Elliot Lipnowski, Doron Ravid and Lars Stole for their invaluable advice and feedback. I am also grateful to Ben Brooks, Modibo Camara, Andy Choi, Alex Frankel, Marina Halac, Daniel Rappoport, Phil Reny, Joseph Root, Christoph Schlom, Kai Hao Yang, and the participants of the micro-theory seminar at The University of Chicago for their helpful comments.

[1] Throughout, I refer to the receiver and the sender using male and female pronouns, respectively.

learn about reliability. In such situations, it need not be the case that the receiver benefits from learning about reliability. For example, while courts can learn about the reliability of witness evidence provided by the parties via cross-examination, parties may produce different witnesses depending on whether and how they expect the witnesses to be cross-examined. Similarly, sellers may make different marketing claims depending on how much they expect the buyers to audit their claims, and politicians may change their arguments depending on their expectations about the opposing politicians' counter-arguments.

In this paper, I study a sender-receiver game of persuasion in which there is doubt about the sender's reliability, and the receiver first decides how much to investigate and learn about the sender's reliability. I show that the receiver can benefit from avoiding learning about the sender's reliability because doing so allows the receiver to trade off information about reliability (which he can obtain himself by investigating) with information that he can only obtain from the sender. To take advantage of the trade-off, however, the receiver must be able to fight his inherent desire to learn. While the receiver can achieve this by committing to ignorance, I also study what the receiver can achieve by delegating investigations to a third party. To that end, I show that, although delegating investigations to an adversary to the sender can be beneficial to the receiver, he is better off when the third party is only partially adversarial to the sender and otherwise partially aligned with the receiver. In fact, I show that the receiver can do just as well with delegation as when he can commit if the third party has such "balanced" preferences.

To understand the implications of my results, let us first take the court context. In many jurisdictions, the court effectively delegates cross-examinations of witnesses to an adversary to the party that calls on the witness to testify. My results suggest that such an adversarial system is more effective than if the court itself (i.e., the judge or the jury) were to directly cross-examine witnesses if the court lacks the ability to commit to cross-examinations that do not seek to discover the truth about the witness' reliability.[2] Moreover, my results also suggest that provisions

[2] One can also interpret my results as suggesting that adversarial cross-examination is unnecessary (and, in fact, weakly less preferred) if the court can commit to conducting cross-examinations that do not seek to discover the truth.

that align the adversarial cross-examiner's interest with that of the court's interests can further increase the efficacy of cross-examination. In the buyer-seller context, my results imply that buyers would be better off with audits that do not always reveal the reliability of the information provided by the seller either by conducting audits themselves (if they can commit to ignorance) or by choosing an auditor with adversarial incentives.

Related literature. This paper contributes to the rich literature on strategic communication.[3] In addition to a payoff-relevant state, I introduce uncertainty about the sender's (manipulative) behaviour in a probabilistic manner akin to models that relax the "commitment assumption" in Bayesian persuasion models (e.g., Frechette, Lizzeri and Perego, 2019; Min, 2021; and Lipnowski, Ravid and Shishkin, 2022).[4] In my model, the receiver is able to induce the sender to provide more information (than without investigations) through a combination of the addition of noise to the sender's communication and the imperfect ability of the receiver to distinguish between truthful and noisy communication. While similar channels have been explored in (mediated) cheap-talk games (e.g., Austen-Smith, 1994; Blume, Board and Kawamura, 2007; Goltsman et al., 2009),[5] a distinguishing feature of my model (in addition to studying a different type of communication) is that, in effect, there is a second sender who adds noise by designing information (only) about the uncertainty regarding the sender's behaviour. The type of information that the second sender provides also sets this paper apart from the existing literature on multiple senders (e.g., Gerardi and Yariv, 2008; Che and Kartik, 2009; Gentzkow and Kamenica, 2017*a*,*b*; Dworczak and Pavan, 2022), games in which the receiver can design information about the payoff-relevant state (Ivanov, 2010*b*; Krähmer, 2021; Ivanov and Sam, 2022), and games in which the receiver can learn about the veracity of the sender's messages (e.g., Dziuda and Salas, 2018; Balbuzanov, 2019; Ederer and Min, 2022; Levkun, 2022; Sadakane and Tam, 2022).[6]

[3] See surveys by Sobel (2013); Özdogan (2016); Kamenica (2019); Bergemann and Morris (2019); Forges (2020).

[4] The commitment assumption in Bayesian persuasion (Kamenica and Gentzkow, 2011) refers to the assumption that the sender will truthfully communicate the realisation of a chosen statistical experiment (equivalently, a communication strategy) to the receiver.

[5] I thank an anonymous referee for pointing out this connection explicitly.

[6] Veracity refers to whether the sender uses messages in a way consistent with an exogenously

That ignorance, or avoidance of information, can be beneficial for strategic reasons has been observed in other contexts (e.g., Taylor and Yildirim, 2011; McAdams, 2012; Roesler and Szentes, 2017; Onuchic, 2022).[7] This paper demonstrates that avoiding information can also be beneficial in a sender-receiver game with unrestricted, costless communication while allowing the receiver to choose his signal (i.e., investigation) based on the sender's choice of an experiment.[8]

Following Schelling (1960), strategic delegation has been studied as a way to mitigate or eliminate commitment issues in many contexts, including industrial organisations (Vickers, 1985; Fershtman and Judd, 1987; Sklivas, 1987) and macroeconomics (Rogoff, 1985). In the context of mediated cheap-talk or disclosure games, Ivanov (2010*a*), Ambrus, Azevedo and Kamada (2013) and Lichtig (2020) find that a mediator or a second sender who is adversarial to the (first) sender can induce the sender to provide more information. I study a different type of strategic communication game and find similar results while also highlighting the importance of the third party having a balanced—and not just adversarial—incentive.

I also bring new arguments based on strategic information considerations to the literature that compares approaches to evidence across legal systems (Shin, 1998; Dewatripont and Tirole, 1999; Posner, 1999), complementing a recent contribution by Lichtig (2020) who studies a disclosure game. The model also brings new insight into how audits can incentivise companies to provide more information in equilibrium.[9] In contrast to existing models (e.g., Townsend, 1979; Mookherjee and Png, 1989; Border and Sobel, 1987), in my model, audits are costless and transfers are not allowed, and audits are about the auditee's reliability type and not about the veracity of the auditee's "messages."

The remainder of the paper is structured as follows. In Section 1, I give an illustrative example that demonstrates the intuitions behind the results and a more detailed explanation of the results. In Section 2, I set out the model and I give characterisations of equilibria when the receiver can and cannot commit to ignorance in

given meanings of messages.

[7] Golman, Hagmann and Loewenstein (2017) provides a recent survey.

[8] In Section 5, I show that the receiver can benefit from a partially revealing investigation even without the ability to vary investigations based on the sender's choice of an experiment .

[9] See Ye (2021) for a survey of economic models that describe the role of audits.

Section 3. Section 4 deals with the delegation case. In Section 5, I provide some interpretations of the results, as well as a discussion of extensions. I give a conclusion in Section 6.

1 An illustrative example

To develop an intuition for the results, consider the following example in which an institutional investor (receiver) is deciding whether to *buy* or *not buy* an asset.[10] There are two states of the world: the asset is either *good* or *bad*. The investor gets utility 1 for making the correct investment decision (*buy* when *good* and *not buy* when *bad*) and 0 for making the wrong decision (*buy* when *bad* and *not buy* when *good*). In contrast, the seller of the asset (sender) wants the investor to buy the asset regardless of the state. Assume she gets utility 1 from the investor buying and 0 from not buying. The investor and the seller share a prior belief 0.3 that the asset is *good* meaning that the investor would not buy the asset under the prior belief.

To persuade the investor to buy, the seller provides an investment appraisal (e.g., as part of the prospectus) to the investor that sets out an appraisal method as well as the result of the appraisal analysis. Formally, an appraisal method is a signal structure, $\xi = (\xi(\cdot|good), \xi(\cdot|bad))$, specifying a distribution over possible results of the appraisal analysis conditional on the state. An appraisal is thus a pair consisting of an appraisal method ξ and the result. Suppose that there are only two possible results of the appraisal: g (meaning good) and b (meaning bad); and that the seller can only choose between three appraisal methods: the "*H*ighly" informative (ξ^H), the "*M*ildly" informative (ξ^M), and the "*L*east" informative (ξ^L) methods, given by

$$\begin{aligned} \xi^H(g|good) &= 1, \quad & \xi^M(g|good) &= 1, \quad & \xi^L(g|good) &= 1, \\ \xi^H(g|bad) &= 0, \quad & \xi^M(g|bad) &= \frac{1}{7}, \quad & \xi^L(g|bad) &= \frac{3}{7}. \end{aligned}$$

Observe that the three methods differ only on their rate of false positives (i.e., the probability that the appraisal result is g when the asset is *bad*). Thus, the in-

[10] This example intentionally borrows from the courtroom example in Kamenica and Gentzkow (2011).

vestor, who prefers appraisal methods that are more informative, strictly prefers ξ^H over ξ^M over ξ^L. In contrast, the seller, who wishes to maximise the probability that the investor buys the asset, prefers to choose the least informative method that can persuade the investor into buying. Suppose that, having chosen a method $\xi \in \{\xi^H, \xi^M, \xi^L\}$, with probability 0.2, the seller is unreliable and can falsify the result of the analysis to her benefit. Because only the result g could persuade the investor to buy the asset, the seller always manipulates the appraisal result to be g when she can. On the other hand, with probability 0.8, the seller is reliable so that the true result of the appraisal method ξ is communicated to the investor. The investor is unable to tell whether the appraisal result he observes has been falsified without conducting an audit.

No audit versus full audit. Without an audit, after the seller has chosen $\xi \in \{\xi^H, \xi^M, \xi^L\}$, the investor believes that the result g was drawn according to ξ with probability 0.8 and chosen independently of the state with probability 0.2. The seller's (ex ante) payoff from a method is the probability that the investor buys the asset, and her payoffs from (ξ^H, ξ^M, ξ^L) without audits are $(0.44, 0.52, 0)$, respectively.[11] Hence, the seller chooses ξ^M when the investor cannot audit the seller. Suppose now that the investor conducts a full audit and finds out whether the seller is reliable. When the investor finds out that the seller is unreliable, he ignores the appraisal and does not buy the asset. Alternatively, when the investor finds out that the seller is reliable, then he knows that the appraisal result was obtained using the stated method and the investor can be persuaded to buy the asset. The seller's payoffs from (ξ^H, ξ^M, ξ^L) with full audits are $(0.24, 0.32, 0.48)$, respectively.[12] Therefore, the seller now chooses ξ^L so that, from the investor's perspective, finding out the seller's reliability leads the seller to choose a worse appraisal method.

[11] The investor's posterior beliefs after seeing g for methods (ξ^H, ξ^M, ξ^L) are $(\frac{15}{22}, \frac{15}{26}, \frac{15}{34})$, respectively. Hence, the investor only buys after seeing g under $\xi \in \{\xi^H, \xi^M\}$. The seller's payoff is zero from ξ^L, and her payoffs from choosing other methods are given by the respective probability that the investor observes g.

[12] The investor's posterior beliefs after seeing g for methods (ξ^H, ξ^M, ξ^L) are $(1, \frac{3}{5}, \frac{1}{2})$, respectively. Hence, the investor buys upon seeing g and finding out that the seller is reliable, and the seller's payoffs are given by the respective probability of this event occurring.

Partially ignorant audits. Consider how the investor can attain the ideal outcome—i.e., finding out whether the seller is reliable after she has chosen ξ^H. Suppose that the investor publicly commits to an *audit strategy* that specifies the audit that will be conducted as a function of the seller's choice of an appraisal method. Note first that the seller's payoff from the investor's ideal outcome is given by the probability that the state is *good* and the sender is reliable; i.e., 0.24. To attain this outcome, the investor must ensure that the seller's payoffs from choosing ξ^M and ξ^L are lower than 0.24. Observe that the investor can simply not audit the seller when ξ^L is chosen in which case the seller's payoff is $0 < 0.24$. To prevent the seller from choosing ξ^M, the investor can conduct an audit that reveals that the seller is unreliable with probability 1, but only reveals that the seller is reliable with some probability $p \in [0,1)$.[13] That $p < 1$ implies that the investor does not always find out that the seller is reliable, and this type of audit makes it less likely that the investor is persuaded to buy the asset. In particular, letting $p = 0.5$ minimises the seller's payoff from choosing ξ^M to $0.16 < 0.24$.[14] Hence, the investor can obtain the ideal outcome by using ignorance—in the form of avoiding learning that the seller is reliable or avoiding learning altogether—as a way to punish the seller for choosing ξ^M and ξ^L.

It is also possible for ignorance to be used as a reward for the seller choosing the "correct" appraisal method. For example, if the probability that the seller is unreliable is 0.1 (instead of 0.2 previously), it turns out it is no longer possible to induce the seller to choose ξ^H while also finding out the seller's reliability.[15] To induce the seller to choose ξ^H, the investor can conduct audits that do not always reveal that the seller is unreliable; i.e., audits that reveal that the seller is reliable

[13] In this example, an audit is a binary-support signal structure about the seller's reliability type. Hence, an audit can be characterised by a pair $(p,q) \in [0,1]^2$ with $p \geq 1-q$, where p (resp. q) is the probability that the audit reveals that the seller is reliable (resp. unreliable) when she is reliable (resp. unreliable). A full audit corresponds to the pair $(1,1)$ and no audit corresponds to the pair $(0.5,0.5)$.

[14] Given ξ^M and $(p,q) = (0.5,1)$, the investor finds it optimal to buy the asset only after seeing g and the audit outcome indicating that the message is reliable. The probability of this event occurring is 0.16.

[15] In this case, the lowest seller payoff that the investor can induce by being strategically ignorant when the seller chooses ξ^M is 0.28, which is strictly higher than the seller's payoff of 0.27 under the investor's ideal outcome.

with probability 1, but reveal that the seller is unreliable with probability $q \in [0,1)$. Then, letting $q \leq \frac{2}{3}$ ensures the seller's payoff from choosing ξ^H is greater than 0.28.[16] Thus, when the seller is less likely to be unreliable, the investor strategically uses ignorance as a way to reward the seller for choosing ξ^H as well as a way to punish the seller for choosing ξ^M and ξ^L. However, when the probability that the seller is unreliable is even lower (e.g., 0.05), it becomes impossible for the investor to ensure that the seller's payoff from choosing ξ^H is greater than her payoff from choosing ξ^M. In this case, the investor's best option is to demand ξ^M and find out whether the seller is reliable (while not auditing ξ^L).

Delegating audits. Once the seller has chosen a method, there is no strategic advantage to the investor from being ignorant. Hence, the sequentially rational audit for the investor is the full audit. Thus, for the investor to directly implement the audit strategy that induces the ideal outcome described in the previous paragraph, the investor must be able to commit to being (sometimes) ignorant. However, even without the ability to commit, the investor can still induce the seller to choose the most informative method, ξ^H, by delegating the audit to a third party who is at least partially adversarial to the seller. For this example, let us consider the case in which the third party is purely adversarial. Given any appraisal method, the sequentially rational audit for a purely adversarial third party is to minimise the probability that the investor buys; i.e., the adversary's sequentially rational audit is maximally punishing. When audits are sequentially rationally conducted by an adversary, the seller's payoffs from methods (ξ^H, ξ^M, ξ^L) are $(0.16, 0.16, 0)$, respectively.[17] Because the seller's payoffs from choosing ξ^H and ξ^M are the same, it is possible for the investor to induce the seller to choose ξ^H by delegating audits to a purely adversarial third party. Notice, however, that the investor is worse off than when he could commit to ignorance because, while he is able to induce the sender to choose ξ^H, he does not always find out the reliability of the seller.

[16] For example, given ξ^H and $(p,q) = (1, \frac{2}{3})$, the investor finds it optimal to buy the asset only after seeing g and the audit outcome that indicating that the message is reliable (in which case he is certain that the seller is reliable). The probability of this event occurring is 0.28.

[17] The maximally punishing audit following ξ^M and ξ^L are as in the audit strategy that induces the ideal outcome as described in the previous paragraph. Maximally punishing audit following ξ^H involves setting $p = \frac{2}{3}$. To be clear, the payoffs correspond to the case when the prior belief that the seller is reliable is 0.2.

Beyond the example. The main sender-receiver model in the paper maintains the assumptions regarding state, actions and preferences while allowing the sender (i.e., the seller in the example) to choose any statistical experiment about the state (i.e., appraisal method) and the receiver (i.e., the investor) to choose any investigation (i.e., audit). It turns out that restricting the sender to choosing from a set of "one-sided" experiments, as in the example, is without loss. Moreover, because such experiments can be parameterised by their rate of false positives (i.e., $\xi(g|bad)$), they are also Blackwell ordered (Blackwell, 1953). Thus, even in the main model, the conflict of interest with respect to the choice of the experiment between the sender and the receiver can be thought of in terms of the rate of false positives of experiments as in the example.

Using this simplification, I show that the result from the example—that the receiver can obtain more information by committing to investigation strategies that involve ignorance as punishments and rewards—continues to hold. I also demonstrate that, without the ability to commit to ignorance, the receiver is no better off with the ability to investigate and learn about the sender's reliability.[18]

I also extend the delegation case from the example and consider a third party who can be partially adversarial to the sender and partially aligned with the receiver; i.e., a third party whose preference is a linear combination of the negative of the sender's payoff and the receiver's payoff. I find that the receiver strictly benefits from delegation investigation to a third party whose preferences are balanced in this manner. This is because such balanced preferences of the third party make it sequentially rational for it to conduct fully revealing investigations (that are never optimal for a pure adversary) whenever the sender's choice of an experiment is more informative than a threshold (in terms of the rate of false positives). At the same time, such a third party finds it sequentially rational to punish the sender for choosing experiments (i.e., appraisal methods) that are less informative than the threshold (which are the experiments that the sender would deviate to). Because the threshold depends on the relative weight the third party places on the

[18] In the example above, the investor was strictly worse off when finding out the seller's reliability. However, if the seller could have also selected $\xi(g|good) = 1$ and $\xi(g|bad) = \frac{2}{7}$, the investor would have been indifferent between finding out the seller's reliability and not finding out.

receiver's preferences, by delegating investigations to a third party with appropriately balanced preferences, the receiver is able to credibly commit to finding out the sender's reliability while still inducing the sender to choose an informative experiment. In fact, I find that, against a sufficiently unreliable sender, the receiver can attain the same outcome with delegation as when he can commit to ignorance.

2 Model

There are two players: a *Sender* (S) and a *Receiver* (R). The Receiver can take one of two actions, denoted $a \in A := \{0,1\}$, and the Receiver's payoff from each action depends on the binary states of the world, $\theta \in \Theta := \{0,1\}$. The preferences are such that the Receiver's optimal action is $a = 1$ (i.e., to *take action*) whenever he believes that the state is $\theta = 1$ with probability at least $\mu^* \in (0,1)$; otherwise, the Receiver's optimal action is to choose $a = 0$ (i.e., to *not take action*). The Sender would like the Receiver to choose $a = 1$ no matter the state. Let $v_{\mathrm{S}} : A \to \mathbb{R}$ and $v_{\mathrm{R}}(a,\theta) : A \times \Theta \to \mathbb{R}$ be the Sender and the Receiver's payoffs, respectively, where[19]

$$v_{\mathrm{R}}(a,\theta) := a\frac{\theta - \mu^*}{\mu^*},\ v_{\mathrm{S}}(a) := a.$$

Let $\mu_0 \in (0,1)$ denote the common prior probability that $\theta = 1$.[20] To make the problem interesting, I assume that the Receiver does not take action under the prior belief; i.e., $\mu_0 < \mu^*$. If this condition does not hold, the Sender has no incentive to provide any information and thus concerns about the reliability of the Sender become moot. Given the normalisations, both the Sender's and the Receiver's default payoffs under the prior beliefs are zero.

To persuade the Receiver to take action, the Sender publicly chooses a signal structure $\xi \in \Xi := (\Delta M)^{\Theta}$, where M is a finite set of messages with at least two elements. I refer to ξ as an *experiment*. The Sender can either be of type $t = r$

[19] I normalise the Receiver's payoff from choosing $a = 0$ to be zero and his payoff from choosing $a = 1$ when $\theta = 0$ to be -1.

[20] Given an arbitrary set X, I use ΔX to denote the set of probability measures on the set X. Given a measure $\nu \in \Delta\Theta$, I denote its support as $\mathrm{supp}(\nu)$ and sometimes abuse notation and use ν to denote $\nu(\{\theta = 1\})$.

(for *reliable*) or u (for *unreliable*), with $T := \{r,u\}$. I let $\rho_0 \in (0,1)$ denote the common prior probability that the Sender is reliable. A reliable Sender truthfully communicates realisations from the experiment ξ to the Receiver. In contrast, an unreliable Sender can communicate any message $m \in M$. The Receiver observes the Sender's message m without observing the Sender's type.

Notice that a reliable Sender can commit to the announced experiment as in Bayesian persuasion models (Kamenica and Gentzkow, 2011).[21] In contrast, the unreliable Sender lacks such commitment power. Thus, the higher the prior belief ρ_0 that the Sender is reliable, the "more" the Sender is committed to the announced experiment. One can also interpret the prior belief ρ_0 as capturing the imperfectness in the enforcement of truthful communication (Min, 2021), or the possibility that the Sender can indirectly alter the realisation of the experiment by influencing the "experimenter" that carries out of the experiment (Lipnowski, Ravid and Shishkin, 2022). More generally, the prior belief ρ_0 can be thought of as capturing the conflicting incentives that an experimenter might have in truthfully communicating the results of the experiment to the Receiver.[22] In addition, the prior ρ_0 can also be interpreted as the probability that the experimenter is competent; i.e., that the experimenter is capable of carrying out the experiment. Another interpretation of the prior belief ρ_0 is that it represents the probability with which the Sender is simply corrupt and alters the result of the experiment.

Importantly, the Receiver can investigate the Sender's type by using any signal structure about the Sender's type. I take the belief-based approach (Kamenica, 2019; Forges, 2020) and express investigations as Bayes-plausible distributions of posterior beliefs about the Sender's type. I assume that the distributions of posterior beliefs have finite support so that an investigation is an element in $\mathscr{I} := \{\iota \in \Delta([0,1]) : \int \rho \mathrm{d}\iota(\rho) = \rho_0, |\mathrm{supp}(\iota)| < \infty\}$ with a typical element ι. The

[21] Forges (2020) describes Bayesian persuasion as the case in which the statistical experiment chosen by the sender is "fully reliable."

[22] For example, on the one hand, the experimenter may have reputational or moral concerns that guide them towards communicating truthfully. On the other hand, they may also have financial or relational concerns (either via explicit payment or implicit payment in the form of future interactions with the Sender) that guide them towards lying on behalf of the Sender. Under this interpretation, the prior ρ_0 captures the Sender's and the Receiver's common uncertainty about the combined effect of these incentives on the experimenter's behaviour.

Receiver's investigation strategy is a mapping from the Sender's choice of an experiment ξ to an investigation ι, which I denote as $i : \Xi \to \mathscr{I}$.[23]

In the next section, I consider two cases: the *commitment case* in which the Receiver can commit to any investigation strategy, and the *no-commitment case* in which the Receiver chooses an investigation after observing the Sender's experiment ξ in a sequentially rational manner.[24]

The timing of the game is as follows. In the commitment case, the Receiver first publicly commits to an investigation strategy $i(\cdot)$. The Sender then publicly chooses an experiment $\xi \in \Xi$. Nature then independently draws the state and the Sender's type, $\theta \sim \mu_0$ and $t \sim \rho_0$, respectively. If the Sender is reliable (i.e., $t = r$), then the Sender truthfully communicates m drawn from $\xi(\theta)$. If the Sender is unreliable (i.e., $t = u$), then the Sender chooses $m \in M$ without observing the realised state.[25] Finally, the Receiver observes the realisation of the investigation ρ drawn from $i(\xi)$ as well as the message from the Sender m, and chooses an action $a \in A$. All players update beliefs using Bayes rule whenever possible. In the no-commitment case, the Sender first publicly chooses an experiment and then the Receiver publicly chooses an investigation $\iota \in \mathscr{I}$. The rest of the play is the same except that the Receiver now observes the realisation of the investigation ρ drawn from ι.

Toward defining an equilibrium in the commitment case, call a weak perfect Bayesian equilibrium (PBE) of the game induced by an investigation strategy i as an i-equilibrium. I define a *commitment equilibrium* as a PBE that maximises the Receiver's payoff with respect to investigation strategy i and i-equilibria. I define a no-commitment equilibrium as a PBE that maximises the Sender's payoff among PBE of games induced by some experiment $\xi \in \Xi$.[26]

[23] More generally, an investigation strategy could also depend on the realisation of the experiment. As discussed in section 5, the Receiver does not benefit from this additional flexibility and I therefore simplify the exposition by assuming that an investigation strategy only depends on the Sender's choice of an experiment.

[24] I also study the *delegation case* in which a third party chooses an investigation after observing ξ in a sequentially rational manner in section 4.

[25] The results do not change materially if the Sender observes her type prior to choosing an experiment or the unreliable-type Sender can observe the realised state (see section 5).

[26] I give the formal definitions of equilibria under the two cases in the appendix.

3 Optimal ignorance about sender's reliability

3.1 No-commitment case

Consider first the case the Receiver chooses an investigation after observing the Sender's choice of an experiment in a sequentially rational manner. In this case, there is no strategic consideration for the Receiver when choosing an investigation. Thus, the standard argument—more information is always better for a decision-maker—means that the Receiver's optimal investigation is fully revealing. Moreover, because the Receiver ignores the Sender's message whenever he learns that the Sender is unreliable (which occurs with probability $1-\rho_0$), the Sender's optimal experiment in the no-commitment case corresponds to the optimal experiment when she is known to be fully reliable (i.e., the Bayesian persuasion case).[27] The following result is then immediate.[28]

Theorem 1. *The Receiver's no-commitment equilibrium payoff is zero. In any no-commitment equilibrium, the Sender chooses her optimal experiment when she is known to be fully reliable and the Receiver finds out the Sender's reliability.*

Consider now the case in which the Receiver cannot investigate the Sender's reliability. In this case, the optimal experiment for the Sender is to provide "just enough" information so as to leave the Receiver always indifferent between taking action and not taking action.[29] Thus, the Receiver's payoff in this case is zero. Combining this observation with the fact that the Receiver's payoff is zero without any information from the Sender gives the corollary below.

[27] That the Receiver is no better off when fully learning relies on the fact that the Receiver's action is binary (see Proposition 5 in Kamenica and Gentzkow, 2011). Moreover, even if there was an upper bound on the informativeness of the Receiver's investigations, the Receiver's no-commitment equilibrium payoff would still be zero—the Receiver without the ability to commit to ignorance would continue to investigate to the full extent possible and the Sender can provide less information to exactly offset the Receiver's benefit from learning about reliability.

[28] I give all proofs of the results in this paper in the appendix. I focus on the Receiver's equilibrium payoff in the main body of the paper. The Sender's equilibrium payoff can be found in the appendix.

[29] The result readily follows from Theorem 1 in Lipnowski, Ravid and Shishkin (2022).

Corollary 1. *The Receiver is indifferent among the following cases: (i) the Sender provides no information; (ii) the Receiver cannot investigate; and (iii) the no-commitment case.*

In particular, Corollary 1 means that, when the Receiver cannot commit to avoiding learning, he does not at all benefit from the ability to learn about the Sender's reliability.

3.2 Commitment case

I now consider the case in which the Receiver is able to commit to an investigation strategy. The main result is the following which establishes that, in contrast to the no-commitment case, the Receiver strictly benefits from the ability to investigate the Sender's reliability. In fact, whenever the Sender is sufficiently unreliable, the Receiver is able to obtain his ideal outcome—i.e., the Sender choosing the fully informative experiment under a fully revealing investigation. To state the result, let $\overline{V}_{\mathrm{R}}$ denote the Receiver's payoff from the ideal outcome, and let $V_{\mathrm{S}}^{\mathrm{maxmin}}$ denote the Sender's maximal payoff against an investigation strategy that minimises the Sender's payoff.

Theorem 2. *For any $\rho_0 \in (0,1)$, the Receiver's commitment equilibrium payoff is*

$$V_{\mathrm{R}}^{*} = \min\left\{\overline{V}_{\mathrm{R}}, \tfrac{\mu_0}{\mu^*}[(1-\mu_0)\rho_0 + V_{\mathrm{S}}^{\mathrm{maxmin}}] - V_{\mathrm{S}}^{\mathrm{maxmin}}, \tfrac{\mu_0}{\mu^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}\right\} > 0, \qquad (1)$$

In particular, the Receiver's commitment equilibrium payoff is strictly positive for every interior prior belief, ρ_0, about the Sender's reliability. Moreover, for sufficiently low ρ_0, the Receiver is able to induce the Sender to choose the fully informative experiment while simultaneously finding out the Sender's reliability on the equilibrium path.

To prove the theorem, I first establish the following lemma which shows that it suffices to focus on experiments that make action recommendations and that are "one-sided" as in the example in the introduction.

Lemma 1. *Any strictly positive commitment equilibrium payoffs are attainable with an experiment $\xi \in \Xi$ such that for some $m_0, m_1 \in M$,* $\operatorname{supp}(\xi) = \{m_0, m_1\}$, $\xi(m_1|1) = 1$, $\xi(m_1|0) \leq 1 - \frac{\mu^* - \mu_0}{\mu^*(1-\mu_0)}$*, and the unreliable Sender always sends m_1.*

In proving the lemma, I establish a revelation principle in games induced by any experiment-investigation pair $(\xi, \iota) \in \Xi \times \mathscr{I}$. Note that one cannot appeal to existing arguments (see, for example, Bergemann and Morris, 2019) to establish Lemma 1 because, in my model, information is effectively being designed by two distinct players. I therefore provide an original proof in the appendix and give a sketch of the proof here. Given any pair $(\xi, \iota) \in \Xi \times \mathscr{I}$, call a tuple (σ, α, μ) an (ξ, ι)*-equilibrium* if it is a PBE of the game induced by (ξ, ι), where σ is the unreliable Sender's messaging strategy, α is the Receiver's action strategy and μ is a belief map. The main step of the proof is to show that any tuple (σ, α, μ) that is a (ξ, ι)-equilibrium can be reduced to a payoff-equivalent tuple $(\tilde{\sigma}, \tilde{\alpha}, \tilde{\mu})$ that is a $(\tilde{\xi}, \iota)$-equilibrium where $\tilde{\xi} \in \Xi$ makes action recommendations, $\tilde{\sigma} \in \Delta M$ always recommends the Receiver to take action, and ι remains optimal for the Receiver against $\tilde{\xi}$. Having established that the restricting attention to experiment with binary messages does not reduce the set of equilibrium payoffs, I show that any equilibrium experiment ensures that the probability of not taking action when the state is $\theta = 1$ is zero, i.e., $\xi(m_1|1) = 1$, because both players' interests are aligned in minimising such an event.

Any experiment in the class of experiments identified in Lemma 1 can be uniquely parameterised by a scalar $\widehat{\rho} \in [\underline{\rho}, 1]$, with $\underline{\rho} := \frac{\mu^* - \mu_0}{\mu^*(1-\mu_0)}$, by defining $\xi_{\widehat{\rho}} : [\underline{\rho}, 1] \times \Theta \to \Delta(\{m_0, m_1\})$ as

$$\xi_{\widehat{\rho}}(m_1|1) := 1, \; \xi_{\widehat{\rho}}(m_1|0) := 1 - \tfrac{1}{\widehat{\rho}}\underline{\rho}.$$

The scalar $\widehat{\rho}$ that identifies the experiment $\xi_{\widehat{\rho}}$ can in fact be interpreted as the threshold for the posterior belief about the Sender's reliability above which the Receiver takes action after observing m_1; i.e., the Receiver takes action if he observes a message (ρ, m_1) with $\rho \geq \widehat{\rho}$. The experiments in Lemma 1 are also totally ordered according to the Blackwell order and this ordering corresponds to the ordering by $\widehat{\rho}$—with lower $\widehat{\rho}$ corresponding to more Blackwell informative experiments. In particular, $\widehat{\rho} = \underline{\rho}$ identifies the fully informative experiment while $\widehat{\rho} = 1$ identifies the Sender-preferred Bayesian persuasion experiment (i.e., the Sender-optimal experiment when the Sender is initially known to be fully reliable). In what follows, I identify the Sender's choice of an experiment from the class of experiments in Lemma 1, $\xi \in \{\xi_{\widehat{\rho}}\}_{\widehat{\rho} \in [\underline{\rho},1]}$, with the associated parameter $\widehat{\rho} \in [\underline{\rho}, 1]$. I also refer to commitment equilibria using $\widehat{\rho}$: with slight abuse of notation, I say that $(\widehat{\rho}, i)$ is a commitment equilibrium if $(\xi_{\widehat{\rho}}, i)$ is part of a commitment equilibrium, where $i : [\underline{\rho}, 1] \to \mathscr{I}$.

Let us now consider how the Receiver can induce the Sender to choose some experiment $\widehat{\rho} \in [\underline{\rho}, 1]$ while conducting an investigation $\iota \in \mathscr{I}$. Clearly, the Receiver must ensure that the Sender's payoff from choosing any $\widehat{\rho}' \neq \widehat{\rho}$ is lower than her payoff from choosing $\widehat{\rho}$ using an appropriate investigation strategy. Moreover, because any investigation following $\widehat{\rho}' \neq \widehat{\rho}$ is off the equilibrium path, it suffices to consider investigation strategies that maximally punish the Sender for choosing any experiment other than $\widehat{\rho}$. To induce the Sender to choose $\widehat{\rho}$, the Receiver must ensure that the Sender's payoff under $(\widehat{\rho}, \iota)$ is at least as high as her best possible deviation against a maximally punishing investigation, which is given by the Sender's *maxmin payoff*, denoted $V_{\mathrm{S}}^{\mathrm{maxmin}}$. The following lemma establishes the Sender's maxmin payoff.

Lemma 2. *The Sender's maxmin payoff is given by*

$$V_{\mathrm{S}}^{\mathrm{maxmin}} = \max\left\{0, \left(1 - \tfrac{\mu^* - \mu_0}{\mu^*} \tfrac{1}{\widehat{\rho}^{\mathrm{maxmin}}}\right) \tfrac{\rho_0 - \widehat{\rho}^{\mathrm{maxmin}}}{1 - \widehat{\rho}^{\mathrm{maxmin}}}\right\},$$

where $\widehat{\rho}^{\mathrm{maxmin}} = \max\left\{\underline{\rho}, \left(1 + \sqrt{\tfrac{\mu_0}{\mu^* - \mu_0} \tfrac{1-\rho_0}{\rho_0}}\right)^{-1}\right\} \in [\underline{\rho}, \rho_0)$.

To prove the result, for each $\widehat{\rho} \in [\underline{\rho}, 1]$, I first define the Sender's *value correspondence* associated with the Sender choosing the experiment $\widehat{\rho}$ as the correspondence $V_{\mathrm{S}}(\cdot|\widehat{\rho}) : [\rho, 1] \rightrightarrows \mathbb{R}$ that maps the Receiver's posterior belief about reliability to the set of payoffs that the Sender can attain given that the unreliable Sender and the Receiver best respond; i.e.,[30]

$$V_{\mathrm{S}}\left(\rho|\widehat{\rho}\right) := \mathrm{co}\left(\left\{v_{\mathrm{S}}(a) : a \in \underset{a' \in A}{\arg\max}\ \tilde{v}_{\mathrm{R}}(a'|\widehat{\rho},\rho))\right\}\right),$$

where $\tilde{v}_{\mathrm{R}}(a|\widehat{\rho},\rho) := \sum_{\theta\in\Theta} v_{\mathrm{R}}(a,\theta)[\rho\xi_{\widehat{\rho}}(m_1|\theta) + (1-\rho)]\mu_0(\theta)$ is the Receiver's interim payoff from choosing action $a \in A$ after the Sender has chosen an experiment $\widehat{\rho} \in [\underline{\rho},1]$ and the Receiver observes (ρ,m_1). Standard arguments (Aumann and Maschler, 1968; Kamenica and Gentzkow, 2011) mean that the minimal payoff for the Sender that can be induced by some investigation $\iota \in \mathscr{I}$ is given by the convex envelope of the function $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$ evaluated at the prior ρ_0,[31] denoted $\mathrm{vex}\underline{V}_{\mathrm{S}}(\widehat{\rho})$.[32] The Sender's maxmin payoff is then obtained by maximising $\mathrm{vex}\underline{V}_{\mathrm{S}}(\widehat{\rho})$ with respect to $\widehat{\rho}$.

Given the arguments above, the Receiver's commitment equilibrium payoff can be obtained by maximising the Receiver's expected payoff with respect to a candidate on-the-equilibrium-path experiment $\widehat{\rho} \in [\underline{\rho},1]$ and an investigation $\iota \in \mathscr{I}$ subject to the Sender attaining at least her maxmin payoff. The following lemma further simplifies the Receiver's problem by reducing the Receiver's choice of an investigation to a choice of a scalar. The lemma also shows that, on the equilibrium path, the Receiver either does not investigate the Sender or conducts a partially informative investigation that always reveals that the Sender is reliable but only sometimes reveals that the Sender is unreliable. Intuitively, the latter type of investigation is the most efficient way for the Receiver to give up information about reliability while increasing the Sender's payoff because the two players are only conflicted in their desires to reveal (or hide) the case when the Sender is unreliable.

[30] I denote the convex hull of a set as $\mathrm{co}(\cdot)$.

[31] The function $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$ is well defined because $V_{\mathrm{S}}(\cdot|\widehat{\rho})$ is non-empty- and compact-valued.

[32] I show in the appendix that the punishing investigation from the illustrative example—in which the Receiver either does not conduct an investigation or conducts a partially informative investigation that always reveals that the sender is unreliable but only sometimes reveal that that the sender is reliable—are indeed the investigations that induce the convex envelope of $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$.

Lemma 3. *Any commitment equilibrium payoff can be obtained with a class of investigations given by* $\{\iota^*(z) : z \in [\rho_0, 1]\} \subseteq \mathscr{I}$*, where*

$$\iota^*(z) := \begin{cases} \frac{\rho_0}{z}\delta_z + \frac{\rho_0}{z-\rho_0}\delta_0 & \textit{if } z \in (\rho_0, 1] \\ \delta_{\rho_0} & \textit{if } z = \rho_0 \end{cases}.$$

The lemma above means that the Receiver's commitment equilibrium payoff is the solution to the following problem:

$$\max_{(\widehat{\rho},z)\in[\underline{\rho},1]\times[\rho_0,1]} \int_0^1 V_{\mathrm{R}}(\rho|\widehat{\rho})\,\mathrm{d}\iota^*(z)(\rho) \tag{2}$$

$$\text{s.t.} \int_0^1 \max V_{\mathrm{S}}(\rho|\widehat{\rho})\,\mathrm{d}\iota^*(z)(\rho) \geq V_{\mathrm{S}}^{\mathrm{maxmin}},$$

where $V_{\mathrm{R}}(\cdot|\widehat{\rho}) := \max_{a'\in A}\ \tilde{v}_{\mathrm{R}}(a'|\widehat{\rho},\rho))$ is the Receiver's value function associated with Sender's experiment $\widehat{\rho} \in [\underline{\rho}, 1]$. In the appendix, I show that the solution to the problem above is given by (1). Moreover, I also show that, for any μ^* and μ_0 with $\mu^* > \mu_0$, there exist cutoffs for the prior belief $\rho_{0,1}, \rho_{0,2} \in [\underline{\rho}, 1]$ with $\rho_{0,1} < \rho_{0,2}$ such that the solution to the Receiver's problem, $(\widehat{\rho}^*, z^*)$, is given by:[33]

$$(\widehat{\rho}^*, z^*) = \begin{cases} \left(\underline{\rho}, 1\right) & \text{if } \rho_0 \in (0, \rho_{0,1}] \\ \left(\underline{\rho}, \frac{\rho_0}{V_{\mathrm{S}}^{\mathrm{maxmin}} + (1-\mu_0)\rho_0}\right) & \text{if } \rho_0 \in (\rho_{0,1}, \rho_{0,2}) \\ \left(\frac{\mu^*-\mu_0}{\mu^*}\frac{\rho_0}{1-V_{\mathrm{S}}^{\mathrm{maxmin}}}, \rho_0\right) & \text{if } \rho_0 \in [\rho_{0,2}, 1) \end{cases} \tag{3}$$

Observe that, when the Sender is sufficiently unreliable (i.e., $\rho_0 \leq \rho_{0,1}$),[34] the Receiver is able to attain his ideal outcome (i.e., the fully informative experiment under the fully revealing investigation). In contrast, when the Sender is sufficiently reliable (i.e., $\rho_0 \geq \rho_{0,2}$), the Receiver is willing to give up learning about the Sender's reliability completely in order to induce the Sender to choose a more informative

[33] In the appendix, I show that there exists $\rho_{0,0} \in (0, \rho_{0,1})$ such that $V_{\mathrm{S}}^{\mathrm{maxmin}} = 0$ for any $\rho_0 \in (0, \rho_{0,0}]$ and $V_{\mathrm{S}}^{\mathrm{maxmin}}$ is strictly positive for all $\rho_0 \in (\rho_{0,1}, 1)$.

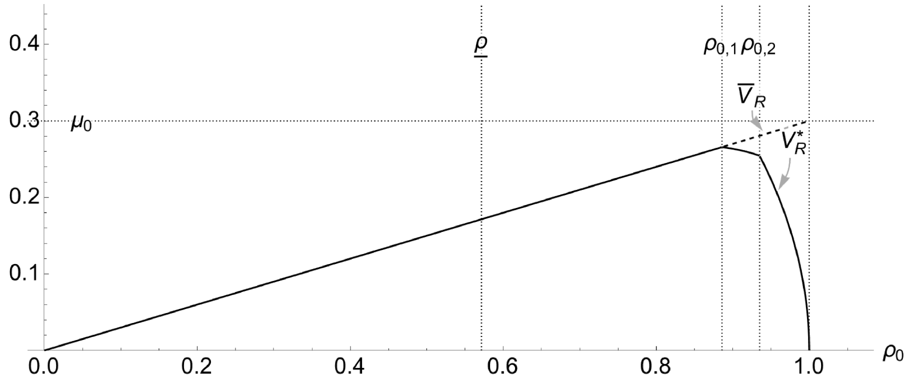
[34] Similar comparative statics results hold if I instead vary μ^* and μ_0: an increase in ρ_0 is analogous to an increase in μ^* and a decrease in μ_0.

experiment. At intermediate prior beliefs (i.e., $\rho_{0,1} < \rho_0 < \rho_{0,2}$), while the Receiver is able to induce the Sender to choose the fully informative experiment, he does so at the cost of not fully learning about the Sender's reliability.

The intuition for Theorem 2, and in particular (3), is the following. When the Receiver cannot investigate, the Sender chooses $\widehat{\rho} = \rho_0 < 1$ because doing so ensures that the Receiver takes action with probability one after observing m_1. Thus, when there is more doubt about the Sender's reliability (i.e., a lower ρ_0), the Sender is willing to choose a more informative experiment even without any investigation. In other words, the Sender has a stronger prior incentive to choose a more informative experiment when ρ_0 is low. Hence, when ρ_0 is low (i.e., $\rho_0 \leq \rho_{0,1}$), the Receiver need not give up the benefit of being able to learn the Sender's type and thus can obtain the ideal payoff. However, when ρ_0 is high (i.e., $\rho_0 > \rho_{0,1}$), the Sender has a weaker prior incentive to choose the fully informative experiment, and the Receiver must give up the benefit of being able to learn the Sender's type to induce the Sender to choose the fully informative experiment. When ρ_0 is sufficiently high (i.e., $\rho_0 > \rho_{0,2}$), the Receiver is willing to give up all the benefits from investigating. It turns out that the Receiver is willing to give up information about the Sender's reliability first—which, after all, is of second-order importance to the Receiver—before giving up information about the payoff-relevant state.

Figure 1 shows how the Receiver's commitment equilibrium varies with the prior belief about reliability ρ_0 when $\mu^* = 0.5$ and $\mu_0 = 0.3$ as in the example from Section 1.

Figure 1: Receiver's commitment equilibrium payoff, V_{R}^*.



Note that the Receiver's commitment equilibrium payoff is necessarily non-monotonic in ρ_0 because the Receiver's equilibrium payoffs when $\rho_0 = 0$ or when $\rho_1 = 1$ are zero. Figure 1 also shows that the Receiver's payoff is concave—i.e., the Receiver's commitment equilibrium payoff is greatest for intermediate prior belief about the Sender's reliability.[35]

4 Implementing ignorance via delegation

Let us now introduce a third player, whom I refer to as the *Third Party* (T, *it*), into the original game, and let the Third Party (instead of the Receiver) choose investigations about the Sender's reliability. I refer to this modified game as the *delegation game.* The timing is as in the no-commitment case of the original game, except that it is now the Third Party that chooses an investigation $\iota \in \mathscr{I}$ having observed the Sender's experiment, and the Receiver simply chooses his action after observing the Sender's message and the realisation of the investigation. The Third Party's preference is a linear combination of the Sender's and the Receiver's payoffs with weights $\lambda_j \in \mathbb{R}$ on player $j \in \{\mathrm{S},\mathrm{R}\}$'s payoff.[36]

I focus on the case where the Third Party is adversarial; i.e., $\lambda_\mathrm{S} < 0$. I therefore normalise the weight on the Sender's preference as $\lambda_\mathrm{S} = -1$ and let $\lambda \geq 0$ denote the weight on the Receiver's preference. I refer to a Third Party whose weights are $(\lambda_\mathrm{S}, \lambda_\mathrm{R}) = (-1, \lambda)$ as a λ*-balanced Third Party*. I refer to a 0-balanced Third Party as being *purely adversarial* and a ∞-balanced Third Party as being *purely Receiver-aligned*.[37] I call an equilibrium of the delegation game with a λ-balanced Third Party a λ*-equilibrium* and define it analogously to the no-commitment equilibrium of the original game. Importantly, the simplification in Lemma 1 applies with respect to equilibrium payoffs in the delegation game.

Let us first compare the two extreme types of a Third Party: the purely Receiver-

[35] One can further show that the prior belief about reliability ρ_0 that maximises the Receiver's commitment equilibrium lies in the interval $[\rho_{0,1}, \rho_{0,2}]$.

[36] With no restrictions on the Third Party's preference, the Receiver can attain the commitment outcome trivially by delegating to a Third Party whose payoff is constant.

[37] See the discussion in section 5 for the case when the Third Party is *purely Sender-aligned* (i.e., $\lambda_\mathrm{S} > 0$ and $\lambda_\mathrm{R} = 0$).

aligned and the purely adversarial Third party. The following proposition establishes that the Receiver is never worse off by delegating investigations to a purely adversarial Third Party; in fact, the Receiver strictly benefits from such adversarial delegation whenever the Sender is sufficiently reliable.

Proposition 1. *The Receiver prefers delegating investigations to a purely adversarial Third Party over a purely Receiver-aligned Third Party—strictly so if* $\rho_0 \in (\underline{\rho}, 1)$.

The proof follows readily from previous results. Notice first that, if the Third Party is purely Receiver aligned, then ∞-equilibrium payoffs correspond to no-commitment equilibrium payoffs in the original game. Thus, the Receiver's delegation equilibrium payoff is zero by Theorem 1. Moreover, since a purely adversarial Third Party maximally punishes the Sender for any choice of experiment, one can compute the Receiver's 0-equilibrium payoff by recalling from the proof of Lemma 2 how the Sender behaves when she expects to be maximally punished.

Can the Receiver do better by delegating to a λ-balanced Third Party with $\lambda \in (0,\infty)$? The next result, in particular, shows that there exists a unique $\lambda^* > 0$ that maximises the Receiver's equilibrium payoff in the delegation game. Moreover, when the Sender is sufficiently unreliable (i.e., ρ_0 is sufficiently low), the Receiver's λ^*-equilibrium coincides with his commitment equilibrium payoff.

Theorem 3. *For any* $\rho_0 \in (0,1)$*, there exists* $\lambda^*(\rho_0) > 0$ *such that the Receiver's and the Sender's* λ*-equilibrium payoff is given by*

$$V_{\mathrm{R}}^{\lambda} = \begin{cases} 0 & \text{if } \lambda < \lambda^*(\rho_0) \\ \frac{1}{1+\lambda^*(\rho_0)} \frac{\mu_0}{\mu^*} \rho_0 & \text{if } \lambda \geq \lambda^*(\rho_0) \end{cases}$$

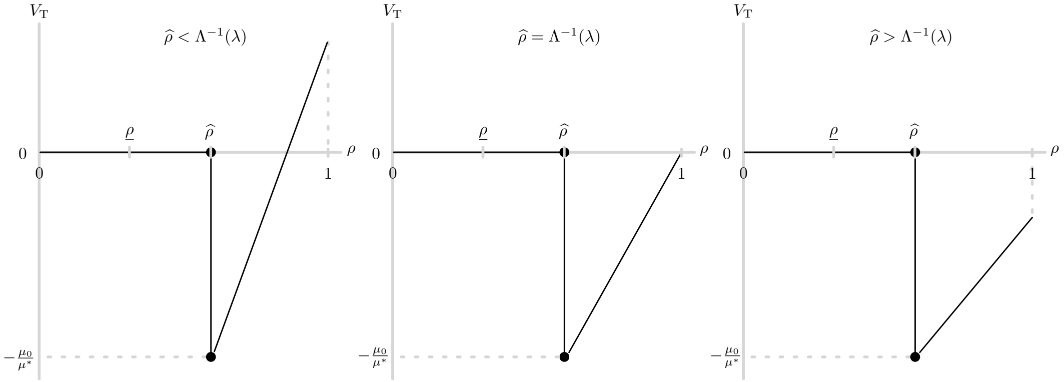
Hence, the Receiver strictly prefers to delegate investigations to a $\lambda^*(\rho_0)$*-balanced Third Party over any other* λ*-balanced Third Party. Moreover, whenever the prior belief that the Sender is reliable,* ρ_0*, is sufficiently low, the Receiver's* λ^**-equilibrium payoff coincides with the Receiver's commitment equilibrium payoff.*

To understand the intuition for the result, define a correspondence $V_{\mathrm{T}}^{\lambda} : [\underline{\rho}, 1] \rightrightarrows \mathbb{R}$ as the λ-balanced Third Party's value correspondence associated with Sender's experiment $\widehat{\rho} \in [\underline{\rho}, 1]$:[38]

$$V_{\mathrm{T}}^{\lambda}\left(\cdot|\widehat{\rho}\right) := -V_{\mathrm{S}}\left(\cdot|\widehat{\rho}\right) + \lambda\left\{V_{\mathrm{R}}\left(\cdot|\widehat{\rho}\right)\right\}.$$

Figure 2 depicts the λ-balanced Third Party's value correspondence.[39] Standard arguments mean that a λ-balanced Third Party's payoff from a sequentially rational investigation is given by the concave envelope of the correspondence $V_{\mathrm{T}}^{\lambda}(\cdot|\widehat{\rho})$. It is immediate from Figure 2 that: if $V_{\mathrm{T}}^{\lambda}(1|\widehat{\rho})$ is nonnegative (resp. nonpositive), then the λ-balanced Third Party's sequentially rational investigation is the fully revealing investigation (resp. the punishing investigation that minimises the Sender's payoffs).

Figure 2: Third Party's value correspondence payoff: $V_{\mathrm{T}}^{\lambda}(\cdot|\widehat{\rho})$.



Importantly, I show that, for any $\lambda > 0$, there exists $\widehat{\rho}_{\lambda}$ such that $V_{\mathrm{T}}^{\lambda}(1|\widehat{\rho}) \geq 0$ if and only if $\widehat{\rho} \leq \widehat{\rho}_{\lambda}$. Consequently, a λ-balanced Third Party finds it sequentially

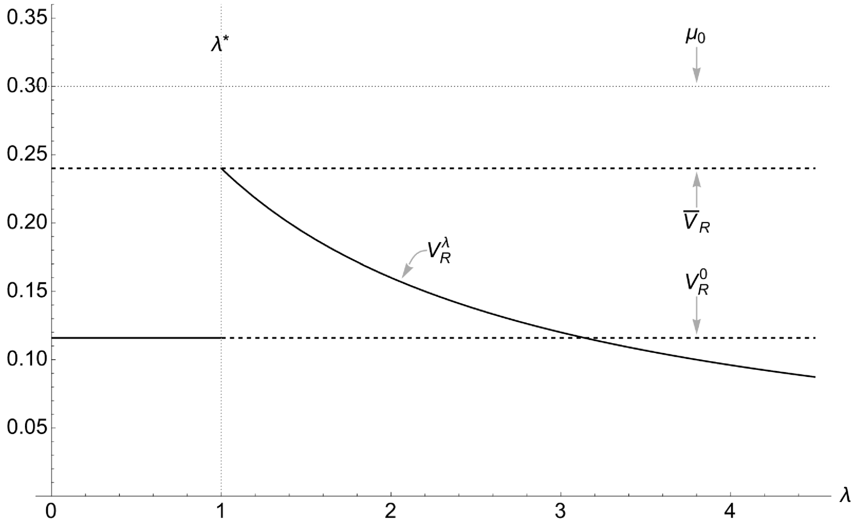
[38] Since V_{S} is a correspondence, the summation in the expression is a Minkowski sum.

[39] To understand the shape of a λ-balanced Third Party's value correspondence, recall that the Receiver strictly prefers to not take action for any posterior belief $\rho < \widehat{\rho}$. Hence, $V_{\mathrm{T}}(\rho|\widehat{\rho}) = \{0\}$ for all $\rho < \widehat{\rho}$. At $\rho = \widehat{\rho}$, the Receiver is indifferent between taking action and not taking action, and the Sender's payoff from the Receiver taking action is given by $\frac{\mu_0}{\mu^*} > 0$. Therefore, $V_{\mathrm{T}}(\widehat{\rho}|\widehat{\rho}) = [-\frac{\mu_0}{\mu^*}, 0]$. Since the Receiver strictly prefers to take action for any $\rho > \widehat{\rho}$, $V_{\mathrm{T}}(\rho|\widehat{\rho})$ is single-valued and, viewed as a function, $V_{\mathrm{T}}(\cdot|\widehat{\rho})$ is right-continuous at $\rho = \widehat{\rho}$. Recalling the definition of $\tilde{v}_{\mathrm{R}}$, both the Sender's and the Receiver's payoffs increase linearly with ρ.

rational to conduct the fully revealing investigation whenever the Sender chooses a more informative experiment than $\widehat{\rho}_\lambda$ and maximally punish the Sender for choosing a less informative experiment than $\widehat{\rho}_\lambda$. Since the Sender prefers a less informative experiment for any given investigation, against a λ-balanced Third Party, the Sender would never choose an experiment that is more informative than $\widehat{\rho}_\lambda$. The proof is completed by showing that the Sender does not find it profitable to deviate to a less informative experiment $\widehat{\rho}_\lambda$ and face maximal punishment. Since $\widehat{\rho}_\lambda$ depends on λ, by delegating investigations to a Third Party with an appropriate weight λ on the Receiver's preference, the Receiver can induce the Sender to choose the most informative experiment under the fully revealing investigation.

Figure 3 shows how the Receiver's λ-equilibrium payoff change with λ when $\mu_0 = 0.3$, $\mu^* = 0.5$ and $\rho_0 = 0.8$.

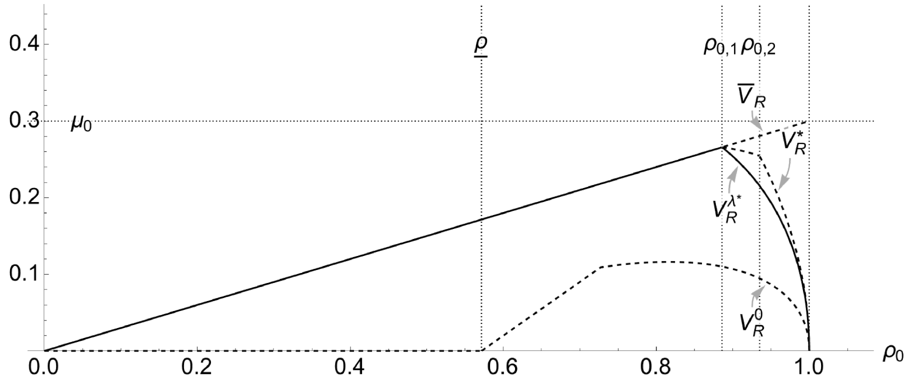
Figure 3: Receiver's λ-equilibrium payoff, $V_{\mathrm{R}}^{\lambda} := V_{\mathrm{R}}(\widehat{\rho}^{\lambda}, i^{\lambda}(\widehat{\rho}^{\lambda}))$.



When the weight on the Receiver's payoff is small ($\lambda < \lambda^* = \lambda^*(\rho_0)$), a λ-balanced third party behaves as a purely adversarial Third Party so that the Receiver's payoff is equal to his 0-equilibrium payoff. However, the Receiver's λ-equilibrium payoff "jumps" up at $\lambda = \lambda^*$ because the Third Party now finds it optimal to conduct the fully revealing investigation on the equilibrium path. Moreover, given the parametric assumptions, the Receiver's λ^*-equilibrium payoff equals the ideal payoff, $\overline{V}_{\mathrm{R}}$, which, in turn, equals the Receiver's commitment equilibrium payoff. As λ increases beyond λ^*, the Receiver's payoff decreases and converges to zero (i.e., the no-commitment equilibrium payoff) as λ becomes large.

Figure 4 compares the Receiver's: λ^*-equilibrium payoff (denoted $V_{\mathrm{R}}^{\lambda^*}$), 0-equilibrium payoff (V_{R}^0), commitment equilibrium payoff (V_{R}^*), and his ideal payoff ($\overline{V}_{\mathrm{R}}$); when $\mu_0 = 0.3$ and $\mu^* = 0.5$.

Figure 4: Receiver's optimal delegation equilibrium payoff, $V_{\mathrm{R}}^{\lambda^*}$.



The Receiver can obtain the commitment equilibrium payoff (that equals his ideal payoff) whenever $\rho_0 \leq \rho_{0,1}$. However, when $\rho_0 > \rho_{0,1}$, the Receiver's payoff is lower in the λ^*-equilibrium because it is not sequentially rational for any λ-balanced Third Party to conduct a partially revealing investigation of the form ι_z^* that the Receiver with the ability to commit to investigation strategies would choose. The difference between V_{R}^* and $V_{\mathrm{R}}^{\lambda^*}$ therefore represents the loss arising from the Third Party's inability to commit to ι_z^* on the equilibrium path. However, notice that the Receiver does better in λ^*-equilibrium than in 0-equilibrium (i.e., when the Third Party is purely adversarial). This difference arises because a purely adversarial Third Party conducts a maximally punishing investigation even on the equilibrium path.

5 Discussion

5.1 Interpretations of the results

Cross-examinations. Courts rely on witnesses to provide information about factual or technical matters concerning cases. A perennial worry, however, is that witnesses do not provide sufficient or reliable information. One prominent way courts deal with this concern is through cross-examinations in which a witness is interrogated by an attorney, usually from the opposing party, whose purported goal is to test the reliability of the witness (and consequently the reliability of the evidence provided by the witness). Cross-examination is an important feature of the US court system and has famously been described as the "greatest legal engine ever invented for the discovery of truth" (Wigmore, 1904).[40] Similarly, in the US, parties can challenge the admissibility of expert evidence through a *Daubert challenge*.[41] The consequence of being found unreliable can range from partial exclusion to a full exclusion of the witness evidence (i.e., impeachment). Between 2000 and 2021, there were 3,342 cases of Daubert challenges specifically against financial expert witnesses, and 43% of these challenges resulted in the partial or full exclusion of the expert (PricewaterhouseCoopers, 2022). The latter statistic, in particular, underscores the fact that outcomes of cross-examinations are not always predictable because they do not always lead to exclusions of the witness. It can also be the case that cross-examinations backfire and lead the court to believe the witness is more reliable than they had initially thought.

The features of cross-examination described above are consistent with how the Receiver in my model learns about the Sender's reliability using investigations. My results therefore highlight the role of cross-examination as not only a way to learn the reliability of witnesses, but also as a way for the court to obtain more information by inducing the parties to select more informative witnesses. In this light, my results have implications for the efficacy of cross-examination as an en-

[40] Wigmore (1904) was the dominant source of US evidentiary law until the codification of the Federal Rules of Evidence in 1975 (Friedman, 2009).

[41] A Daubert challenge is a type of motion to exclude expert witness testimony (scientific or otherwise) on the basis that it represents unqualified evidence (*Daubert v. Merrell Dow Pharmaceuticals, Inc.,* 509 U.S. 579, 1993; *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 1999).

gine for the discovery of truth. Theorem 1 implies, perhaps surprisingly, that a cross-examination conducted by a judge may not help the court obtain additional information. Theorems 1 and 2 together suggest that, for the judge to be an effective cross-examiner, he must paradoxically be able to commit to not discovering the truth about the witness' reliability.

In case the judge cannot commit, Proposition 1 suggests that delegating cross-examinations to an adversarial third party, as is in fact done in the US and other jurisdictions, allows the court to circumvent this commitment issue. Moreover, Theorem 3 suggest that there is a significant benefit in ensuring that the cross-examiner is not only adversarial but also cares about the discovery of truth. To this end, in the US, for example, prosecutors have a dual role as advocates seeking a conviction and as "ministers of justice" (Fisher, 1988), and the courts have also recognised that prosecutors have a special duty not to impede the truth (Gershman, 2001). In some jurisdictions (e.g., Germany), the court system is described as being inquisitorial (as opposed to adversarial as in the US) meaning that judges often direct the debates by asking questions. To the extent that a combination of the opposing party and an "inquisitorial judge" can be considered a balanced third party, Theorem 3 gives a reason to prefer an inquisitorial legal system over an adversarial system. Of course, the lesson from Theorem 1 applies—the judges must refrain from always finding out the truth about the reliability of witnesses.

The model predicts different kinds of cross-examinations depending on the cross-examiner's ability to commit—and in the case without commitment—the weight λ the cross-examiner places on the court arriving at just decisions. For example, the ideal cross-examiner who can commit would not cross-examine the witness when the prior belief that the witness is reliable is sufficiently high. While this prediction appears reasonable, the model also predicts that such a cross-examiner would be willing to hide that the witness is unreliable (with intermediate beliefs about reliability), which is perhaps less reasonable. If we instead assume that the cross-examiner cannot commit and is strongly adversarial (i.e., $\lambda < \lambda^*$), then we should expect the cross-examiner to be less willing to provide evidence that suggests that the witness is reliable. In contrast, a cross-examiner who also cares more about the court arriving at just decisions (i.e., $\lambda \geq \lambda^*$) would be willing to provide such evidence. While we sometimes observe that prosecutors provide evidence that supports the defendant's innocence, there appear to be fewer instances in which the prosecutor provides evidence that supports the reliability of the defendant's witness. To the extent that my model, which is a significant simplification of the complex legal system in real life, has some empirical content, the discussion above suggests that reality might be most consistent with the case in which the cross-examiner does not have the ability to commit and is strongly adversarial.

Audits. An audit of a piece of information, such as financial statements or investment appraisals, involves an examination of whether a particular method was followed to produce the information at hand. By interpreting the sender's choice of an experiment as a choice of such a method, an investigation into the sender's reliability can be thought of as a type of audit. Importantly, in my model, auditing is costless and is not about the veracity of the sender's message but rather about the sender's reliability type. Theorem 1 suggests that an unfettered audit when conducted by the receiver (or by an auditor whose incentives are aligned with that of the receiver) might not be beneficial because such an audit can result in the sender choosing a less informative method that negates the receiver's benefit from being able to identify unreliable information. Theorem 2 characterises how audits that are not always fully revealing can induce the sender to adopt a more informative method that provides the receiver with more information in equilibrium. Finally, Theorem 3 suggests how the receiver can implement such an auditing strategy by ensuring that the auditor balances his preference for the receiver and his antagonism toward the sender appropriately. I also note that delegating audits may alleviate the coordination problem that might arise if a group of investors (as opposed to a single investor) is considering whether to buy the seller's asset.

***Ad hominem* arguments.** Because investigations are about the nature of the sender, they can be thought of as examples of *ad hominem* (i.e., "to the person") counter-arguments against the Sender's *ad rem* (i.e., "to the point") arguments. Under this interpretation, one can think of the Sender as being, for example, a politician making statements about an issue, and the investigations as being about the politician (e.g., whether the politician is a flip-flopper) and not about the political issue itself. Such uses of *ad hominem* counter-arguments are prevalent in many debates.

In this light, one can interpret the results in this paper as concerning: (i) the extent to which *ad hominem* arguments are effective as counter-arguments by looking at the effect of investigations on the Sender's payoff, and (ii) the extent to which *ad hominem* arguments are productive by looking at the effect of investigations on the Receiver's payoff. The results from the delegation game show that *ad hominem* counter-arguments can be both effective and productive. The latter is perhaps surprising given that *ad hominem* arguments are often criticised for being fallacious. In the political context, one can think of a media outlet that opposes the politician as an example of an adversarial third party. My results demonstrate how such a media outlet can help the voters by inducing the politician to speak more truthfully. Moreover, the results also suggest that rules that prevent outlets from making *ad hominem* arguments may, in fact, harm the voters.

5.2 Extensions

Investigation strategies that can depend on the realisation of the Sender's experiment. In some cases, an investigation is chosen after the Receiver has observed the Sender's message; e.g., a cross-examiner is able to read the witness statement before cross-examining the witness. Formally, such a case corresponds to the investigation strategy being a function of both the Sender's choice of an experiment as well as the Sender's message. In the online appendix,[42] I show that the Receiver is no better off with this additional flexibility. The result follows from the fact that it remains sufficient for the Sender to make action recommendations even against investigation strategies that can be conditioned on message realisations.

Character witnesses. The US Federal Rules of Evidence 608 states that: "A witness's credibility may be [...] supported by testimony about the witness's reputation for having a character for truthfulness or untruthfulness..." One can think of testimony in support of the witness' credibility as information provided by a purely Sender-aligned Third Party; i.e., a (λ_S, λ_R)-balanced Third Party with $\lambda_S > 0$ and $\lambda_R = 0$. In the online appendix,[43] I show that while the Sender can benefit from having such a Third Party (strictly so in cases the Receiver would ignore the Sender's

[42] See Section 1 in the online appendix.

[43] See Section 2 in the online appendix.

message without any information, i.e., $\rho_0 < \underline{\rho}$), the Receiver does not benefit and his payoff is zero.

Sender who observes her type before choosing an experiment. Suppose the Sender observes her type before choosing her experiment. In this case, the Sender can potentially signal her type by choosing experiments according to her realised type. However, because the Sender cannot benefit from being identified as the unreliable type by the Receiver, the unreliable Sender would never choose an experiment that differs from the one that the reliable Sender chooses. It follows that there always exists an outcome-equivalent pooling equilibrium of the signalling version of the game. Thus, whether the sender knows her type before choosing an experiment would not affect the results.

Unreliable Sender who can observe the state. In some situations, it may be reasonable to assume that the unreliable Sender observes the realisation of the state before choosing how to manipulate:[44] e.g., a product reviewer might observe the quality of the product before deciding whether/how to (mis)communicate to the buyers about the quality for the manufacturer's benefit.

If the unreliable Sender can observe the state, then her messaging strategy is now a mapping $\sigma : \Theta \to \Delta M$. The change allows for the possibility that an (ξ, ι)-equilibrium exists in which σ is informative. While this complicates the analysis,[45] one can show that, for any (ξ, ι)-equilibrium in which σ is informative, there exists a slight perturbation of ι, $\iota' \in \mathscr{I}$, such that σ is not part of any (ξ, ι')-equilibrium. Consequently, even if the Sender is able to improve her payoff by choosing an informative σ for some ι, the Receiver can prevent such a strategy from being part of an equilibrium. Thus, it follows that both the Sender's and the Receiver's payoffs must be higher in any equilibrium in which σ is informative. But the same argument as when investigations can depend on the messages means that such an equilibrium does not exist. Thus, whether the Sender can observe θ is unimportant for the results.

[44] When the unreliable Sender can observe the state before choosing a message, she behaves exactly as the sender in Crawford and Sobel (1982).

[45] For example, it is no longer guaranteed that the unreliable Sender would only send messages that would induce the Receiver to take action when the message is known to have been drawn from ξ. Moreover, ξ and σ are no longer necessarily Blackwell ordered.

Receiver with limited commitment. In some situations, it may not be possible for the Receiver to condition the investigation on the Sender's experiment. For example, this might be because the Receiver must commit to an investigation before the Sender chooses an experiment (e.g., a regulator committing to a rule that applies to all regulated entities) or because the Sender's experiment is unobservable (e.g., the experiment is the Sender's private communication strategy). One can model these situations by assuming that the Receiver must commit to a constant investigation strategy. Whether the Receiver benefits from this limited commitment depends on whether the Sender can observe the result of the investigation prior to choosing the experiment.

To see why, suppose that the Receiver has committed to an investigation $\iota \in \mathscr{I}$. Consider first the timing in which the Sender observes the result of the investigation before choosing an experiment. Then, every possible ρ in the support of ι induces a subgame in which the Receiver cannot investigate the Sender, and the prior belief that the Sender is reliable is ρ. By Corollary 1, the Receiver's payoff in such a game is always zero. It follows that the Receiver is unable to benefit from limited commitment under this timing.

Suppose now that the Sender does not observe the result of the investigation before choosing an experiment. Since the Receiver can always implement a constant investigation strategy in the commitment case, the Receiver's payoff in this case must be weakly lower than the Receiver's commitment equilibrium payoff. A pertinent question is thus whether the Receiver benefits from investigations when he must commit to a single investigation. In the online appendix,[46] I show that the Receiver's equilibrium payoff, even in this case, is strictly positive for any interior prior belief about reliability, and that the optimal investigation has either two or three beliefs in its support.

Costly investigations. The results are robust to the addition of fixed cost of investigations to the model. Specifically, with a fixed cost of investigations, either the Receiver would not conduct an investigation (so that his payoff is zero) or the Receiver investigates as characterised in the case with costless investigations (and his payoff is lower by the fixed cost of investigating). With variable costs of invest-

[46] See Section 3 in the online appendix.

igations, the outcome would depend, in general, on the specification of the costs. In particular, it is possible to specify costs that make the Receiver's optimal investigation strategy (in the commitment case) sequentially rational for the Receiver.[47] Thus, my assumption that investigations are costless can be viewed as a way to ensure that investigation costs do not give the Receiver the ability to commit to ignorance.

6 Conclusion

I show in this paper that a decision-maker should not always learn the reliability of a strategic information source using a persuasion game in which the conflicted source of information is sometimes unreliable. Strategically committing to avoiding learning about reliability allows the decision-maker to obtain more information by trading off information about reliability, which is of second-order importance, with information about the payoff-relevant state, which is of first-order importance. Even when the decision-maker is unable to commit to ignorance, I show that he can obtain more information—sometimes as much as when he can commit—by delegating investigations to someone partially adversarial to the sender and partially aligned with the receiver. My results shed light on the efficacy of cross-examination, audits, and *ad hominem* arguments.

[47] For example, this is possible using costs that depend both on the choice of the sender's experiment and the investigations.

References

Ambrus, Attila, Eduardo M. Azevedo, and Yuichiro Kamada. 2013. "Hierarchical cheap talk." *Theoretical Economics*, 8: 233–261.

Aumann, Robert J., and Michael Maschler. 1968. "Repeated games of incomplete information: the zero-sum extensive case." *Mathematica*.

Austen-Smith, David. 1994. "Strategic Transmission of Costly Information." *Econometrica*, 62(4): 955.

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

Bergemann, Dirk, and Stephen Morris. 2019. "Information Design: A Unified Perspective." *Journal of Economic Literature*, 57(1): 44–95.

Blackwell, David. 1953. "Equivalent Comparisons of Experiments." *The Annals of Mathematical Statistics*, 24: 265–272.

Blume, A., O.J. Board, and K. Kawamura. 2007. "Noisy talk." *Theoretical Economics*, 2(4): 395–440.

Border, Kim C., and Joel Sobel. 1987. "Samurai Accountant: A Theory of Auditing and Plunder." *The Review of Economic Studies*, 54: 525.

Che, Yeon Koo, and Navin Kartik. 2009. "Opinions as incentives." *Journal of Political Economy*, 117: 815–860.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica*, 50(6): 1431–1451.

Dewatripont, Mathias, and Jean Tirole. 1999. "Advocates." *Journal of Political Economy*, 107: 1–39.

Dworczak, Piotr, and Alessandro Pavan. 2022. "Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion." *Econometrica*, 90: 2017–2051.

Dziuda, Wioletta, and Christian Salas. 2018. "Communication with Detectable Deceit." *SSRN Electronic Journal*.

Ederer, Florian, and Weicheng Min. 2022. "Bayesian Persuasion with Lie Detection." National Bureau of Economic Research Working Paper 30065.

Fershtman, Chaim, and Kenneth L Judd. 1987. "Equilibrium Incentives in Oligopoly." *American Economic Review*, 77: 927–940.

Fisher, Stanley Z. 1988. "In Search of the Virtuous Prosecutor: A Conceptual Framework." *American Journal of Criminal Law*, 197.

Forges, Françoise. 2020. "Games with Incomplete Information: From Repetition to Cheap Talk and Persuasion." *Annals of Economics and Statistics*, 137: 3–30.

Frechette, Guillaume R., Alessandro Lizzeri, and Jacopo Perego. 2019. "Rules and Commitment in Communication: An Experimental Analysis." *SSRN Electronic Journal*.

Friedman, Richard D. 2009. *The Yale Biographical Dictionary of American Law.* Yale University Press.

Fudenberg, Drew, and Jean Tirole. 1991. "Perfect Bayesian equilibrium and sequential equilibrium." *Journal of Economic Theory*, 53: 236–260.

Gentzkow, Matthew, and Emir Kamenica. 2017*a*. "Bayesian persuasion with multiple senders and rich signal spaces." *Games and Economic Behavior*, 104: 411–429.

Gentzkow, Matthew, and Emir Kamenica. 2017*b*. "Competition in persuasion." *Review of Economic Studies*, 84: 300–322.

Gerardi, Dino, and Leeat Yariv. 2008. "Costly expertise." *American Economic Review*, 98: 187–193.

Gershman, Bennett L. 2001. "The Prosecutor's Duty to Truth." *Gerogetown Journal of Legal Ethics*, 14: 309–354.

Golman, Russell, David Hagmann, and George Loewenstein. 2017. "Information avoidance." *Journal of Economic Literature*, 55: 96–135.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani. 2009. "Mediation, arbitration and negotiation." *Journal of Economic Theory*, 144: 1397–1420.

Ivanov, Maxim. 2010*a*. "Communication via a strategic mediator." *Journal of Economic Theory*, 145: 869–884.

Ivanov, Maxim. 2010*b*. "Informational control and organizational design." *Journal of Economic Theory*, 145: 721–751.

Ivanov, Maxim, and Alex Sam. 2022. "Cheap talk with private signal structures." *Games and Economic Behavior*, 132: 288–304.

Kamenica, Emir. 2019. "Bayesian Persuasion and Information Design." *Annual*

Review of Economics, 11: 249–272.

Kamenica, Emir, and Matthew Gentzkow. 2011. “Bayesian persuasion.” *American Economic Review*, 101(6): 2590–2615.

Krähmer, Daniel. 2021. “Information Design and Strategic Communication.” *American Economic Review: Insights*, 3: 51–66.

Levkun, Aleksandr. 2022. “Communication with Strategic Fact-Checking.”

Lichtig, Avi. 2020. “Adversarial Disclosure.”

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin. 2022. “Persuasion via Weak Institutions.” *Journal of Political Economy*, 130(10): 2705–2730.

McAdams, David. 2012. “Strategic ignorance in a second-price auction.” *Economics Letters*, 114: 83–85.

Min, Daehong. 2021. “Bayesian persuasion under partial commitment.” *Economic Theory*, 72: 743–764.

Mookherjee, Dilip, and Ivan Png. 1989. “Optimal Auditing, Insurance, and Redistribution.” *The Quarterly Journal of Economics*, 104: 399.

Onuchic, Paula. 2022. “Advisors with Hidden Motives.”

Özdogan, Ayça. 2016. “A Survey of strategic communication and persuasion.” *Bogazici Journal*, 30.

Posner, Richard A. 1999. “An Economic Approach to the Law of Evidence.” *Law Review*, 51: 1477–1546.

PricewaterhouseCoopers. 2022. “Daubert challenges to financial expert.”

Roesler, Anne-Katrin, and Balázs Szentes. 2017. “Buyer-Optimal Learning and Monopoly Pricing.” *American Economic Review*, 107: 2072–2080.

Rogoff, Kenneth. 1985. “The Optimal Degree of Commitment to an Intermediate Monetary Target.” *The Quarterly Journal of Economics*, 100: 1169–1189.

Sadakane, Hitoshi, and Yin Chi Tam. 2022. “Cheap talk and Lie detection.” *Working Paper*.

Schelling, T.C. 1960. *The Strategy of Conflict.* Harvard University Press.

Shin, Hyun Song. 1998. “Adversarial and Inquisitorial Procedures in Arbitration.” *The RAND Journal of Economics*, 29: 378.

Sklivas, Steven D. 1987. “The Strategic Choice of Managerial Incentives.” *The RAND Journal of Economics*, 18: 452.

Sobel, Joel. 2013. "Giving and Receiving Advice." *Advances in Economics and Econometrics: Tenth World Congress* Vol. 1, Chapter 10, 305–341. Cambridge University Press.

Taylor, Curtis R., and Huseyin Yildirim. 2011. "Subjective performance and the value of blind evaluation." *Review of Economic Studies*, 78: 762–794.

Townsend, Robert M. 1979. "Optimal contracts and competitive markets with costly state verification." *Journal of Economic Theory*, 21: 265–293.

Vickers, John. 1985. "Delegation and the Theory of the Firm." *The Economic Journal*, 95: 138.

Wigmore, John Henry. 1904. *Treatise on the Anglo-American System of Evidence in Trials at Common Law.* Vol. II. 1st ed., Little, Brown, and Co.

Ye, Minlei. 2021. "Theory of Auditing Economics: A Review of Analytical Auditing Research." *SSRN Electronic Journal.*

A Appendix

A.1 Formal definitions of commitment and no-commitment equilibria

Let $\sigma : \Xi \times \mathscr{I} \to \Delta M$ denote the unreliable Sender's messaging strategy, $\alpha : \Xi \times \mathscr{I} \times [0,1] \times M \to \Delta A$ denote the Receiver's action rule, and $\mu : \Xi \times \mathscr{I} \times [0,1] \times M \to \Delta\Theta$ denote the Receiver's belief map. Given a tuple $(\xi, \iota, \sigma_{\xi,\iota}, \alpha_{\xi,\iota})$,[48] let $V_j(\cdot)$ denote player $j \in \{\mathrm{S},\mathrm{R}\}$'s associated ex ante payoff; i.e.,

$$V_j(\cdot) := \sum_{\theta\in\Theta} \int_0^1 \sum_{m\in M} \sum_{a\in A} v_j(\cdot)\alpha_{\xi,\iota}(a|\rho,m)[\rho\xi(m|\theta) + (1-\rho)\sigma_{\xi,\iota}(m)]\mathrm{d}\iota(\rho)\mu_0(\theta).$$

For the commitment case, let us call the game that follows after the Receiver has chosen an investigation strategy i as an i-commitment game. A PBE of a i-commitment game is a tuple $(\xi, \sigma_{\cdot,i(\cdot)}, \alpha_{\cdot,i(\cdot)}, \mu_{\cdot,i(\cdot)})$ that satisfies the following conditions: (i) for each $\xi' \in \Xi$, the belief map $\mu_{\xi',i(\xi')}(\cdot) : [0,1] \times M \to \Delta\Theta$ is derived by updating μ_0 using the signal structure $\rho\xi' + (1-\rho)\sigma_{\xi',i(\xi')} : \Theta \to \Delta M$ via Bayes rule whenever possible, i.e., for all $(\rho, m) \in \mathrm{supp}(i(\xi')) \times M$,

$$\mu_{\xi',i(\xi')}(\cdot|\rho,m) = \frac{\left[\rho\xi'(m|\cdot) + (1-\rho)\,\sigma_{\xi',i(\xi')}(m)\right]\mu_0(\cdot)}{\sum_{\theta\in\Theta}\left[\rho\xi'(m|\theta) + (1-\rho)\,\sigma_{\xi',i(\xi')}(m)\right]\mu_0(\theta)}$$

whenever the denominator is strictly positive, and otherwise $\mu_{\xi',i(\xi')}(\cdot|\rho,m) = \mu_0(\cdot)$;[49] (ii) the Receiver's action rule $\alpha_{\cdot,i(\cdot)}$ is optimal given $\mu_{\cdot,i(\cdot)}$, i.e., for all $(\xi', \rho, m) \in \Xi \times \mathrm{supp}(i(\xi')) \times M$,

$$\mathrm{supp}\left(\alpha_{\xi',i(\xi')}(\rho,m)\right) \subseteq \arg\max_{a\in A} \sum_{\theta\in\Theta} v_{\mathrm{R}}(a,\theta)\,\mu_{\xi',i(\xi')}(\theta|\rho,m);$$

(iii) the Unreliable Sender's messaging strategy $\sigma_{\cdot,i(\cdot)}$ is incentive compatible given

[48] I define $\sigma_{\xi,\iota}(\cdot) := \sigma(\xi,\iota)$; $\alpha_{\xi,\iota}$ and $\mu_{\xi,\iota}$ are defined analogously.

[49] The assumption that off-equilibrium-path belief equals the prior belief reflects the idea of "no signalling what you don't know" (Fudenberg and Tirole, 1991) since off-equilibrium-path messages can only be sent by the unreliable Sender who neither observes the realised state nor the realisation of the experiment.

$\alpha_{\cdot,i(\cdot)}$, i.e., for all $\xi' \in \Xi$,

$$\operatorname{supp}\left(\sigma_{\xi',i(\xi')}\right) \subseteq \underset{m \in M}{\arg\max} \int_0^1 \sum_{a \in A} v_{\mathrm{S}}\left(a\right) \alpha_{\xi',i(\xi')}\left(a|\rho,m\right) \frac{1-\rho}{1-\rho_0} \mathrm{d}i\left(\xi'\right)\left(\rho\right); \qquad \text{(IC)}$$

(iv) the Sender's experiment is sequentially rational given i, $\sigma_{\cdot,i(\cdot)}$ and $\alpha_{\cdot,i(\cdot)}$, i.e., ξ maximises $V_{\mathrm{S}}(\xi', i(\xi'), \sigma_{\xi',i(\xi')}, \alpha_{\xi,i(\xi')})$ with respect to $\xi' \in \Xi$. I refer to a tuple $(\xi, i, \sigma_{\cdot,i(\cdot)}, \alpha_{\cdot,i(\cdot)}, \mu_{\cdot,i(\cdot)})$ as a *commitment equilibrium* if $(\xi, \sigma_{\cdot,i(\cdot)}, \alpha_{\cdot,i(\cdot)}, \mu_{\cdot,i(\cdot)})$ is a PBE of the i-commitment equilibrium that maximises Receiver's ex ante payoff among any PBE of any i'-commitment game.[50]

For the no-commitment case, call the game that follows after the Sender chooses an experiment $\xi \in \Xi$ a *ξ-no-commitment game.* A PBE of a ξ-no-commitment game is a tuple $(\iota, \sigma_{\xi,\cdot}, \alpha_{\xi,\cdot}, \mu_{\xi,\cdot})$ that satisfies the following conditions: (i) for each $\iota' \in \mathscr{I}$, the belief map $\mu_{\xi,\iota'}(\cdot) : [0,1] \times M \to \Delta\Theta$ is derived by updating μ_0 using the signal structure $\rho\xi + (1-\rho)\sigma_{\xi,\iota'} : \Theta \to \Delta M$ via Bayes rule whenever possible, and otherwise $\mu_{\xi,\iota'}(\cdot|\rho,m) = \mu_0(\cdot)$; (ii) Receiver's action rule $\alpha_{\xi,\cdot}$ is optimal given $\mu_{\xi,\cdot}$; (iii) the Unreliable Sender's messaging strategy $\sigma_{\xi,\cdot}$ is incentive compatible given $\alpha_{\xi,\cdot}$; (iv) the Receiver's investigation ι is sequentially rational given $\sigma_{\xi,\cdot}$ and $\rho_{\xi,\cdot}$, i.e., ι maximises $V_{\mathrm{R}}(\xi, \iota', \sigma_{\xi,\iota'}, \alpha_{\xi,\iota'})$ with respect to $\iota' \in \mathscr{I}$. I refer to a tuple $(\xi, \iota, \sigma_{\xi,\cdot}, \alpha_{\xi,\cdot}, \mu_{\xi,\cdot})$ as a *no-commitment equilibrium* if $(\iota, \sigma_{\xi,\cdot}, \alpha_{\xi,\cdot}, \mu_{\xi,\cdot})$ is a PBE of the ξ-no-commitment game that maximises Sender's ex ante payoff among any PBE of any ξ'-no-commitment game.

Observe that the Receiver's commitment-equilibrium payoff is an upper bound on the Receiver's no-commitment-equilibrium payoff.[51]

A.2 Proof of Theorem 1

I first establish that given any Sender's experiment and the unreliable Sender's strategy, a more informative investigation results in a mean-preserving spread of

[50] Allowing the Receiver to select an i-equilibrium given any $i(\cdot)$ (as implied by the definition above) ensures that a solution to the problem above exists.

[51] To see this, take any arbitrary set of ξ'-no-commitment equilibria $(\iota_{\xi'}, \sigma_{\xi',\cdot}, \alpha_{\xi',\cdot}, \mu_{\xi',\cdot})$ for each $\xi' \in \Xi$ and construct an investigation strategy as $i(\xi') = \iota_{\xi'}$ for all $\xi' \in \Xi$. Observe that the Receiver's commitment equilibrium payoff must be weakly higher than in any i-commitment equilibrium.

induced beliefs. In proving the lemma, I allow the unreliable Sender's strategy to potentially depend on θ; i.e., the unreliable Sender can observe the realisation of state before choosing the message.

Lemma 4. *Fix any $\xi, \sigma \in \Xi$ and $\iota, \iota' \in \mathscr{I}$ such that ι is a mean-preserving spread of ι'. Then, the distribution of posterior beliefs about the state induced by (ξ, σ, ι) is a mean-preserving spread of that induced by (ξ, σ, ι').*

Proof. Let N be a finite message space about the Sender's reliability that is sufficiently rich and let $\eta, \eta' : T \to \Delta N$ be the signals that induce distributions of posterior beliefs ι and ι' respectively. By Blackwell's theorem, η' is a garbling of η; i.e., there exists $g : N \to \Delta N$ such that

$$\eta'(n|t) = \sum_{n'} g\left(n|n'\right) \eta(n'|t).$$

Given $\xi, \sigma \in \Xi$ and η, the Receiver's posterior joint belief about the state and Sender's type is given by

$$\begin{aligned}
\nu(\theta, r|m,n) &= \frac{\xi(m|\theta)\, \eta(n|r)\, \rho_0 \mu_0(\theta)}{\sum_{\theta' \in \Theta} \left[\xi(m|\theta')\, \eta(n|r)\, \rho_0 + \sigma(m|\theta')\, \eta(n|u)\,(1-\rho_0)\right] \mu_0(\theta')}, \\
\nu(\theta, u|m,n) &= \frac{\sigma(m|\theta)\, \eta(n|u)\,(1-\rho_0)\, \mu_0(\theta)}{\sum_{\theta' \in \Theta} \left[\xi(m|\theta')\, \eta(n|r)\, \rho_0 + \sigma(m|\theta')\, \eta(n|u)\,(1-\rho_0)\right] \mu_0(\theta')}.
\end{aligned}$$

Thus, the Receiver's marginal belief about the state given any (m,n) in the support is

$$\begin{aligned}
\mu(\theta|m,n) &:= \sum_{t \in T} \nu(\cdot, t|m,n) \\
&= \frac{\left[\xi(m|\theta)\, \eta(n|r)\, \rho_0 + \sigma(m|\theta)\, \eta(n|u)\,(1-\rho_0)\right] \mu_0(\theta)}{\sum_{\theta' \in \Theta} \left[\xi(m|\theta')\, \eta(n|r)\, \rho_0 + \sigma(m|\theta')\, \eta(n|u)\,(1-\rho_0)\right] \mu_0(\theta')}.
\end{aligned}$$

Hence, beliefs about the state are updated after observing (m,n) as if the pair was drawn according to signal structure $\pi^{\xi,\sigma,\eta} \in \Xi$ such that

$$\pi^{\xi,\sigma,\eta}(m,n|\theta) := \xi(m|\theta)\, \eta(n|r)\, \rho_0 + \sigma(m|\theta)\, \eta(n|u)\,(1-\rho_0).$$

Using the fact that η' is a garbling of η,

$$\begin{aligned}
&\pi^{\xi,\sigma,\eta'}(m,n|\theta,t)\\
&=\xi(m|\theta)\left(\sum_{n'\in N} g\left(n|n'\right)\eta\left(n'|r\right)\right)\rho_0+\sigma(m|\theta)\left(\sum_{n'\in N} g\left(n|n'\right)\eta\left(n'|u\right)\right)(1-\rho_0)\\
&=\sum_{n'\in N} g\left(n|n'\right)\underbrace{\left[\xi(m|\theta)\eta\left(n'|r\right)\rho_0+\sigma(m|\theta)\eta\left(n'|u\right)(1-\rho_0)\right]}_{=\pi^{\xi,\sigma,\eta}(m,n'|\theta)}.
\end{aligned}$$

If we let $f: M\times N\to\Delta(M\times N)$ be

$$f\left(m',n'|m,n\right):=\mathbb{1}_{\{m'=m\}}g\left(n'|n\right)$$

we realise that $\pi^{\xi,\sigma,\eta'}$ is a garbling of $\pi^{\xi,\sigma,\eta}$ via f. ■

Theorem 1 follows almost immediately from the previous lemma.

Theorem 1. *In any no-commitment equilibrium, the Sender chooses her optimal experiment when she is known to be fully reliable, $\widehat{\rho}^*=1$, and the Receiver always conducts the fully revealing investigation. The Sender's and Receiver's no-commitment equilibrium payoffs are $\rho_0\frac{\mu_0}{\mu^*}$ and zero, respectively.*

Proof of Theorem 1. The previous lemma, together with Blackwell's theorem, implies that the sequentially rational investigation for the Receiver is always fully revealing in any no-commitment equilibrium; i.e., $i(\cdot)=\overline{\imath}:=\rho_0\delta_1+(1-\rho_0)\delta_0$. By condition (i),

$$\mu\left(1|\xi',\overline{\imath},0,\cdot\right)=\mu_0,\ \mu\left(1|\xi',\overline{\imath},1,\cdot\right)=\mu^{\xi}(1|\cdot).$$

Moreover, condition (iii) is moot because the unreliable Sender's payoff is always zero. Moreover,

$$V_{\mathrm{S}}\left(\xi',\overline{\imath},\sigma,\alpha\right)=\rho_0\sum_m\sum_{\theta'}\alpha(1|\xi,\overline{\imath},1,m)\,\xi\left(m|\theta'\right)\mu_0\left(\theta'\right).$$

Observe that the Sender's problem given above is equivalent to the Sender's problem in the case when $\rho_0=1$ except for the coefficient ρ_0 in Sender's payoff.

Thus, the Sender-optimal-fully-reliability experiment, $\widehat{\rho} = 1$, is optimal for the Sender. ■

A.3 Proof of Lemma 1

Given any $\pi \in \Xi$, let μ^{π} denote the posterior belief induced by signal structure π. First, observe that $\mu^{\sigma} = \mu_0$ because the unreliable Sender's strategy cannot depend on the realised state θ. It follows that the unreliable Sender would only send messages that would induce the Receiver to take action if the message is known to have been sent by the reliable Sender; i.e., $\text{supp}(\sigma) \subseteq M_1^{\xi} := \{m \in M : \mu^{\xi}(m) \geq \mu^*\}$. Because $\mu^{\sigma}(m) = \mu_0 < \mu^*$, for any $m \in M_1^{\xi}$, there exists a threshold belief about the Sender's type, $\overline{\rho}_m \in [0,1]$, at which the Receiver is indifferent between the two actions after observing m, and would only be willing to take action if $\rho \geq \overline{\rho}_m$. The unreliable Sender's payoff is (weakly) decreasing in $\overline{\rho}_m$. Moreover, the unreliable Sender's incentive compatibility requires that $\iota([\overline{\rho}_m, \overline{\rho}_{m'}]) = 0$ for all such $m, m' \in M_1^{\xi}$. Because pooling messages in M_1^{ξ} (in both ξ and σ) results in a threshold that is a weighted average of the cutoffs $\{\overline{\rho}_m\}_{m \in M_1^{\xi}}$, it follows that the unreliable Sender's payoff remains unchanged. This, in turn, implies that both the Sender's and the Receiver's ex ante payoffs are unaffected when pooling messages in M_1^{ξ}. That the unreliable Sender never sends messages in $\{m \in M : \mu^{\xi}(m) < \mu^*\}$ means that these messages can also be pooled without affecting payoffs. Let $\tilde{\xi}$ and $\tilde{\sigma}$ denote the strategies after pooling. The proof is completed by showing that the pooling of messages does not affect the choice of investigation. Specifically, I show that if there exists $\tilde{\iota} \in \mathscr{I}$ such that the Receiver's $(\tilde{\xi}, \tilde{\iota})$-equilibrium payoff is different from his $(\tilde{\xi}, \iota)$-equilibrium payoff, then one can construct a $(\xi, \tilde{\iota})$-equilibrium in which the Receiver's payoff is the same as in the $(\tilde{\xi}, \tilde{\iota})$-equilibrium.

Recall that, given any $(\xi, \iota) \in \Xi \times \mathscr{I}$, a tuple $(\sigma_{\xi,\iota}, \alpha_{\xi,\iota}, \mu_{\xi,\iota})$ is a (ξ, ι)-equilibrium if it is a PBE of the game induced by (ξ, ι). For brevity, I write $(\sigma, \alpha, \mu) \equiv (\sigma_{\xi,\iota}, \alpha_{\xi,\iota}, \mu_{\xi,\iota})$. Fix some $(\xi, \iota) \in \Xi \times \mathscr{I}$. Denote the unreliable and reliable Sender's interim payoffs from sending message $m \in M$ given (σ, α), respectively, as follows:

$$V_u(m|\xi,\iota,\sigma,\alpha) := \int_{[0,1]} \alpha(1|\xi,\iota,\rho,m) \frac{1-\rho}{1-\rho_0} \mathrm{d}\iota(\rho),$$
$$V_r(m|\xi,\iota,\sigma,\alpha) := \int_{[0,1]} \alpha(1|\xi,\iota,\rho,m) \frac{\rho}{\rho_0} \mathrm{d}\iota(\rho),$$

where $\frac{1-\rho}{1-\rho_0}\iota(\rho)$ (resp. $\frac{\rho}{\rho_0}\iota(\rho)$) is the probability that posterior belief ρ is induced when the Sender is unreliable (resp. reliable). These two interim payoffs combine to give the Sender's ex ante payoff from (σ,α):

$$\begin{aligned} &V_{\mathrm{S}}(\xi,\iota,\sigma,\alpha) \\ &= \rho_0 \sum_{m\in M} V_r(m|\xi,\iota,\sigma,\alpha)\,\xi(m) + (1-\rho_0)\int V_u(m|\xi,\iota,\sigma,\alpha) \sum_{m\in M} \sigma(m|\xi,\iota), \end{aligned}$$

where $\xi(m) := \sum_{\theta\in\Theta} \xi(m|\theta)\mu_0(\theta)$. The Receiver's payoff is

$$\begin{aligned} &V_{\mathrm{R}}(\xi,\iota,\sigma,\alpha) \\ &= \frac{1-\mu^*}{\mu^*}\mu_0 \sum_{m\in M} \int_{[0,1]} \alpha(1|\rho,\sigma,m)\,\rho x^{\xi}(m)\,\mathrm{d}\iota(\rho) \\ &\quad + \frac{1-\mu^*}{\mu^*}\mu_0 \sum_{m\in M} \int_{[0,1]} \alpha(1|\rho,\sigma,m)(1-\rho)\frac{\mu_0-\mu^*}{\mu_0(1-\mu^*)}\sigma(m)\,\mathrm{d}\iota(\rho), \end{aligned}$$

where

$$x^{\xi}(\cdot) := \xi(\cdot|1) - \xi(\cdot|0)\frac{1-\mu_0}{\mu_0}\frac{\mu^*}{1-\mu^*} \in \Delta M.$$

Define

$$M_1^{\xi} := \left\{ m \in \operatorname{supp}(\xi) : x^{\xi}(m) \geq 0 \right\},\ M_0^{\xi} := \left\{ m \in \operatorname{supp}(\xi) : x^{\xi}(m) < 0 \right\}.$$

Observe that, for any $m \in \operatorname{supp}(\xi)$,

$$\mu^{\xi}(m) \geq \mu^* \Leftrightarrow x^{\xi}(m) \geq 0;$$

and M_a^{ξ} represents the set of messages that can induce the Receiver to choose action

$a \in A$ when the Receiver's belief about the Sender's reliability is $\rho = 1$. Define

$$\overline{\rho}\left(m|\xi,\sigma\right) := \frac{\frac{\mu^{*}-\mu_0}{\mu_0(1-\mu^{*})}\sigma\left(m\right)}{\frac{\mu^{*}-\mu_0}{\mu_0(1-\mu^{*})}\sigma\left(m\right)+x^{\xi}\left(m\right)}.$$

Then, for any $m_1 \in M_1^{\xi}$,

$$\mu^{\rho\xi+(1-\rho)\sigma}\left(1|m_1\right) \geq \mu^{*} \Leftrightarrow \rho \geq \overline{\rho}\left(m|\xi,\sigma\right);$$

and for any $m_0 \in M_0^{\xi}$,

$$\mu^{\rho\xi+(1-\rho)\sigma}\left(1|m_0\right) < \mu^{*}\ \forall \rho \in [0,1].$$

It follows that, for any $m_1 \in M_1^{\xi}$ and $m_0 \in M_0^{\xi}$,

$$\begin{aligned}
&V_u\left(m_1|\xi,\iota,\sigma,\alpha\right) && (4)\\
&= \int_{[0,1]} \left[\mathbb{1}_{(\overline{\rho}(m_1|\xi,\sigma),1]}\left(\rho\right) + \alpha\left(1|\overline{\rho}\left(m_1|\xi,\sigma\right),m_1\right)\mathbb{1}_{\{\overline{\rho}(m_1|\xi,\sigma)\}}\left(\rho\right)\right]\frac{1-\rho}{1-\rho_0}\mathrm{d}\iota\left(\rho\right),\\
&V_r\left(m_1|\xi,\iota,\sigma,\alpha\right) && (5)\\
&= \int_{[0,1]} \left[\mathbb{1}_{(\overline{\rho}(m_1|\xi,\sigma),1]}\left(\rho\right) + \alpha\left(1|\overline{\rho}\left(m_1|\xi,\sigma\right),m_1\right)\mathbb{1}_{\{\overline{\rho}(m_1|\xi,\sigma)\}}\left(\rho\right)\right]\frac{\rho}{\rho_0}\mathrm{d}\iota\left(\rho\right),\\
&V_u\left(m_0|\xi,\iota,\sigma,\alpha\right)\\
&= 0 = V_r\left(m_0|\xi,\iota,\sigma,\alpha\right).
\end{aligned}$$

The following lemma shows that pooling messages do not affect equilibrium payoffs.

Lemma 5. *Fix $(\xi,\iota) \in \Xi \times \mathscr{I}$. Suppose (σ,α,μ) is a (ξ,ι)-equilibrium with strictly positive Sender ex ante payoff. There exists a tuple $(\sigma^{*},\alpha^{*},\mu^{*})$ that is a*

$(\tilde{\xi},\iota)$-equilibrium such that, for some $\tilde{m}_1 \in M_1^\xi$ and $\tilde{m}_0 \in M_0^\xi$, we have

$$\tilde{\xi}\left(m|\cdot\right) := \begin{cases} \xi\left(M_1^\xi|\cdot\right) & \text{if } m=\tilde{m}_1 \\ \xi\left(M_0^\xi|\cdot\right) & \text{if } m=\tilde{m}_0 \\ 0 & \text{otherwise} \end{cases}, \ \sigma^*\left(m|\cdot\right) := \mathbb{1}_{\{m=\tilde{m}_1\}},$$

and

$$V_j\left(\xi,\iota,\sigma,\alpha\right) = V_j\left(\tilde{\xi},\iota,\sigma^*,\alpha^*\right)\ \forall j \in \{\mathrm{S},\mathrm{R}\}.$$

Proof. That the Sender's ex ante payoff is strictly positive implies that the unreliable Sender's interim payoff must be strictly positive, $M_1^\xi, M_0^\xi \neq \varnothing$, and $\mathrm{supp}(\sigma) \subseteq M_1^\xi$. Fix some $\tilde{m}_1 \in \mathrm{supp}(\sigma) \cap M_1^\xi$ and define $\tilde{\xi}$ and σ^* as given in the statement of the lemma. Define

$$\overline{\rho}_{\min}\left(\xi,\sigma\right) := \min_{m\in\mathrm{supp}(\sigma)\cap M_1^\xi} \overline{\rho}\left(m|\xi,\sigma\right),\ \overline{\rho}_{\max}\left(\xi,\sigma\right) := \max_{m\in\mathrm{supp}(\sigma)\cap M_1^\xi} \overline{\rho}\left(m|\xi,\sigma\right).$$

Let $m_{\min}$ and $m_{\max}$ be such that $\overline{\rho}_{\min} = \overline{\rho}_{m_{\min}}$ and $\overline{\rho}_{\max} = \overline{\rho}_{m_{\max}}$.

Consider the case when $\mathrm{supp}(\sigma) = M_1^\xi$. Then, pooling messages in M_1^ξ to $\tilde{m}_1$ results in a threshold reliability belief, $\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*)$, that is a weighted average of the thresholds under (ξ,σ); i.e.,

$$\begin{aligned}\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*) &= \sum_{m_1\in M_1^\xi} \frac{\frac{\mu^*-\mu_0}{\mu_0(1-\mu^*)}\sigma\left(m_1\right)+x^\xi\left(m_1\right)}{\sum_{m_1'\in M_1^\xi}\left(\frac{\mu^*-\mu_0}{\mu_0(1-\mu^*)}\sigma\left(m_1'\right)+x^\xi\left(m_1'\right)\right)}\overline{\rho}\left(m_1|\xi,\sigma\right) \\ &\in \left[\overline{\rho}_{\min}\left(\xi,\sigma\right),\overline{\rho}_{\max}\left(\xi,\sigma\right)\right].\end{aligned}$$

For any $m_1, m_1' \in M_1^\xi$ such that $\overline{\rho}_{m_1} = \overline{\rho}_{m_1'} < 1$, the unreliable Sender's incentive compatibility implies that

$$\begin{aligned}0 &= V_u\left(m_1|\xi,\iota,\sigma,\alpha\right) - V_u\left(m_1'|\xi,\iota,\sigma,\alpha\right) \\ &= \int_{[0,1]}\left(\alpha\left(1|\overline{\rho}_{m_1},m_1\right)-\alpha\left(1|\overline{\rho}_{m_1'},m_1'\right)\right)\frac{1-\overline{\rho}_{m_1}}{1-\rho_0}\iota\left(\{\overline{\rho}_{m_1}\}\right).\end{aligned}$$

Hence, if $\iota(\overline{\rho}_{m_1}) > 0$, we must have $\alpha(1|\xi,\iota,\overline{\rho}_{m_1},m_1) = \alpha(1|\xi,\iota,\overline{\rho}_{m_1'},m_1')$. If, instead, we have $\overline{\rho}_{m_1'} > \overline{\rho}_{m_1}$, then the unreliable Sender's incentive compatibility implies that

$$\begin{aligned} 0 &= V_u\left(m_1|\xi,\iota,\sigma,\alpha\right) - V_u\left(m_1'|\xi,\iota,\sigma,\alpha\right) \\ &= \int_{\left(\overline{\rho}_{m_1},\overline{\rho}_{m_1'}\right)} \frac{1-\rho}{1-\rho_0}\mathrm{d}\iota\left(\rho\right) + \alpha\left(1|\xi,\iota,\overline{\rho}_{m_1},m_1\right)\frac{1-\overline{\rho}_{m_1}}{1-\rho_0}\iota\left(\left\{\overline{\rho}_{m_1}\right\}\right) \\ &\quad + \left[1-\alpha\left(1|\overline{\rho}_{m_1'},m\right)\right]\frac{1-\overline{\rho}_{m_1'}}{1-\rho_0}\iota\left(\left\{\overline{\rho}_{m_1'}\right\}\right). \end{aligned}$$

Above implies that (i) $\iota((\overline{\rho}_{m_1},\overline{\rho}_{m_1'}))=0$, (ii) if $\iota(\{\overline{\rho}_{m_1}\})>0$, then $\alpha(1|\xi,\iota,\overline{\rho}_{m_1},m_1)=0$; (iii) if $\overline{\rho}_{m_1'}<1$ and $\iota(\{\overline{\rho}_{m_1'}\})>0$, then $\alpha(1|\xi,\iota,\overline{\rho}_{m_1'},m)=1$.

If $\overline{\rho}_{\min}(\xi,\sigma)=\overline{\rho}_{\max}(\xi,\sigma)$, then $\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*)=\overline{\rho}_{\min}(\xi,\sigma)$. It cannot be that $\overline{\rho}_{\min}(\xi,\sigma)=1$ in this case because that would imply that $x^{\xi}(m_1)=0$ for all $m_1\in M_1^{\xi}$ so that the unreliable Sender's incentive compatibility requires $\iota(1)=1$; but this contradicts that $\iota\in\mathscr{I}$. Hence,

$$\alpha^*\left(1|\cdot,\tilde{m}_1\right)=\alpha^*\left(1|\cdot,m_{\min}\right)$$

and clearly, players payoffs remain unchanged. Now suppose that $\overline{\rho}_{\min}(\xi,\sigma)<\overline{\rho}_{\max}(\xi,\sigma)$. Then, we must have

$$\iota\left(\left(\overline{\rho}_{\min}\left(\xi,\sigma\right),\overline{\rho}_{\max}\left(\xi,\sigma\right)\right)\right)=0$$

and that if $\iota(\overline{\rho}_{\min})>0$, then $\alpha(1|\xi,\iota,\overline{\rho}_{\min},m_1)=0$ for any $m_1\in M_1^{\xi}$ such that $\overline{\rho}_{m_1}=\overline{\rho}_{\min}$. Since $\overline{\rho}_{\max}<1$ (otherwise $V_u(m_{\max}|\xi,\iota,\sigma,\alpha)=0$, which contradicts the earlier observation that the unreliable Sender's interim payoff is strictly positive), if $\iota(\overline{\rho}_{\max})>0$, then $\alpha(1|\xi,\iota,\overline{\rho}_{\max},m_1)=1$ for any $m_1\in M_1^{\xi}$ such that $\overline{\rho}_{m_1}=\overline{\rho}_{\max}$. Hence, I can let

$$\alpha^*\left(1|\rho,\tilde{m}_1\right)=\mathbb{1}_{\left\{\rho>\overline{\rho}\left(\tilde{m}_1|\tilde{\xi},\sigma^*\right)\right\}}\left(\rho\right)$$

so that

$$
\begin{aligned}
& V_u\left(\tilde{m}_1|\tilde{\xi},\iota,\sigma^*,\alpha^*\right) \\
& = \int_{(\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*),1]} \frac{1-\rho}{1-\rho_0}\mathrm{d}\iota\left(\rho\right) \\
& \quad + \underbrace{\alpha^*\left(1|\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*),\tilde{m}_1\right)}_{=0}\frac{1-\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*)}{1-\rho_0}\iota\left(\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*)\right) \\
& = \int_{(\overline{\rho}(\tilde{m}_1|\tilde{\xi},\sigma^*),1]} \frac{1-\rho}{1-\rho_0}\mathrm{d}\iota\left(\rho\right) = \int_{(\overline{\rho}_{\min},1]} \frac{1-\rho}{1-\rho_0}\mathrm{d}\iota\left(\rho\right) \\
& = V_u\left(m_{\min}|\xi,\iota,\sigma,\alpha\right);
\end{aligned}
$$

i.e., players payoffs are again unchanged from pooling messages in $\mathrm{supp}(\sigma)\cap M_1^{\xi}$ to $\tilde{m}_1$.

Suppose now there exists $m'' \in M_1^{\xi}\backslash\mathrm{supp}(\sigma)$ such that $x^{\xi}(m'')>0$. Then, by the unreliable Sender's incentive compatibility, it must be that $V_u(m)=1$ for all $m\in\mathrm{supp}(\sigma)$. By the argument above, pooling messages in $\mathrm{supp}(\sigma)\cap M_1^{\xi}$ would not affect the Sender's incentives. Moreover, putting weights on $M_1^{\xi}\backslash\mathrm{supp}(\sigma)$ to $\tilde{m}_1$ can only lower $\tilde{\rho}$, which, in turn, can only weakly improve the payoffs. Since V_u is already ideal at one, it follows that pooling messages would not alter the players' payoff. Finally, suppose there exists $m''\in M_1^{\xi}\backslash\mathrm{supp}(\sigma)$ such that $x^{\xi}(m'')=0$. Because pooling m'' would not alter $\tilde{\rho}$, players' payoffs remain unchanged. To ensure that $(\sigma^*,\alpha^*,\mu^*)$ is a $(\tilde{\xi},\iota)$-equilibrium, I can specify off-path μ^* to equal μ_0 and ensure that α^* is optimal for the Receiver given μ^*. ■

Remark 1. Suppose (σ,α,μ) and (σ',α',μ') are both (ξ,ι)-equilibria, the players' equilibrium payoffs are equal if

$$
\overline{\rho}_{\min}(\xi,\sigma)=\overline{\rho}_{\min}\left(\xi,\sigma'\right),\ \overline{\rho}_{\max}(\xi,\sigma)=\overline{\rho}_{\max}\left(\xi,\sigma'\right).
$$

I now show that pooling messages using the lemma above would not affect the choice of an investigation.

Lemma 6. *Fix $(\xi,\iota)\in\Xi\times\mathscr{I}$. Suppose (σ,α,μ) is a (ξ,ι)-equilibrium and* *let $(\sigma^*, \alpha^*, \mu^*)$ be a $(\tilde{\xi}, \iota)$-equilibrium derived via the previous lemma. Suppose there exists $\tilde{\iota} \in \mathscr{I}$ such that players' $(\tilde{\xi}, \tilde{\iota})$-equilibrium payoffs strictly positive and are different from their payoff under $(\sigma^*, \alpha^*, \mu^*)$. Then, there exists a $(\xi, \tilde{\iota})$-equilibrium in which the players' payoffs are the same as in the $(\tilde{\xi}, \tilde{\iota})$-equilibrium.*

Proof. By the previous lemma,

$$V_{\mathrm{S}}(\xi, \iota, \sigma, \alpha) = V_{\mathrm{S}}\left(\tilde{\xi}, \iota, \sigma^*, \alpha^*\right),\ V_{\mathrm{R}}(\xi, \iota, \sigma, \alpha) = V_{\mathrm{R}}\left(\tilde{\xi}, \iota, \sigma^*, \alpha^*\right).$$

Suppose there exists $\tilde{\iota} \in \mathscr{I}$ and a tuple $(\tilde{\sigma}, \tilde{\alpha}, \tilde{\mu})$ that is a $(\tilde{\xi}, \tilde{\iota})$-equilibrium such that

$$V_{\mathrm{R}}\left(\tilde{\xi}, \tilde{\iota}, \tilde{\sigma}, \tilde{\alpha}\right) \neq V_{\mathrm{R}}\left(\tilde{\xi}, \iota, \sigma^*, \alpha^*\right) \text{ or } V_{\mathrm{S}}\left(\tilde{\xi}, \iota, \sigma^*, \alpha^*\right) \neq V_{\mathrm{S}}\left(\tilde{\xi}, \tilde{\iota}, \tilde{\sigma}, \tilde{\alpha}\right) > 0.$$

The goal is to construct a tuple $(\widehat{\sigma}, \widehat{\alpha}, \widehat{\mu})$ that is a $(\xi, \tilde{\iota})$-equilibrium such that

$$V_j(\xi, \tilde{\iota}, \widehat{\sigma}, \widehat{\alpha}) = V_j\left(\tilde{\xi}, \tilde{\iota}, \tilde{\sigma}, \tilde{\alpha}\right)\ \forall j \in \{\mathrm{S}, \mathrm{R}\}.$$

First, observe that it must be that $\tilde{\sigma}(\cdot) = \sigma^*(\cdot)$ so that

$$\overline{\rho}_{\tilde{m}_1}\left(\tilde{\xi}, \tilde{\sigma}\right) = \overline{\rho}_{\tilde{m}_1}\left(\tilde{\xi}, \sigma^*\right).$$

Writing $\overline{\rho} \equiv \overline{\rho}_{\tilde{m}_1}(\tilde{\xi}, \tilde{\sigma})$, for each $m_1 \in M_1^{\xi}$, let

$$\begin{aligned} \widehat{\sigma}(m_1) &= \frac{x^{\xi}(m_1)}{\frac{1-\overline{\rho}}{\overline{\rho}} \frac{\mu^* - \mu_0}{\mu_0(1-\mu^*)}} \\ \widehat{\alpha}_{(\overline{\rho}, \widehat{\sigma})}(m_1) &= \tilde{\alpha}_{(\overline{\rho}, \tilde{\sigma})}(\tilde{m}_1), \end{aligned}$$

where I write $\alpha_{(\rho,\sigma)}(m) \equiv \alpha(1|\rho, \sigma, m_1)$. By construction, I have $\overline{\rho}_{m_1}(\xi, \hat{\sigma}) = \overline{\rho}$

for all $m_1 \in M_1^{\xi}$. Consider the Sender's payoff first:

$$\begin{aligned}
&V_{\mathrm{S}}\left(\xi, \tilde{\imath}, \widehat{\sigma}, \widehat{\alpha}\right)\\
&= \sum_{m_1 \in M_1^{\xi}} \int_{[0,1]} \left[\mathbb{1}_{(\overline{\rho},1]}\left(\rho\right) + \widehat{\alpha}_{(\overline{\rho},\widehat{\sigma})}(m_1)\mathbb{1}_{\{\overline{\rho}\}}\left(\rho\right)\right]\left[\xi\left(m_1\right)\rho + \left(1-\rho\right)\right]\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&= \int_{[0,1]} \mathbb{1}_{(\overline{\rho},1]}\left(\rho\right)\left[\xi\left(M_1^{\xi}\right)\rho + \left(1-\rho\right)\right]\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&\qquad + \sum_{m_1 \in M_1^{\xi}} \widehat{\alpha}_{(\overline{\rho},\widehat{\sigma})}(m_1)\left[\xi\left(m_1\right)\overline{\rho} + \left(1-\overline{\rho}\right)\right]\tilde{\imath}\left(\{\overline{\rho}\}\right)\\
&= \int_{(\overline{\rho},1]} \left[\tilde{\xi}\left(\tilde{m}_1\right)\rho + \left(1-\rho\right)\right]\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&\qquad + \tilde{\alpha}_{(\overline{\rho},\tilde{\sigma})}\left(\tilde{m}_1\right)\left[\tilde{\xi}\left(\tilde{m}_1\right)\overline{\rho} + \left(1-\overline{\rho}\right)\right]\tilde{\imath}\left(\{\overline{\rho}\}\right)\\
&= V_{\mathrm{S}}\left(\tilde{\xi}, \tilde{\imath}, \tilde{\sigma}, \tilde{\alpha}\right).
\end{aligned}$$

Now consider the Receiver's payoff:

$$\begin{aligned}
&V_{\mathrm{R}}\left(\xi, \tilde{\imath}, \hat{\sigma}, \widehat{\alpha}\right)\\
&= \frac{\mu_0\left(1-\mu^*\right)}{\mu^*} \int_{[0,1]} \sum_{m_1 \in M_1^{\xi}} \widehat{\alpha}_{(\rho,\widehat{\sigma})}\left(m_1\right)\rho x^{\xi}\left(m_1\right)\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&\qquad + \frac{\mu_0\left(1-\mu^*\right)}{\mu^*} \int_{[0,1]} \sum_{m_1 \in M_1^{\xi}} \widehat{\alpha}_{(\rho,\widehat{\sigma})}\left(m_1\right)\frac{\left(1-\rho\right)\left(\mu_0-\mu^*\right)}{\mu_0\left(1-\mu^*\right)}\widehat{\sigma}\left(m_1\right)\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&= \frac{1-\mu^*}{\mu^*}\mu_0 \int_{(\overline{\rho},1]} \left(\rho x^{\tilde{\xi}}\left(\tilde{m}_1\right) + \left(1-\rho\right)\frac{\mu_0-\mu^*}{\mu_0\left(1-\mu^*\right)}\right)\mathrm{d}\tilde{\imath}\left(\rho\right)\\
&\qquad + \frac{1-\mu^*}{\mu^*}\mu_0\tilde{\alpha}_{(\overline{\rho},\tilde{\sigma})}\left(\tilde{m}_1\right)\left(\overline{\rho}x^{\tilde{\xi}}\left(\tilde{m}_1\right) + \left(1-\overline{\rho}\right)\frac{\mu_0-\mu^*}{\mu_0\left(1-\mu^*\right)}\right)\tilde{\imath}\left(\{\overline{\rho}\}\right)\\
&= V_{\mathrm{R}}\left(\tilde{\xi}, \tilde{\imath}, \tilde{\sigma}, \tilde{\alpha}\right).
\end{aligned}$$

Observe that I can appropriately define $\widehat{\sigma}$, $\widehat{\alpha}$ and $\widehat{\mu}$ to ensure that $(\hat{\sigma}, \widehat{\alpha}, \widehat{\mu})$ is a $(\xi, \tilde{\imath})$-equilibrium that yields the same payoffs for the players as $(\tilde{\xi}, \tilde{\imath}, \tilde{\sigma}, \tilde{\alpha})$. ■

Let us now prove Lemma 1.

Lemma 1. *Any strictly positive commitment equilibrium payoffs are attainable* *with an experiment $\xi \in \Xi$ such that for some $m_0, m_1 \in M$,* $\operatorname{supp}(\xi) = \{m_0, m_1\}$*,* $\xi(m_1|1) = 1$*,* $\xi(m_1|0) \leq 1 - \frac{\mu^* - \mu_0}{\mu^*(1-\mu_0)}$*, and the unreliable Sender always sends m_1.*

Proof. By the previous two lemmata, given any (ξ, ι)-equilibrium with strictly positive Sender payoff, there exists a payoff equivalent $(\tilde{\xi}, \iota)$-equilibrium. Moreover, if there exists a $(\xi, \tilde{\iota})$-equilibrium with different payoffs (such that Sender's payoff is still strictly positive), there exists $(\tilde{\xi}, \tilde{\iota})$-equilibrium that obtains the same payoffs. Thus, it is without loss to focus on the equivalence class of Ξ given by $\tilde{\xi}$, which, in turn, implies that I can focus on $\operatorname{supp}(\xi) = \{m_0, m_1\}$ for some $m_0, m_1 \in M$ and σ such that $|\operatorname{supp}(\sigma)| = 1$. We may assume that the unreliable Sender always sends m_1 and $M_1^\xi = \{m_1\}$. Then, for any $\iota \in \mathscr{I}$,

$$\begin{aligned}
&V_S(\xi, \iota, \overline{\sigma}, \alpha) \\
&= \int \left[\mathbb{1}_{\left(\overline{\rho}_{m_1}, 1\right]}(\rho) + \mathbb{1}_{\left\{\overline{\rho}_{m_1}\right\}}(\rho)\, \alpha\left(1 | \overline{\rho}_{m_1}, \overline{\sigma}\right) \right] \left[\rho \xi(m_1) + (1-\rho)\right] \mathrm{d}\iota(\rho),
\end{aligned}$$

where

$$\overline{\rho}_{m_1} = \frac{\frac{1-\mu_0}{\mu_0} \frac{\mu^*}{1-\mu^*} - 1}{\frac{1-\mu_0}{\mu_0} \frac{\mu^*}{1-\mu^*} - 1 + \xi(m_1|1) - \xi(m_1|0) \frac{1-\mu_0}{\mu_0} \frac{\mu^*}{1-\mu^*}}.$$

Since $M_1^\xi = \{m_1\}$, it must be that

$$x^\xi(m_1) > 0 \Leftrightarrow \xi(m_1|1) > \xi(m_1|0) \frac{1-\mu_0}{\mu_0} \frac{\mu^*}{1-\mu^*}.$$

Observe that $\xi(m_1) = \xi(m_1|1)\mu_0 + \xi(m_1|0)(1-\mu_0)$ is increasing in $\xi(m_1|1)$ and $\overline{\rho}_{m_1}$ is decreasing in $\xi(m_1|1)$, so that the Sender's payoff is increasing in $\xi(m_1|1)$ and, moreover, larger $\xi(m_1|1)$ relaxes the constraint on $\xi(m_1|0)$. It follows that $\xi(m_1|1) = 1$. This, together with the fact that $M_1^\xi = \{m_1\}$ implies $\xi(m_1|0) \leq 1 - \underline{\rho}$. ■

Remark 2. Lemma 6 means that the simplification applies leaves the payoffs of a Third Party whose preference is a linear combination of the Sender's and the Receiver's preferences unchanged. Hence, 1 is applicable to delegation equilibrium payoffs.

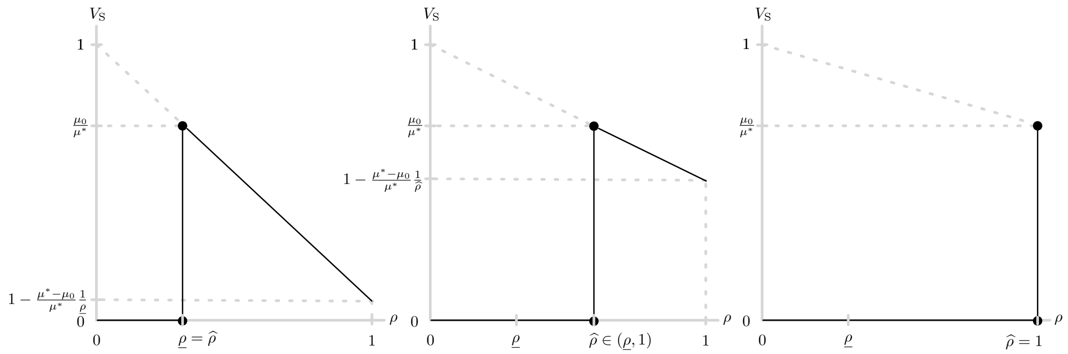
A.4 Proof of Lemma 2

The Sender's value correspondence given any experiment $\widehat{\rho} \in [\underline{\rho}, 1]$ is given by

$$V_{\mathrm{S}}(\rho|\widehat{\rho}) = \begin{cases} \{0\} & \text{if } \rho < \widehat{\rho} \\ \left[0, \frac{\mu_0}{\mu^*}\right] & \text{if } \rho = \widehat{\rho} \\ \left\{1 - \frac{\mu^* - \mu_0}{\mu^*} \frac{\rho}{\widehat{\rho}}\right\} & \text{if } \rho > \widehat{\rho} \end{cases}.$$

Figure 5 depicts $V_{\mathrm{S}}(\cdot|\widehat{\rho})$ for three different types of experiments: the fully informative experiment (i.e., $\widehat{\rho} = \underline{\rho}$), a partially informative experiment (i.e., $\widehat{\rho} \in (\underline{\rho}, 1)$), and the Sender-preferred Bayesian persuasion experiment (i.e., $\widehat{\rho} = 1$).

Figure 5: Sender's value correspondence given experiment $\widehat{\rho}$: $V_{\mathrm{S}}(\cdot|\widehat{\rho})$.



Observe that the Sender's value correspondence is single-valued except at $\rho = \widehat{\rho}$ and is otherwise affine. Standard arguments (Aumann and Maschler, 1968; Kamenica and Gentzkow, 2011) mean that the minimal payoff for the Sender is given by the convex envelope of $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$ (evaluated at the prior ρ_0). The minimal payoff is induced by the investigation, $i^{\min} : [\underline{\rho}, 1] \to \mathscr{I}$, that I call the *punishing investigation strategy*, and is given by

$$i^{\min}(\widehat{\rho}) := \begin{cases} \frac{\rho_0 - \widehat{\rho}}{1-\widehat{\rho}} \delta_1 + \frac{1-\rho_0}{1-\widehat{\rho}} \delta_{\widehat{\rho}} & \text{if } \rho_0 \geq \underline{\rho} \text{ and } \widehat{\rho} \in (\underline{\rho}, \rho_0) \\ \delta_{\rho_0} & \text{otherwise} \end{cases}. \tag{6}$$

As in the introductory example, in the punishing investigation strategy, the Receiver either chooses not to investigate or conducts a partially revealing investigation that always reveals that the Sender is unreliable but only sometimes reveal that the Sender is reliable. Let us now prove Lemma 2.

Lemma 2. *The Sender's maxmin payoff is given by*

$$V_{\mathrm{S}}^{\text{maxmin}} = \max\left\{0, \left(1 - \frac{\mu^*-\mu_0}{\mu^*}\frac{1}{\widehat{\rho}^{\text{maxmin}}}\right)\frac{\rho_0-\widehat{\rho}^{\text{maxmin}}}{1-\widehat{\rho}^{\text{maxmin}}}\right\},$$

where $\widehat{\rho}^{\text{maxmin}} = \max\left\{\underline{\rho}, \left(1+\sqrt{\frac{\mu_0}{\mu^*-\mu_0}\frac{1-\rho_0}{\rho_0}}\right)^{-1}\right\} \in [\underline{\rho}, \rho_0)$.

Proof. Standard arguments (Aumann and Maschler, 1968; Kamenica and Gentzkow, 2011) mean that the minimal payoff for the Sender that can be induced by some investigation is given by the convex envelope of the function $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$ evaluated at the prior ρ_0.[52] given by

$$\operatorname{vex}\min V_{\mathrm{S}}(\cdot|\widehat{\rho})(\rho_0) = \begin{cases} \left(1 - \frac{\mu^*-\mu_0}{\mu^*}\frac{1}{\widehat{\rho}}\right)\frac{\rho_0-\widehat{\rho}}{1-\widehat{\rho}} & \text{if } \rho_0 \in (0,\widehat{\rho}] \\ 0 & \text{if } \rho_0 \in (\widehat{\rho},1) \end{cases}$$

Suppose $\rho_0 < \underline{\rho}$. Then, not investigating ensures zero payoff for the Sender from choosing for any $\widehat{\rho} \in [\underline{\rho}, 1]$. Hence, $V_{\mathrm{S}}^{\text{maxmin}} = 0$. Suppose instead that $\rho_0 \geq \underline{\rho}$. If $\widehat{\rho} \geq \rho_0$, once again, not investigating ensures that Sender's payoff is zero. If $\widehat{\rho} \in [\underline{\rho}, \rho_0)$, then the Sender's problem is

$$\max_{\widehat{\rho}\in[\underline{\rho},\rho_0)} \left(1 - \frac{\mu^*-\mu_0}{\mu^*}\frac{1}{\widehat{\rho}}\right)\frac{\rho_0-\widehat{\rho}}{1-\widehat{\rho}},$$

which is solved by $\widehat{\rho}^{\text{maxmin}}$ given in the lemma. ■

Remark 3. If $\rho_0 \leq \rho_{0,0} := \frac{\mu^*-\mu_0}{\mu^*}\frac{1}{1-(2-\mu^*)\mu_0}$, then $\widehat{\rho}^{\text{maxmin}} = \underline{\rho}$. The maxmin experiment, $\widehat{\rho}^{\text{maxmin}}$, is strictly increasing in $\rho_0 \in (\rho_{0,1}, 1]$ while $V_{\mathrm{S}}^{\text{maxmin}}$ is also strictly increasing in $\rho_0 \in [\underline{\rho}, 1]$. Moreover, $\lim_{\rho_0\to 1}\widehat{\rho}^{\text{maxmin}} = 1$ and $\lim_{\rho_0\to 1} V_{\mathrm{S}}^{\text{maxmin}} = \frac{\mu_0}{\mu^*}$.

[52] The function $\min V_{\mathrm{S}}(\cdot|\widehat{\rho})$ is well defined because $V_{\mathrm{S}}(\cdot|\widehat{\rho})$ is non-empty- and compact-valued.

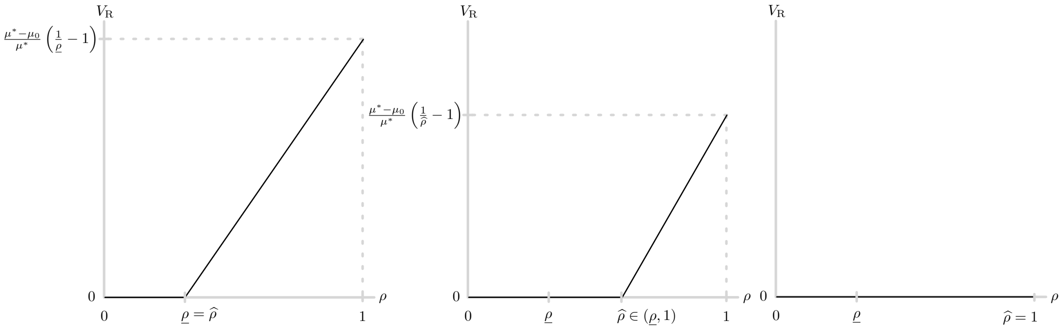
A.5 Lemma 3

The Receiver's value function given any experiment $\widehat{\rho} \in [\underline{\rho}, 1]$ is given by

$$V_{\mathrm{R}}(\rho|\widehat{\rho}) = \begin{cases} 0 & \text{if } \rho \leq \widehat{\rho} \\ \frac{\mu^* - \mu_0}{\mu_0}\left(\frac{\rho}{\widehat{\rho}} - 1\right) & \text{if } \rho > \widehat{\rho} \end{cases}.$$

Figure 6 depicts $V_{\mathrm{R}}(\cdot|\widehat{\rho})$ for the three types of experiments: $\widehat{\rho} = \underline{\rho}$, $\widehat{\rho} \in (\underline{\rho}, 1)$, and $\widehat{\rho} = 1$. The figure shows that $V_{\mathrm{R}}(\cdot|\widehat{\rho})$ is continuous and convex, and equals zero for all $\rho \leq \widehat{\rho}$ and increases linearly thereafter.

Figure 6: Receiver's value function an experiment $\widehat{\rho} \in [\underline{\rho}, 1]$: $V_{\mathrm{R}}(\cdot|\widehat{\rho})$.



Lemma 3. *Any commitment equilibrium payoff can be obtained with the class of investigations given by* $\{\iota^*(z) : z \in [\rho_0, 1]\} \subseteq \mathscr{I}$*, where*

$$\iota^*(z) := \begin{cases} \frac{\rho_0}{z}\delta_z + \frac{\rho_0}{z - \rho_0}\delta_0 & \textit{if } z \in (\rho_0, 1] \\ \delta_{\rho_0} & \textit{if } z = \rho_0 \end{cases}.$$

Proof. Fix some $\widehat{\rho} \in [\underline{\rho}, 1]$ and consider the following problem:

$$\max_{\iota \in \mathscr{I}} \int_0^1 V_{\mathrm{R}}(\rho|\widehat{\rho})\, \mathrm{d}\iota(\rho) \;\text{ s.t. } \int_0^1 \max V_{\mathrm{S}}(\rho|\widehat{\rho})\, \mathrm{d}\iota(\rho) \geq V_{\mathrm{S}}^{\mathrm{maxmin}}.$$

Because the fully revealing investigation (i.e., $z = 1$) maximises the Receiver's expected payoff, if the Sender's payoff under a fully revealing investigation is greater than her maxmin payoff, i.e., $V_{\mathrm{S}}(1|\widehat{\rho})\rho_0 \geq V_{\mathrm{S}}^{\mathrm{maxmin}}$, then the fully revealing investigation solves the problem above. So suppose instead that $V_{\mathrm{S}}(1|\widehat{\rho})\rho_0 < V_{\mathrm{S}}^{\mathrm{maxmin}}$.

Consider first the case in which $\rho_0 \geq \widehat{\rho}$. Then, the greatest payoff that the Receiver can induce for the Sender is given by not investigating the Sender (i.e., $z = \rho_0$). If the Sender's maxmin payoff strictly exceeds the maximum Sender payoff that the Receiver can induce with an investigation, then $\widehat{\rho}$ is not a feasible solution to the Receiver's problem. If they are equal, then the solution to the problem is $(\widehat{\rho}, i^*(\rho_0))$. So suppose that $\max V_{\mathrm{S}}(\rho_0|\widehat{\rho}) > V_{\mathrm{S}}^{\mathrm{maxmin}}$. Let us now argue that a solution to the problem is the investigation with support at $\{0, z^*\}$, where $\int_0^1 \max V_{\mathrm{S}}(\cdot|\widehat{\rho})\mathrm{d}\iota^*(z^*) = V_{\mathrm{S}}^{\mathrm{maxmin}}$. Note first that z^* is well-defined by the intermediate value theorem because $\tilde{v}_{\mathrm{S}}(z) := \int_0^1 \max V_{\mathrm{S}}(\cdot|\widehat{\rho})\mathrm{d}\iota^*(z)$ is continuous in z and $\max V_{\mathrm{S}}(1|\widehat{\rho})\rho_0 < V_{\mathrm{S}}^{\mathrm{maxmin}} < \max V_{\mathrm{S}}(\rho_0|\widehat{\rho})$. Moreover, z^* is unique and $z^* > \widehat{\rho} \geq \rho_0$ because $\tilde{v}_{\mathrm{S}}(\cdot)$ is strictly increasing. Now take any investigation $\iota \in \mathscr{I}$ that satisfies the constraint. Because $\max V_{\mathrm{S}}(\cdot|\widehat{\rho})$ and $V_{\mathrm{R}}(\cdot|\widehat{\rho})$ are both affine on $[0, \widehat{\rho})$ and $[\widehat{\rho}, 1]$, we may collapse the mass ι puts on the interval $[0, \widehat{\rho})$ to a single point in $[0, \widehat{\rho})$ and also collapse the mass on the interval $[\widehat{\rho}, 1]$ to a single point in $[\widehat{\rho}, 1]$ via mean-preserving contraction. The procedure gives us an investigation ι' with supports at some $z_1 \in [0, \widehat{\rho})$ and some $z_2 \in [\widehat{\rho}, 1]$. Moreover, we must have $z_2 \leq z^*$ for ι to have satisfied the constraint. But observe that $\iota^*(z^*)$ obtained by spreading the mass at z_1 and z_2 to $\{0, z^*\}$ via a mean-preserving spread increases the Receiver's payoff (because $V_{\mathrm{R}}(\cdot|\widehat{\rho})$ is convex) while ensuring that the constraint is satisfied (in fact, binding). In other words, any investigation $\iota \in \mathscr{I}$ that satisfies the constraint can be transformed into $i^*(z^*)$ that still satisfies the constraint but leads to weakly greater payoff for the Receiver.

Now consider the case in which $\rho_0 < \widehat{\rho}$. Then, the greatest payoff that the Receiver can induce for the Sender is given by an investigation $i^*(\widehat{\rho})$. If Sender's maxmin payoff strictly exceeds this maximum Sender payoff that the Receiver can induce, then $\widehat{\rho}$ is not a feasible solution to the Receiver's problem. If they are equal, then the solution to the problem is $(\widehat{\rho}, i^*(\widehat{\rho}))$. So suppose that $\int_0^1 \max V_{\mathrm{S}}(\cdot|\widehat{\rho})\mathrm{d}\iota^*(\widehat{\rho}) > V_{\mathrm{S}}^{\mathrm{maxmin}}$. Once again, there exists $z^* > \rho_0$ such that $\int_0^1 \max V_{\mathrm{S}}(\cdot|\widehat{\rho})\mathrm{d}\iota^*(z^*) = V_{\mathrm{S}}^{\mathrm{maxmin}}$. Moreover, since $\max V_{\mathrm{S}}(\cdot|\widehat{\rho}) = 0$ on $[0, \widehat{\rho})$, we must have $z^* \geq \widehat{\rho}$. In fact, $z^* > \widehat{\rho}$ because $\int_0^1 \max V_{\mathrm{S}}(\cdot|\widehat{\rho})\mathrm{d}\iota^*(\widehat{\rho}) > V_{\mathrm{S}}^{\mathrm{maxmin}}$. But then the same procedure as described above means that starting from any investigation $\imath \in \mathscr{I}$ that satisfies the constraint, one can argue that $\imath^*(z^*)$ leads to an improvement for the Receiver while still satisfying the constraint. ■

A.6 Proof of Theorem 2

We are now ready to prove Theorem 2.

Theorem 2. *For any $\rho_0 \in (0,1)$, the Receiver's commitment equilibrium payoff is given by*

$$V_{\mathrm{R}}^* = \min\left\{\overline{V}_{\mathrm{R}}, \tfrac{\mu_0}{\mu^*}[(1-\mu_0)\rho_0 + V_{\mathrm{S}}^{\mathrm{maxmin}}] - V_{\mathrm{S}}^{\mathrm{maxmin}}, \tfrac{\mu_0}{\mu^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}\right\} > 0. \tag{7}$$

In particular, the Receiver's commitment equilibrium payoff is strictly positive for all interior prior belief, ρ_0, about the Sender's reliability. Moreover, for sufficiently low ρ_0, the Receiver is able to induce the Sender to choose the fully informative experiment while simultaneously finding out the Sender's reliability on the equilibrium path.

Proof of Theorem 2. Let $\overline{V}_{\mathrm{R}} := V_{\mathrm{R}}(1|\underline{\rho})\rho_0 = \frac{(1-\mu^*)\mu_0\rho_0}{\mu^*}$ the Receiver's payoff from the ideal outcome (i.e., fully informative experiment and fully revealing investigation). As noted in Remark 3, the Receiver can induce the Sender to choose the fully informative experiment, $\widehat{\rho} = \underline{\rho}$, using a punishing investigation strategy whenever $\rho_0 \leq \rho_{0,0}$. Consider an investigation strategy i^+ that maximally punishes the Sender for choosing $\widehat{\rho} \neq \underline{\rho}$ and otherwise conducting a fully revealing investigation; i.e., $i^+(\underline{\rho}) = \rho_0\delta_1 + (1-\rho_0)\delta_0$ and $i^+(\widehat{\rho}) = i^{\min}(\widehat{\rho})$ for all $\widehat{\rho} \neq \underline{\rho}$. Since a fully revealing investigation is never part of a punishing investigation strategy, the investigation strategy i incentivises the Sender to choose the fully informative experiment *a fortiori* compared to the punishing investigation strategy. Thus, the Receiver can obtain $\overline{V}_{\mathrm{R}}$ whenever $\rho_0 \leq \rho_{0,0}$ with i. In fact, if $\rho_0 \in (\rho_{0,0}, \rho_{0,1}]$, where $\rho_{0,1} := (1 + \frac{\mu_0}{\mu^*-\mu_0}(1-\sqrt{\mu^*})^2)^{-1} \in (\rho_{0,0}, 1)$, the fully revealing experiment can still be induced by i^+ because the investigation guarantees that $V_{\mathrm{S}}(1|\underline{\rho})\rho_0 \geq V_{\mathrm{S}}^{\mathrm{maxmin}}$—i.e. the Sender's payoff from choosing the fully informative experiment and the Receiver conducting a fully revealing investigation is greater than her maximum payoff from deviating. Thus, for any $\rho_0 \in (0, \rho_{0,1}]$, the Receiver can obtain $\overline{V}_{\mathrm{R}}$.

Suppose now that $\rho_0 > \rho_{0,1}$. Define $\widehat{\rho}^{+}$ as the most informative experiment that still ensures that the Sender attains her maxmin payoff when the Receiver fully learns the Sender's reliability; i.e.,

$$\widehat{\rho}^{+} := \min\left\{\widehat{\rho} \in [\underline{\rho}, 1] : V_{\mathrm{S}}\left(1|\widehat{\rho}\right)\rho_0 \geq V_{\mathrm{S}}^{\mathrm{maxmin}}\right\} = \frac{\mu^* - \mu_0}{\mu^*}\frac{\rho_0}{\rho_0 - V_{\mathrm{S}}^{\mathrm{maxmin}}}.$$

That $\rho_0 > \rho_{0,1}$ implies $\widehat{\rho}^{+} > \underline{\rho}$. Note that the Receiver can induce the Sender to choose any experiment that is less informative than $\widehat{\rho}^{+}$ (i.e., any experiment $\widehat{\rho} \geq \widehat{\rho}^{+}$) using i^{+} but among all such experiment, the Receiver clearly prefers $\widehat{\rho}^{+}$. To induce the Sender to choose any experiment that is more informative than $\widehat{\rho}^{+}$ (i.e., any experiment $\widehat{\rho} \in [\underline{\rho}, \widehat{\rho}^{+})$), the Receiver must provide more incentive to the Sender to choose such an experiment by conducting a partially revealing investigation of the form 3. The commitment equilibrium payoff for the Receiver is given by the solution to the following

$$\max_{(\widehat{\rho}, z) \in [\underline{\rho}, \widehat{\rho}^{+}] \times [\rho_0, 1]} \int_0^1 V_{\mathrm{R}}\left(\rho|\widehat{\rho}\right) \mathrm{d}\iota^*\left(z\right)\left(\rho\right)$$
$$\text{s.t.} \int_0^1 \max V_{\mathrm{S}}\left(\rho|\widehat{\rho}\right) \mathrm{d}\iota^*\left(z\right)\left(\rho\right) \geq V_{\mathrm{S}}^{\mathrm{maxmin}}.$$

I can substitute the parametric expressions for V_{R}, $\max V_{\mathrm{S}}$ and ι^* to write the Sender's problem as

$$\max_{\widehat{\rho} \in [\underline{\rho}, \widehat{\rho}^{+}], z \in [\rho_0, 1]} \frac{\mu^* - \mu_0}{\mu^*}\rho_0\left(\frac{1}{\widehat{\rho}} - \frac{1}{z}\right) \text{ s.t. } \widehat{\rho} \geq \frac{\mu^* - \mu_0}{\mu^*}\frac{\rho_0}{\frac{\rho_0}{z} - V_{\mathrm{S}}^{\mathrm{maxmin}}}.$$

Because the objective is strictly decreasing in $\widehat{\rho}$, at any optimal, either $\widehat{\rho} = \underline{\rho}$ or the constraint must bind so that

$$\widehat{\rho}^{*} = \max\left\{\underline{\rho}, \frac{\mu^* - \mu_0}{\mu^*}\frac{\rho_0}{\frac{\rho_0}{z^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}}\right\}$$

If the constraint is binding at the optimal, then the objective is given by $\frac{\rho_0}{z^*}\frac{\mu_0}{\mu^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}$, which is strictly decreasing in z. Moreover, we have

$$\frac{\mu^* - \mu_0}{\mu^*}\frac{\rho_0}{\frac{\rho_0}{z} - V_{\mathrm{S}}^{\mathrm{maxmin}}} \geq \underline{\rho} \Leftrightarrow z \geq \frac{\rho_0}{V_{\mathrm{S}}^{\mathrm{maxmin}} + (1-\mu_0)\,\rho_0}.$$

Hence, if the constraint is binding at the optimal, then z^* must satisfy the inequality above with equality or equal to ρ_0; i.e.,

$$z^* = \max\left\{\rho_0, \frac{\rho_0}{V_{\mathrm{S}}^{\mathrm{maxmin}} + (1-\mu_0)\,\rho_0}\right\} = \begin{cases} \frac{\rho_0}{V_{\mathrm{S}}^{\mathrm{maxmin}}+(1-\mu_0)\rho_0} & \text{if } \rho_0 \in (\rho_{0,1}, \rho_{0,2}) \\ \rho_0 & \text{if } \rho_0 \in [\rho_{0,2}, 1] \end{cases},$$

where $\rho_{0,2} := \frac{\mu_0}{\mu^*(2-\mu_0) - 2\sqrt{\mu^*(1-\mu_0)(\mu^*-\mu_0)}} \in (\rho_{0,1}, 1)$. ■

Remark 4. Table 1 gives the on-equilibrium-path experiment and investigation as well as the Receiver's and Sender's payoffs in the commitment equilibrium for various regions of prior belief about reliability ρ_0 identified in the proof above.

Table 1: Commitment equilibrium.

ρ_0	$\widehat{\rho}^*$	z^*	V_{R}^*	V_{S}^*
$= 0$	n/a	n/a	0	0
$\in (0, \rho_{0,0}]$	$\underline{\rho}$	1	$\overline{V}_{\mathrm{R}} = \frac{1-\mu^*}{\mu^*}\mu_0\rho_0$	$\mu_0\rho_0$
$\in (\rho_{0,0}, \rho_{0,1}]$	$\underline{\rho}$	1	$\overline{V}_{\mathrm{R}}$	$\mu_0\rho_0$
$\in (\rho_{0,1}, \rho_{0,2}]$	$\underline{\rho}$	$\frac{\rho_0}{V_{\mathrm{S}}^{\mathrm{maxmin}}+(1-\mu_0)\rho_0}$	$\frac{\mu_0[(1-\mu_0)\rho_0+V_{\mathrm{S}}^{\mathrm{maxmin}}]}{\mu^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}$	$V_{\mathrm{S}}^{\mathrm{maxmin}} > 0$
$\in (\rho_{0,2}, 1)$	$\frac{\mu^*-\mu_0}{\mu^*}\frac{\rho_0}{1-V_{\mathrm{S}}^{\mathrm{maxmin}}}$	ρ_0	$\frac{\mu_0}{\mu^*} - V_{\mathrm{S}}^{\mathrm{maxmin}}$	$V_{\mathrm{S}}^{\mathrm{maxmin}} > 0$
$= 1$	1	n/a	0	$\frac{\mu_0}{\mu^*}\rho_0$

A.7 Proof of Proposition 1

Proposition 1. *The Receiver prefers delegating investigations to a purely adversarial Third Party over a purely Receiver-aligned Third Party—strictly so if* $\rho_0 \in [\underline{\rho}, 1)$.

Proof. The Receiver's ∞-equilibrium payoff corresponds to the Receiver's no-commitment equilibrium payoff which is zero by Theorem 1. A 0-balanced Third Party's sequential rational investigation is the punishing investigation strategy, (6). Recall from the proof of Lemma (2) that the Receiver does not investigate the Sender under the punishing investigation strategy if $\rho_0 \in (0, \underline{\rho})$, and both players' payoffs are zero because no experiment is able to induce the Receiver to take action. If, instead, $\rho_0 \in [\underline{\rho}, 1)$, the Sender chooses $\widehat{\rho}^{\text{maxmin}} < \rho_0$ so that the Receiver's payoff is strictly positive because any investigation must induce a posterior belief that yields the Receiver some strictly positive payoff. ■

A.8 Proof of Theorem 3

A λ-balanced Third Party's value correspondence given any experiment $\widehat{\rho} \in [\underline{\rho}, 1]$ is given by

$$V_{\mathrm{T}}^{\lambda}\left(\rho|\widehat{\rho}\right) = \begin{cases} \{0\} & \text{if } \rho < \widehat{\rho} \\ \left[0, -\frac{\mu_0}{\mu^*}\right] & \text{if } \rho = \widehat{\rho} \\ \left\{-\left(1 - \frac{\rho}{\widehat{\rho}}\frac{\mu^* - \mu_0}{\mu^*}\right) + \lambda \frac{\mu^* - \mu_0}{\mu^*}\left(\frac{\rho}{\widehat{\rho}} - 1\right)\right\} & \text{if } \rho > \widehat{\rho} \end{cases}.$$

In particular, for any $\widehat{\rho} \in [\underline{\rho}, 1)$, note that

$$V_{\mathrm{T}}^{\lambda}\left(1|\widehat{\rho}\right) \geq 0 \Leftrightarrow \lambda \geq \frac{\frac{\mu^* - \mu_0}{\mu^*}\widehat{\rho} - 1}{1 - \widehat{\rho}} =: \Lambda\left(\widehat{\rho}\right)$$

and $\Lambda(\cdot)$ is strictly increasing so that its inverse, Λ^{-1}, is well defined.

Theorem 3. *For any $\rho_0 \in (0,1)$, there exists $\lambda^*(\rho_0) > 0$ such that the Receiver's λ-equilibrium payoff is given by*

$$V_{\mathrm{R}}^{\lambda} = \begin{cases} 0 & \text{if } \lambda < \lambda^*\left(\rho_0\right) \\ \frac{1}{1+\lambda^*(\rho_0)}\frac{\mu_0}{\mu^*}\rho_0 & \text{if } \lambda \geq \lambda^*\left(\rho_0\right) \end{cases}$$

Hence, the Receiver strictly prefers to delegate investigations to a $\lambda^*(\rho_0)$-balanced Third Party over any other λ-balanced Third Party. Moreover, whenever the prior belief that the Sender is reliable, ρ_0, is sufficiently low, the Receiver's λ^*-equilibrium payoff coincides with the Receiver's commitment equilibrium payoff.

Proof. Fix $\lambda \in (0,\infty)$. Suppose first that $\lambda < \Lambda(\underline{\rho}) = \frac{\mu^*}{1-\mu^*}$. Then, $V_{\mathrm{T}}^{\lambda}(1|\widehat{\rho}) \leq 0$ for all $\widehat{\rho} \in [\underline{\rho},1]$. Since $\Lambda(\underline{\rho}) \leq \Lambda(\widehat{\rho})$ for all $\widehat{\rho} \in [\underline{\rho},1]$, this means that for all possible choice of experiment, the Third Party's value correspondence looks like the last case in Figure 2. Consequently, the (λ-balanced) Third Party's sequentially rational investigation strategy (which concavifies V_{T}) corresponds to the punishing investigation strategy $i^{\min}(\cdot)$. Thus, in this case, λ-equilibrium payoffs corresponds to 0-equilibrium payoffs.

Suppose instead that $\lambda \geq \Lambda(\underline{\rho})$. For any $\widehat{\rho} \in (\Lambda^{-1}(\lambda),1]$, the sequentially rational investigation for the Third Party would be $i^{\min}(\widehat{\rho})$ as in the previous case. However, for any $\widehat{\rho} \in [\underline{\rho},\Lambda^{-1}(\lambda))$, $V_{\mathrm{T}}^{\lambda}(1|\widehat{\rho}) > 0$ (the first case in Figure 2) so that the sequentially rational investigation for the Third Party is fully revealing. Finally, if $\widehat{\rho} = \Lambda^{-1}(\lambda)$, $V_{\mathrm{T}}^{\lambda}(1|\widehat{\rho}) = 0$ (the second case in Figure 2) so that the Third Party is indifferent between the fully revealing investigation and $i^{\min}(\widehat{\rho})$—let us suppose that Third Party conducts a fully revealing investigation in this case.

Given the Third Party's best response to the Sender's experiment described above, if $\widehat{\rho}^{\mathrm{maxmin}} \leq \Lambda^{-1}(\lambda) \Leftrightarrow \lambda \geq \Lambda(\widehat{\rho}^{\mathrm{maxmin}})$, the Sender can attain payoffs that are weakly lower than $V_{\mathrm{S}}^{\mathrm{maxmin}}$ by choosing any $\widehat{\rho} > \Lambda^{-1}(\lambda)$ or attain payoffs that are weakly greater than $V_{\mathrm{S}}^{\mathrm{maxmin}}$ associated with choosing any $\widehat{\rho} \in [\underline{\rho},\Lambda^{-1}(\lambda)]$ under the fully revealing investigation. Since the Sender's payoff is increasing in $\widehat{\rho}$ given an investigation, the Sender would choose the experiment $\widehat{\rho} = \Lambda^{-1}(\lambda)$ to attain a payoff of $V_{\mathrm{S}}(1|\Lambda^{-1}(\lambda))\rho_0$. If, instead, $\widehat{\rho}^{\mathrm{maxmin}} > \Lambda^{-1}(\lambda)$, the Sender can attain payoffs associated with any experiment $\widehat{\rho} \in (\Lambda^{-1}(\lambda),1]$ under the punishing investigation—the highest being $V_{\mathrm{S}}^{\mathrm{maxmin}}$— or attain payoffs associated with any experiment $\widehat{\rho} \in [\underline{\rho},\Lambda^{-1}(\lambda)]$ under the fully revealing investigation—the highest being $V_{\mathrm{S}}(1|\Lambda^{-1}(\lambda))\rho_0$. Hence, the Sender chooses $\widehat{\rho} = \Lambda^{-1}(\lambda)$ if and only if $V_{\mathrm{S}}(1|\Lambda^{-1}(\lambda))\rho_0 \geq V_{\mathrm{S}}^{\mathrm{maxmin}}$.

Recall from Remark 3 that $\widehat{\rho}^{\mathrm{maxmin}} = \underline{\rho}$ if $\rho_0 \in (0,\rho_{0,0}]$ and $\widehat{\rho}^{\mathrm{maxmin}} \in (\underline{\rho},\rho_0)$ if $\rho_0 \in (\rho_{0,0})$. If $\rho_0 \in [\underline{\rho},\rho_{0,0}]$, since $\lambda > \Lambda(\underline{\rho})$ by hypothesis, on the (λ-)equilibrium path, the Sender chooses $\widehat{\rho}^* = \Lambda^{-1}(\lambda)$ and the Third Party conducts a fully revealing investigation. If, instead, $\rho_0 \in (\rho_{0,0},1)$, so that $\widehat{\rho}^{\mathrm{maxmin}} > \Lambda^{-1}(\lambda) > \underline{\rho}$, then the

Sender chooses $\widehat{\rho}^* = \Lambda^{-1}(\lambda)$ if and only if

$$V_{\mathrm{S}}(1|\Lambda^{-1}(\lambda))\rho_0 \geq V_{\mathrm{S}}^{\mathrm{maxmin}} \Leftrightarrow \Lambda^{-1}(\lambda) \geq \widehat{\rho}^+.$$

Recall that when $\rho_0 \in [\rho_{0,0}, \rho_{0,1}]$, $\widehat{\rho}^+ \leq \underline{\rho}$ so that the inequality above holds; i.e., on equilibrium path, the Sender chooses $\Lambda^{-1}(\lambda)$ and the Third Party conducts a fully revealing investigation. Note that, for any $\rho_0 \in (\rho_{0,0}, 1)$,

$$\widehat{\rho}^{\mathrm{maxmin}} \leq \Lambda^{-1}(\lambda) \Leftrightarrow \lambda \geq \tfrac{1-\widehat{\rho}^{\mathrm{maxmin}}}{\widehat{\rho}^{\mathrm{maxmin}}} = \sqrt{\tfrac{\mu_0}{\mu^*-\mu_0}\tfrac{\rho_0}{1-\rho_0}} - 1$$

and $\widehat{\rho}^+ < \widehat{\rho}^{\mathrm{maxmin}}$. Hence, for any $\lambda \geq \sqrt{\tfrac{\mu_0}{\mu^*-\mu_0}\tfrac{\rho_0}{1-\rho_0}} - 1$ and $\rho_0 \in (\rho_{0,0}, 1)$, we have $\Lambda^{-1}(\lambda) \geq \widehat{\rho}^+$. Moreover, for any $\rho_0 \in (\rho_{0,1}, 1)$, there exists a unique $\lambda^+ \in [\tfrac{\mu^*}{1-\mu^*}, \sqrt{\tfrac{\mu_0}{\mu^*-\mu_0}\tfrac{\rho_0}{1-\rho_0}} - 1]$ such that $\Lambda^{-1}(\lambda) \geq \widehat{\rho}^+$ for all $\lambda \geq \lambda^+$ and $\Lambda^{-1}(\lambda) < \widehat{\rho}^+$ for all $\lambda \in [\tfrac{1-\mu^*}{\mu^*}, \lambda^+)$.

The argument above means that for any $\lambda \geq \lambda^* := \max\{\tfrac{\mu^*}{1-\mu^*}, \lambda^+\}$, the Receiver's λ-equilibrium payoff is given by $V_{\mathrm{R}}(1|\Lambda^{-1}(\lambda))\rho_0 = \tfrac{1}{1+\lambda}\rho_0$. Hence, the Receiver's λ-equilibrium payoff is greater for when $\lambda = \lambda^*$. Finally, observe that for any $\rho_0 \in (0, \rho_{0,0})$, the Receiver's λ^*-equilibrium payoff equals $\overline{V}_{\mathrm{R}}$ which, in turn, is also equal to the Receiver's commitment equilibrium payoff. ■

Remark 5. Table 2 gives the on-equilibrium-path experiment and investigation as well as the Receiver's and Sender's payoffs in the λ-equilibrium when $\lambda \geq \lambda^*$ for various regions of prior belief about reliability ρ_0. Recall that if $\lambda < \lambda^*$, then λ-equilibrium payoffs are the same as in 0-equilibrium.

Table 2: Delegation equilibrium if $\lambda \geq \lambda^*$.

ρ_0	$\widehat{\rho}^*$	ι^*	V_{R}^{λ}	V_{S}^{λ}
$= 0$	n/a	n/a	0	0
$\in (0, \rho_{0,0}]$	$\underline{\rho}$	Fully revealing	$\overline{V}_{\mathrm{R}} = \frac{1-\mu^*}{\mu^*}\mu_0\rho_0$	$\mu_0\rho_0$
$\in (\rho_{0,0}, 1]$	$\Lambda^{-1}(\lambda)$	Fully revealing	$\frac{1}{1+\lambda}\frac{\mu_0}{\mu^*}\rho_0$	$\frac{\lambda}{1+\lambda}\frac{\mu_0}{\mu^*}\rho_0$
$= 1$	1	n/a	0	$\frac{\mu_0}{\mu^*}\rho_0$