THEMA
théorie économique,
modélisation et applications

THEMA Working Paper n°2021-02
CY Cergy Paris Université, France

# Designing Interrogations

Alessandro Ispano, Peter Vida

January 2021

# Designing Interrogations

Alessandro Ispano Péter Vida

January 2021

**Abstract**

We provide an equilibrium model of interrogations with two-sided asymmetric information. The suspect knows his status as guilty or innocent and the likely strength of law enforcers' evidence, which is informative about the suspect's status and may also disprove lies. We study the evidence strength standards for interrogating and for drawing adverse inferences from silence that minimize prosecution errors. We consider scenarios where interrogations can be delegated. We describe the optimal mechanism under full commitment and a dynamic interrogation with two-sided information revelation implementing the optimum in equilibrium.

*Keywords*: lie, evidence, leniency, questioning, confession, law, prosecution, two-sided asymmetric information
*JEL classifications*: D82, D83, C72, K40

Alessandro Ispano: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, alessandro.ispano@gmail.com; Peter Vida: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, vidapet@gmail.com. We thank Helmuts Āzacis, Eric Danan, Harry Di Pei, Lucie Menager, Marcus Pivato, Régis Renault, Karl Schlag, Johannes Schneider, Joel Sobel, Egor Starkov as well as seminar participants at the THEMA theory meetings, the Workshop on Signaling in Markets Auctions and Games 2019, University of Copenhagen, Paris Applied Theory Day 2019, and Paris II University for useful comments. Financial support from Labex MME-DII is gratefully acknowledged.

# 1 Introduction

In most legal systems, the interrogation of a suspect is an important resource in the investigation phase that may lead to his prosecution. A body of law is therefore in place to ensure both the respect of the suspect's rights and the admissibility of the information law enforcers obtain. This paper develops a theoretical framework that describes how interrogations unfold based on essential features of the legal system. It then uses this framework to determine which institutions enhance information revelation from the suspect and yield to more accurate prosecution decisions.

An interrogation can be represented as a game of two-sided asymmetric information between the suspect (henceforth he), who aims to convince the law enforcer of his innocence and be let go, and the law enforcer (henceforth she), who aims to obtain truthful information to minimize some weighted sum of type I (prosecuting an innocent) and type II (letting a guilty go) errors.[1] The suspect's private information is his status as guilty or innocent and the likely strength of the incriminating evidence. The law enforcer's private information is the actual evidence gathered.

To fix ideas, imagine the suspect is selling a product whose purity allegedly falls short of a commonly known legal requirement. While the suspect knows the exact purity of the product, the law enforcer observes the result of a test that provides an upper bound. The evidence is directly informative about the suspect's status. Moreover, the evidence may disprove his eventual false claims. For his part, the suspect enjoys the right to silence but also some leniency for confessing[2] (the claimed purity is less than the legal requirement) relative to when he denies and he is caught in a lie (the claimed purity exceeds the upper bound), or he stays silent and the evidence is unambiguously incriminating (the upper bound is less than the legal requirement).

[1] In this flexible specification, the law enforcer's exact objective may derive from fundamental features of the legal system, e.g. adversarial or inquisitorial, and her precise role, e.g. police officer or prosecutor. Throughout, we abstract from details about this separation of roles, which is fuzzier than traditionally thought (Abel, 2016), and assume the law enforcer directly takes prosecution decisions even when she does not have formal authority. The model can equivalently apply to any other decision that the law enforcer would want to base on the suspect's status, e.g. an arrest, and generates disutility to the suspect irrespectively.

[2] Leniency may take many different forms. On a first level, for minor crimes some legal systems explicitly provide suspects with the possibility to avoid prosecution in exchange of an admission of guilt and face only lighter consequences, as in the case of police cautions in the United Kingdom. Besides, the suspect's confession may be part of a plea bargain with the prosecutor. Finally, by cooperating early on the suspect may obtain better detainment conditions or a more favorable sentence even in the absence of an explicit agreement.

In our baseline model, communication is one-shot and unidirectional from the suspect to the law enforcer - we will allow for both gradual and two-sided information revelation later on. The suspect's message must be interpreted as claim about his private information or "type" (e.g. the product purity) in reply to the law enforcer's inquiry. Under an equilibrium refinement that gives prominence to honesty due to Hart et al. (2017), innocent types and confessors claim the truth. Conversely, some sufficiently unsuspicious guilty types necessarily lie and deny. Clear predictions then obtain on the equilibrium outcome of the interrogation both in terms of players' strategies (proposition 1) and payoffs (corollary 1).

With minimization of prosecution errors as underlying objective, we use this baseline model to explore the optimal design of three distinct instruments regulating interrogations in the legal system, namely the right to silence, the standard for interrogating, and delegation. These instruments work as partial commitment devices for the law enforcer's behavior. Maintaining the same information structure as in the baseline model, we also describe the optimal mechanism under full commitment and a dynamic game with two-sided information revelation that implements the optimum in equilibrium without any commitment.

**The right to silence.** The suspect's right to refuse to answer law enforcers' questions is recognized in most legal system. Still, important differences remain in the level of protection this right entails and, in particular, on whether an adverse inference, i.e. a conclusion pointing at the suspect's guilt, can be drawn.[3] In our model, these differences are parametrized by the evidence strength standard required to prosecute a suspect who stays silent. In accordance with the logic behind adverse inferences, a silent suspect is necessarily guilty since innocents are honest. Moreover, higher protection of silence in the form of a more stringent standard may dissuade a guilty from confessing. Nonetheless, the optimal level of protection of silence (proposition 2) is not always minimal. Indeed, even though having to let a silent, hence guilty, suspect go by law due to insufficiently strong evidence is suboptimal ex-post, the suspect's need to lie also decreases. From an ex-ante perspective, the latter effect may dominate and overall yield to more accurate prosecution decisions. Thus, our results also provide a purely information based justification for

[3] See for instance the debate among legal scholars (O'Reilly, 1994; Ingraham, 1995) on the 1994 reform limiting the right to silence in the United Kingdom. Their arguments make explicit reference to the impact on the suspect's strategy in the interrogation.

adverse inferences being insufficient to trigger prosecution without additional supporting evidence.[4] Besides, other than protecting the suspect against self-incrimination, the compulsory reminder of the right to silence at the start of the interrogation may serve informational purposes (corollary 2).

**The standard for interrogating.** As in the case of other restraints of individual freedom such as searches and seizures, law enforcers may be required to hold sufficiently strong evidence to be able to interrogate the suspect. As pointed out by Reinganum (1988) in the context of arrests, this type of standard also conveys information. Indeed, if the suspect is interrogated, he then knows the law enforcer's evidence must meet the standard. Under a more stringent standard a guilty suspect is less inclined to lie and more inclined to confess. Thus, if on the one hand the suspect must necessarily be let go when the evidence is too weak, on the other hand interrogations become more informative. The optimal standard for interrogating (proposition 3) often prescribes to let the suspect go when the evidence is sufficiently weak.

**Delegation.** The previous analysis assumed full alignment between the designer's and the law enforcer's preferences, parametrized by the weight of type I and type II errors used in evaluating the accuracy of prosecution decisions. From a normative perspective, these preferences can be thought of as reflecting the social costs of each error.[5] If the weights of the law enforcer in charge of the interrogation can differ, however, her ideal preferences from the point of view of the designer (proposition 4) are always biased. The direction of the law enforcer's ideal bias may be towards prosecution, i.e. a higher relative weight than the designer attached to type II errors, but also towards dismissal. In the former case, more guilties confess instead of lying. In the latter, liars resort to less lies, so that innocents are more easily set apart. Prosecution errors further decrease if the bias can be made contingent on the strength of the evidence and, in particular, tilted towards prosecution when evidence is strong and towards dismissal when evidence is weak. These

[4] For instance, in the United Kingdom, ss 34 of the CJPOA 1994 establishes that adverse inferences can be drawn from the accused's failure to mention facts when questioned under caution, i.e. having being warned about his right to silence. At the same time, ss 38 states that "A person shall not have the proceedings against him transferred to the Crown Court for trial, have a case to answer or be convicted of an offence solely on such a failure or refusal."

[5] For the determinants of the preferences of the social planner over the entire prosecution process, which in particular also incorporate the suspect's disutility, see Grossman and Katz (1983), Reinganum (1988) and Siegel and Strulovici (2018). While Siegel and Strulovici (2018) adopt a mechanism design approach, both Grossman and Katz (1983) and Reinganum (1988) maintain that the prosecutor's and the social planner's preferences are perfectly aligned.

insights on the delegation of interrogations equally apply at a more micro level, e.g. to the appointment of law enforcers within a police department.

**The optimal mechanism.** Overall, the previous instruments can be effective because they alleviate the law enforcer's lack of commitment over prosecution decisions inherent in the equilibrium. While the legal system can be designed from an ex-ante perspective, some discretion plausibly remains for all agents involved in its functioning due to institutional constraints, incompleteness of the law, and other informational frictions. Law enforcers will then act upon this discretion in their own interest and, for example, will not let go a confessor who they know to be surely guilty. We complement the analysis with the alternative mechanism design approach, which assumes full commitment on the outcome of the interrogation based on the suspect's claim and the evidence.[6] The optimal mechanism (proposition 5) has a close link with the equilibrium of our baseline model. The suspect's payoff is the same, while the accuracy of prosecution decisions increases thanks to a reduction in type II errors.

**Implementation without commitment.** The optimal mechanism can be implemented in equilibrium of a natural sequential game (proposition 6) built on our baseline model. This game combines features of standards for interrogating and of delegation, even though law enforcers' behavior and the associated information revelation about the evidence to the suspect is now entirely dictated by equilibrium considerations rather than fixed by law. The suspect gives away some information in the first round of the interrogation anticipating if the evidence is weak relative to his claim he will be let go. When this is not the case, a maximally tough interrogator, i.e. who only cares about avoiding type II errors, will continue. A guilty suspect will step back on his lie, which will be forgiven, and an innocent type will stick to his story, which may lead to either prosecution or dismissal depending on the strength of the evidence. The equilibrium in this second round is reminiscent of screening outcomes in plea bargaining (Grossman and Katz, 1983; Reinganum, 1988) and the optimal judicial mechanism of Siegel and Strulovici (2018), in which only innocents reject the plea and the decision to proceed to trial is based on the strength of the evidence not because of its informational content

[6] Siegel and Strulovici (2018) adopt this approach in a general framework which encompasses the entire prosecution process. We discuss in some more detail how our results relate to theirs below. See also the discussion about the design of the legal system in Hart et al. (2017), who consider a class of persuasion games with one-sided asymmetric information in which the outcome with and without commitment on behalf of the uninformed party is the same.

but purely for screening purposes. At the same time, in our setting more screening takes place both within and across innocents and guilties due to additional heterogeneity in the strength of the evidence they expect.

We conclude by considering some additional questions on the conduct and regulation of interrogations our framework can address. In particular, throughout we maintained that all aspects of the strategic environment, other than the players' private information, are common knowledge and outside the law enforcer's control. These are determined by institutional features of the legal system, which the suspect should be familiar with, especially if assisted by an attorney. At the same time, the elements of arbitrariness in interrogations are a major cause of criticism and an important reason behind the general movement towards their mandatory recording.[7] Our framework can easily incorporate asymmetric information between the law enforcer and the suspect about the legal environment. It also directly allows to identify the direction of the misleading efforts the law enforcer would want to engage in if these are tolerated by law or go undetected and the suspect is prone to deception. In accordance with the logic behind common interrogations tactics, the law enforcer would always want to overstate the benefits of confession, exaggerate the strength of the incriminating evidence and misrepresent her true preferences over type I and type II errors (proposition 7). While surely objectionable on other grounds, when successful these deceptive tactics improve information elicitation. Also, this improvement does not come at the cost of extorting false confessions.

The paper is structured as follows. After a discussion of the related literature here below, section 2 presents the baseline model and section 3 describes its equilibria. Section 4 explores design instruments. Section 5 presents the optimal mechanism and its implementation. Section 6 concludes. Additional material and proofs are in the appendix (section A and B) and the online appendix (section C, D and E), which in particular contains proofs involving tedious calculations.

**Related Literature.** While the entire judicial process is a prominent field of application of the strategic communication literature, suspects' interrogations have received little explicit attention.[8] In parallel, the law and economics literature generally studies

[7] See for instance Sullivan (2005).

[8] A notable exception unrelated to this paper is Baliga and Ely (2016), who study the interrogator's commitment problems inherent to torture.

the judicial process assuming prosecution is already undergoing. This paper instead focuses explicitly on how communication in the interrogation contributes to the decision to prosecute and the overall strength of the case.[9] Thus, capturing essential features of the judicial process only in stylized form,[10] our framework accounts for important specificities that arise in interrogations due to the information structure and the nature of communication. First, asymmetric information about the incriminating evidence is presumably more pervasive than further down the judicial process, where the prosecution is typically subject to mandatory disclosure requirements and discovery occurs.[11] Thus, taking two-sided asymmetric information between the suspect and law enforcers a step further than previous work on plea bargaining, our model allows for heterogeneity not only between guilties and innocents, but also within guilties and innocents, in the strength of the incriminating evidence they expect. This heterogeneity explains why different guilty types prefer different strategies and, for instance, some "break" and others do not. Moreover, we allow for communication from the suspect as opposed to screening or signaling by the prosecutor only. The information different parties acquire and present in front of a judicial officer, including the defendant's claims, is typically modeled as hard evidence (Milgrom, 1981), i.e. it can be disclosed or withheld but not misreported.[12] To allow for the possibility of plain lying that is intrinsic to interrogations, in our model the suspect's claims are soft information, i.e. the set of his available messages is independent from the truth. At the same time, the suspect's claims are not pure cheap talk (Crawford and Sobel, 1982) since these might be contradicted by the law enforcer's evidence, entailing a cost. Differently from models of strategic communication with lying costs (Kartik, 2009) and detectable deceit (Dziuda and Salas, 2018; Balbuzanov, 2019), the detectability of a lie and its cost for the suspect derive explicitly from the law enforcer's private

[9] See Redlich et al. (2018) for an empirical account of how confessions in interrogations correlate with plea bargaining and sentencing outcomes.

[10] In particular, we take as given that guilt entails some punishment and confession entails some leniency without considering the complex determinants of plea bargaining (Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezzetti, 2001; Daughety and Reinganum, 2020) and sentencing (Siegel and Strulovici, 2018, 2019). We thereby ignore considerations on crime deterrence (and chilling of socially desirable behavior (Kaplow, 2011)), commensurate punishment, endogenous evidence acquisition and deployment of resources in prosecution.

[11] The plea bargaining literature features alternative assumptions on when this source of asymmetric information exactly resolves, i.e. if already at the plea bargaining stage (Grossman and Katz, 1983) or after (Reinganum, 1988). Daughety and Reinganum (2018, 2020) explore the prosecutor's incentives to comply with the disclosure requirements established by the Supreme Court of the United States in Brady v. Maryland (1963).

[12] See for instance Shin (1994), Bhattacharya and Mukherjee (2013), and Hart et al. (2017).

information, which in particular naturally implies that the detectability of a lie increases with its size.[13] Moreover, our framework allows studying the effects of revelation of the law enforcer's private information to the suspect, as it occurs in the case of evidence strength standards for interrogating and in the dynamic interrogation that implements the optimal mechanism. Our model hence also joins the growing literature on strategic communication that, departing from seminal works, considers two-sided asymmetric information between the sender and the receiver.[14] It differs in players' incentives and the information structure as well as in the main questions of interest. A recurrent theme in this literature is that the receiver may sometimes be hurt from her information since as a result the sender may reveal less. In our setting, absent the possibility that the suspect may be caught in a lie or proven guilty thanks to the law enforcer's evidence, the interrogation would be completely uninformative. Finally, the signaling considerations between liars and innocents are reminiscent of the ones that arise in the completely different context of electoral competition of Kartik and McAfee (2007), where candidates driven by holding office want to pass for ones with intrinsic preferences for their campaign platform.

## 2 A model of interrogations

**Information structure.** There are two players: a suspect (he), denoted by $S$, and a law enforcer (she), denoted by $R$. At the initial stage, $S$ privately observes his type $y$, which is drawn uniformly from $[0, 1]$. $S$'s status $\mathcal{Y} \in \{0, 1\}$ depends on whether $y$ is above some fixed cutoff $t \in (0, 1)$: $\mathcal{Y} = 0$, i.e. $S$ is **guilty**, when $y < t$ and $\mathcal{Y} = 1$, i.e. $S$ is **innocent**, when $y \geq t$. Thus, $t$ also represents the prior probability that $S$ is guilty. $R$'s private information is determined as follows. Independently of $y$, $z$ is drawn uniformly from $[0, 1]$. If $y \geq z$ $R$ does not observe anything and the game ends.[15] If $y < z$, instead,

[13] Kartik (2009) assumes a lie entails a direct cost that increases with its size and he invokes penalties upon lying detection as a possible interpretation. In both Dziuda and Salas (2018) and Balbuzanov (2019), instead, any lie has an equal exogenous chance of being detected and the cost is endogenously determined by the receiver's response. Perez-Richet and Skreta (2020) consider general lying (falsification) cost functions in a setting in which the receiver has no private information. Instead, in Ioannidis et al. (2020) the message of the sender determines the costly investigation technology of the receiver.

[14] See de Barreda (2010), Chen (2012), Lai (2014), Ishida and Shimizu (2016), and Pei (2017) for models of soft information and Ispano (2016) and Frenkel et al. (2020) for models of hard information.

[15] A first interpretation is that in this case $R$ is not even aware of $S$. Alternatively, as $S$'s guilt becomes less likely than under the prior, $R$ does not find it worth or has no legal ground to go after $S$. Our framework can easily accommodate the alternative scenario in which $R$ can still interrogate $S$ in

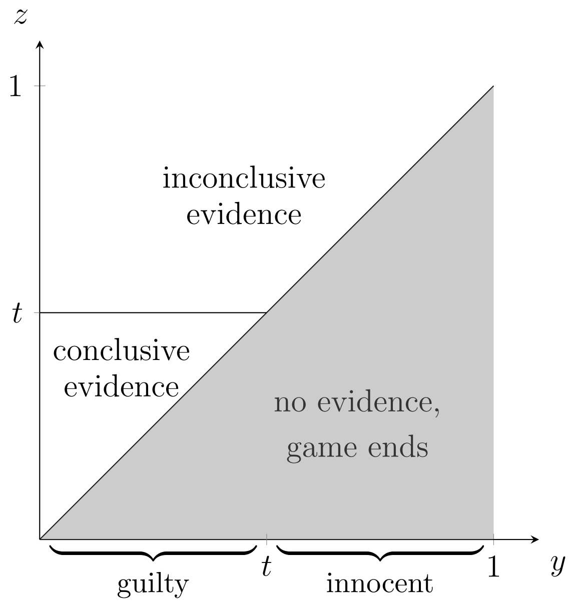


Figure 1  The sample space and the evidence

$R$ privately observes $z$ and hence knows that $y < z$. In particular, we say that evidence is **conclusive** if it proves with probability one that $S$ is guilty, i.e. if $z \leq t$.

Figure 1 displays the three possibilities based on realizations $y$ and $z$. Each point within the unit square is equally likely and the relevant region for our game is above the 45-degrees line, where $y < z$. While ensuring tractability, our information structure respects two important general features inherent to the evidence identified by Reinganum (1988). First, stronger evidence makes the suspect's guilt more likely. Second, a guilty suspect expects stronger evidence than an innocent one. Additionally, in our setting there is heterogeneity both within guilties and innocents in the strength of the evidence they expect.

**Moves.** After $y$ and $z$ have been drawn and the information structure determined accordingly, provided $y < z$, $S$ sends a message $m \in \mathcal{M} = [0, 1] \cup \{s\}$ to $R$, who then takes an action $a \in \{0, 1\}$ and payoffs realize as described below.

$S$'s message must be interpreted as a literal claim about his type and $m = s$ represents his choice to stay **silent**. Provided $S$ does not stay silent, we say that he **lies** when $m \neq y$, that he **is honest** when $m = y$, that he **confesses** when $m < t$, and he **denies** when $m \geq t$. Also, we say that he is **caught in a lie** when $R$'s evidence contradicts his claim, i.e. when $m \geq z$. $R$'s action can be interpreted as a decision on whether $S$ should

the hope of obtaining an admission of guilt but must let him go otherwise.

be prosecuted, i.e. $a = 0$, or let go freely without charges, i.e. $a = 1$. We impose $R$ mechanically chooses $a = 0$ when evidence is conclusive and/or $S$ confesses and/or $S$ is caught in a lie. Likewise, $R$ mechanically chooses action $a = 0$ upon silence if allowed by law. Namely, if we let $Z_s \in (t, 1]$ be the evidence standard required to prosecute $S$ upon silence in the legal system, then

$$a(s, z) = \begin{cases} 0 & \text{if } z \leq Z_s \\ 1 & \text{if } z > Z_s. \end{cases} \tag{1}$$

Instead, when $S$ denies and he is not caught in a lie, i.e. in the set $D = \{(m, z) | m \in [t, 1], z > m\}$, $R$ is free to choose her action, and we say that she has **discretion**.

**Payoffs.** $R$'s loss (i.e. the negative of her payoff) is

$$e(a, \mathcal{Y}) = \alpha \, a \mathbb{1}_{\mathcal{Y}=0} + (1 - \alpha) \, (1 - a) \, \mathbb{1}_{\mathcal{Y}=1}, \tag{2}$$

where $\mathbb{1}_{\mathcal{Y}=0}$ and $\mathbb{1}_{\mathcal{Y}=1}$ are indicator functions for $S$'s status as guilty and innocent, respectively, and $\alpha \in (0, 1)$ a commonly known parameter. That is, $R$ aims at prosecuting a guilty suspect and letting an innocent suspect go and $\alpha$ measures the relative importance of a type II error over a type I error. As for $S$, we normalize his payoff from being prosecuted and being let go respectively to 0 and 1 and we distinguish the following cases

$$\pi(y, m, z, a) = \begin{cases} 0 & S \text{ confesses and he is not caught in a lie} \\ -b & S \text{ is caught in a lie} \\ & \text{or he is silent and evidence is conclusive} \\ a(s, z) & S \text{ is silent and evidence is inconclusive} \\ a(m, z) & S \text{ denies and he is not caught in a lie (i.e. } (m, z) \in D). \end{cases}$$

By confessing (honestly) $S$ saves $b > 0$ relative to when he lies and he is caught or to when he stays silent and he is directly proven guilty. Thus, $b$ measures the punishment for reticence or, equivalently, the **leniency** that confession entails. Instead, if $S$ remains silent and $R$ has only inconclusive evidence, his payoff depends on whether the strength standard for prosecuting is met as described at equation (1). Finally, when $S$ denies and there is inconclusive evidence that does not contradict his claim, his payoff is determined by $R$'s action upon discretion.

Thus, taking into account $R$'s mechanical moves, $S$'s payoff in each instance is equal to $R$'s action net of the eventual cost $b$. In section 6.1, we discuss how we can dispense with the simplifying assumption of $R$'s mechanical moves, which are indeed sequentially rational since innocents are honest. There, we allow $S$'s payoff when he confesses and he is not caught in a lie to depend explicitly on $R$'s action. We also allow $S$'s cost when he is silent and evidence is conclusive to be lower than when he is caught in a lie.

# 3 Equilibrium

Throughout, we concentrate on pure strategies and restrict our attention to equilibria in which innocent types and confessors are honest.[16] Therefore, $S$'s strategy is fully described by a partition of the set of guilty types into the sets $Y_c$, $Y_s$ and $Y_\ell$, denoting the set of guilty types who confess, remain silent and lie, respectively, and a **lying function** $\boldsymbol{\ell}$ which associates to each type $y \in Y_\ell$ a lie $\ell(y) \in [t, 1]$. We impose that $\boldsymbol{\ell}$ is measurable and we denote its range $\boldsymbol{\ell}(Y_\ell)$ by $L$. $R$'s strategy $\boldsymbol{a} : D \to \{0, 1\}$ specifies an action $a(m, z)$ for each message $m$ and evidence realization $z$. We impose that for all $m \geq t$ the Lebesgue integral $A(m) \equiv \int_{\{z:z>m\}} a(m, z)/(1-m)\mathrm{d}\lambda$ exists, which is also the expected payoff of innocent type $m$ from being honest. $R$'s belief system $\boldsymbol{\mu} : D \to [0, 1]$ specifies a probability $\mu(m, z) = \mathbb{P}(\mathcal{Y} = 1|m, z)$ that $S$ is innocent. The relevant solution concept (henceforth: equilibrium) is weak perfect Bayesian equilibrium, that is, a triple $\langle \boldsymbol{\ell}, \boldsymbol{a}, \boldsymbol{\mu} \rangle$ together with the sets $Y_c, Y_s, Y_\ell$ such that:

(i) the message of each type of $S$ is optimal given $R$'s strategy;

(ii) $R$'s action after each message $m$ and evidence $z$ is optimal given her belief;

(iii) $R$'s belief system is consistent with (a generalized version of) Bayes' rule.[17]

[16] In section C of the online appendix we derive this restriction as a result in a more general game using an extremely weak (possibly mixed) equilibrium concept requiring only a very weak form of truth-leaning adapted from Hart et al. (2017). We also show that the equilibrium singled out in proposition 1 satisfies truth-leaning. All of our results, including our derivation that innocents and confessors are honest in any truth-leaning equilibrium, go through also with distributional strategies as defined in Milgrom and Weber (1985).

[17] More precisely, we require that beliefs are derived from a regular conditional probability (see appendix A for details). In addition to consistency with Bayes' rule, among other things this requirement implies that for zero probability but on the equilibrium path messages: almost surely $\mu(m, z) = 1$ if $m$ is only sent by an innocent type; almost surely $\mu(m, z)$ does not depend on $z$. Purely for ease of exposition, we are going to assume that these two conditions hold exactly for each $m$.

**Proposition 1** (Equilibrium)**.** *There is an equilibrium in which:*

*(i)* $Y_c = [0, y_c), Y_s = [y_c, y_\ell)$ *and* $Y_\ell = [y_\ell, t)$ *or* $Y_c$ *and/or* $Y_s$ *are empty;*

*(ii)* $\boldsymbol{\ell} : Y_\ell \to L$ *with* $L = [t, \bar{y})$*,* $\bar{y} \in (t, 1)$ *and* $\ell(y) = t + \frac{\alpha}{1-\alpha}(y - y_\ell)$*;*

*(iii)* $\mu(m, z) = \frac{1}{1+\ell^{-1\prime}(m)} = \alpha$ *for all* $(m, z) \in D$ *if* $m \in L$*, so that* $R$ *is indifferent between actions;*

*(iv)* $A(m) = \frac{1-\bar{y}-b(\bar{y}-m)}{1-m}$ *for* $m \in L$ *and* $A(m) = 1$ *for* $m \geq \bar{y}$*, so* $A(m)$ *is continuous, increasing and such that all types in* $Y_s \cup Y_\ell$ *are indifferent to any message* $m \in L$ *and, provided* $Y_s$ *is non-empty, staying silent;*

*(v) for all* $(m, z) \in D$*:*

$$a(m, z) = \begin{cases} 0 & \text{if } z \leq \bar{z}(m) \\ 1 & \text{if } z > \bar{z}(m) \end{cases}, \tag{3}$$

*where* $\bar{z}(m) = 1 - A(m)(1 - m)$*.*

*Moreover, in any other equilibrium point (iii) and (iv) hold,* $Y_c$ *and* $L$ *are the same (except these intervals can be right closed),* $\lambda(Y_\ell)$ *and* $\lambda(Y_s)$ *are the same and* $A(m)$ *is the same.*

*Proof.* See section B.1 in the appendix. □

Thus, lying is an indissoluble part of the interrogation. Indeed, if sufficiently low denying claims were only sent by innocent types and hence be fully persuasive of $S$'s innocence, these would be too tempting for sufficiently high guilty types. As the distribution of messages of innocent types is atomless, so must be the one of liars for them to disguise effectively. Over the lying region $L$, $R$'s expected action upon not detecting a lie must increase with $m$ to compensate liars for the higher risk of detection that higher lies entail. Simultaneously, the strategy of liars must ensure that $R$ indeed finds it rational to choose her actions accordingly, which can only be the case if she is indifferent between prosecuting $S$ and letting him go. In particular, her belief about $S$'s innocence does not depend on the evidence since knowing that message $m$ must have been sent either by innocent type $m < z$ or guilty type $\ell^{-1}(m) < z$ contains infinitely more information than knowing that $y < z$. Once $S$'s incentives to confess or stay silent are also taken into account, these observations pin down the equilibrium fraction of guilty types who confess, are silent and lie as well as the lies sent and $R$'s expected action upon each $m$. Equilibria may only differ in the exact identity of silent types and liars, in the exact shape of their lying function and in the exact action policy of $R$ that induces each expected action. In the equilibrium singled out in the proposition the set of liars is an interval, the lying function is increasing, and $R$'s action policy takes a rather natural cutoff form, in that she lets $S$ go when the evidence is sufficiently weak.

The model generates some intuitive comparative statics that are common to all equilibria and follow from simple inspection of closed-form solutions for the respective objects of interest, which can all be found in the appendix. Weakly more types confess when confession entails higher leniency, when $R$ is tougher as measured by a higher weight she attaches to a type II error, when the prior likelihood of innocence is lower and when protection of silence is weaker as measured by a less stringent prosecution standard (i.e. a higher $Z_s$). Likewise, a guilty type may resort to his right to silence only if doing so entails enough protection. Higher leniency also reduces the lying region, so that a smaller claim suffices to convince $R$ of $S$'s innocence. Conversely, the lying region is larger with a tougher $R$, which can be intuitively understood as that she requires more convincing to let $S$ go.

Moreover, the model yields unique welfare predictions, in that players' expected payoffs are the same in every equilibrium. In particular, $R$'s ex-ante expected loss is

$$(1-\alpha)\underbrace{\int_L (1-y)\left(1-A\left(y\right)\right)\mathrm{d}\lambda}_{\text{type I errors}}+\alpha\underbrace{\int_{Y_\ell}(1-\ell\left(y\right))\,A\left(\ell\left(y\right)\right)\mathrm{d}\lambda}_{\text{type II errors on liars}}+\alpha\underbrace{(1-Z_s)\lambda(Y_s).}_{\text{type II errors on silents}} \tag{4}$$

It turns out that the expression only depends on the measures of liars and silent types, which are identical across equilibria. Also, for both $S$ and $R$, payoff equivalence not only holds from an ex-ante perspective, i.e. before $S$ has observed $y$ and $R$ has observed $z$, but also ex-post.

**Corollary 1** (Payoff equivalence)**.** *Every equilibrium is payoff equivalent for $S$ and $R$ both from an ex-ante and an ex-post perspective.*

*Proof.* See section B.2 in the appendix. □

Figure 2 displays the equilibrium payoff of $S$ and the associated type I and type II errors $R$ makes based on the realization of $y$ and $z$, where no type is silent given the

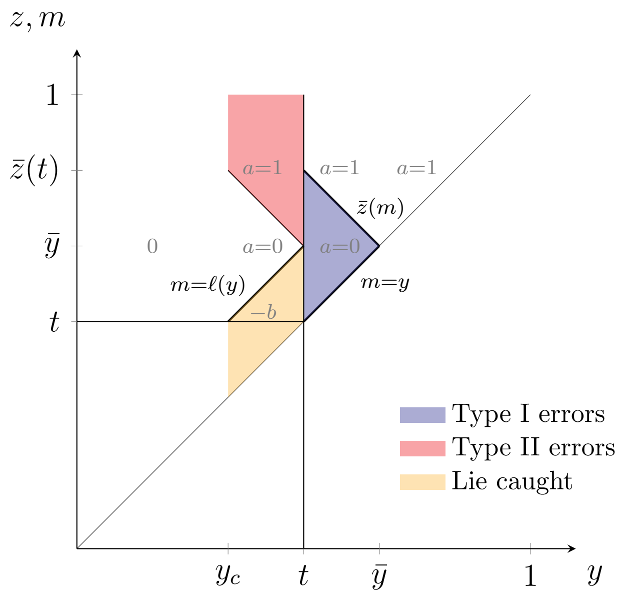


Figure 2 Equilibrium payoffs
($t = 1/2$, $b = 1$, $\alpha = 1/2$, $Z_s \geq 5/6$)

parameter configuration chosen. Separating guilty types get 0 and separating innocent types get $a = 1$, so that $R$ makes no errors. As for pooling types, $R$'s action is $a = 1$ if $z \geq \bar{z}(m)$ and $a = 0$ otherwise. A guilty type above $y_c$ is caught in a lie when $z \leq \ell(y)$ and in this case he gets $-b$. Provided he is not caught, he gets $a = 1$ when $z$ is above $\bar{z}(\ell(y))$, so that $R$ makes a type II error, and $a = 0$ otherwise. Likewise, an innocent type below $\bar{y}$ gets $a = 1$ when $z \geq \bar{z}(y)$ and $a = 0$ otherwise, and in the latter case $R$ makes a type I error.

# 4 Design instruments

## 4.1 Protection of the right to silence

In our framework, the level of protection of silence that the legal system grants is determined by the evidence strength standard $Z_s$ required to prosecute a silent suspect (equation (1)). Naturally, no type will resort to his right to silence when $Z_s$ is sufficiently high. When instead $Z_s$ is sufficiently low to induce a positive measure of types to remain silent, since a silent type is necessarily guilty, $R$ is sometimes forced to suboptimally let $S$ go due to insufficiently strong evidence. However, $R$ may still welcome such a level of protection once the aggregate effects on $S$'s strategy are considered.

**Proposition 2** (Optimal level of protection of silence)**.** *Let $Z_s^\star$ denote $R$'s optimal standard for prosecuting a silent $S$ (i.e. she prosecutes iff $z \leq Z_s$).*

- *If $\lambda(Y_c) = 0$ under $Z_s = 1$, then $\lambda(Y_s) > 0$ under $Z_s^\star$.*
- *If instead $\lambda(Y_c) > 0$ under $Z_s = 1$:*
  - *$\lambda(Y_s) = 0$ under $Z_s^\star$ if $t$, $\alpha$ and $b$ are large;*[18]
  - *if under $Z_s^\star$ it is the case that $\lambda(Y_s) > 0$, then it is also the case that $\lambda(Y_c) = 0$.*

*Proof.* See section D.1 in the online appendix. □

An effective level of protection of silence, i.e. a level that induces a positive measure of types to remain silent, may be optimal for $R$ because, if on the one hand it entails a type II error upon silence when evidence is weak, on the other hand it reduces the fraction of liars and hence the pooling of innocents and guilties. When all guilty types would lie even without any protection, an effective level is always optimal since the loss introduced on silent types is initially negligible relative to the benefits of increased separation. Instead, higher protection has less clear benefits for $R$ when it also discourages some guilty types from confessing. As shown in the proof of the proposition, this negative effect always dominates at the margin. If the extent of voluntary confession absent any protection is large and type II errors are rather costly for $R$, i.e. if $t$, $\alpha$ and $b$ are large, $R$'s expected loss is always increasing in the level of protection of silence. Otherwise, $R$'s expected loss is non-monotone and an effective level of protection can be optimal if it is large enough so only liars and silent types, but no confessors, remain.

While the proposition is expressed in terms of $R$'s welfare, it also implies that an increase from an ineffective to an effective level of protection of silence is sometimes unambiguously Pareto improving since $S$ always favors higher protection. Indeed, the expected payoff from not confessing for guilty types and from being honest for innocent types who pool increases, and more innocent types separate since the lying region decreases. Besides, the proposition also speaks to the effects of the common legal requirement that $S$ must be explicitly reminded of his right to silence for the interrogation to be admissible in court. In line with its most apparent intent, the reminder may limit

[18] That is, there exists a known cutoff $\hat{t}(b, \alpha) > 0$ such that under $Z_s^\star$ we have that $\lambda(Y_s) = 0$ if $t > \hat{t}(b, \alpha)$, where $\hat{t}(b, \alpha)$ is decreasing in $b$ and $\alpha$.

self-incrimination by an otherwise unaware $S$. However, sometimes it may also help $R$ due to informational reasons.

**Corollary 2** (Reminder of the right to silence)**.** *If $Z_s^\star$ is such that $\lambda(Y_s) = 0$, $R$ weakly benefits from concealing the right to silence to an unaware $S$. If $Z_s^\star$ is such that $\lambda(Y_s) > 0$, instead, $R$ benefits from reminding it provided $|Z_s^\star - Z_s|$ is sufficiently small.*

*Proof.* If $S$ is unaware of the possibility to stay silent, the game is observationally equivalent to one where $S$ is aware but $Z_s$ is such that $\lambda(Y_s) = 0$. Therefore, the result follows directly from proposition 2. □

From now on, unless stated otherwise, we prevent the possibility that $S$ might stay silent by restricting his message space to $\mathcal{M} = [0, 1]$, which is equivalent to assume that the level of protection of silence is sufficiently low not to be effective. This assumption allows ignoring the rather subtle effects of changes in the fraction of silent types on $R$'s payoff described above. Besides, it ensures that $R$ always weakly benefits from interrogating before taking a decision, which may not necessarily be the case if the level of protection of silence is disproportionately high.

**Assumption** (NS)**.** *Henceforth, $\mathcal{M} = [0, 1]$, i.e. $S$ is never silent.*

## 4.2 The evidence strength standard for interrogating

In our baseline model, $R$ always interrogates $S$ as long as she has some evidence, no matter her belief about $S$'s status as guilty or innocent. Suppose instead $R$ can interrogate $S$ only if his guilt is sufficiently likely based on the available evidence and must let him go otherwise, as figure 3 illustrates. The posterior probability of innocence $\mathbb{P}(\mathcal{Y} = 1 \,|\, z)$ given $R$'s evidence is 0 if $z \leq t$ and $\frac{z-t}{z}$ if $z > t$. A rule such as "reasonable suspicion" or "probable cause" requiring the probability of innocence to be sufficiently low (i.e. below the horizontal red line in figure 3) maps into an evidence strength standard $Z$ such that $R$ interrogates only if $z \leq Z$. Since $Z$ is observable by law, when $S$ is interrogated he knows $R$ must have sufficiently strong evidence, i.e. that indeed $z \leq Z$.

The equilibrium analysis of our baseline model, which corresponds to $Z = 1$, easily generalizes to describe the outcome of the interrogation under any standard $Z \in (t, 1]$.[19]

[19] If $Z \leq t$, when the interrogation occurs $R$ knows $S$ is surely guilty and all types will confess. Our analysis for $Z > t$ encompasses $Z = t$ as limit case and demonstrates that such a stringent standard is naturally always suboptimal for $R$.

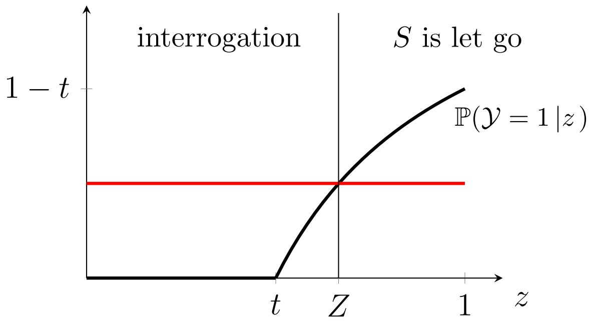


Figure 3 An evidence strength standard for interrogating

A more stringent standard has the effect to incentivize confession and discourage lying due to $S$'s increased pessimism about $R$'s evidence. Thus, the introduction of the standard entails a trade-off for $R$. On the negative side, $R$ gives up the chance to interrogate $S$ upon weak evidence, which may entail a loss of information transmission and introduce a type II error for types that would have confessed anyway. On the positive side, $R$ can conduct more informative interrogations upon strong evidence.

**Proposition 3** (Optimal standard)**.** *Let $Z^\star$ denote $R$'s optimal standard for interrogating (i.e. she interrogates iff $z \leq Z^*$).*

- *If $\lambda\left(Y_c\right)=0$ for $Z=1$, then $Z^\star<1$.*
- *If $\lambda\left(Y_c\right)>0$ for $Z=1$, then:*
    - *$Z^\star<1$ if and only if $t$, $b$ and $\alpha$ are small;*[20]
    - *in particular, $Z^\star=1$ if $\lambda\left(Y_c\right)>0$ for any $b$.*

*Proof.* See section D.2 in the online appendix. □

The general message of the proposition is that a stringent interrogation standard is optimal when the extent of voluntary confession is otherwise low. A rough intuition behind this result is that confession has a major informational benefit that is always worth reaping. However, the result is also driven by the importance of the loss from not interrogating that the standard generates. When the extent of voluntary confession is low this loss is low because interrogations would be rather uninformative anyway. It is also low because of the conditions that explain low confession in the first place, namely a

[20] That is, there exists a known cutoff $\bar{t}\left(b,\alpha\right)>0$ such that $Z^\star<1$ if and only if $t<\bar{t}\left(b,\alpha\right)$, where $\bar{t}\left(b,\alpha\right)$ is decreasing in $b$ and $\alpha$.

low weight $\alpha$ $R$ attaches to a type II error and a low prior $t$ that $S$ is guilty. As shown in the proof, starting from a situation in which no type confesses, a more stringent standard entails no loss if it leaves the set of confessors still empty. Then, making the standard sufficiently stringent to induce some types to confess is always beneficial for $R$.

## 4.3 Delegation and the objective of law enforcers

Suppose now $R$ can choose to delegate the interrogation to an interrogator whose loss is still given by equation (2) but with an arbitrary, possibly different relative weight for type II errors $\alpha \in [0, 1]$.[21] The interrogator's preferences are observable to $S$ and the interrogation then plays out as in the baseline model, except that it is now the interrogator who takes prosecution decisions.

**Proposition 4** (Optimal delegation). *Let $\alpha^\star$ denote $R$'s optimal delegation choice.*

- *It is always the case that $\alpha^\star \neq \alpha$.*
- *When $\lambda (Y_c) > 0$ absent delegation, $\alpha^\star > \alpha$ and, in particular, $\alpha^\star < 1$ if and only if $\alpha < 2/3$.*
- *When $\lambda (Y_c) = 0$ absent delegation then $\alpha^\star \in (0, \alpha)$ if $\alpha$ is sufficiently small.*

*Proof.* See section D.3 in the online appendix. □

$R$'s choice to delegate to a tougher interrogator affects $R$'s expected loss in three ways. First, it yields to suboptimal decisions biased towards prosecution. Second, it disciplines $S$ to favor confession over lying. Third, it induces types who still elect not to confess to also use bigger lies, i.e. the lying region increases. Starting from a situation in which the set of confessors would be non-empty without delegation, the informational benefit of increased confession at least initially dominates the two other negative effects. Indeed, the set of confessors increases at a faster rate than the size of the lying region. Moreover, $R$ would take the same optimal action as the interrogator both upon confession and

[21] When the interrogator has extreme preferences, i.e. she only cares about type I or type II errors, we suppose the limit respectively for $\alpha \to 0$ and $\alpha \to 1$ of any given equilibrium at section 3 obtains. The limit of this equilibrium is indeed an equilibrium, even though there may be others that are not payoff equivalent for $R$. In the limit for $\alpha = 0$ the distribution of lies is all concentrated at $t$ and, upon observing $m = t$ and having discretion, the interrogator sometimes lets $S$ go even though she is sure he is guilty. Instead, in the limit for $\alpha = 1$ the measure of liars is zero and upon observing a message in the lying region, the interrogator sometimes prosecutes $S$ even though she is sure he is innocent.

upon the detection of a lie. Thus, in this case $R$ always finds it optimal to delegate to a tougher interrogator. In particular, the interrogator should be maximally biased towards minimizing type II errors if letting a guilty $S$ go is already rather costly for $R$. Instead, when given $R$'s preferences the set of confessors would be empty without delegation, the minimal interrogator's toughness required to benefit from increased confession may be too far off. In this case, $R$ prefers a less tough interrogator because, in spite of the suboptimal decisions biased towards dismissal, the lying region decreases, enhancing separation of innocents and guilties.

**Conditional delegation.** Suppose $R$ can further condition the interrogator's preferences on the strength of the evidence. That is, $R$ chooses a delegation policy $\boldsymbol{\alpha} : [0,1] \to [0,1]$ which associates to each evidence realization $z$ the interrogator's preference parameter $\alpha(z)$. $S$ observes the delegation policy but not the actual preferences of the interrogator, which may otherwise convey information about the evidence, and the interrogation unfolds as before. For concreteness and brevity, we address the issue of how $R$ can benefit from conditional delegation by means of an example in the online appendix (section E). Still, the insights we develop are completely general and may be summarized as follows:

- for any unconditional delegation policy that is not extreme, i.e. $\boldsymbol{\alpha} \equiv \alpha_{\text{const}} \in (0,1)$, there always exists a strictly loss-reducing conditional policy that prescribes delegating to a nicer interrogator, i.e. with $\alpha_{\text{nice}} < \alpha_{\text{const}}$, when the evidence is sufficiently weak and to a tougher interrogator, i.e. with $\alpha_{\text{tough}} > \alpha_{\text{const}}$, otherwise;
- this policy can always be chosen so as to reduce type II errors while leaving type I errors unaffected hence it is preferred by $R$ independently of her actual preferences;
- the loss reduction from this policy is maximal when the preferences of the nicer and the tougher interrogator are as extreme as possible, i.e. $\alpha_{\text{nice}} = 0$ and $\alpha_{\text{tough}} = 1$.

# 5 The optimal interrogation

## 5.1 Optimal mechanism

In this section we suppose $R$ can commit to her action $a(m,z)$ based on the message $m$ received from $S$ and her evidence $z$. We are interested in $R$'s lowest attainable ex ante expected loss in a deterministic direct mechanism[22] in which $S$ only receives payoffs $a(m,z) \in \{0,1\}$ but, as before, detected lies are punished at a level of $-b$. We can further restrict our attention to cutoff mechanisms $\hat{\boldsymbol{z}} : [0,1] \to [0,1]$ which specify for each message $y$ a cutoff level $\hat{z}(y) \in [y,1]$ such that $a(y,z) = 1$ if and only if $z \geq \hat{z}(y)$.

Accordingly, the optimal direct mechanism minimizes

$$\alpha \int_0^t (1-\hat{z}(y))\mathrm{d}y + (1-\alpha)\int_t^1 (\hat{z}(y)-y)\,\mathrm{d}y. \tag{5}$$

subject to the constraint that each type finds it weakly optimal to be honest, i.e. for every $y, y' \in [0,1]$ such that $y < y'$

$$1-\hat{z}(y) \geq 1-\hat{z}(y') - b\,(y'-y)\,.^{23} \tag{6}$$

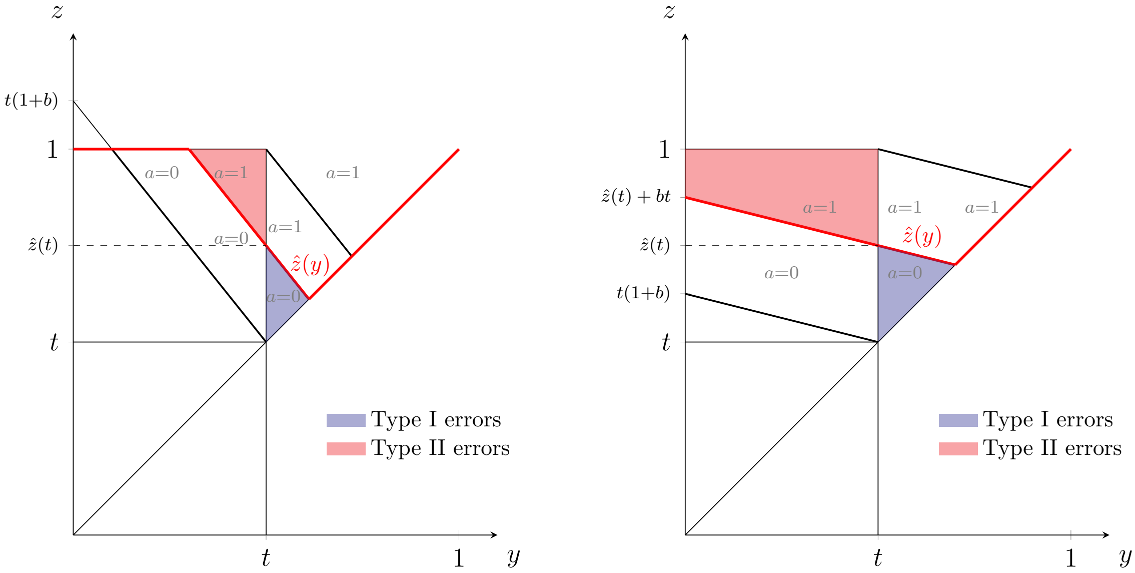
This constraint can be rewritten as $\hat{z}(y) - \hat{z}(y') \leq b\,(y'-y)$, which clarifies that if $y$ pretends to be $y' > y$ then he can get an additional measure $\hat{z}(y) - \hat{z}(y')$ of $a = 1$ if $z > y'$ but he will be caught in a lie when $z \in (y, y']$ and receive punishment $-b$.

Technically, this is an optimal control problem with a jump in the state variable $\mathcal{Y}$ at $t$. However, its solution is extremely simple. Candidate solutions can be indexed by $\hat{z}(t) \in [t,1]$ and the constraint must bind for types sufficiently close to $t$. Thus, in that region $\hat{z}(y)$ is linear with slope $-b$, as figure 4 demonstrates. We distinguished two possible cases depending on whether only sufficiently high types obtain a positive expected payoff (figure 4a) or all types do so (figure 4b).

Interestingly, the optimal mechanism $\hat{\boldsymbol{z}}^\star$ has a close relation with the equilibrium of

[22] As shown in section D.4.3 of the online appendix, when looking for the optimal mechanism all the imposed restrictions on the class of mechanisms are without loss of generality.

[23] We could introduce further constraints. First, we could require that the mechanism is immune to downward lies. Second, we could require that a participation constraint also holds assuming the possibility of silence. As we will see, however, downward lies will be clearly suboptimal in our mechanism. Also, provided that the level of protection of silence is sufficiently low, the participation constraint will also be satisfied.



**(a)** Non binding constraint for low guilties **(b)** Binding constraint for all guilties

Figure 4 Determination of the optimal mechanism

the baseline model (section D.4.1 in the online appendix provides detailed intuitions for the interested reader). Let us use the notation introduced at section 3 and take $y_c$, $\bar{y}$ and $\bar{z}(m)$ at their equilibrium levels (with $y_c = 0$ by convention if $\lambda(Y_c) = 0$). For types $y \in [y_c, \bar{y})$, i.e. types who pool in the equilibrium of the baseline model, let $\hat{z}^\star(y)$ be such that $1 - \hat{z}^\star(y) = 1 - \bar{z}(m(y)) - (m(y) - y)\, b = 1 - \bar{z}(y)$. That is, $R$ still uses the same cutoff strategy as in equilibrium and extends it to honest confessions of types who lied. Besides, $\hat{z}^\star(y) = 1$ for $y < y_c$ and $\hat{z}^\star(y) = y$ for $y \geq \bar{y}$, i.e. guilty types and innocent types who separate in the equilibrium of the baseline model still get always 0 and 1, respectively.

**Proposition 5** (Optimal mechanism)**.** *Mechanism* $\hat{\boldsymbol{z}}^\star$*, described above, is optimal. Accordingly:*

- *the expected payoff of each type of* $S$ *is the same as in equilibrium;*
- $R$*'s expected loss is strictly lower than in equilibrium due to the decreased amount of type II errors, while type I errors are the same.*

*Proof.* See section D.4 in the online appendix. □

As already implied indirectly by proposition 2, 3 and 4, $R$ suffers from her lack of commitment power over actions. In particular, the optimal mechanism requires that some guilty confessors are sometimes let go. In the next section, we show how $R$'s expected loss under the optimal mechanism can be replicated in equilibrium of a natural game built on the baseline model that does not exhibit this somehow unnatural property.

## 5.2 Implementation without commitment

Consider the following game with three players, $S$, $R$ and $I$, where $I$ is a maximally tough interrogator (i.e. whose loss is given by equation (2) with $\alpha = 1$):

- **Stage 0** $S$ and $R$ observe their private information as in the baseline model. Additionally, $I$ also observes $R$'s private information;
- **Stage 1** $R$ and $I$ interrogate $S$, i.e., $S$ sends them a public message $m \in \mathcal{M}$;
- **Stage 2** based on $S$'s message $m$ and the evidence $z$, $R$ can either take a prosecution decision $a \in \{0, 1\}$, in which case the game ends and payoffs realize as in the baseline model, or choose to delegate the continuation of the interrogation to $I$, so that stage 3 is reached;
- **Stage 3** $I$ interrogates $S$ again by specifying a set of messages $\mathcal{M}_I \subseteq \mathcal{M}$ from which $S$ can send a new message $m_2$;[24]
- **Stage 4** based on $S$'s new message $m_2$ as well as on $z$ and $m$, $I$ takes a prosecution decision $a_2 \in \{0, 1\}$, the game ends and payoffs realize as in the baseline model with $a_2$ replacing $a$.

**Proposition 6** (Implementation without commitment)**.** *There is an equilibrium of this game in which R's and S's expected payoffs are as in the optimal mechanism and S's behavioral strategy in stage 1 is as in the equilibrium at proposition 1.*

*Proof.* See section B.3 in the appendix. □

The structure of the equilibrium is intuitive. $R$ immediately takes the correct action for separating types. Instead, for pooling types, $R$ lets $S$ go if the evidence is weak relative to the received message, i.e. if $z \geq Z(m)$, and delegate the continuation of the

[24] This modeling choice is just for simplicity. $S$ could equivalently send any arbitrary message. Then $S$ would get $-b$ if caught in a lie as in the baseline model. Additionally, $I$ could just impose a punishment $-b$ in case "she does not get an answer to her question".

interrogation to $I$ otherwise. In particular, $m < Z(m)$, i.e. a liar is never let go when caught ($z \leq m$), and $Z(m)$ is decreasing, i.e. higher messages require stronger evidence for the interrogation to continue.[25] It is the increasing chance of being let go that allows screening among guilty types unwilling to confess in stage 1. $R$ finds it optimal not to deviate from her delegation policy due to the disciplining off the equilibrium path behavior of $I$, who would then prosecute $S$ with probability one. When the interrogation continues, $I$ chooses $\mathcal{M}_I = \{\ell^{-1}(m), m\}$, i.e. asks $S$ the question: "Are you $m$ or $\ell^{-1}(m)$?". Guilty type $\ell^{-1}(m)$, who sent the pooling message $m$, hence gets a second chance to confess. The appropriate choice of $Z(m)$ now induces type $\ell^{-1}(m)$ to do so, as he learned that the evidence is strong from the fact that the interrogation continued. In order not to leave to $S$ any unnecessary surplus and to reach payoffs as in the optimal mechanism, $Z(m)$ must be chosen to make type $\ell^{-1}(m)$ exactly indifferent between confessing and sticking to his stage 1 story $m$. Equilibrium lies in stage 1 are hence forgiven. Instead, innocent type $m$ sticks to his stage 1 story, i.e. he sends $m_2 = m$, and $I$ may either prosecute him or let him go depending on the evidence.[26]

# 6 Discussion

We provided a theoretical framework to analyze interrogations and derived several implications for the design of the legal system. We now cover some of the extensions and additional questions our framework allows to address.

## 6.1 More general moves and payoffs

Consider the same information structure, action set $\mathcal{M} = [0, 1] \cup \{s\}$ of $S$ and payoff of $R$ as in the baseline model. However, we now let $R$ freely choose her action $a \in \{0, 1\}$ for any possible $(m, z)$, i.e. her strategy is $\boldsymbol{a} : \mathcal{M} \times [0, 1] \to \{0, 1\}$. $R$'s action upon silence may still be determined mechanically by the legal system as per equation (1), in which

[25] This last property may not hold for different equilibrium lying functions than the monotonically increasing one singled out in proposition 1, which is in fact the only one for which the delegation policy of $R$ is continuous in $m$.

[26] Notice that on the equilibrium path $I$ will know that an $S$ who is sticking to his story is surely innocent. Nonetheless, as $I$ is maximally tough she is indifferent between prosecuting $S$ or letting him go. If there is no access to a maximally tough interrogator, the equilibrium above obtains in the limit as $I$ gets tougher and tougher. The difference is that, to make $I$ indifferent, guilty types must now confess randomly with a probability approaching 1 as $I$'s toughness becomes maximal.

case there will simply be no sequential rationality requirement for $a(s,\cdot)$. Moreover, $S$'s payoff is now

$$\pi(y,m,z,a) = \begin{cases} -b_\ell & S \text{ is caught in a lie} \\ -b_s & S \text{ is silent and evidence is conclusive} \\ a & \text{otherwise,} \end{cases}$$

where $b_\ell \geq b_s \geq 0$ with at least a strict inequality.

In section C of the online appendix we show that, under our truth-leaning equilibrium refinement, innocents and confessors are honest in this more general specification. Therefore, $R$'s mechanical actions and the payoff of a confessor who is not caught in a lie in our baseline model are consistent with our selected equilibria. Moreover, if $b_\ell > b_s$, then in any equilibrium the set of silent types and liars will necessarily have the same interval form and order as in the equilibrium singled out in proposition 1. All points of the proposition still hold with $b_\ell$ replacing $b$ and the difference that, provided $\lambda(Y_s) > 0$, the values of $y_c$, $y_\ell$, $\bar{y}$ and $A(t)$ will adjust based on $b_s$. In particular, the smallest liar will be indifferent between lying at $m = t$ and staying silent, i.e. $(1-t)A(t)-b_\ell(t-y_\ell) = (1-t)A(s)-b_s(t-y_\ell)$, where $A(s) = \int_{[t,1]} a(s,z)/(1-t)\mathrm{d}\lambda$.

The equilibrium also remains unaffected if $R$ can choose her action continuously from $[0,1]$. Indeed, for any message $m \in L$, upon not catching $S$ in a lie $R$ would again be indifferent among actions. If simultaneously $R$'s payoff changes to a quadratic loss, e.g. to $e(a,\mathcal{Y}) = 1/2(a-\mathcal{Y})^2$ when $\alpha = 1/2$ (the argument easily adapts to any $\alpha \in (0,1)$), then the equilibrium construction changes only in that when $R$ receives the message $m$ she should believe that $S$ is innocent with probability $A(m)$ and take action $a = A(m)$. $R$'s indifference can be achieved by choosing a lying function for which $\mu(m,z) = \frac{1}{1+\ell^{-1\prime}(m)} = A(m)$. Of course, the value of $A(m)$ might change now as well as the measure of liars and the measure of the lying region $L$ accordingly. In both models higher undetected lies are more rewarding in that they induce a higher expected action. In this alternative specification they are also more credible in that $R$ becomes more convinced of $S$'s innocence.

## 6.2 Deceptive interrogation tactics

Without explicitly modeling asymmetric information about the legal system (see section 6.3), we can investigate how the law enforcer would want to mislead the suspect about several parameters of interest. Formally, departing from the full rationality benchmark, we suppose $S$ plays according to what he considers as equilibrium behavior given his perception while $R$ best responds given the true environment. Proposition 2 together with corollary 2 already indirectly cover the effects of deception about the right to silence. Minimization and maximization tactics consist respectively, but ultimately equivalently, in downplaying the severity of legal consequences upon confession and overstating the legal consequences upon no confession,[27] and can both be thought of as increasing $S$'s perception of $b$. The exaggeration of the strength of incriminating evidence, which is another dimension on which maximization operates, can be captured by a more stringent perceived interrogation standard (i.e. a lower $Z$ as defined in section 4.2). Finally, in the same vein of the well-known "Good Cop Bad Cop" tactic, $R$ may deceive $S$ about her preference parameter $\alpha$, e.g. pretend that she simply aims to prosecute $S$ no matter his guilty status. As the next proposition clarifies, $R$ would always want to engage in such tactics.

**Proposition 7** (Deceptive tactics)**.** *$R$'s expected loss is weakly decreasing in $S$'s perception of the leniency $b$ and of the strength of the evidence as measured by a more stringent interrogation standard (i.e. a lower $Z$). Besides, $R$'s expected loss is concave in $S$'s perception of her toughness $\alpha$, minimal when $S$'s perception is extreme (i.e. $\alpha = 0$ or $\alpha = 1$) and maximal when $S$'s perception is correct.*

*Proof.* See section D.5 in the online appendix. □

Under higher perceived leniency $S$ is simultaneously more inclined to confess and less inclined to lie. A stronger perceived interrogation standard has similar effects. The benefits of $S$'s misperception of $\alpha$ are related to those of delegation, except that there is no downside for $R$ as she retains decision rights. Looking tougher or nicer are two effective ways to achieve the same objective. When $S$ perceives $R$ as tougher, he confesses more, even though $R$ will in fact always let $S$ go upon discretion. At the extreme belief $\alpha = 1$, $R$ eliminates errors completely since the measure of liars shrinks to zero. When $S$ perceives

[27] See for instance Kassin and McNall (1991).

$R$ as nicer, instead, he uses smaller lies anticipating these will suffice to be let go, even though $R$ will in fact always prosecute $S$ upon a pooling message. At the extreme belief $\alpha = 0$, $R$ again eliminates errors completely since the lying region shrinks to zero.

Overall, if successful, these deceptive tactics increase the accuracy of prosecution decisions. A recurrent source of criticism against discretion in interrogations is the possibility that law enforcers may extort false confessions.[28] No matter how persuasive these deceptive tactics may be, in our setting the only possibility for these to induce innocent types to depart from honesty is if they generate a shift away from truth-leaning (see section C in the online appendix).

## 6.3 Asymmetric information about the legal system

Parameters $b$, $Z_s$ and $Z$ can be directly thought of as expectations of $S$ about random variables whose realizations are private information of $R$. Without fully describing the equilibrium, we now show how also the parameter $\alpha$ can be made $R$'s private information. In the equilibrium of the baseline model, when $R$ has discretion she has to be indifferent between actions upon each pooling message and there is a cutoff $\bar{z}(m)$ for each message $m$ above which $R$ chooses $a = 1$ and below which $R$ chooses $a = 0$, resulting in the appropriate $A(m)$. Now the role of these cutoffs is played by cutoffs $\bar{\alpha}(m)$ such that nicer types of $R$ choose $a = 1$ and tougher types of $R$ choose $a = 0$ and the measures of nicer and tougher types induce the appropriate $A(m)$. The indifference between actions for type $\bar{\alpha}(m)$ can be ensured by choosing a lying function for which $\frac{1}{1+\ell^{-1\prime}(m)} = \bar{\alpha}(m)$. This differential equation now pins down the lying region and the set of liars. As a next step, one could investigate whether the deceptive tactics at section 6.2 would remain effective in a persuasion framework where $S$ is fully rational but uninformed.

## 6.4 Evidence revelation

We did not consider comprehensively communication about the evidence to the suspect and laws that govern it. Full disclosure of $R$'s private information seems clearly detrimental to the informativeness of interrogations. Indeed, provided the evidence is not conclusive, $S$ would then know how to tailor his lies and be less inclined to confess.

[28] See for instance Kassin et al. (2005).

Still, from a Bayesian persuasion perspective (Kamenica and Gentzkow, 2011), the optimal evidence revelation policy may take a different form than an evidence strength cutoff as we considered in the case of the standard for interrogating. If $R$'s claims about the evidence are soft information, instead, new interesting strategic considerations arise due to the possibility that $S$ may in turn catch $R$ in a lie, e.g. know that she is exaggerating the strength of the evidence. If the cost of doing so is sufficiently large for $R$ due to the risk of legal action or the inadmissibility of the interrogation in court, then evidence revelation becomes at least partially credible. Relatedly, $R$'s choice to interrogate $S$, and whether by means of a casual conversation or a formal interrogation, can be made itself a strategic variable. For example, by officially marking the start of a formal interrogation, the legal requirement that $S$ is explicitly notified of his right to silence may also implicitly convey additional information.[29]

## 6.5 Long and more articulated interrogations

We considered inter-temporal screening of suspects only in the game that implements the optimal mechanism. It would be interesting to determine if $R$ alone can improve upon the baseline model by gradually giving away the strength of her evidence in a dynamic interrogation. The answer highly depends on the extent of $R$'s commitment power over the stopping of the interrogation, which the law can affect. For example, if the maximum period of detention increases with the strength of incriminating evidence, continuing the interrogation credibly signals to $S$ that the evidence is sufficiently strong. Conversely, if $S$ must be formally charged with an offense as soon as there is sufficiently strong evidence, continuing the interrogation signals the lack thereof. Likewise, in the spirit of Glazer and Rubinstein (2004), one may investigate whether $R$ could benefit from formulating different questions to $S$ other than "what is your type?". The results of Perez-Richet and Skreta (2020) suggest that $R$ may prefer to obtain garbled information from $S$.

[29] This notification is referred to as Miranda warning in the United States and cautioning in the United Kingdom, where code C of the CJPOA 1994 states that "A person whom there are grounds to suspect of an offence, see Note 10A, must be cautioned before any questions about an offence, or further questions if the answers provide the grounds for suspicion, are put to them if either the suspect's answers or silence, (i.e. failure or refusal to answer or answer satisfactorily) may be given in evidence to a court in a prosecution". Thus, a suspect being cautioned should infer that circumstances at Note 10A apply i.e. "There must be some reasonable, objective grounds for the suspicion, based on known facts or information which are relevant to the likelihood the offence has been committed and the person to be questioned committed it."

# References

**Abel, Jonathan**, "Cops and pleas: Police officers' influence on plea bargaining," *Yale Law Journal*, 2016, *126*, 1730.

**Baker, Scott and Claudio Mezzetti**, "Prosecutorial resources, plea bargaining, and the decision to go to trial," *Journal of Law, Economics, and Organization*, 2001, *17* (1), 149–167.

**Balbuzanov, Ivan**, "Lies and consequences," *International Journal of Game Theory*, 2019, pp. 1–38.

**Baliga, Sandeep and Jeffrey C Ely**, "Torture and the commitment problem," *The Review of Economic Studies*, 2016, *83* (4), 1406–1439.

**Bhattacharya, Sourav and Arijit Mukherjee**, "Strategic information revelation when experts compete to influence," *The RAND Journal of Economics*, 2013, *44* (3), 522–544.

**Chen, Ying**, "Value of public information in sender–receiver games," *Economics Letters*, 2012, *114* (3), 343–345.

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

**Daughety, Andrew F and Jennifer F Reinganum**, "Evidence Suppression by Prosecutors: Violations of the Brady Rule," *The Journal of Law, Economics, and Organization*, 2018, *34* (3), 475–510.

**_ and _** , "Reducing Unjust Convictions: Plea Bargaining, Trial, and Evidence Disclosure," *The Journal of Law, Economics, and Organization*, 2020, *36* (2), 378–414.

**de Barreda, Ines Moreno**, "Cheap talk with two-sided private information," *Working paper*, 2010.

**Dziuda, Wioletta and Christian Salas**, "Communication with detectable deceit," *Working paper*, 2018.

**Frenkel, Sivan, Ilan Guttman, and Ilan Kremer**, "The effect of exogenous information on voluntary disclosure and market quality," *Journal of Financial Economics*, 2020.

**Glazer, Jacob and Ariel Rubinstein**, "On Optimal Rules of Persuasion," *Econometrica*, 2004, *72* (6), pp. 1715–1736.

**Grossman, Gene M and Michael L Katz**, "Plea bargaining and social welfare," *The American Economic Review*, 1983, *73* (4), 749–757.

**Hart, Sergiu, Ilan Kremer, and Motty Perry**, "Evidence games: Truth and commitment," *American Economic Review*, 2017, *107* (3), 690–713.

**Ingraham, Barton L**, "The right of silence, the presumption of innocence, the burden of proof, and a modest proposal: A reply to O'Reilly," *Journal of Criminal Law and Criminology*, 1995, *86*, 559.

**Ioannidis, Konstantinos, Theo Offerman, and Randolph Sloof**, "Lie detection: A strategic analysis of the Verifiability Approach," *Working paper*, 2020.

**Ishida, Junichiro and Takashi Shimizu**, "Cheap talk with an informed receiver," *Economic Theory Bulletin*, 2016, *4* (1), 61–72.

**Ispano, Alessandro**, "Persuasion and receiver's news," *Economics Letters*, 2016, *141*, 60–63.

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

**Kaplow, Louis**, "On the optimal burden of proof," *Journal of Political Economy*, 2011, *119* (6), 1104–1140.

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

_ **and R. Preston McAfee**, "Signaling Character in Electoral Competition," *American Economic Review*, June 2007, *97* (3), 852–870.

**Kassin, Saul M and Karlyn McNall**, “Police interrogations and confessions,” *Law and Human Behavior*, 1991, *15* (3), 233–251.

**_ , Christian A Meissner, and Rebecca J Norwick**, ““I’d know a false confession if I saw one”: A comparative study of college students and police investigators,” *Law and Human Behavior*, 2005, *29* (2), 211–227.

**Lai, Ernest K**, “Expert advice for amateurs,” *Journal of Economic Behavior & Organization*, 2014, *103*, 1–16.

**Milgrom, Paul R.**, “Good News and Bad News: Representation Theorems and Applications,” *Bell Journal of Economics*, Autumn 1981, *12* (2), 380–391.

**Milgrom, Paul R and Robert J Weber**, “Distributional strategies for games with incomplete information,” *Mathematics of operations research*, 1985, *10* (4), 619–632.

**O’Reilly, Gregory W**, “England limits the right to silence and moves towards an inquisitorial system of justice,” *Journal of Criminal Law and Criminology*, 1994, *85*, 402.

**Pei, Harry**, “Uncertainty about Uncertainty in Communication,” *Working paper*, 2017.

**Perez-Richet, Eduardo and Vasiliki Skreta**, “Test Design under Falsification,” *Working paper*, 2020.

**Redlich, Allison D, Shi Yan, Robert J Norris, and Shawn D Bushway**, “The influence of confessions on guilty pleas and plea discounts.,” *Psychology, Public Policy, and Law*, 2018, *24* (2), 147.

**Reinganum, Jennifer F**, “Plea bargaining and prosecutorial discretion,” *The American Economic Review*, 1988, pp. 713–728.

**Shin, Hyun Song**, “The Burden of Proof in a Game of Persuasion,” *Journal of Economic Theory*, 1994, *64* (1), 253 – 264.

**Siegel, Ron and Bruno Strulovici**, “Judicial mechanism design,” *Working paper*, 2018.

**_ and _** , “The Economic Case for Probablity-Based Sentencing,” *Working paper*, 2019.

**Sullivan, Thomas P**, "Electronic Recording of Custodial Interrogations: Everybody Wins," *Journal of Criminal Law and Criminology*, 2005, *95* (3), 1127.

# Appendix

## A Lying and equilibrium updating

Given the lying function $\boldsymbol{\ell}$ of guilty types with range $L$, let the inverse lying correspondence $\boldsymbol{g} = \boldsymbol{\ell^{-1}} : L \to Y_\ell$ associate to each lie $m$ the set of guilty types in $Y_\ell$ for which $\ell(y) = m$. We allow $\boldsymbol{g}$ to also take sets as arguments, i.e. $g(A) = \{y \in Y_\ell : \ell(y) \in A\}$ for any set $A \subseteq L$. We impose the following restriction on equilibrium beliefs.

(RCP) $R$'s equilibrium beliefs must be derived form a regular conditional probability.

**Lemma A.1.** *Under restriction RCP, R's equilibrium belief $\mu(m, z)$ in $D$ is such that*

($\mu.i$) *$\mu(m, z)$ obtains from Bayes' rule whenever $\lambda(\boldsymbol{g}(m)) > 0$, so that then $\mu(m, z) = 0$;*

($\mu.ii$) *$\lambda \circ \boldsymbol{g}$-almost surely $\mu(m, z) = 1$ if $m \notin L$;*

($\mu.iii$) *$\lambda \circ \boldsymbol{g}$-almost surely $\mu(m, z) = \mu(m, z')$ for all $z, z' \geq m$.*

*Proof.* Point ($\mu.i$) and ($\mu.ii$) are trivial. For point ($\mu.iii$), given the strategy of liars $\boldsymbol{\ell}$ and evidence $z$ such that $R$ has discretion, the total pushforward measure of the Lebesgue measure $\lambda$ (i.e. both by the liars and the innocents) on the messages $[t, z)$ is $\lambda \circ \boldsymbol{g} + \lambda$, since innocent types are honest and no liar is excluded by $z$ from $g([t, z))$. If $\mu(., z, .)$ is a regular conditional probability then, for every $m \in [t, z)$, for every measurable $A \subseteq [t, z)$ and $B \in [0, 1]$ we have that:

$$\lambda(B \cap (g(A) \cup A)) = \int_A \mu(m, z, B) \mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda). \tag{7}$$

Choosing $B = [t, z)$ we have that $\mu(m, z, B) = \mu(m, z)$ and for any $A \subseteq B$ we have that $\lambda(A) = \int_A \mu(m, z) \mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda)$. For any other $z' > z$ we have that for all measurable $A \subseteq B$

$$\int_A \mu(m, z, B) \mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda) = \int_A \mu(m, z', B') \mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda),$$

where $B' = [t, z')$ and $\mu(m, z') = \mu(m, z', B')$ hence $\lambda \circ \boldsymbol{g}$-almost surely $\mu(m, z) = \mu(m, z')$ for $m \in [t, z)$. □

Throughout, for ease of exposition we assume that restriction $\mu.ii$ and $\mu.iii$ hold for each $m$ rather than only $\lambda \circ \boldsymbol{g}$-almost surely, that is,

($\mu.ii*$) $\mu(m, z) = 1$ if $m \notin L$;

($\mu.iii*$) $\mu(m, z) = \mu(m, z')$ for all $z, z' \geq m$.

# B Main proofs

## B.1 Proof of proposition 1

First, we identify some properties that must hold in any equilibrium (section B.1.1). Then, we distinguish three possible cases (all guilty types lie, some guilty types lie and the rest confess, some guilty types are silent) and show that in each case the set of confessors, the lying region and the measure of liars and silent types are uniquely pinned down (section B.1.2). Next, we show that the three cases do not overlap and span the whole parameter space (section B.1.3). Finally, we show that in each case the equilibrium singled out in proposition 1 indeed exists (section B.1.4).

### B.1.1 Preliminary observations

In any equilibrium, the expected payoff of any type[30] from confessing with a message $m \leq y$ is $\pi_c = 0$, while confessing with a message $m > y$ yields $0 - b(m - y) < \pi_c$. The expected payoff of type $y$ from denying by lying upward, i.e. from sending a message such that $m > y$ and $m \geq t$ is

$$\pi_\ell(m; y) = \underbrace{(1 - m)A(m)}_{\text{lie not detected}} - \underbrace{(m - y)b}_{\text{lie detected}}. \tag{8}$$

[30] Formally, the expected payoff of each type $y$ is defined conditional on $R$ having evidence, i.e. $z > y$. From the perspective of type $y$, this conditioning amount to a payoff normalization (a division by $1 - y$), which we can hence ignore. Throughout, with a slight abuse of terminology, we simply refer to $\int_y^1 \pi\Big(y, m, z, a\,(m, z)\Big)\mathrm{d}z$ as to the expected payoff of type $y$.

The expected payoff of type $y$ from remaining silent when he is guilty is

$$\pi_{s,g}(y) = \underbrace{1 - Z_s}_{\text{inconclusive evidence}} - \underbrace{(t-y)b}_{\text{conclusive evidence}} \tag{9}$$

and when he is innocent is

$$\pi_{s,i}(y) = 1 - Z_s \mathbb{1}_{y \leq Z_s}. \tag{10}$$

The expected payoff of innocent type $y$ from denying by lying downward, i.e. form sending a message $m \in [t, y)$ is

$$\pi_{d\ell,i}(m; y) = \int_y^1 a(m,z)\mathrm{d}z. \tag{11}$$

Finally, the expected payoff of innocent type $y$ from being honest is simply equation (8), or equivalently equation (11), evaluated in $m = y$

$$\pi_{h,i}(y) = (1-y)A(y). \tag{12}$$

Let us say that type $y$ separates if no other type sends $m = y$.

**Lemma B.1.** *In any equilibrium:*

*(i) there exists a $\bar{y} \in (t, 1)$ such that innocent types $y > \bar{y}$ separate and innocent types $y < \bar{y}$ do not, i.e. $L = [t, \bar{y})$ or $L = [t, \bar{y}]$;*

*(ii) each $m \in L$ is sent by a set of guilty types of measure zero;*

*(iii) each $m \in L$ gives a liar the same payoff;*

*(iv) $A(m)$ is continuous and increasing and converges to 1 in $\bar{y}$. In particular, for $m \in L$*

$$A(m) = \frac{1 - \bar{y} - b\,(\bar{y} - m)}{1-m}. \tag{13}$$

*Proof.* Point (i):

- **A sufficiently high innocent type separates**. Consider message $m_\epsilon = 1-\epsilon > t$, where $\epsilon > 0$ is arbitrarily small. This message is sent by innocent type $y = m_\epsilon$. Suppose there is a guilty type $y_\epsilon$ who also sends this message. Using equation (8),

$y_\epsilon$ earns $\epsilon A(m_\epsilon) - (1 - \epsilon - y_\epsilon)b$, which letting $\epsilon$ go to 0 converges to $-(1 - y_\epsilon)b$. There is therefore an arbitrary small $\epsilon$ such that type $y_\epsilon$ could profitably deviate to confessing honestly.

- **If an innocent type separates, so do higher types.** Suppose by contradiction that innocent type $y$ separates but innocent type $y' > y$ does not. As $A(y) = 1$ by restriction $\mu.ii*$, from equation (8) it is apparent that guilty types strictly prefer $m = y$ to $m = y'$ and hence also type $y'$ must separate.

- **A sufficiently low innocent type does not separate**. Suppose innocent type $t$ separates so that, by restriction $\mu.ii*$, $A(t) = 1$ and consider type $t_\epsilon^- = t - \epsilon$, where $\epsilon > 0$ is arbitrarily small. From equation (8) it is apparent that the expected payoff of type $t_\epsilon^-$ from lying to $m \geq t$ is decreasing in $m$. As he is not sending $t$, he must then earn the maximum between the expected payoff from confessing honestly, i.e. 0, and staying silent, i.e. $\pi_{s,g}(t_\epsilon^-) = 1 - Z_s - b\epsilon$ by equation (9). The expected payoff of type $t_\epsilon^-$ from deviating to $m = t$ is $1 - t - b\epsilon$, which for $\epsilon$ arbitrarily small is arbitrarily close to $1 - t$ so that the deviation is profitable.

Point (ii): Each message $m \in L$ is sent by innocent type $y = m$, a set that has zero measure. By point (i), it is also sent by at least a guilty type. If the set of guilty types who send $m$ has positive measure, by restriction $\mu.i$, i.e. Bayes' rule, $A(m) = 0$. Then, comparing equation (8) and (9) clarifies that each guilty type sending $m$ could profitably deviate to staying silent or confessing honestly.

Point (iii): From equation (8) it is apparent that the expected payoff difference $\pi_\ell(m; y) - \pi_\ell(m'; y)$ from any two messages $m$ and $m'$ such that $m > m' \geq y$ is independent from $y$. Therefore, if in equilibrium $m$ and $m'$ are sent by two distinct types $y \leq m'$ and $y' \leq m'$, any type $y'' \leq m'$ is indifferent between the two messages.

Point (iv): By point (i) and (iii), $\pi_\ell(m; y)$, which is defined in equation (8) and represents the expected payoff of guilty type $y$ who lies and denies, must be must constant in $m$ over $L$. Solving $\pi_\ell(m; y) = k$ with respect to $A(m)$, where $k$ is a constant, yields $A(m) = (k + b(m - y))/(1 - m)$, from which it is apparent that $A(m)$ is continuous and strictly increasing in $m$ over $L$. Also, $A(m)$ must converge to 1 at $m = \bar{y}$ since, by point (i) and restriction $\mu.ii*$, $A(m) = 1$ for each $m > \bar{y}$, so that if $A(\bar{y}) < 1$ type $y$ could profitably deviate to $m = \bar{y} + \epsilon$ for $\epsilon > 0$ arbitrarily small. As $A(m)$ is also differentiable, so is $\pi_\ell(m;y)$. Thus, differentiating equation (8) with respect to $m$ and setting the expression equals to zero, as required by the indifference of type $y$ to any $m \in L$, yields

$$\underbrace{(1-m)\,A'(m)}_{\text{benefit of increase in action if lie undetected}} = \underbrace{b + A(m)}_{\text{cost of higher chance of lie being detected}}.$$

Solving this differential equation with terminal condition $A(\bar{y}) = 1$ yields equation (13).

□

**Lemma B.2.** *In any equilibrium:*

*(i) whenever $A(m) \in (0,1)$ $R$ is indifferent between actions, i.e.*

$$\mu(m,z) = \alpha; \tag{14}$$

*(ii) it must be that $(1-\alpha)\lambda = \alpha(\lambda \circ \boldsymbol{g})$ over measurable subsets of $L$ and hence, in particular,*

$$\frac{\lambda\,(Y_\ell)}{\lambda\,(L)} = \frac{1-\alpha}{\alpha}. \tag{15}$$

*Proof.* Point (i) follows directly from equation (2) and restriction *$\mu.iii*$*. For point (ii), by the previous point and point (iv) of lemma B.1, $\mu(m,z,L) = \mu(m,z) = \alpha$ for all $m \in L$ except possibly $m = t$ (if $A(t) = 0$) and $m = \bar{y}$ (if $L = [t, \bar{y}]$). Choosing $z = 1$ and $B = L$ in equation (7) yields $\lambda(A) = \alpha\lambda(\boldsymbol{g}(A)) + \alpha\lambda(A)$ and hence the result. Equation (15) follows from choosing $A = L$. □

Evaluating equation (13) in $t$ shows that there must be a one to one relationship between $A(t)$ and $\bar{y}$, i.e.

$$A(t) = \frac{1 - \bar{y} - b\,(\bar{y} - t)}{1-t} \quad \text{or, equivalently,} \tag{16}$$

$$\bar{y} = \frac{1 - (1-t)A(t) + bt}{1+b}. \tag{17}$$

Also, as by lemma B.1 at least some guilty type $y$ must send $t$ and $y$ must be indifferent to any pooling lie, evaluating equation (8) in $t$ yields the expected payoff from lying for type $y$

$$\pi_\ell(y) = (1-t)A(t) - (t-y)b. \tag{18}$$

Let us denote by $v$ the expected payoff for a guilty type from remaining silent conditional on evidence being inconclusive, i.e., from equation (1),

$$v \equiv \mathbb{P}\left(z > Z_s \,|\, \text{inconclusive evidence}, \mathcal{Y} = 0\right) = \frac{1 - Z_s}{1 - t}. \tag{19}$$

**Lemma B.3.** *In any equilibrium, lying yields a guilty type at least the same expected payoff as staying silent, i.e.*

$$A(t) \geq v, \tag{20}$$

*with equality if the set of silent types $Y_s$ is non-empty.*

*Proof.* Replacing equation (19) in (9) and subtracting from equation (18) yields $\pi_\ell(y) - \pi_{s,g}(y) = A(t) - v$. If inequality (20) was violated, no type would lie as required by lemma B.1. Likewise, a guilty type can find it optimal to stay silent only if the inequality is not strict. □

**Lemma B.4.** *In any equilibrium, if non-empty the set of confessors $Y_c$ is $[0, y_c)$ or $[0, y_c]$, where*

$$y_c \equiv \frac{bt - (1-t)A(t)}{b}, \tag{21}$$

*and $Y_c$ has positive measure if and only if*

$$b > \frac{1-t}{t} A(t). \tag{22}$$

*Proof.* By lemma B.4, $\pi_\ell(y)$, which is defined in equation (18), represents the equilibrium expected payoff for a guilty type who does not confess. As the expression is strictly increasing in $y$ and positive for $y = t$, while confessing (honestly) yields 0, $y_c$ as defined in equation (21) is the unique solution to $\pi_\ell(t; y) = 0$. Also, $y_c > 0$ if and only if equation (22) holds. □

### B.1.2 Possible cases

Throughout superscripts index the respective cases. Also, we innocuously ignore subcases in which the set of confessors $Y_c$ and/or of silent types $Y_s$ are non-empty but have zero measure.

**Case I: some guilty types lie and the rest confess.** Suppose $\lambda\left(Y_s^I\right) = 0$ but $\lambda\left(Y_c^I\right) > 0$. It must then be that $y_c^I$ is as in equation (21) and $\lambda\left(Y_\ell^I\right) = t - y_c$ so that, by equation (15), (16) and (17), $\bar{y}^I = \frac{\alpha+bt}{\alpha+b}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I = \frac{(1+b)t-(1-\alpha)}{b+\alpha}$. By equation (22), this case can only occur if $b > \frac{1-t-\alpha}{t}$.

**Case II: all guilty types lie.** Suppose $\lambda\left(Y_s^I\right) = 0$ and $\lambda\left(Y_c^I\right) = 0$, so that $\lambda\left(Y_\ell^{II}\right) = t$. By equation (15), (16) and (17) it then follows that $\bar{y}^{II} = \frac{t}{1-\alpha}$ and that $A^{II}(t) = \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}$. This case can only occur if $b \leq \frac{1-t-\alpha}{t}$, so that equation (22) is violated (then, indeed $\bar{y}^{II} < 1$ and $A^{II}(t) > 0$).

**Case III: some guilty types are silent.** Suppose $\lambda\left(Y_s^I\right) > 0$. It must then be that $A^{III}(t) = v$ by equation (20), so that, by equation (17), $\bar{y}^{III} = \frac{1-(1-t)v+bt}{1+b}$ (where, using equation (19), indeed $\bar{y}^{III} < Z_s$). By equation (15), $\lambda\left(Y_\ell^{III}\right) = \frac{(1-\alpha)(1-t)(1-v)}{\alpha(b+1)}$. To determine $\lambda\left(Y_s^{III}\right)$, we must then distinguish two subcases depending on whether $\lambda\left(Y_c\right) > 0$. From equation (22), $\lambda\left(Y_c\right) > 0$ if and only if

$$v < v_{IIIa} \equiv \frac{tb}{(1-t)}. \tag{23}$$

- **Case IIIa: some types confess**. When $v < v_{IIIa}$, from equation (21), $\lambda\left(Y_c^{III}\right) = y_c^{III} = \frac{bt-(1-t)v}{b} > 0$, so that $\lambda\left(Y_s^{III}\right) = t - \lambda\left(Y_\ell^{III}\right) - \lambda\left(Y_c^{III}\right) = \frac{(1-t)(v\alpha - b(1-v-\alpha))}{b(1+b)\alpha}$.
- **Case IIIb: no types confess**. When $v \geq v_{IIIa}$, $\lambda\left(Y_s^{III}\right) = t - \lambda\left(Y_\ell^{III}\right) = t - \frac{(1-\alpha)(1-t)(1-v)}{\alpha(b+1)}$.

### B.1.3 Equilibrium regions

**Lemma B.5.** *Every equilibrium has the same $\lambda\left(Y_c\right)$, $\lambda\left(Y_s\right)$ and $\lambda\left(Y_\ell\right)$. In particular,*

- $\lambda\left(Y_s\right) > 0$ *iff*

$$\frac{1-Z_s}{1-t} > max\left\{\frac{(1-\alpha)b}{b+\alpha}, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\}; \tag{24}$$

- $\lambda\left(Y_c\right) > 0$ *iff*

$$b > \frac{1-t}{t} max\left\{\frac{1-Z_s}{1-t}, \frac{(1-\alpha)b}{b+\alpha}\right\}, \tag{25}$$

*which if* $\lambda(Y_s) = 0$ *simplifies to*

$$b > \frac{1 - t - \alpha}{t}. \tag{26}$$

*Proof.* In each of the three cases described at section B.1.2, $\lambda(Y_c)$, $\lambda(Y_s)$ and $\lambda(Y_\ell)$ are uniquely pinned down. To prove the statement, hence it suffices to show the three cases do not overlap and span the whole parameter space. Case I and case II do not overlap since case I requires $b > \frac{1-t-\alpha}{t}$ and case II the reverse inequality. However, case III cannot overlap with case I and II either. Consider for instance case I (the argument for case II is analogous). If $A^{III}(t) = v > A^{I}(t)$, the candidate equilibrium at case I cannot exist by lemma B.3. Suppose instead $A^{III}(t) = v \leq A^{I}(t)$ and both equilibria exist. From equation (21), it must be that $y_c^I \leq y_c^{III}$, so that $\lambda\left(Y_c^I\right) \geq \lambda\left(Y_c^{III}\right)$. Moreover, $\lambda\left(Y_\ell^I\right) = t - \lambda\left(Y_c^I\right) > \lambda\left(Y_\ell^{III}\right) = t - \lambda\left(Y_c^{III}\right) - \lambda\left(Y_s^{III}\right)$. Since in equilibrium $\bar{y}$ is strictly increasing in $\lambda(Y_\ell)$ by lemma B.2, it follows that $\bar{y}^I > \bar{y}^{III}$ and, from equation (16), that $A^I(t) > A^{III}(t)$, yielding a contradiction. It follows that the prevalence of case I, II or III is uniquely determined by $t$, $b$, $v$ and $\alpha$:

- when $b \leq \frac{1-t-\alpha}{t}$: case II occurs if $v \leq A^{II}(t)$ and case III otherwise;
- when $b > \frac{1-t-\alpha}{t}$: case I occurs if $v \leq A^{I}(t)$ and case III otherwise.

$A(t)$ can hence be written as

$$A(t) = \begin{cases} max\left\{v, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\} & \text{if}\, b \leq \frac{1-t-\alpha}{t} \\ max\left\{v, \frac{(1-\alpha)b}{b+\alpha}\right\} & \text{if}\, b > \frac{1-t-\alpha}{t}. \end{cases} \tag{27}$$

Using the definition of $v$ (equation (19)), equation (24) is simply the condition for the prevailing of case III after noting that $A^{II}(t) > A^{I}(t)$ if and only if $b < \frac{1-t-\alpha}{t}$. Using again the definition of $v$, equation (25) obtains by evaluating equation (22) using equation (27) and noting that $\lambda(Y_c) = 0$ by construction whenever case II occurs. Finally, equation (26) obtains by taking $max\left\{\frac{1-Z_s}{1-t}, \frac{(1-\alpha)b}{b+\alpha}\right\} = \frac{(1-\alpha)b}{b+\alpha}$ in equation (25), which by the previous observations must necessarily be the case when $\lambda(Y_s) = 0$. □

### B.1.4 Existence

The previous observations clarify that the strategy of confessors is optimal and that no other type prefers to confess. The strategy of a guilty type who does not confess is also optimal. Indeed, he is indifferent between sending any lie $m \in [t, \bar{y}]$, he prefers doing so rather than to stay silent whenever no type is silent in equilibrium and he is indifferent to remain silent otherwise. Conversely, any $m > \bar{y}$ is strictly dominated for him since, as $A(m) = 1$, it is apparent from equation (8) that his expected payoff is strictly decreasing in $m$ in that region. For the same reasons, an innocent type $y \in [t, \bar{y})$ is indifferent between being honest and sending any lie $m \in [y, \bar{y})$ and he strictly prefers to be honest rather than to send any lie $m > \bar{y}$. From a comparison of equation (11) and (12) and the fact that $A(m)$ is increasing, he also strictly prefers to be honest rather than to deny with a message $m < y$. From a comparison of equation (10) and (12) and the fact that $A(t) \geq v$ and $A(m)$ is increasing, he also strictly prefers to be honest rather than to be silent (except type $t$, who might be indifferent). Finally, as $A(m) = 1$ for each $m \geq \bar{y}$ by lemma B.1, by being honest an innocent type $y \geq \bar{y}$ earns the maximum attainable payoff.

To conclude, in any of the three cases described at section B.1.2, one can always take $Y_\ell = [y_\ell, t)$, where $y_\ell = \frac{t-(1-\alpha)\bar{y}}{\alpha}$, so that equation (15) is satisfied and $\lambda(Y_\ell) + \lambda(Y_c) + \lambda(Y_s) = t$. Also, the lying function at point (ii) of the proposition has image $L = [t, \bar{y})$ and satisfies restriction RCP and equation (14). Indeed, by equation (7), $\mu(m, z) = \frac{1}{1+\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m)}$, where $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}$ denotes the Radon-Nikodym derivative, and $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m) = \frac{1}{\ell'(y)} = \frac{1-\alpha}{\alpha}$ and hence $\mu(m, z) = \alpha$. Finally, since $\mu(m, z)$ is independent from $z$ for any $m$, there are multiple choices of $a(m, z)$ for which $A(m)$ satisfies equation (13) over $L$ as required by lemma B.1. In particular, so does the strategy described at point (v) of the proposition with $\bar{z}(m) \equiv \bar{y} + b(\bar{y} - m)$ for $m \in L$ and $\bar{z}(m) \equiv m$ otherwise.

## B.2 Proof of corollary 1

Suppose there are silent types in the equilibrium singled out in proposition 1. Then equilibria can only differ in the identity of liars and silent types, and in the exact shape of the lying function. However, by construction, these types are always indifferent between any pooling lie or staying silent. If there are no silent types in the equilibrium singled out in proposition 1, equilibria can only differ in the exact shape of the lying function. But again, liars are indifferent between any pooling lie. Hence, for $S$, the result follows directly from these indifference conditions.

As for $R$, her ex-ante expected loss (equation (4)) can be rewritten as

$$(1-\alpha)\int_L (1-y)\,(1-A\,(y))\,\mathrm{d}\lambda + \alpha\frac{1-\alpha}{\alpha}\int_L (1-y)\,(A\,(y))\,\mathrm{d}\lambda + \alpha(1-Z_s)\lambda(Y_s) \tag{28}$$

$$=(1-\alpha)\int_L (1-y)\,\mathrm{d}\lambda + \alpha\lambda\,(Y_s)\,(1-Z_s) \tag{29}$$

$$=\alpha\lambda\,(Y_\ell)\left(1-t-\frac{\alpha\lambda\,(Y_\ell)}{2(1-\alpha)}\right)+\alpha\lambda\,(Y_s)\,(1-Z_s). \tag{30}$$

Equation (28) obtains by a change of variables under pushforward integrability where, by lemma B.2, $(1-\alpha)\lambda = \alpha\,(\lambda\circ\boldsymbol{g})$. Equation (29) obtains because the first term of equation (28) is zero whenever the second term is one. Intuitively, this can be understood as $R$ being indifferent between always choosing $a=1$, which only generates type II errors, and always choosing $a=0$, which only generates type I errors. Since $L$ and $\lambda\,(Y_s)$ are the same in every equilibrium, the result obtains. Equation (30) is for future use and, after some rearranging, obtains by substituting back the length of the pooling interval as a function of the measure of liars using equation (15).

Ex-post, i.e. once $z$ has realized, if $z \le t$, $R$'s loss is zero. Otherwise, using analogous simplifications as above, $R$'s expected loss is

$$(1-\alpha)\int_{L:y<z}(1-a\,(y,z))\,\mathrm{d}\lambda + \alpha\int_{Y_\ell:\ell(y)<z} a\,(\ell\,(y)\,,z)\,\mathrm{d}\lambda + \alpha\lambda\,(Y_s)\,\mathbb{1}_{z>Z_s}$$

$$=(1-\alpha)\int_{L:y<z}\mathrm{d}\lambda + \alpha\lambda\,(Y_s)\,\mathbb{1}_{z>Z_s},$$

which is again identical across equilibria.

## B.3 Proof of proposition 6

We first describe players' equilibrium strategies and then verify sequential rationality along the equilibrium path (since beliefs are free off the path, we can always make sure that decisions specified there are sequentially rational). Throughout, all specified beliefs are consistent with restriction RCP, and $g(m) \equiv \ell^{-1}(m)$.

**Candidate equilibrium strategies.** Let $\bar{y}$, $\bar{z}(m)$ and $S$'s behavioral strategy in stage 1 be as in proposition 1 (with $y_c = 0$ by convention if $\lambda(Y_c) = 0$). $R$ always chooses $a = 0$ if $m < t$ and $a = 1$ if $m \geq \bar{y}$ (provided $S$ is not caught in a lie, otherwise off the equilibrium path $R$ again automatically chooses $a = 0$ and $S$ gets $-b$). Instead, for $m \in [t, \bar{y})$, $R$ chooses $a = 1$ if $z \geq Z(m)$ and delegate to $I$ if $z < Z(m)$ where

$$\begin{aligned} Z(m) &= \bar{z}(m) + b(m - g(m)) = \bar{z}\left(g\left(m\right)\right) \ \in \left(\bar{z}(m), 1\right) \\ &= \bar{y} - b(t - \bar{y} + y_c) + \frac{bt}{\alpha} - \frac{b(1-\alpha)}{\alpha} m. \end{aligned} \tag{31}$$

Consider now stage 3 after message $m$ was sent and $R$ delegated in accordance with the strategy above. Then, $I$ chooses $\mathcal{M}_I = \{g(m), m\}$ and $S$ sends $m_2 = g(m)$ if guilty and $m_2 = m$ if innocent. $I$ chooses $a_2 = 0$ if $m_2 = g(m)$ (and off the equilibrium path also if $S$ is caught in a lie, in which case $S$ gets $-b$ as in the baseline model). If $m_2 = m$ then $I$ follows $\bar{z}(m)$. Finally, assume that if $R$ delegates when she should not given $Z(m)$, $I$ always chooses $a_2 = 0$.

**Sequential rationality.** $R$'s strategy upon a pooling message is sequentially rational:

- if $S$ is caught in a lie, $R$ believes that $S$ is surely guilty and anticipates he will confess honestly to $I$ who will choose $a_2 = 0$;
- if $S$ is not caught in a lie, $R$ believes $S$ is innocent with probability $\alpha$ and:
  - when $z \geq Z(m)$, she is hence indifferent to any action or delegate to $I$, who will choose $a_2 = 0$;
  - when $\bar{z}(m) \leq z < Z(m)$, she strictly prefers to delegate since she will make no error at all since $I$ will choose $a_2 = 0$ if $S$ is guilty and $a_2 = 1$ if $S$ is innocent;
  - when $z < \bar{z}(m) < Z(m)$, $R$ knows that $I$ will choose $a_2 = 0$, no matter if $S$ is guilty or innocent. Given that $R$ believes $S$ is innocent with probability $\alpha$ she is again just indifferent between delegating and choosing $a = 1$.

$I$'s strategy is also sequentially rational together with the belief that $S$ is surely innocent in the only instance in which she does not choose $a_2 = 0$.

Finally, consider $S$'s strategy. When interrogated by $I$, the strategy of innocent type $m$ is clearly optimal. As for a guilty type $g(m)$, given his belief that $z < Z(m)$, by construction he is now indifferent between confessing honestly, which yields 0, and sending $m_2 = m$, since his expected payoff from doing so is $-b\,(m - g\,(m)) + Z(m) - \bar{z}\,(m) = 0$. Consider now stage 1 and notice that for each type $y$ the joint on the equilibrium path behavior of $R$ and $I$ is in expectation equivalent to the optimal mechanism $\hat{\boldsymbol{z}}^{\star}$. In particular, for pooling innocent types $\bar{z}(y) = \hat{z}^{\star}(y)$ and for pooling guilty types $\bar{z}(g(m)) = \bar{z}(y) = \hat{z}^{\star}(y)$. It follows that no type $y$ can benefit from playing as if he was some other type $y''$ throughout the game otherwise she would do so in the optimal mechanism as well. Finally, no type can profit from deviating at stage 1 to some pooling message $m'$ and then send $m_2 = m'$ in stage 2. Indeed the choice of $Z(m)$ is such that it is as if this type was deviating in the equilibrium of the baseline model, where it is also the case that $a(m, z) = 1$ whenever $z \geq Z(m)$. In short, $S$ can either behave as if he was another type or lie and stick to his stage 1 story. In the first case, it is as if he was playing in the optimal mechanism, hence this type of deviation is not profitable. In the second case, it is exactly as he was playing in the equilibrium of the baseline model, so that this type of deviation is again not profitable.

# Online Appendix

## C Truth-leaning and honesty

Consider the more general game described at section 6.1. We say that $\langle \boldsymbol{M}, \boldsymbol{a} \rangle$ is a **quasi-equilibrium** if $\boldsymbol{M} = (M_y)_{y \in [0,1]}$ and for all $y \in [0,1] : \emptyset \neq M_y \subseteq \mathcal{M}$ and for every $m \in M_y$, $m$ is optimal for $y$ given $\boldsymbol{a}$. This equilibrium notion can be easily applied to any perturbed game, possibly with infinite type and action sets, and, after adjusting notation slightly, an equilibrium as defined in the main body is a quasi-equilibrium.

We think of $M_y$ as the support of the behavioral strategy of type $y$ (which is a regular conditional probability if strategies are distributional), while $A(m)$ as defined in section 3 represents the expected payoff of type $m$ from choosing message $m$ determined by $\boldsymbol{a}$. For $m = s$, $A(s)$ may depend on $y$ and we write $A_y(s) = \int_{[y,1]} a(s,z)/(1-y) \mathrm{d}\lambda$ for $y \geq t$ and $A(s) = A_y(s) = \int_{[t,1]} a(s,z)/(1-t)\mathrm{d}\lambda$ for $y < t$. Also, if some type $y$ sends a message $m < y$ we write $A_y(m) = \int_{[y,1]} a(m,z)/(1-y)\mathrm{d}\lambda$.

A **truth-leaning test sequence** of the baseline game with $(\varepsilon_y^n)_{y \in [0,1]} \to 0$ is a sequence of games such that each type $y \in [0,1]$ obtains an extra $\varepsilon_y^n > 0$ when choosing $m = y$ relative to its baseline game payoff and if for each $m \in [0,1]$ we have that: if $t \leq m \notin \cup_{y<t} M_y^n$ then $A^n(m) = 1$ and if $t > m \notin \cup_{y \geq t} M_y^n$ then $A^n(m) = 0$. Namely, there is a small reward for honesty and if a denying message is never sent by guilty types or a confessing message is never sent by innocent types then $R$'s action, provided that $S$ is not caught in a lie, is 1 and 0, respectively.

$\langle \boldsymbol{M}, \boldsymbol{a} \rangle$ is a **truth-leaning quasi-equilibrium** if there is a truth-leaning test sequence and a corresponding sequence $\langle \boldsymbol{M}^n, \boldsymbol{a}^n \rangle$ of quasi-equilibria of the perturbed games such that for all $y \in [0,1], M_y \subseteq \limsup_{n \to \infty} M_y^n = \{m \in \mathcal{M} | \exists (n_k)_{k \in \mathbb{N}}, m_{n_k} \in M_y^{n_k} : \lim_{k \to \infty} m_{n_k} = m\}$ and for all $m : A^{n_k}(m) \to A(m)$.

**Proposition C.1.** *In any truth-leaning quasi-equilibrium $M_y = \{y\}$ for $y \geq t$, i.e. innocents are honest, and if $t > m \in M_y$ then $M_y = \{y\}$, i.e. confessors are honest. Moreover, the equilibrium in proposition 1 is a truth-leaning (quasi-)equilibrium.*

*Proof.* Consider a truth-leaning quasi-equilibrium and a test sequence justifying it. First, we show that innocents are honest. Suppose by contradiction that $\exists y \geq t, m \neq y : m \in M_y$ but $m^n \neq y$ in $M_y^n$ for some $n$. There must hence be some $t > y' : y \in M_{y'}^n$. We have to distinguish four cases: $m^n > y > y'$, $m^n = s$, $y' \leq m^n < y$ and $m^n \leq y' < y$. For each case there are two weak inequalities that must hold in the quasi-equilibrium of the nearby game, the first saying that $y$ weakly prefers $m^n$ to $y$ and the second saying that $y'$ weakly prefers $y$ to $m^n$, yielding respectively:

$$\begin{aligned}
(1-y)A^n(y) + \varepsilon_y^n &\leq (1-m)A^n(m^n) - b_\ell(m^n - y) && \text{and}\\
(1-m^n)A^n(m^n) - b_\ell(m^n - y') &\leq (1-y)A^n(y) - b_\ell(y-y');\\
(1-y)A^n(y) + \varepsilon_y^n &\leq (1-y)A_y^n(s) && \text{and}\\
(1-t)A^n(s) - b_s(t-y') &\leq (1-y)A^n(y) - b_\ell(y-y');\\
(1-y)A^n(y) + \varepsilon_y^n &\leq (1-y)A_y^n(m^n) && \text{and}\\
(1-m^n)A^n(m^n) - b_\ell(m^n - y') &\leq (1-y)A^n(y) - b_\ell(y-y');\\
(1-y)A^n(y) + \varepsilon_y^n &\leq (1-y)A_y^n(m^n) && \text{and}\\
(1-y')A_{y'}^n(m^n) &\leq (1-y)A^n(y) - b_\ell(y-y').
\end{aligned}$$

For each case we can decrease both sides of the inequality for $y$ by $b_\ell(y-y')$ and get a contradiction by showing that the RHS is less than or equal to the LHS of the inequality for $y'$. One just have to arrange the $b_s, b_\ell$ terms separated on one side and compare them to the difference of the measures of $z$ where $a(m,z) = 1$ (in the second case, one must also use that $b_\ell \geq b_s$). The intuition is that $y$ must be indifferent between $y$ and $m^n$ as otherwise $y'$ would strictly prefer $m^n$ to $y$. But even if $y$ is indifferent, $y'$ will strictly prefer $m$ to $y$ because $y'$ does not obtain the extra $\varepsilon_y^n$.

Given that innocents are honest, by the definition of truth-leaning test sequence, in the quasi-equilibrium of the nearby game after confession $R$'s action upon not catching $S$ in a lie must be 0, so that confessors obtain 0 and honest confessors additionally obtain $\varepsilon^n$. It follows that in the limit confessors indeed are honest and obtain 0 as we assumed in the body of the paper.

Finally, the second statement is trivial if there is a positive measure of silent types or there are no confessors. To be precise, if there are confessors and silent types we must choose $Y_c = [0, y_c]$ and $Y_s = (y_c, y_\ell)$. Then, for any perturbation there is a nearby quasi-equilibrium which differs from the equilibrium in proposition 1 only in that some silent types become confessors. If instead there are no confessors then there is a quasi-equilibrium of the nearby game which is identical to the equilibrium in proposition 1, no matter whether there are silent types or not.

We hence consider the case in which there are no silent types but there are confessors and we have the strictly increasing lying function from $Y_\ell = [y_c, t)$ to $L = [t, \bar{y})$. For simplicity, we describe the proof for a fixed parameter combination, which can be then easily generalized. Consider the case with parameter values $\alpha = t = 1/2$, $b_\ell = b_s = 1$ and $Z_s \geq 5/6$. In this case $y_c = 1/3, \bar{y} = 2/3$, and $A(1/2) = 1/3$. The rest of the equilibrium description clearly follows from these values given that the lying function is strictly increasing. Fix an $\eta^n > 0$. Consider a perturbation such that $\varepsilon_y^n = (\eta^n/2 + (y - y_c))/(1 - y)$ for $y \in [y_c, y_c + \eta^n]$ and set $\varepsilon_y^n$ sufficiently small otherwise. There is an equilibrium of the perturbed game with strictly increasing lying function which is similar in its structure to the one in proposition 1, with the difference that the values which pin down the equilibrium are: $A^n(1/2) = 1/3 + \eta^n, y_c^n = y_c + \eta^n$, and $\bar{y}^n = \bar{y} - \eta^n$. These equilibria almost justify the truth-leaning property of the equilibrium in proposition 1. The only problem is that for $y_c$ we have that $t \notin \limsup_{n \to \infty} M_{y_c}^n = \{y_c\}$. But given that $y_c$ is indifferent between $m = y_c$ and $m = t$ in the nearby equilibria of the nearby games, we can choose $M_{y_c}^n = \{t\}$ instead of $M_{y_c}^n = \{y_c\}$ without violating any of the equilibrium conditions. This sequence now completely justifies the truth-leaning property of the limit equilibrium which is the one described in proposition 1.[31] □

# D Remaining proofs

## D.1 Proof of proposition 2

Throughout, let the level of protection of silence be defined in terms of $v = \frac{1 - Z_s}{1 - t}$ rather than $Z_s$ as per equation (19). We first prove the first statement. Note first of all that requiring $\lambda(Y_c) > 0$ even without any protection of silence, i.e. for $v = 0$, is equivalent

[31] Notice that in the justifying sequence the nearby quasi-equilibria are in fact equilibria. Nevertheless, one could still object that this notion of convergence is too weak. In this case, instead of the strictly increasing lying function of the equilibrium in proposition 1, one can consider another lying function, induced by a distributional strategy, in which liars choose messages according to the uniform distribution over $[t, \bar{y}]$ and where $y_c$ confesses (for simplicity). One can then choose the same perturbation as before. Observe that any $y \in (y_c^n, t)$ is indifferent between any lie $m \in [t, \bar{y}^n]$ in the nearby equilibrium of the nearby game described above. Hence, each type $y > y_c^n$ can choose now messages uniformly from $[t, \bar{y}^n]$ in the nearby equilibrium and one gets convergence in distribution (or setwise convergence of the corresponding measures) for each type $y$.

to condition $b \leq \frac{1-t-\alpha}{t}$ by inequality (26), since inequality (24) for having $\lambda\left(Y_s\right) > 0$ is violated at $v = 0$ (and hence for any $v$). Thus, suppose that indeed $b \leq \frac{1-t-\alpha}{t}$. Using the results of section B.1.2 and B.1.3,

- if $v \leq A^{II}(t)$, case II of section B.1.2 obtains, i.e. all guilty types lie;
- if $v > A^{II}(t)$ case IIIb of section B.1.2 obtains, i.e. some guilty types lie and the rest are silent.

Assume case IIIb obtains (case II then obtains by continuity for $v = A^{II}(t)$ and $R$'s expected loss is independent from $v$ for any $v \leq A^{II}(t)$ since then $\lambda(Y_s) = 0$). Replacing the equilibrium measures of $\lambda\left(Y_\ell\right)$ and $\lambda(Y_s)$ in equation (30), $R$'s expected loss is

$$\begin{aligned} E(v) =& (1-\alpha)\frac{(1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)}{1+b} \\ &+\alpha(1-t)v\left(t-\frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \end{aligned} \tag{32}$$

As $E'(v)|_{v=A^{II}(t)} = -\frac{\alpha b(1-t)t}{b+1} < 0$, $E'(v)|_{v=1} = \alpha(1-t)t > 0$ and $E''(v) = \frac{(1+2b)(1-t)^2(1-\alpha)}{(1+b)^2} > 0$, the FOC gives a unique minimum

$$\tilde{v} = \frac{1-t-\alpha-b^2t\alpha+2b(1-t-\alpha)}{(1+2b)(1-t)(1-\alpha)} \in \left(A^{II}(t), 1\right). \tag{33}$$

Thus, $R$'s optimal protection level is always such that $\lambda\left(Y_s\right) > 0$ and given by $\tilde{v}$.

Suppose now that $b > \frac{1-t-\alpha}{t}$, instead, so that the value of $v$ can affect whether $\lambda\left(Y_c\right) > 0$. By the results of section B.1.2 and B.1.3, case II of section B.1.2 cannot occur since some guilty types will necessarily confess or be silent. Thus,

- if $v \leq A^{I}(t)$, case I of section B.1.2 obtains, i.e. some guilty types lie and the rest confess;
- otherwise, recalling that $v_{IIIa} > A^{I}(t)$ as defined in equation (23) represents the level of protection of silence above which $\lambda\left(Y_c\right) = 0$,
  - if $v \in \left(A^{I}(t), v_{IIIa}\right)$ case IIIa of section B.1.2 obtains, i.e. some guilty types lie, some are silent and some confess;
  - if $v \geq v_{IIIa}$, case IIIb obtains, i.e. some guilty types lie and the rest are silent.

Consider first the region of case IIIa (case I then obtains by continuity for $v = A^I(t)$ and $R$'s expected loss is independent from $v$ for any $v \leq A^I(t)$ since $\lambda(Y_s) = 0$). Replacing the equilibrium measures of $\lambda(Y_\ell)$ and $\lambda(Y_s)$ in equation (30), $R$'s expected loss is

$$\begin{aligned} E(v) =&(1-\alpha)\frac{(1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)}{1+b} \\ &+\alpha(1-t)v\left(t-\frac{bt-(1-t)v}{b}-\frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \end{aligned} \tag{34}$$

Since $E(v)$ is convex[32] and $E'(v)|_{v=A^I(t)} = \frac{(1-t)^2(1-\alpha)\alpha}{(1+b)(b+\alpha)} > 0$, in this region $R$'s expected loss is minimized at $v = A^I(t)$, i.e. for a level of protection such that $\lambda(Y_s) = 0$, yielding

$$E(v)\,|_{v=A^I(t)} = \frac{(1-t)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2}. \tag{35}$$

Consider now the region $v \geq v_{IIIa}$. $R$'s expected loss in this region is again given by equation (32) which, as seen above, is convex and, absent the constraint $v \geq v_{IIIa}$, it is uniquely minimized in $\tilde{v}$ as defined in equation (33). Thus, if $\tilde{v} \leq v_{IIIa}$, i.e. if

$$t \geq \hat{t}(\alpha, b) \equiv \frac{(1+2b)(1-\alpha)}{(1+b)(1+b(2-\alpha))},$$

where $\hat{t}(\alpha, b)$ is strictly decreasing in its arguments,[33] $E'(v)\,|_{v=v_{IIIa}} \geq 0$ and $R$'s global optimal level of protection is $v = A^I(t)$, i.e. it is such that $\lambda(Y_s) = 0$. Indeed, $R$'s expected loss is always continuous in $v$ (i.e. equation (32) and (34) coincide when $v = v_{IIIa}$) and it is then increasing in $v$ for any $v \geq A^I(t)$. If instead $\tilde{v} > v_{IIIa}$, so that $E'(v)\,|_{v=v_{IIIa}} < 0$, $R$'s optimal level of protection is either $v = A^I(t)$, i.e. such that $\lambda(Y_s) = 0$, or it is such that $\lambda(Y_s) > 0$ and equal to $\tilde{v}$, depending on whether equation (32) evaluated at $\tilde{v}$ is greater or lower than equation (35) (and there exist parameter combinations for which the optimal level is such that $\lambda(Y_s) > 0$, as for instance $t = 17/64$, $\alpha = 1/2$ and $b = 1$).

[32]
$$E''(v) = \frac{(1-t)^2\left(b+2b^2+2\alpha+3b\alpha\right)}{b(1+b)^2} > 0.$$

[33]
$$\frac{\partial\hat{t}\,(b,\alpha)}{\partial b} = -\frac{(1-\alpha)((1-\alpha)+2b(1+b)(2-\alpha))}{(1+b)^2(1-b(\alpha-2))^2} < 0 \quad \frac{\partial\hat{t}\,(b,\alpha)}{\partial\alpha} = -\frac{1+2b}{(1+b(2-\alpha))^2} < 0.$$

## D.2 Proof of proposition 3

Throughout this section, let us modify the definition of $A(m)$ in the baseline model to

$$A(m) = \frac{\int_{z\in(m,Z]} a(m,z)\mathrm{d}\lambda}{Z-m}.$$

For any given standard $Z \in (t,1]$, the analysis at section B.1 easily generalizes to characterize the equilibrium when $S$ is indeed interrogated. In particular, lemma B.1 still holds with $\bar{y} < Z$ replacing $\bar{y} < 1$, where $\bar{y}$ hence the lying region $L$ may now depend on $Z$, and

$$A(m) = \frac{Z - \bar{y}(Z) - b\left(\bar{y}(Z) - m\right)}{Z-m} \tag{36}$$

replacing equation (13). Equation (16) and (17) become respectively

$$A(t) = \frac{Z - \bar{y}(Z) - b\left(\bar{y}(Z) - t\right)}{Z-t} \qquad \text{and} \tag{37}$$

$$\bar{y}(Z) = \frac{Z - (Z-t)A(t) + bt}{1+b}. \tag{38}$$

Also, equation (18), i.e. the expected payoff from lying in the lying region for a guilty type, is

$$\pi_\ell(y) = (Z-t)A(t) - (t-y)b,$$

so that equation (21), i.e. the highest confessor (if any), becomes

$$y_c(Z) \equiv \frac{bt - (Z-t)A(t)}{b} \tag{39}$$

and the necessary and sufficient condition (22) to have that $\lambda\left(Y_c\right) > 0$ becomes

$$b > \frac{Z-t}{t}A(t). \tag{40}$$

By assumption NS, no type is ever silent. Similar to section B.1.2, we distinguish two possible cases.

**Case I: some guilty types lie and the rest confess.** If $\lambda\left(Y_c\right) > 0$, $\lambda\left(Y_\ell^I\right) = t - y_c(Z)$, where $y_c(Z)$ is defined in equation (39). Using equation (15) and (37), $\bar{y}^I(Z) = \frac{bt+Z\alpha}{b+\alpha}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I(Z) = \frac{t+bt-(1-\alpha)Z}{b+\alpha}$. This case can only occur if $b > \frac{(1-\alpha)Z-t}{t}$, i.e. if $Z < \frac{(1+b)t}{1-\alpha}$ (which is satisfied for any $b$ and $Z$ if $t \geq 1-\alpha$), so that condition (40) holds (this then automatically implies that $\bar{y}^I < Z$).

**Case II: all guilty types lie.** If $\lambda\left(Y_c\right) = 0$, the measure of liars $Y_\ell^{II}$ is then $\lambda\left(Y_\ell^{II}\right) = t$. Using equation (15) and (37), it then follows that $\bar{y}^{II} = \frac{t}{1-\alpha}$ and $A^{II}(t) = \frac{(1-\alpha)Z-t(1+b\alpha)}{(Z-t)(1-\alpha)}$. By equation (40), this case can only occur if $b \leq \frac{(1-\alpha)Z-t}{t}$ (this then automatically implies that $\bar{y}^{II} < Z$ and $A^{II}(t) > 0$).

For any $Z$ the two cases do not overlap and span the whole parameter space. The proof of the existence of the equilibrium and of payoff irrelevance of any multiplicity is omitted as it follows analogous steps as section B.1.4 and B.2. $R$'s expected loss under interrogation standard $Z$ is then

$$E(Z) = (1-\alpha)\int_{L(Z)} (Z-y)\mathrm{d}\lambda + \alpha\, t(1-Z) \tag{41}$$

$$= \alpha\lambda\left(Y_\ell\left(Z\right)\right)\left(Z - t - \frac{\alpha\lambda\left(Y_\ell\left(Z\right)\right)}{2(1-\alpha)}\right) + \alpha\, t(1-Z) \tag{42}$$

The first term in equation (41) obtains by analogous simplifications as at equation (29), while the second term is due to the fact that when $z > Z$ now $R$ takes action $a = 1$ and hence makes a type II error when facing a guilty type. Equation (42) follows from some rearranging using equation (15).

Assume for the moment that given $Z$ case I above obtains, i.e. $\lambda\left(Y_c\right) > 0$. Then, using that $\lambda\left(Y_\ell\left(Z\right)\right) = \frac{(1-\alpha)(Z-t)}{\alpha+b}$, equation (42) becomes

$$E(Z) = \frac{(t-Z)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2} + t(1-Z)\alpha. \tag{43}$$

As $E''(Z) = \frac{(1-\alpha)\alpha(2b+\alpha)}{(b+\alpha)^2} > 0$, $E'(Z)|_{Z=t} = -t\alpha < 0$ and $E'(Z)|_{Z=1} = -t\alpha + \frac{(1-t)(1-\alpha)\alpha(2b+\alpha)}{(b+\alpha)^2}$, the optimal $Z$, denoted by $Z^\star$, is interior if and only if $E'(Z)|_{Z=1} > 0$, i.e. if and only if

$$t < \bar{t}\left(b,\alpha\right) \equiv \frac{(1-\alpha)(2b+\alpha)}{2b+\alpha+b^2}, \tag{44}$$

where $\bar{t}\,(b,\alpha)$ is strictly decreasing in its arguments.[34] In such a case, the FOC gives

$$\tilde{Z} = \frac{t(b(2+b)+\alpha)}{(1-\alpha)(2b+\alpha)}. \tag{45}$$

Now, if case I above obtains even at $Z = 1$, condition (44) is necessary and sufficient for an interior standard to be optimal. If $\lambda\,(Y_c) = 0$ for $Z = 1$, instead, equation (43) on which the minimization was taken over represents $R$'s expected loss only in the region $Z < \hat{Z} \equiv \frac{(1+b)t}{1-\alpha}$. For $Z \geq \hat{Z}$, instead, all guilty types lie and hence, replacing $\lambda\,(Y_\ell) = t$ in equation (42), $R$'s expected loss is

$$\frac{t(2-t(2-\alpha)-2\alpha)\alpha}{2(1-\alpha)},$$

which is independent of $Z$. As $E'(Z)|_{Z=\hat{Z}} = \frac{\alpha bt}{\alpha+b} > 0$, the optimal standard is always interior, i.e. condition (44) always holds, and it is given by equation (45).

## D.3 Proof of proposition 4

Throughout, let $\alpha_0$ and $\alpha$ denote the preference parameter of $R$ and of the interrogator, respectively. For any $\alpha \in (0,1)$ (our analysis allows for $\alpha = 0$ and $\alpha = 1$ as limit cases given that $R$'s expected loss varies continuously with the choice of $\alpha$) the equilibrium of the interrogation is as at proposition 1 and, by assumption NS, no type is silent. The only difference with respect to the baseline model is $R$'s expected loss, which is now

$$E(\alpha) = (1-\alpha_0)\int_{L(\alpha)} (1-y)\,(1-A\,(y))\,\mathrm{d}\lambda + \alpha_0\frac{1-\alpha}{\alpha}\int_{L(\alpha)} (1-y)\,A\,(y)\,\mathrm{d}\lambda \tag{46}$$

$$= (1-\alpha_0)\frac{1}{2}(1+b)(\bar{y}\,(\alpha)-t)^2 + \alpha_0\frac{1-\alpha}{\alpha}\frac{1}{2}(2-b(\bar{y}\,(\alpha)-t)-2\bar{y}\,(\alpha))(\bar{y}\,(\alpha)-t). \tag{47}$$

Equation (46) differs from equation (28) since $\alpha$ may now differ from $R$'s $\alpha_0$. Equation (47) obtains by replacing the definition of $A(y)$ (equation (13)) and integrating. Since $\bar{y}(\alpha)$ differs depending on whether $\lambda\,(Y_c) > 0$, i.e. on whether $\alpha > \bar{\alpha} \equiv \max\{1-(1+b)t, 0\}$

[34] $$\frac{\partial \bar{t}\,(b,\alpha)}{\partial b} = -\frac{2b(1-\alpha)(b+\alpha)}{(b(2+b)+\alpha)^2} < 0 \quad \frac{\partial \bar{t}\,(b,\alpha)}{\partial \alpha} = -\frac{(b+\alpha)(b(3+2b)+\alpha)}{(b(2+b)+\alpha)^2} < 0.$$

by equation (26) (we simply rewrote the cutoff in terms of $\alpha$ rather than $b$), we consider $R$'s optimal choice separately in the two cases (keeping in mind case $\lambda(Y_c) = 0$ can only occur if $\bar{\alpha} > 0$). Letting the subscripts $c$ and $nc$ indicate respectively the region with and without confessors throughout, we solve for the optimal choices in the two regions, denoted respectively $\alpha_c^\star$ and $\alpha_{nc}^\star$, and then compare $E(\alpha_c^\star)$ and $E(\alpha_{nc}^\star)$.

**Some types confess.** When $\alpha > \bar{\alpha}$, replacing $\bar{y}(\alpha) = \frac{\alpha+bt}{\alpha+b}$ (see case I at section B.1.2) in equation (47) yields

$$E_c(\alpha) = \frac{(1-t)^2\left(\alpha^2(1+b-\alpha_0) + 2b\alpha_0 - 3b\alpha\alpha_0\right)}{2(b+\alpha)^2} \tag{48}$$

The FOC gives a unique solution

$$\tilde{\alpha}_c = \frac{\alpha_0(3b+4)}{\alpha_0+2b+2} > \alpha_0$$

and the SOC is verified.[35] If $\alpha_0 \geq 2/3$, $\tilde{\alpha}_c \geq 1$ and, since $E'(\alpha) < 0$ for all $\alpha \in (0,1]$, $R$'s expected loss is minimized for $\alpha_c^\star = 1$ (as $\bar{\alpha} < 1$, the constraint $\alpha > \bar{\alpha}$ is then non-binding). When instead $\alpha_0 < 2/3$, $R$'s expected loss is minimized for $\alpha_c^\star = \tilde{\alpha}_c$ provided $\tilde{\alpha}_c > \bar{\alpha}$, i.e. if $\alpha_0 > \frac{2-(b+1)2t}{t+3} = \frac{2}{3+t}\bar{\alpha}$, and for $\alpha_c^\star = \bar{\alpha}$, i.e. at the boundary, otherwise.

**No types confess.** When $\alpha \leq \bar{\alpha}$, replacing $\bar{y}(\alpha) = \frac{t}{1-\alpha}$ (see case II at section B.1.2) in equation (47) yields

$$E_{nc}(\alpha) = \frac{t\left(2\alpha_0(1-\alpha)^2 + (1+b)t\alpha^2 - t\alpha_0\left(2-(2-b-\alpha)\alpha\right)\right)}{2(1-\alpha)^2} \tag{49}$$

The FOC gives a unique solution

$$\tilde{\alpha}_{nc} = \frac{(2+b)\alpha_0}{2+2b-b\alpha_0} \in (0, \alpha_0)$$

[35]
$$E_c^{''}(\alpha)\big|_{\alpha=\tilde{\alpha}_c} = \frac{b(1-t)^2(\alpha_0+2b+2)^4}{16(b+1)^3(2\alpha_0+b)^3} > 0.$$

and the SOC is verified.[36] Hence, $E_{nc}(\alpha)$ is minimized for $\alpha^\star_{nc} = \tilde{\alpha}_{nc}$ if $\tilde{\alpha}_{nc} < \bar{\alpha}$, i.e. if $\alpha_0 < \frac{2-(b+1)2t}{2-bt} = \frac{2}{2-bt}\bar{\alpha}$ and for $\alpha^\star_{nc} = \bar{\alpha}$, i.e. at the boundary, otherwise.

As $\alpha^\star_c > \alpha_0$ and $\alpha^\star_{nc} < \alpha_0$, $R$'s global optimum $\alpha^*$ differs from $\alpha_0$, which proves the first statement. If $\bar{\alpha} = 0$, $\alpha^* = \alpha^\star_c$. Suppose instead that $\bar{\alpha} > 0$. The previous considerations and the fact that $R$'s expected loss is continuous in $\alpha$ with $E_c(\bar{\alpha}) = E_{nc}(\bar{\alpha})$ imply that whenever the minimum of a given case obtains at the boundary $\alpha = \bar{\alpha}$, the minimum of the other case is strictly lower. Indeed, if $\alpha_0 \leq \frac{2}{3+t}\bar{\alpha}$, $E_c(\alpha)$ is increasing in the whole $\alpha > \bar{\alpha}$ region and hence $\alpha^\star = \tilde{\alpha}_{nc}$, which proves the third statement. Likewise, if $\alpha_0 \geq \frac{2}{2-bt}\bar{\alpha}$, $E_{nc}(\alpha)$ is increasing in the whole $\alpha \leq \bar{\alpha}$ and hence $\alpha^\star = \alpha^*_c$. Conversely, in the region $\alpha_0 \in \left(\frac{2}{3+t}\bar{\alpha}, \frac{2}{2-bt}\bar{\alpha}\right)$, $\alpha^*$ may be either $\alpha^\star_c$ or $\alpha^\star_{nc}$. Still, we now prove that $\alpha^* = \alpha^\star_c$ whenever $\alpha_0 > \bar{\alpha}$, i.e. the second statement.

Consider hence the case $\alpha_0 \in \left(\bar{\alpha}, \frac{2}{2-bt}\bar{\alpha}\right)$, or equivalently, $t \in (\underline{t}, \bar{t})$, where $\underline{t} \equiv \frac{1-\alpha_0}{b+1}$ and $\bar{t} \equiv \frac{2(1-\alpha_0)}{2+2b-b\alpha_0}$. Also, let $\Delta \equiv E_{nc}(\alpha^\star_{nc}) - E_c(\alpha^\star_c)$ be the difference in $R$'s expected loss in the case without and with confessors given $R$'s respective locally optimal choices, where $\alpha^\star_{nc} = \hat{\alpha}_{nc}$ necessarily since $\alpha_0 < \frac{2}{2-bt}\bar{\alpha}$. When $\alpha_0 < 2/3$, using that $\alpha^\star_c = \tilde{\alpha}_c$,

$$\Delta = \frac{\alpha_0}{8(1+b)}\left(\frac{t(8(1+b)(1-t) - (8-4t+b(8-(4-b)t))\alpha_0)}{1-\alpha_0} - \frac{(1-t)^2(8+8b-8\alpha_0-9b\alpha_0)}{b+2\alpha_0}\right).$$

The expression is strictly positive, so that $\alpha^* = \alpha^\star_c$, since it is concave in $t$ and strictly positive in the two extrema (the symbol $\propto$ means "has the same sign as"):

$$\Delta\,\Big|_{t=\underline{t}} \propto \alpha_0^2 b\left(4b+4-2\alpha_0-3\alpha_0 b\right)^2 > 0,$$
$$\Delta\,\Big|_{t=\bar{t}} \propto \left(2-\alpha_0\right)^2 + \alpha_0 b^2 + 2\left(2-\left(2-\alpha_0\right)\alpha_0\right)b > 0.$$

When $\alpha_0 \geq 2/3$, instead, using that $\alpha^\star_c = 1$,

$$\Delta = \frac{8t(1-\alpha_0)(1+b\alpha_0) - (t^2(4+b\alpha_0(8-(4-b)\alpha_0)) + 4(1-\alpha_0)^2)}{8(1+b)(1-\alpha_0)}.$$

Again, the expression is strictly positive, so that $\alpha^* = \alpha^\star_c$, as it is concave in $t$ and strictly positive at the two extrema. Indeed, $\Delta\,\Big|_{t=\underline{t}} \propto 4\alpha_0^2 - b\alpha_0^2 + 8b\alpha_0 - 4b$ which is increasing

[36]
$$E''_{nc}(\alpha)\big|_{\alpha=\tilde{\alpha}_{nc}} = \frac{t^2(2+b(2-\alpha_0))^4}{16(1+b)^3(1-\alpha_0)^3} > 0.$$

in $\alpha_0$ and equal to $\frac{16+8b}{9} > 0$ in $\alpha_0 = 2/3$. Likewise, $\Delta\,|_{t=\bar{t}} \propto 2\alpha_0^2 + b\,(3\,(2-\alpha_0)\,\alpha_0 - 2)$ which is increasing in $\alpha_0$ and equal to $\frac{2b}{3} + \frac{8}{9} > 0$ in $\alpha_0 = 2/3$.

## D.4 Proof of proposition 5

After providing an intuition for the relation between the optimal mechanism and the equilibrium (section D.4.1), we prove its optimality in the class of mechanisms considered in the main body of the paper (section D.4.2). We then show how arbitrary mechanisms offer no improvement (section D.4.3).

### D.4.1 Intuition

As pointed out in the body of the paper, in the optimal mechanism the truth-telling constraint must be binding for types sufficiently close to $t$. Clearly, for any given value of $\hat{z}(t)$, for types to the right of $t$ one minimizes type I errors by having the constraint binding till $\hat{\boldsymbol{z}}$ reaches the diagonal $z = y$. Likewise, for types to the left of $t$ one minimizes type II errors by having the constraint binding till the line $z = 1$ (the case of figure 4a) or the vertical axis (the case of figure 4b). The exact counterpart of the truth-telling constraint in the equilibrium of the baseline model is that pooling types are indifferent between any lie. The optimal choice of $\hat{z}(t)$ is then determined by the fact that $R$ is trading off type I and type II errors. Suppose one increases $\hat{z}(t)$. In the (interior) optimum the marginal increment of type I errors weighted by $(1-\alpha)$ must be equal to the marginal decrement of type II errors weighted by $\alpha$. These are measured by the appropriately weighted lengths of the $\hat{\boldsymbol{z}}$ line from $\hat{z}(t)$ respectively to the right of $t$ (till the diagonal $z = y$) and to the left of $t$ (till the line $z = 1$ or till the vertical axis). The exact equilibrium counterpart of this constraint is the required indifference of $R$ conditional on discretion, i.e. the condition at lemma B.2 relating the measure of liars with the measure of the lying region. Thus, when projecting the optimal $\hat{\boldsymbol{z}}^{\star}$ onto the horizontal axis, given linearity, one obtains exactly the lying region and the set of liars with the equilibrium measures of the baseline model as required by lemma B.2. It follows that each type obtains the same payoff as in equilibrium and only type II errors become smaller in the optimal mechanism.

### D.4.2 Optimality

Let $y_c\left(\hat{z}\left(t\right)\right)$ and $\bar{y}\left(\hat{z}\left(t\right)\right)$ denote respectively the smallest guilty type and the largest innocent type for which constraint (6) binds. The line with slope $-b$ passing through the point $(t, \hat{z}(t))$ has equation $-by + \hat{z}(t) + bt$, so that $\bar{y}\left(\hat{z}(t)\right) = \frac{\hat{z}(t)+bt}{1+b}$. Also,

$$y_c\left(\hat{z}(t)\right) = max\left\{\frac{\hat{z}(t) + bt - 1}{b}, 0\right\}$$

and $y_c(t) > 0$, i.e the case of figure 4a, obtains if and only if $(1+b)t > 1$. Suppose first this is indeed the case. Then $R$'s expected loss, i.e. equation (5), becomes

$$\begin{aligned} E_{y_c>0}(\hat{z}(t)) &= \alpha \int_{\frac{\hat{z}(t)+bt-1}{b}}^{t} (1 - (-by + \hat{z}(t) + bt))\mathrm{d}y + (1-\alpha)\int_{t}^{\frac{\hat{z}(t)+bt}{1+b}} \left(-by + \hat{z}(t) + bt - y\right)\mathrm{d}y \\ &= \alpha\frac{(1-\hat{z}(t))^2}{2} + (1-\alpha)\frac{\left(\hat{z}(t) - t\right)^2}{2(1+b)}. \end{aligned}$$

As $E''_{y_c>0}(\hat{z}(t)) = \frac{b+\alpha}{b(1+b)} > 0$, i.e. $E_{y_c>0}(\hat{z}(t))$ is convex, with $E'_{y_c>0}(t) = -\frac{(1-t)\alpha}{b} < 0$ and $E'_{y_c>0}(1) = \frac{(1-\alpha)(1-t)}{b+1} > 0$, the FOC identifies the unique minimizer

$$\hat{z}^{\star}_{y_c>0}(t) = \frac{\alpha + b(t + \alpha - t\alpha)}{b+\alpha}. \tag{50}$$

Suppose now that $(1+b)t \leq 1$, instead, so that $y_c(t) = 0$, i.e the case of figure 4b obtains. Then, $R$'s expected loss is as before if $\hat{z}(t) > 1 - bt$, while if $\hat{z}(t) \leq 1 - bt$, it is

$$\begin{aligned} E_{y_c=0}(\hat{z}(t)) &= \alpha \int_{0}^{t} (1 - (-by + \hat{z}(t) + bt))\mathrm{d}y + (1-\alpha)\int_{t}^{\frac{\hat{z}(t)+bt}{1+b}} \left(-by + \hat{z}(t) + bt - y\right)\mathrm{d}y \\ &= \alpha\frac{t(2 - 2\hat{z}(t) - bt)}{2} + (1-\alpha)\frac{\left(\hat{z}(t) - t\right)^2}{2(1+b)}. \end{aligned}$$

As $E''_{y_c=0}(\hat{z}(t)) = \frac{1-\alpha}{1+b}$, $E_{y_c=0}$ is again convex and, moreover, $E'_{y_c=0}(t) = -t\alpha$. It follows that the minimizer differs from the one at equation (50) if and only if $E'_{y_c=0}(\hat{z}(t)) \mid_{\hat{z}(t)=(1-bt)} = \frac{1-t-bt-\alpha}{1+b} \geq 0$, i.e. if and only if $b \leq \frac{1-t-\alpha}{t}$. In such a case, it is uniquely identified by the FOC, which gives

$$\hat{z}^{\star}_{y_c=0}(t) = \frac{t + bt\alpha}{1-\alpha}.$$

Thus, to summarize, the optimum is

$$\hat{z}^{\star}(t)=\begin{cases}\hat{z}^{\star}_{y_c=0}(t) & \text{if } b\leq\frac{1-t-\alpha}{t}\\ \hat{z}^{\star}_{y_c>0}(t) & \text{otherwise.}\end{cases}$$

It follows that conditions for the optimal mechanism to yield that $y_c\left(\hat{z}^{\star}(t)\right)>0$ are identical to the equilibrium conditions for which the measure of confessors is positive. One can also easily verify that $y_c\left(\hat{z}^{\star}(t)\right)=y_c$ and $\bar{y}\left(\hat{z}^{\star}(t)\right)=\bar{y}$ as in the equilibrium, so that guilty types for which the constraint does not bind get respectively 0 and 1 in both cases. Finally, define $\hat{z}^{\star}(y)=\hat{z}(y)|_{\hat{z}(t)=\hat{z}^{\star}(t)}$. Using the equilibrium value of $\bar{z}(m)$, guilty types for which the constraint binds get $1-\hat{z}^{\star}(y)=1-\bar{z}(t)-(t-y)b$ as in equilibrium. Likewise, innocent types for which the constraint binds get $1-\hat{z}^{\star}(y)=1-\bar{z}(y)$ as in equilibrium.

### D.4.3 Arbitrary mechanisms

Let us consider some arbitrary, possibly non-direct and random, mechanism in which $S$ can receive payoffs of $-b$, 0 and 1 with the single constraint that the expected payoff $p(y)$ of each type $y$ must be weakly larger than 0 (and can be calculated). We assume, as before, that when $-b$ is given to a guilty suspect $R$ makes no error. However, when $-b$ is given to an innocent suspect $R$ makes an error of size $1+b$. Fix the resulting expected loss of $R$ when each type sends only messages which are optimal for him given the mechanism (which can be calculated and is the lowest in case of multiplicity). Consider now the deterministic cutoff direct mechanism using only 0s and 1s which, in expectation with respect to $z$, gives each type $y$ exactly $p(y)$ when $y$ reports that his type is $y$. First, note that this direct mechanism satisfies the constraint

$$1-\hat{z}(y)\geq 1-\hat{z}(y')-b\left(y'-y\right)$$

for each $y,y'$ for which $y'\geq y$. Indeed,

$$p(y)=\frac{1-\hat{z}(y)}{1-y}\geq p(y')\frac{1-y'}{1-y}-b\frac{y'-y}{1-y}=\frac{1-\hat{z}(y')}{1-y'}\frac{1-y'}{1-y}-b\frac{y'-y}{1-y},$$

where the first inequality follows from the fact that, conditional on $z \geq y'$, it must be that $y'$ expects $p(y')$ and so does any other type $y \leq y'$ and that $y$ does not strictly prefer to play as if he was $y'$ in the original mechanism, where the worst possibility is that $y$ expects $-b$ conditional on that $y' \geq z \geq y$.

Clearly, type I errors are the same in both mechanisms while type II errors may only decrease when using the direct mechanism. It is not necessarily true though that this direct mechanism is immune to downward deviations, i.e. some type $y$ now may prefer to report that he is type $y' < y$. Thus, all we can say is that the obtained direct mechanism is weakly better for $R$ than the original mechanism in an environment where downward deviations are not feasible for $S$. However, our optimal direct mechanism is also optimal in the environment where downward deviations are not possible. Moreover, of course, our optimal mechanism is also immune to such deviations. Therefore, our restrictions are without loss of generality.

## D.5 Proof of proposition 7

By the results of section B.2, $R$'s expected expected loss can be written as

$$E(b) = (1-\alpha)\int_{L(b)} (1-y)\mathrm{d}\lambda.$$

By the results of section B.1.2 and B.1.3, $L(b)$ is continuous and decreasing in $b$, and strictly so if $b > \frac{1-\alpha-t}{t}$, which proves the result on the effect of $S$'s perceived $b$.

Consider now the effect of $S$'s perceived $Z$, possibly different from the true standard $Z_0$. Using the results of section D.2 and adjusting equation (41) for the fact that $z \leq Z$ (otherwise the interrogation would not be occurring) yields

$$E(Z) = (1-\alpha)\int_{L(Z):y\leq min\{\bar{y}(Z),Z_0\}} (Z_0 - y)\,\mathrm{d}\lambda.$$

The expression $min\{\bar{y}(Z), Z_0\}$ is due to the fact that if $\bar{y}(Z) > Z_0$ then $R$ catches liars with probability one when $m \in [Z_0, \bar{y}(Z)]$. As shown at section D.2, $\bar{y}(Z)$ is continuous and decreasing in $Z$, and strictly so whenever $Z \leq \frac{(1+b)}{1-\alpha}$, so that the result follows.

Finally, consider the effect of $S$'s perceived $\alpha$, possibly different from $R$'s true preference parameter $\alpha_0$. Using again the results of section B.2, $R$'s expected loss when $\alpha = \alpha_0$

can be written as

$$E(\alpha_0) = \underbrace{(1-\alpha_0)\int_{L(\alpha)}(1-y)\mathrm{d}\lambda}_{\text{type I error}} = \underbrace{\alpha_0\,\frac{1-\alpha}{\alpha}\int_{L(\alpha)}(1-y)\mathrm{d}\lambda}_{\text{type II error}}.$$

$R$ is hence indifferent between always doing only type I errors and always doing only type II errors. If $\alpha < \alpha_0$, instead, $R$ now finds it strictly optimal to always choose $a = 0$ and makes only type I errors, so that her expected loss is

$$E_{\alpha<\alpha_0}(\alpha) = (1-\alpha_0)\int_{L(\alpha)}(1-y)\mathrm{d}\lambda.$$

As $L(\alpha)$ is strictly increasing and continuous in $\alpha$, it follows that $E(\alpha)_{\alpha<\alpha_0}$ is increasing, and minimized and equal to zero in $\alpha = 0$ where $L(\alpha)$ converges to $t$. If $\alpha > \alpha_0$, instead, $R$ now finds it strictly optimal to always choose $a = 1$ and makes only type II errors, so that her expected loss is

$$E_{\alpha>\alpha_0}(\alpha) = \alpha_0\frac{1-\alpha}{\alpha}\int_{L(\alpha)}(1-y)\mathrm{d}\lambda.$$

When $\lambda(Y_c) = 0$, replacing the equilibrium value of $L(\alpha)$ and differentiating yields.

$$E'_{\alpha>\alpha_0}(\alpha) = -\alpha_0\frac{t^2}{2(1-\alpha)^2} < 0$$

Likewise, when $\lambda(Y_c) > 0$

$$E'_{\alpha>\alpha_0}(\alpha) = -\alpha_0\frac{(1-t)^2(b(3+2b)+\alpha)}{2(b+\alpha)^3} < 0.$$

Thus, $E_{\alpha>\alpha_0}(\alpha)$ is continuous and decreasing and minimized in $\alpha = 1$, where it is equal to zero since, while $L(\alpha)$ converges to $\left[t, \frac{1+bt}{1+b}\right)$, $\lambda(Y_\ell)$ converges to zero.

# E The benefits of conditional delegation

Fix $t = 1/2$, $b = 1$ and $\alpha_{\text{const}} = 1/2$. Using the results of section 3, in equilibrium $y_c = 1/3$ and $\bar{y} = 2/3$, so that the measure of liars and of lies sent are both equal to 1/6. Upon discretion the interrogator is indifferent and chooses $a = 1$ when $m \geq$

$\bar{z}(m) = 4/3 - m$ and $a = 0$ otherwise. The equilibrium and the resulting type I and type II errors were displayed in figure 2. Suppose instead $R$ delegates to an interrogator with preference parameter $\alpha_{\text{nice}} = 1/4$ when $z \geq \tilde{z} = 19/24$ and to an interrogator with preference parameter $\alpha_{\text{tough}} = 3/4$ when $z < \tilde{z}$. Suppose also (we will ensure that this is indeed sequentially rational) that in her respective region of competence each interrogator still follows the same cutoff strategy $\bar{z}(m)$ as in the equilibrium under unconditional delegation. Then, $S$'s incentives are completely unaffected, so that $y_c$ is the same and, in particular, liars are still indifferent to any lie in $[t, \bar{y})$. It is then possible to construct a lying function with image $[t, \bar{y})$ such that each interrogator finds it optimal to follow $\bar{z}(m)$.

To see this, let us denote by $\tilde{m} = 13/24$ the message such that $\tilde{z} = \bar{z}(\tilde{m})$, as represented in figure 5. Looking from the perspective of the horizontal axis after projecting $\tilde{m}$ onto it, when $m \in [t, \tilde{m}]$ the nice interrogator takes both action $a = 0$ and $a = 1$ based on $z$. This is only possible if she is indifferent, i.e. if equation (14) holds for $\alpha = \alpha_{\text{nice}}$. In turn, this implies that the tough interrogator indeed finds it strictly optimal to always choose $a = 0$. Likewise, when $m \in (\tilde{m}, \bar{y})$, the tough interrogator takes both action $a = 0$ and $a = 1$, which again is only possible if she is indifferent, i.e. if equation (14) holds for $\alpha = \alpha_{\text{tough}}$. The nice interrogator then indeed finds it strictly optimal to always choose $a = 1$. Easy calculations show that these two indifference conditions hold for the following lying function

$$\ell(y) = \begin{cases} 7/18 + 1/3y & \text{if } y \in [y_c, \tilde{y}) \\ 17/18 - 3y & \text{if } y \in [\tilde{y}, t)\,, \end{cases}$$

where $\tilde{y} = 11/24$ is the guilty type who sends message $\tilde{m}$.

This lying function is depicted in red in figure 5. A comparison with figure 2 illustrates how type I errors, as well as type II errors for $z \geq \bar{z}(t) = 5/6$, remain unaffected relative to unconditional delegation. Instead, type II errors for $z \in (t, \bar{z}(t))$ decrease of the green region, whose area has size $1/144$. For an intuition, notice that, in order to maintain incentive compatibility for liars as dictated by equilibrium, by construction those errors are also equal to the utility loss of liars caught when evidence is inconclusive (the area represented in yellow in the figure, given that $b = 1$). The change in the lying function

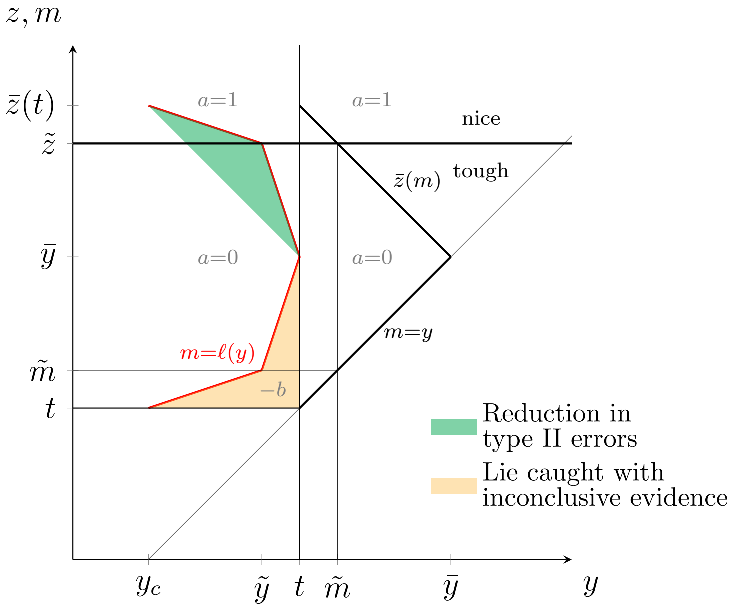


Figure 5 The benefits of conditional delegation

induced by conditional delegation shifts the distribution of lies towards lower messages. As a result, $S$ is caught in a lie and hence punished *less* often, so that upon discretion $R$ also chooses $a = 1$ less often. From these arguments, one can first of all see that the chosen delegation policy is actually the optimal one among the ones $\alpha(z) : [0, 1] \to [1/4, 3/4]$ that leave $y_c$ unaffected. Indeed, $R$ aims to make the lying function as flat as possible before the kink and as steep as possible after the kink - this is in fact how we computed $\tilde{z}$ in the first place. Besides, as the preferences of the nicer and tougher interrogator gets more extreme, lies get more and more concentrated around $t$. In the limit, these type II errors entirely disappear since lies are never caught by inconclusive evidence (yet all lies will still be caught with conclusive evidence, so that the $-b$ area does not disappear completely as in the optimal mechanism). Upon observing $m = t$, the nice interrogator is now sure to face a guilty type but, as type II errors yield her no disutility, her indifference condition is preserved.