THEMA

théorie économique,
modélisation et applications

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

Designing Interrogations

Alessandro Ispano, Peter Vida

March 2023

CY CERGY PARIS UNIVERSITÉ

cnrs

Designing Interrogations

Alessandro Ispano Péter Vida

March 2023

Abstract

We provide a model of interrogations with two-sided asymmetric information. The suspect knows his status as guilty or innocent and the likely strength of the law enforcer's evidence, which is informative about the suspect's status and may also disprove lies. We compare prosecution errors in the equilibrium of the one-shot interrogation and in the optimal mechanism under full commitment. We describe a "back and forth" interrogation with disclosure of the evidence and discretionary forgiveness of lies that implements the optimum in equilibrium without any commitment.

Keywords: lie, evidence, questioning, confession, law, prosecution, disclosure, persuasion, 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; Péter Vida: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, and Corvinus University of Budapest, Corvinus Institute for Advanced Studies and Institute of Economics vidapet@gmail.com. We thank Helmuts Āzacis, Eric Danan, Harry Di Pei, Sidartha Gordon, Jeanne Hagenbach, Frédéric Koessler, Laurent Linnemer, Lucie Menager, Alfonso Montes, Marco Ottaviani, Marcus Pivato, Régis Renault, Karl Schlag, Johannes Schneider, Joel Sobel, Egor Starkov as well as seminar participants at the THEMA theory meetings, Workshop on Signaling in Markets Auctions and Games, University of Copenhagen, Paris Applied Theory Day, Paris II University, Paris School of Economics, Annual Conference of the European Association of Law and Economics in Barcelona, European Summer Meeting of the Econometric Society in Copenhagen, Hungarian Academy of Science, University of Gratz and University of East Anglia for useful comments. Financial support from Labex MME-DII and ANR grant 19-CE26-0010-03 is gratefully acknowledged.

1 Introduction

Beyond arousing collective imagination and ensuring the fortunes of many detective stories,[1] the interrogation of a suspect is an important investigative resource for law enforcers in most legal systems. In this paper, we propose a model of interrogations that we use to explore several questions on their conduct and regulation to determine which institutions enhance information revelation and yield to more accurate decisions.[2]

Interrogations exhibit two distinctive features that our model seek to capture. First, as both common sense and empirical studies indicate (see section 4.3), the strength of the incriminating evidence as perceived by law enforcers and by the suspect is key. And while law enforcers are typically better informed about the evidence gathered, the suspect is better informed about how strong he expects this evidence to be. For example, a meticulous criminal will be more confident than a clumsy one to have left no trace behind. Likewise, an innocent suspect who was many miles away from the crime scene should anticipate that law enforcers' case will be speculative at best. In spite of the complexity resulting from the correlation between the private information of the two parties, we provide a handy information structure that incorporates this dimension.[3] Besides, interrogations typically involve a dynamic process whereby not only the suspect is asked questions but also law enforcers may give away information about the evidence, inducing the suspect to revise his expectation and, possibly, his strategy, e.g. "break" and confess. We formalize this not always transparent persuasion process and shed light on its effectiveness.

We represent the interrogation as a game of two-sided asymmetric information between a suspect (he) and a law enforcer (she). The suspect's private information, or type, can be thought of as the lawfulness of his behavior measured on a vertical scale, e.g. the care he put into driving or the distance he stayed from his ex-wife who obtained a restraining order. The suspect is guilty when his lawfulness falls short of a known threshold, e.g. the carefulness required to avoid vehicular homicide or the distance specified by the restraining order. The law enforcer's

[1] At the moment of writing, a search based on the keyword "interrogation" on the popular IMDB internet movie database yields 4308 entries (`https://www.imdb.com/search/keyword/?keywords=interrogation`).

[2] While we focus on law enforcement as the leading application, interrogations, or comparable situations, arise in many other contexts ranging from private litigation, e.g. the assessment of an employee's misconduct, to fraud in academia, e.g. the investigation of cheating in an exam, and daily life, e.g. the determination of a spouse's betrayal.

[3] For instance, in the related context of plea bargaining, Reinganum (1988) writes:

> A more difficult task is to incorporate the discovery process. One way to do this is to assume that the defendant receives a signal which is (imperfectly) correlated with the strength of the case. If the prosecutor also observes this signal, then this is basically an exercise in updating priors. [...] If the signal is private information for the defendant, matters could become considerably more complicated.

private information is a piece of evidence that provides a bound on the suspect's lawfulness, e.g. a speed his car surely exceeded or a distance at which he was spotted. Thus, the higher the suspect's lawfulness, the greater his confidence that the law enforcer's evidence is weak. After interacting with the suspect, the law enforcer must decide whether to prosecute him.[4]

In our baseline model, the suspect makes a claim, interpreted as a reply to the law enforcer's inquiry about his type, and then the law enforcer makes a decision. The suspect additionally incurs some cost if he is caught in a lie, i.e. if his claim is disproved by the evidence. In equilibrium, innocents are honest. Possibly, some unconfident guilties are honest as well, i.e. they confess, to avoid the risk of getting caught in a lie. Instead, sufficiently confident guilties necessarily lie and mimic unconfident innocents. This baseline model yields clear predictions on players' equilibrium strategies (proposition 1) and payoffs (corollary 1). Besides, it sheds light on the relation between the usefulness of interrogating and the elicitation of a confession since the law enforcer's payoff is higher than when she relies on the evidence alone if and only if some guilties indeed confess (remark 1).

We complement the equilibrium analysis with the mechanism design approach, which assumes the law enforcer can commit to a decision based on the suspect's claims and the evidence. A comparison between the equilibrium and the optimal mechanism (proposition 2), which can also be made immune to the law enforcer's incentives to misrepresent the evidence (remark 2), identifies a commitment problem inherent to interrogations. The law enforcer would benefit from committing to prosecute unconfident innocents when the evidence is strong and to let go confident guilties when the evidence is weak. Besides, while the threat of catching and punishing lies makes eliciting information from the suspect possible in the first place, the fact that lying occurs in equilibrium harms the law enforcer and overall efficiency, since too many guilties are let go relative to the optimal mechanism.

We then investigate the scope for information revelation about the evidence and richer interrogation protocols to compensate for the law enforcer's lack of commitment over decisions. Initially, we suppose the law enforcer can credibly commit to an arbitrary revelation policy, or persuasion rule, and, after the suspect observes a signal accordingly, the interrogation unfolds as in the baseline model. There exists a persuasion rule such that players' payoffs are as in the optimal mechanism (proposition 3). However, in any optimal persuasion rule there are evidence realizations such that the law enforcer would gain from deviating and understating the strength of the actual evidence (remark 3). Under natural restrictions on evidence revelation, notably

[4] The model equivalently applies to any other decision that the law enforcer would want to base on the suspect's guilt, e.g. an arrest, and generates disutility to the suspect irrespectively. Until section 4.1.2, we abstract from details about the separation of roles in the legal system.

that the strength of the evidence cannot be overstated but can be understated, this deviation would always be possible, so that longer communication is needed to attain the optimum.

Indeed, the optimal mechanism can be implemented without any commitment in a "back and forth" variation of the baseline model (proposition 4) based on the idea of letting the suspect provide his own account and only then challenging him with the evidence accordingly. The law enforcer must be able to disclose information about the evidence, possibly vaguely, as a function of the suspect's initial claim, who can then reply back. The law enforcer must also have discretion on whether to punish or forgive lies, which she will use as "carrot and stick". When the evidence is sufficiently strong relative to the suspect's initial claim, the law enforcer proves it rather than immediately taking a decision. In the second round, a guilty suspect will step back on his lie, which will be forgiven, and an innocent type will stick to his story. The equilibrium behavior 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 (forthcoming), in which only an innocent rejects the plea. However, while those models require the court to sometimes convict a suspect known to be surely innocent, in our game the law enforcer's decisions are sequentially rational at each information set. In particular, in the second round an innocent suspect is always let go. Since the equilibrium exhibits an implicit promise of leniency for confession, this result also supports the view that these kinds of agreements, on which the law is often blurry or controversial, should be allowed.

In the remaining part of the paper, we first discuss our main modeling assumptions and the robustness of our main findings under some natural variations. In particular, we consider the case in which the suspect bears no cost for being caught in a lie but enjoys some leniency for confessing. Also, we consider the possibility that the prosecution process continues after the interrogation and the suspect's and the law enforcer's payoffs are determined by a future decision of a third party, e.g. conviction or acquittal in court, which may also depend on the uncovering of new evidence. Next, we demonstrate how some recurrent legal institutions governing interrogations, namely protection of the suspect's right to silence (proposition 5) and evidence strength standards required to interrogate (proposition 6), can alleviate the law enforcer's commitment problem above and improve decisions. We then describe empirical foundations of our assumptions and results drawing on literature from other disciplines such as criminology and psychology. We conclude by discussing important considerations we left aside and avenues for future research, in particular the use of deceptive interrogations tactics and the strategic choice to interrogate the suspect. The appendix contains all proofs, while we relegate more technical material and some examples to the online appendix.

Relation to the literature. While the judicial process is a prominent field of application of information economics, suspects' interrogations have received only limited attention. A notable exception is Baliga and Ely (2016), who study the interrogator's commitment problems inherent to torture. More closely related is Seidmann (2005), who focuses on how protection of the suspect's right to silence affects his communication.[5] As detailed in section 4.1.2, we simultaneously confirm how his results extend to our setting and we offer some new insights on the issue. Recently, Bull (2022) shows how non-disclosure of the evidence can be superior to full disclosure at incentivizing confession in a framework in which the interrogator also has some unverifiable information about the suspect's guilt. This result also holds in our setting, which additionally demonstrates how partial disclosure contingent on the suspect's message can be even better.

Otherwise, the law and economics literature generally studies the judicial process assuming prosecution is already undergoing. This paper instead focuses on how interrogating the suspect contributes to the decision to prosecute. To this end, our framework captures essential features of the judicial process only in stylized form but accounts for key specificities of interrogations in the information structure and the nature of communication.[6] 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.[7] Thus, taking two-sided asymmetric information between the suspect and the law enforcer a step further than previous work, our model allows for heterogeneity not only between a guilty and an innocent, but also within guilties and innocents, in the strength of the incriminating evidence they expect. This heterogeneity explains why different guilties prefer different strategies. Besides, the information the different parties present to a judicial officer is typically modeled as hard evidence (Milgrom, 1981), i.e. it can be disclosed or withheld but not misreported.[8] To allow for the possibility of plain lying intrinsic to interrogations, in our model the suspect's claims are soft information, i.e. the set of his available messages is independent

[5] Leshem (2010) extends the analysis to a setting in which even innocent suspects may prefer to exercise the right to silent as their honest claims may be disproved.

[6] In particular, we take as given that guilt and reticence entail some punishment without considering the complex determinants of plea bargaining (Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezzetti, 2001) and sentencing (Siegel and Strulovici, 2019, forthcoming). 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.

[7] 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). Cuellar (2020) studies plea bargaining outcomes when the prosecutor can acquire and disclose evidence over time.

[8] See for instance Shin (1994), Mialon (2005), Bhattacharya and Mukherjee (2013) and Hart et al. (2017).

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 theoretical models of lying (e.g. Kartik (2009), Dziuda and Salas (2018), Balbuzanov (2019) and Jehiel (2021)), the detectability of a lie and its cost for the suspect derive from the law enforcer's private information, which in particular implies that the detectability of a lie increases with its size.[9] Moreover, our framework allows studying communication protocols with two-sided information revelation and, in particular, the effects of revelation of the law enforcer's evidence to the suspect. Our model hence also joins the growing theoretical literature on strategic communication that, departing from seminal works, considers two-sided asymmetric information between the sender and the receiver.[10] 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 disproved by the law enforcer's evidence, the interrogation would be completely uninformative. Likewise, our model is related to the literature on Bayesian persuasion (Kamenica and Gentzkow, 2011), in particular of a privately informed receiver (Kolotilin et al., 2017). Our back and forth game resembles the idea of a persuasion mechanism in Kolotilin et al. (2017) in that the way information is disclosed to the suspect depends on the information he sends. Finally, our model is also connected to the literature on improvements from multistage communication (e.g. Aumann and Hart (2003), Krishna and Morgan (2004) and Goltsman et al. (2009)).

2 A simple model of interrogation

2.1 Model

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 and R privately observes her type, or evidence, z. These are drawn from $\left\{(y,z)\in[0,1]^2 : y<z\right\}$

[9] 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 exogenous chance of being detected and the cost is endogenously determined by the receiver's response. Jehiel (2021) considers a repeated communication setting in which a sender who lies then forgets his message and may hence later contradict himself. Relatedly, in Ioannidis et al. (2022) the message of the sender determines the receiver's costly investigation technology. Perez-Richet and Skreta (2022) consider general cost functions for the sender from manipulating a test that the receiver designs. Also, see Sobel (2020) for a general definition of lying.

[10] 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.

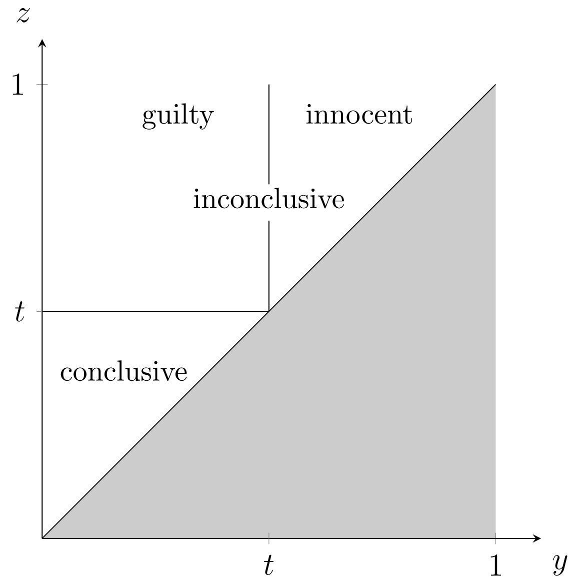


Figure 1 The sample space, the suspect's status and the evidence

according to joint density $\boldsymbol{f}$ of the form $f(y,z) = k\,h(y)g(z)$, where k is a normalizing constant and functions $\boldsymbol{h}$ and $\boldsymbol{g}$ are densities on $[0,1]$ satisfying mild regularity conditions.[11] Throughout, $H(y) \equiv \int_0^y h(y)\mathrm{d}y$ and $G(z) \equiv \int_0^z g(z)\mathrm{d}z$. S is **guilty** when $y < t$ and **innocent** when $y \geq t$, where $t \in (0,1)$ is a commonly known parameter.

R's evidence is a signal about S's type proving that $y < z$, the lower the z the stronger the evidence since S's guilt becomes more likely. In particular, when $z \leq t$ we say that the evidence is **conclusive** since R knows that S is surely guilty (see figure 1). Likewise, in addition to his status as guilty or innocent, S's type determines the strength of the evidence he expects R to possess, which is stronger the lower the y as he knows that $y < z$. Note that $\boldsymbol{f}$ can be thought of as resulting from conditioning on the event that $y < z$ after y and z have been drawn independently according to $\boldsymbol{h}$ and $\boldsymbol{g}$, respectively. It can also be thought of as resulting from first drawing y from $[0,1)$ (according to density $kh(y)(1-G(y))$) and then drawing z conditional on y from $(y,1]$ (according to density $g(z)/(1-G(y))$). In section 4.1.1, we discuss properties of this tractable information structure, the restrictions it entails and generalizations.

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

S's message is a literal claim about his type. We say that he **lies** when $m \neq y$, that he is **honest** when $m = y$, that he **confesses** when $m < t$, and that he **denies** when $m \geq t$. Also,

[11] Namely, $\boldsymbol{h}$ and $\boldsymbol{g}$ are Lipschitz continuous, bounded from above and bounded away from 0 and, additionally, $\boldsymbol{g}$ is differentiable with continuous derivative.

we say that he is caught in a lie when R's evidence contradicts his claim, i.e. when $m \geq z$. In section 4.2.1, we consider S's possibility to stay silent. R's action can be interpreted as a decision on whether S should be prosecuted, i.e. $a = 0$, or let go freely, i.e. $a = 1$.

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

$$\alpha\, a \mathbb{1}_{y<t} + (1-\alpha)\,(1-a)\,\mathbb{1}_{y\geq t}, \tag{1}$$

where $\mathbb{1}_{y<t}$ and $\mathbb{1}_{y\geq t}$ are indicator functions for S's status as guilty and innocent, respectively, and $\alpha \in (0,1)$ a commonly known parameter. S's payoff is

$$a - b\mathbb{1}_{m\geq z}, \tag{2}$$

where $b > 0$ is a commonly known parameter and $\mathbb{1}_{m\geq z}$ the indicator function for when S is caught in a lie.

R aims at prosecuting a guilty and letting an innocent go and $\alpha/(1-\alpha)$ measures the relative importance of a type II error over a type I error. Therefore, α is the threshold probability of innocence below which R finds it optimal to prosecute. S aims at being let go regardless of whether he is innocent or guilty but also incurs some cost b for being caught in a lie. In section 4.1.2, we discuss interpretations of the incentive structure and alternative specifications.

2.2 Equilibrium

Throughout, we restrict our attention to Nash equilibria in pure strategies in which innocent types are honest.[12] Also, after any message which is sent only by guilty types, after a detected lie and when R has conclusive evidence she must infer that S is surely guilty and prosecute. Likewise, upon any message only sent by an innocent type R must infer that S is surely innocent and always let him go. Instead, after messages which are sent by both a guilty and an innocent, R's action must be sequentially rational using a generalized version of Bayes rule (see footnote 13 and section B.1 in the online appendix for details).

To ease exposition, we further focus on a particular class of candidate equilibrium strategy profiles and then explain why doing so is without loss of generality. Accordingly, a strategy of S is described by a strictly increasing and everywhere differentiable lying function $\boldsymbol{\ell} : [y_c, t) \to [t, \bar{y})$ which associates to each guilty type $y \in [y_c, t)$ a lie $\ell(y) \in [t, \bar{y})$ for some $y_c \in [0, t)$, with the

[12] In a previous version of the paper (Ispano and Vida, 2021), we derive this restriction as a result under a truth-leaning equilibrium refinement adapted from Hart et al. (2017).

understanding that types $y < y_c$ (if any, i.e. if $y_c > 0$) confess honestly. We refer to the range $[t, \bar{y})$ of the lying function as to the **lying region** and to S's types sending messages in this region as to pooling types. The behavior of R is described by a **cut-off strategy** $\boldsymbol{z} : [0, 1] \to [t, 1]$, differentiable on the lying region, which specifies for each message $m \in [0, 1]$ and $z > m$, i.e. when S is not caught in a lie, a cut-off $z(m) \in [t, 1]$ such that $a(m, z) = 1$ if and only if $z \geq z(m)$, i.e. a level of weakness of the evidence above which S is let go. Naturally, $z(m) = 1$ for $m < y_c$ and $z(m) = m$ for $m \geq \bar{y}$, i.e. guilties and innocents who do not pool are respectively always prosecuted and always let go. An equilibrium is a pair $\langle \boldsymbol{\ell}, \boldsymbol{z} \rangle$ such that the message of each type of S, including innocents, is optimal given R's strategy and R's action upon each message and evidence realization is optimal given S's strategy.

Specifically, given S's strategy, to compute her expected payoff after a message $\ell(y) \in [t, \bar{y})$ such that $\ell(y) < z$, by Nash equilibrium R must believe that S is innocent with probability

$$\frac{f(\ell(y), z)}{f(\ell(y), z) + \frac{f(y,z)}{\ell'(y)}}, \tag{3}$$

which, given that $f(y, z) = kh(y)g(z)$, does not depend on z. Intuitively, given that $\boldsymbol{f}$ is a product, 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 $y < z$. It then immediately follows from R's sequential rationality that if in equilibrium $z(\ell(y)) \in (\ell(y), 1)$, i.e. if upon message $m = \ell(y)$ R is neither always prosecuting S nor always letting him go, this belief must be exactly α, so that she is indifferent between actions.[13] Moreover, R's ex-ante expected loss in equilibrium is proportional to

$$(1-\alpha) \underbrace{\int_t^{\bar{y}} \left(G\left(z\left(y\right)\right) - G\left(y\right)\right) h\left(y\right) \mathrm{d}y}_{\text{type I errors}} + \alpha \underbrace{\int_{y_c}^{t} \left(1 - G\left(z\left(\ell\left(y\right)\right)\right)\right) h(y) \mathrm{d}y}_{\text{type II errors}} =$$

$$= (1-\alpha) \int_t^{\bar{y}} (1 - G(y)) h(y) \mathrm{d}y. \tag{4}$$

[13] Formally, by Nash equilibrium, R's indifference in any rectangle is given by

$$(1-\alpha) \int_c^d \int_a^b f(m, z) \mathrm{d}m\mathrm{d}z = \alpha \int_c^d \int_{\ell^{-1}(a)}^{\ell^{-1}(b)} f(y, z) \mathrm{d}y\mathrm{d}z = \alpha \int_c^d \int_a^b f(\ell^{-1}(m), z) \ell^{-1\prime}(m) \mathrm{d}m\mathrm{d}z$$

by substituting $\ell^{-1}(m) = y$. Hence, it must be that

$$(1-\alpha) f(m, z) = \alpha f(\ell^{-1}(m), z) \ell^{-1\prime}(m) = \alpha f(\ell^{-1}(m), z) \frac{1}{\ell'(\ell^{-1}(m))},$$

which rearranged with respect to α gives expression (3) by writing $m = \ell(y)$.

Equation (4) obtains since, given her indifference, R would obtain the same payoff by always prosecuting S upon a message in the lying region and hence make only type I errors.

We now establish an indifference condition for S that in turn R's equilibrium strategy will have to satisfy. We say that S is indifferent among denying messages given $\boldsymbol{z}$ if for any two messages $m, m' \in [t, \bar{y})$ such that $m < m'$ and for any $y \in [y_c, m]$

$$1 - G(z(m)) - b(G(m) - G(y)) = 1 - G(z(m')) - b(G(m') - G(y)).$$

This condition means that type y is indifferent between claiming that he is type m or m'. Differentiating with respect to m, the condition simplifies to

$$g(z(m))z'(m) = -bg(m). \tag{5}$$

From equation (5) it is apparent that $\boldsymbol{z}$ must be strictly decreasing over the lying region, i.e. higher messages require weaker evidence for S to be let go, to compensate liars for the higher risk of detection that higher lies entail.

It will become handy to choose $\boldsymbol{z}$ so that this indifference condition and equation (5) extend to every $m \in [y_c, \bar{y})$, i.e. including unexpected (honest) confessions of types $y \in [y_c, \bar{y})$, who hence have no strict incentive to do so.[14] Since a type $y \in [y_c, \bar{y})$ who sends some $m \in [y, \bar{y})$ is then also indifferent to send y, it becomes apparent that his payoff (multiplied by $1 - G(y)$) can be written as

$$1 - G(z(m)) - b(G(m) - G(y)) = 1 - G(z(y)), \tag{6}$$

which is strictly increasing in y. Therefore, confessors, if any, are necessarily low types, who expect the evidence to be stronger and hence have a higher probability of being caught in a lie and incur cost b if they deny. Also, if $y_c > 0$ then $1 - G(z(y_c)) = 0$, i.e. if some types confess the smallest liar must be indifferent between lying and confessing. These observations pin down the equilibrium.

Proposition 1 (Characterization of equilibrium). *There exists an equilibrium $\langle \boldsymbol{\ell}^*, \boldsymbol{z}^* \rangle$, which is uniquely determined by the indifference condition of S and the indifference condition of R together with the appropriate initial and terminal conditions. Namely, there exist a unique $y_c^* \in [0, t)$, a unique $\bar{y}^* \in (t, 1)$ and unique functions $\boldsymbol{\ell}^* : [y_c^*, t) \to [t, \bar{y}^*)$ and $\boldsymbol{z}^* : [0, 1] \to [t, 1]$ such that:*

[14] This particular choice for R's off the equilibrium path behavior simplifies the exposition but is immaterial. Equivalently, R could always prosecute S upon such messages.

(i) $\boldsymbol{\ell}^*$ *is the solution of differential equation*

$$\frac{h(\ell(y))}{h(\ell(y))+\frac{h(y)}{\ell'(y)}}=\alpha, \tag{7}$$

with initial condition $\ell(y_c^*)=t$ *and* $\lim_{y\to t}\ell(y)=\bar{y}^*$*;*

(ii) for $m\in[y_c^*,\bar{y}^*)$ $\boldsymbol{z}^*$ *is the solution of differential equation* $g(z(m))z'(m)=-bg(m)$ *with terminal condition* $z(\bar{y}^*)=\bar{y}^*$ *and* $\min\{y_c^*,1-G(z(y_c^*))\}=0$.

As an illustration, we report closed formed solutions for the equilibrium when (y,z) is uniformly distributed.

Example 1 (Uniform case)**.** *Let* $f(y,z)=2$. *Then,* $\ell^*(y)=\frac{\alpha}{1-\alpha}y+\bar{y}^*-\frac{\alpha}{1-\alpha}t$ *and* $z^*(m)=\bar{y}^*+b(\bar{y}^*-m)$*, where*

- *if* $b\leq\frac{1-t-\alpha}{t}$*, then* $y_c^*=0$ *and* $\bar{y}^*=\frac{t}{1-\alpha}$*;*
- *while if* $b>\frac{1-t-\alpha}{t}$*, then* $y_c^*=\frac{(1+b)t-(1-\alpha)}{b+\alpha}$ *and* $\bar{y}^*=\frac{\alpha+bt}{\alpha+b}$.

Figure 2 displays the payoff of S and the associated type I and type II errors R makes based on the realization of y and z in the uniform case. Separating guilty types, i.e. types below y_c, get $a=0$ and separating innocent types, i.e. types above $\bar{y}$, get $a=1$, so that R makes no errors. As for pooling types, R's action is $a=1$ if $z\geq z(m)$ 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 $z(\ell(y))$, so that R makes a type II error, and $a=0$ otherwise. Likewise, an innocent type below $\bar{y}$ gets $a=1$ when $z\geq z(y)$ and $a=0$ otherwise, and in the latter case R makes a type I error.

We have focused on strategies in which S's lying is increasing and R's action policy takes a natural cut-off form in that she lets S go when the evidence is sufficiently weak. Once one allows for arbitrary (measurable) strategies, there may exist other equilibria. Nevertheless, the lying region, the set of confessors and of lying types, and R's expected action upon each message remain the same. Importantly, players' expected payoffs are therefore also the same, both ex-ante, i.e. before S has observed y and R has observed z, and ex-post.

Corollary 1 (Payoff equivalence)**.** *Any other equilibrium is payoff equivalent for S and R to the one at proposition 1 both from an ex-ante and an ex-post perspective.*

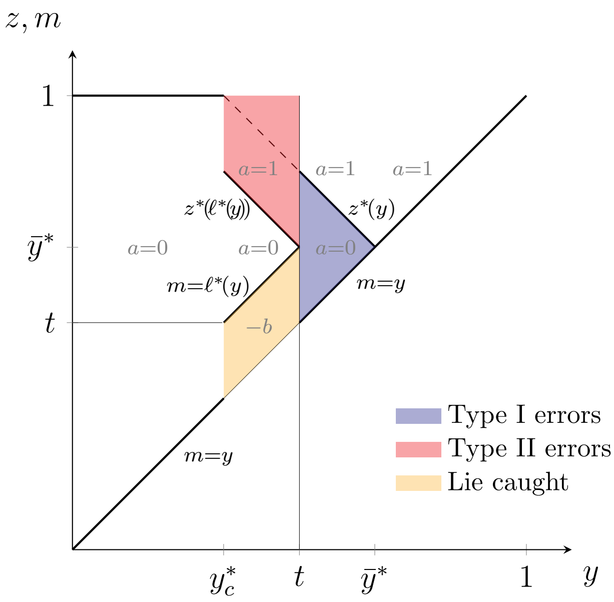


Figure 2 Equilibrium payoffs in the uniform case ($t = 1/2$, $b = 1$, $\alpha = 1/2$)

Due to uniformity, players' strategies are linear. The thick increasing lines $m = y$, $m = \ell^*(y)$, and $m = y$ represent S's strategy. Line $\ell^*(y)$ has 45° slope only because $\alpha = 1/2$, which is also why the interval of liars and the lying region have equal size (again, uniformity is also important). Decreasing thick lines $z^*(\ell^*(y))$ and $z^*(y)$ represent R's cutoff strategy for pooling messages, whereby S is let go in the region above. There are two lines of identical height for the same m since R, not knowing whether S is guilty or innocent, must choose identical actions. Line $z^*(y)$ has −45° slope only because $b = 1$ and line $z^*(\ell^*(y))$ only because $b = 1$ and $\alpha = 1/2$ but the two lines are always decreasing since R lets S go more often upon higher messages. Finally, the dotted line represents our particular selection for R's off the equilibrium path behavior upon unexpected confessions.

The model generates some intuitive comparative statics, which are direct consequence of the following equation which must hold in any equilibrium

$$\alpha \int_{y_c}^{t} h(y)\mathrm{d}y = (1-\alpha) \int_{t}^{\bar{y}} h(y)\mathrm{d}y. \tag{8}$$

Indeed, in the lying region R must be indifferent on average as well given that she is indifferent upon any m. Thus, weakly more types confess (strictly if $y_c^* > 0$) when S's cost for being caught in a lie b is higher, when R is tougher as measured by a higher weight α she attaches to type II errors, and when the prior likelihood of innocence is lower as measured by a higher t or a downward shift of density $\boldsymbol{h}$ in the first-order stochastic dominance sense. Also, some types confess, i.e. $y_c^* > 0$, if and only if

$$G^{-1}(\frac{1}{1+b}) < H^{-1}(\frac{H(t)}{1-\alpha}), \tag{9}$$

which is more easily satisfied when the evidence is stronger as measured by a downward shift of density $\boldsymbol{g}$ in the first-order stochastic dominance sense. A higher cost also strictly reduces the lying region, so that a smaller claim suffices to convince R of S's innocence. Conversely and less trivially, the lying region is larger with a tougher R, which can be intuitively understood as that she requires more convincing to let S go.[15]

Finally, in the context of this simple model, we can ask whether interrogating is useful for R relative to when she relies only on the evidence to make a decision. In that case, R finds it strictly optimal to prosecute if and only if the evidence is lower than the RHS of equation (9), which is also the equilibrium value of $\bar{y}$ when no type confesses, so that the following remark obtains.

Remark 1. *R strictly benefits from interrogating (relative to relying only on the evidence to make decisions) if and only if some types confess, i.e. if and only if $G^{-1}(\frac{1}{1+b}) < H^{-1}(\frac{H(t)}{1-\alpha})$.*

3 The optimal interrogation

In this section, we first consider the mechanism design problem of maximizing R's payoff with full commitment (and with commitment only over actions). Next, we consider the case of commitment only over information revelation, i.e. persuasion. Finally, we show how to implement the optimum without any commitment with back and forth communication.

3.1 Optimal mechanism

In this section, differently from section 2.2, we suppose R can commit to her actions based on S's message and her evidence. We are interested in R's highest attainable ex-ante expected payoff with full commitment and how it compares to R's equilibrium payoff. When allowing for arbitrary mechanisms, the notion of confessing and being caught in lie ($m \geq z$) are not well-defined because S's message space needs not coincide with $\mathcal{M} = [0, 1]$. Therefore, we assume that R can now choose actions $-b$, 0, and 1 and S's payoff is simply equal to R's action a. Besides, since in the baseline model S always has the option to obtain 0 by confessing, we restrict our attention to mechanisms in which each type y of S can guarantee himself a non-negative expected payoff. R's loss is still as at equation (1) with the modification that when R chooses $-b$ it is zero when S is guilty, just as in the baseline model.[16]

[15] Technically, as α increases, equation (8) can be satisfied either with a higher $\bar{y}$ or with a higher y_c. The result is hence obvious if $y_c^* = 0$. In case $y_c^* > 0$, so that y_c^* increases, type y_c^* can be indifferent between confessing and lying, in particular at $\bar{y}^*$, only if $\bar{y}^*$ increases as well. Equivalently, a tougher R decreases the payoffs of all types of S, except those who were already confessing and those who still separate.

[16] If R's loss when she takes action $-b$ and S is guilty was $-b$, i.e. R enjoyed utility from punishing a guilty, then the baseline equilibrium would already yield R's highest attainable expected payoff. Instead, note that R suffers disutility from punishing an innocent since in that case when she chooses action $-b$ her loss is $1 + b$.

While the space of such arbitrary, possibly random, mechanisms is large (see the proof of lemma 1 below for a formal definition), to determine R's optimum it suffices to consider a very simple class. A **direct deterministic cut-off mechanism** $\boldsymbol{z} : [0,1] \to [0,1]$ specifies for each message $y \in [0,1]$ a cut-off level $z(y) \in [y,1]$ such that R's action is $a(y,z) = 1$ if $z \geq z(y)$ and $a(y,z) = 0$ otherwise. Additionally, such a mechanism satisfies the **truth-telling constraint** if for every $y, y' \in [0,1]$ such that $y < y'$

$$1 - G(z(y)) \geq 1 - G(z(y')) - b\left(G(y') - G(y)\right). \tag{10}$$

Lemma 1 (Revelation principle)**.** *For any mechanism, there exists a direct deterministic cut-off mechanism which satisfies the truth-telling constraint and yields R a weakly higher ex-ante expected payoff.*

Thus, the optimal mechanism minimizes

$$\alpha \int_0^t (1 - G(z(y)))h(y)\mathrm{d}y + (1-\alpha) \int_t^1 \left(G(z(y)) - G(y)\right) h(y)\mathrm{d}y \tag{11}$$

subject to the truth-telling constraint (equation (10)). Candidate solutions can be indexed by $z(t) \in [t,1]$ and the constraint must bind for types sufficiently close to t, as figure 3 demonstrates for the uniform case. More precisely, the constraint must bind for all values of y for which $y < z(y) < 1$, yielding

$$g(z(y))z'(y) = -bg(y). \tag{12}$$

This constraint is just S's equilibrium indifference condition, i.e. equation (5). It turns out that the optimal mechanism exactly coincides with the decision rule of R in equilibrium (section A.5.1 in the appendix provides detailed intuitions), which hence differs only in S's behavior.

Proposition 2 (Optimal mechanism)**.** *R's equilibrium strategy $\boldsymbol{z}^*$ at proposition 1 is an optimal mechanism. The sole difference is that S's types who lie in equilibrium instead confess honestly. Hence, type II errors strictly decrease while type I errors remain the same.*

A comparison of the optimal mechanism (figure 3) and the equilibrium (figure 2) clarifies the effects of R's lack of commitment over decisions. R would benefit from committing to sometimes prosecute some low innocent types and to sometimes let go some high guilty types. However, in equilibrium, R can find it sequentially rational to do so only if those types pool in the lying region. Lying creates an inefficiency because R must let liars go more often than in the optimal mechanism to compensate them for the cost of getting caught. It turns out that this is the only

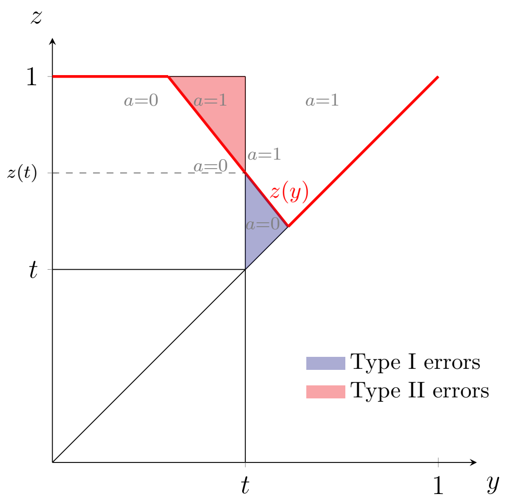


Figure 3 Determination of the optimal mechanism in the uniform case

Given $z(t)$, for guilties and innocents close to t one minimizes respectively type II and type I errors by having the decreasing portion of $z(y)$ (which is linear due to uniformity) as steep as possible, which would be vertical at t in R's first best. Thus, to the right of t constraint (10) binds till $z(y)$ reaches the diagonal $z = y$, after which $z(y) = y$, i.e. R always chooses $a = 1$. Likewise, to the left of t constraint (10) binds till $z(y)$ reaches line $z = 1$ (or the vertical axis when the constraint binds for all guilties, e.g. if $z(y)$ is very flat), after which $z(y) = 1$, i.e. R always chooses $a = 0$. The optimal $z(t)$ trades off type I and type II errors made on S's types for which the constraint binds.

source of inefficiency. Indeed, all types of S get the same payoff as in the optimal mechanism, which implies that type I errors are the same and only type II errors are higher.

Optimal mechanism $\boldsymbol{z}^*$ hinges on R's commitment to act as promised based on S's message m and on the true realized evidence z because R may benefit from misrepresenting her evidence. It turns out that $\boldsymbol{z}^*$ can be made immune to such deviations. Therefore, to reach R's optimum, it suffices that R can commit to a mechanism which implements possibly random actions depending on S's and R's private reports about their types.[17]

Remark 2. *There exists a direct mechanism such that both R and S find it optimal to truthfully report their types and their ex-ante expected payoffs are as in the optimal mechanism.*

3.2 Implementation with persuasion

In this section, as a counterpart of remark 2, using a Bayesian persuasion perspective we investigate whether R can compensate for lack of commitment over actions with commitment over information revelation about the evidence. Namely, R commits to an experiment, or

[17] Such a mechanism, instead of implementing actions, could give recommendations to R about how to act depending on the private reports of S and R and then R could freely choose an action. Mechanisms in which R and S report truthfully and R obeys the recommendations in equilibrium are called communication equilibria (see Forges (1986) and Myerson (1986). It is easy to see that R cannot attain her optimal payoff in a communication equilibrium. The reason is that the threat to punish lies of S by action $-b$ becomes less effective because otherwise R could simply report that her type is t and could always tell apart an innocent (receiving recommendation $-b$) and a guilty (receiving recommendations 0 or 1.)

persuasion rule, which specifies for each z a distribution over signals. S, after observing his type y and the realized signal, sends a message m to R, who then chooses an action and payoffs realize as described in section 3.1. Accordingly, the baseline model is a special case in which R reveals no information about z. Another special case is when R perfectly reveals z but, in line with common intuition, doing so would always be detrimental to R, since then the interrogation would necessarily be uninformative.

We can concentrate on simple persuasion rules, which following Elliott et al. (2021) we refer to as deterministic matchings, in which each signal is sent by two types of R. More precisely, any such rule specifies a signal realization space $D \subseteq [0, 1]$ and a matching function $\boldsymbol{z} : D \to [0, 1]$ such that each signal $d \in D$ is sent by type d and $z(d)$, while unmatched types do not send any signal (e.g. send a signal $\emptyset \notin D$).

Proposition 3 (Optimal Persuasion)**.** *The deterministic matching with $D = [y_c^*, \bar{y}^*]$ and matching function $\boldsymbol{z}^*$ as at proposition 1 is an optimal persuasion rule. The resulting ex-ante expected payoffs of R and S are as in the optimal mechanism.*

Thus, provided R can commit to reveal information about the evidence, and with a sufficiently rich signal space, R can completely dispense with commitment over actions. At the same time, the effectiveness of this persuasion rule hinges on R's commitment not to understate the strength of the evidence. Indeed, any type $z \in (t, z^*(t))$ would gain from deviating and mimicking type $z^*(t)$, whose signal t always allows to set a guilty and an innocent apart. With substantially more work, one can show that a similar issue would arise under any other optimal persuasion rule so that, in contrast to what established at remark 2 for the optimal mechanism, optimal persuasion cannot be made immune to such deviations.

Remark 3. *In any optimal, possibly random, persuasion rule there is necessarily some type z of R who would strictly benefit from sending a signal sent by some weaker type $z' > z$.*

Naturally, under any effective persuasion rule, it is also the case that some types of R would want to overstate the strength of their evidence, i.e. send signals sent by stronger types. Still, as discussed at the next section, doing so will be plainly impossible under the natural assumption that R's private information is hard, i.e. it can be disclosed, possibly vaguely, or withheld but not fabricated.

3.3 Implementation without commitment

In this section, we show how R's expected payoff under the optimal mechanism can be replicated without *any* commitment in a simple game built on the baseline model that combines “back and forth” communication between S and R and discretion for R on whether to punish detected lies. In modeling R's information revelation about the evidence z to S, we suppose that R cannot make false statements, for instance due to the risk of legal action or the inadmissibility of the interrogation in court. Equivalently, S only believes claims backed up by physical proof. Still, R can disclose information vaguely and understate the strength of the evidence, i.e. prove that her evidence is stronger than any given level $\zeta \geq z$. Technically, we assume that any type z of R can send a signal $\zeta \in [z, 1]$ to S. Under these assumptions, remark 3 implies that R cannot attain her optimal payoff in a “short” game in which first R, then S, reveal information.

We consider the following **back and forth** game:

- **stage 0** S and R privately observe their types y and z as in the baseline model;
- **stage 1** S sends a message $m \in [0, 1]$;
- **stage 2** R either immediately chooses an action $a(m, z) \in \{-b, 0, 1\}$, so that the game ends, or sends a signal $\zeta_m \in [z, 1]$ and the game continues to stage 3;
- **stage 3** S sends a new message $m' \in \{0, m\}$;
- **stage 4** R chooses an action $a'(m, m', z, \zeta_m) \in \{-b, 0, 1\}$.

R's action in stage two or four then determines payoffs as in section 3.1. S's message in stage three will be interpreted in equilibrium as an answer to R's question “Are you guilty or you stick to your original story m?” and message 0 as a confession.

Proposition 4 (Implementation without commitment)**.** *There is an equilibrium of the back and forth game in which ex-ante expected payoffs are as in the optimal mechanism.*

The structure of the equilibrium is intuitive. S's behavioral strategy in stage one is as in the equilibrium at proposition 1. R immediately takes the correct action for separating types. Instead, for pooling types, R's continues the interrogation provided the evidence is sufficiently strong relative to S's claim so that by confronting him with the evidence she will be able to persuade a guilty type to withdraw his lie. R does so using the threat of more severe expected punishment. More in detail, when $z > \zeta_m$, where ζ_m is an evidence strength level contingent on m, she immediately makes a decision. As usual, she lets S go if the evidence is sufficiently weak and prosecutes S if the evidence is strong (but still above ζ_m). Instead, when $z \leq \zeta_m$, she discloses this information to S and offers him a second opportunity to confess. The appropriate choice of ζ_m now makes type $\ell^{-1}(m)$ barely willing to do so, while innocent type m sticks to his stage one story anticipating he will be let go. The equilibrium construction entails a combination of "carrot and stick" in the treatment of lies. Small lies caught, i.e. equilibrium lies in stage one are withdrawn in stage three and are forgiven. Conversely, big lies, i.e. off the equilibrium path lies of a guilty type who mimics an innocent type higher than he is supposed to, would still be punished immediately in stage two if caught, i.e. when $z \leq \ell^{-1}(m)$.

4 Discussion

4.1 Modeling assumptions and extensions

In this section, we discuss our main modeling assumptions concerning the information structure and the incentive structure and we cover some extensions, namely, more general prior distributions, no cost for lies caught, leniency for confession, and continuation of the prosecution process after the interrogation.

4.1.1 Properties, interpretations and generalizations of the information structure

Consider first the joint distribution of S's status as guilty or innocent and of the evidence z. Seeing z as a signal about S's status, the monotone likelihood ratio property holds since the probability that S is innocent $\mathbb{P}(y \geq t|z)$ is increasing in z, which clarifies in which sense a higher z represents weaker evidence. Likewise, the distribution of z conditional on S's innocence dominates in the first order stochastic dominance sense the corresponding one conditional on S's guilt. The information structure allows for heterogeneity within guilty and innocent suspects and ensures these natural properties extend to the conditional distribution of y given z and of z given y.

The information structure can also be derived from very simple stories.[18] The fact that the evidence excludes some of S's types - technically the conditional distribution of y given z does not have full support - allows for a natural definition of lie catching. The fact that excluded types are the high ones, e.g. R's evidence can never prove S's innocence, captures the idea that whistleblowing, anonymous tips and voluntary reports to law enforcement authorities are inherently incriminating, which may also explain why S qualifies as a suspect and is interrogated in the first place.[19]

[18] Consider the following examples in addition to the ones in the introduction:

- A jeweler is selling gold rings whose purity y allegedly falls short of the declared purity t. A forensic test provides an upper bound z on the purity level;
- A telephone operator left work at time y, allegedly before the time of the end of his shift t. An unanswered call occurred at time z, proving he had left by then.

[19] The information structure can be equivalently thought of as arising within the isomorphic model in which

Finally, the specific family of joint distributions of y and z that we consider has the convenient property that the private information of each party does not contain information about the private information of the other beyond the fact that $z > y$. In section B.2 of the online appendix, we provide more general conditions on the prior under which the model remains tractable and explain how our main insights generalize.

4.1.2 Interpretations of the incentive structure and alternative specifications

In our baseline model, S incurs a cost $b > 0$ when he is caught in a lie. Depending on the context, this cost have multiple, possibly concurrent, interpretations. First of all, it may be direct and explicit, e.g. if lying entails a penalty or constitutes an independent offense. Besides, it may capture reputation costs, e.g. if S's loss of credibility compromises his position in other investigations. It may also incorporate psychological costs associated with S's dishonesty being exposed. And as formalized in section 3, it may be due to R taking, or threatening to take, a more unfavorable action when S is proving uncooperative. It has long been recognized, and criticized, that law enforcers may resort to such discretion even when they do not have formal authority, which can take many different forms and also be exercised indirectly by influencing the decisions of other parties along the prosecution process.[20]

Our results admit the possibility that S incurs no cost or R has no such discretion, i.e. $b = 0$, as limit case. However, the model becomes one of pure cheap-talk and, due to the conflict of interest between S and R, interrogating is then never useful for R, regardless of the game being played. Also, except for factors outside the model such as surrendering to psychological pressure or relief for admitting guilt, there would then be no rational reasons for S to change his communication strategy over time.

Of course, the source of partial alignment of interests between S and R, and of R's discretion, may also take different, related, forms. Suppose that S incurs no cost but if he confesses he obtains a premium $u \in (0, 1)$, e.g. a plea bargaining deal. When S obtains u, we specify that R makes no error if S is guilty and an error of size $1 - u$ if S is innocent. Let us refer to this incentive structure as to the "leniency setup" to contrast it to the "punishment setup" of our model. After an appropriate parametrization, i.e. $u = b/(1 + b)$, R's baseline equilibrium payoff

y and z are drawn independently from $\boldsymbol{h}$ and $\boldsymbol{g}$ and R, knowing z and whether $z \leq y$ or $z > y$, only interrogates when $z > y$. When $z \leq y$, the evidence is exculpatory in that it excludes some guilty types (all of them if $z \geq t$) and S's innocence becomes more likely than under the prior.

[20] See for instance the discussion in Kassin and McNall (1991) and Bull (2022) and the evidence in Baldwin (1993) and Pearse and Gudjonsson (1999). Also, see Abel (2016) for the involvement of police officers in plea bargaining. Finally, see the guidelines that the training company John E. Reid and Associates provides police officers on promises they can, or cannot make (https://reid.com/resources/investigator-tips/interrogation-procedures-promises-of-leniency).

in the leniency setup is *the same* as in the punishment setup. This comparison illustrates in which sense parameter b can also be thought as capturing leniency for confession. As detailed in section B.3 of the online appendix, the optimal mechanism in the leniency setup can be calculated with similar steps as in the punishment setup. Likewise, there is a correspondent version of the back and forth interrogation in which R grants leniency to types who withdraw their lie and confess. Thanks to the possibility to disclose information about the evidence and discretion on whether to forgive lies by being lenient with a late confessor, R's payoff again improves relative to the baseline model. The main difference is that now, to attain her optimal payoff, R must also be lenient with pooling types of S when the interrogation stops early and the evidence is weak.

Besides, we have assumed that R's loss can be represented as a weighted sum of type I and type II errors. In this flexible, reduced-form, specification, R's exact objective may derive from features of the legal system, e.g. adversarial or inquisitorial, her preferences, e.g. inclined to prosecute or mostly concerned with avoiding to detain an innocent, and her precise role, e.g. police officer or prosecutor. Our analysis can also encompass the case in which payoffs are determined by a future, uncertain decision of a third party, e.g. conviction at trial, which may also depend on the strength of the evidence.

To see this, suppose R only cares about judicial errors, i.e. R's loss is still given by equation (1) but the court's decision c to convict ($c = 0$) or acquit ($c = 1$) S replaces R's action a. If R lets S go, then S faces no trial and is acquitted. If R prosecutes S, then S goes to court, where he will be acquitted with some given probability $\sigma(z, m) \in [0, 1]$ and convicted otherwise. Letting μ represent R's belief that S is innocent, R now prefers to prosecute S if and only if $(1 - \mu)\alpha\sigma(z, m) + \mu(1 - \alpha)(1 - \sigma(z, m)) \leq \alpha(1 - \mu)$. The inequality simplifies to $\mu \leq \alpha$, i.e. the same decision rule of R as in our baseline model, so that equation (1) still captures her incentives. Moreover, once S's incentives are also taken into account, the equilibrium construction of the baseline model easily adjusts so that the equilibrium becomes one of the game under discussion.[21] Likewise, this extended model can accommodate that the probability of conviction also depends on S's true status as guilty or innocent rather than only

[21] Suppose S's payoff is given by equation (2) but the court's decision c replaces a and, for simplicity, that $\sigma(z, m) = 0$ whenever S confesses or is caught in a lie. The equilibrium is as at proposition 1 except that, when S denies and is not caught in a lie, R, which is again indifferent between actions, should now let S go if and only if $z \geq z_\sigma(m)$, where $z_\sigma(m)$ solves $1 - G(z^*(m)) = \int_m^{z_\sigma(m)} \sigma(z, m)g(z)\mathrm{d}z + 1 - G(z_\sigma(m))$. A solution $z_\sigma(m) \in [z(m), 1)$ exists provided $\sigma(z, m)$ is on average not too high. Then, S's incentives to follow the equilibrium strategy are completely unaffected and R obtains the same equilibrium payoff given that her decision may differ from the one in the equilibrium at proposition 1 only when she is indifferent. If $\sigma(z, m)$ is on average too high, instead, the model is essentially equivalent to the case in which when S does not confess he must be let go when the evidence is too weak (see section 4.2.1).

on the evidence, e.g. because new evidence is likely to be uncovered in the future. In this case, it is essentially as if R became tougher.[22] The model can also accommodate that R's preferences depend directly on the evidence, e.g. because a tougher interrogator is appointed when the evidence is stronger (see Ispano and Vida (2022)). Finally, while we have assumed that R cares about S's communication purely for its informational content, this extended model can account for why R may attach additional value to some messages of S, most notably to confessions. This would be the case if there are instances in which R is sure of S's guilt, for example because the evidence is conclusive or S has been caught in a lie, but unless S confesses his conviction remains uncertain.

4.2 Partial commitments

In this section, we show how protection of the suspect's right to silence and an evidence strength standard for interrogating can effectively alleviate the law enforcer's commitment problem identified in section 3.1.

4.2.1 Protection of 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, there is a longstanding debate on whether an adverse inference, i.e. a conclusion pointing at the suspect's guilt, can be drawn (see for instance Seidmann and Stein (2000), Seidmann (2005), and the discussion between O'Reilly (1994) and Ingraham (1995)).

Let us consider the baseline model but augment S's message space to include the possibility to stay silent. Given that innocents are honest, R always finds it optimal to prosecute a silent S. And if doing so is always possible for R, no guilty type has ever an incentive to stay silent in the first place. Suppose instead that, upon silence, R can prosecute S only if $z \leq Z_s$, in which case S's payoff is $-b_s$, while if $z > Z_s$ then she must necessarily let S go, where $Z_s \in (t, 1]$ and $b_s \in [0, b]$ are commonly known parameters.

[22] Let $\beta = \mathbb{P}(c = 1|y < t, m, z)$ and $\gamma = \mathbb{P}(c = 0|y \geq t, m, z)$ be the probability that the court makes an error conditional on the suspect being guilty and innocent, respectively, with $\beta = 0$ and $\gamma = 1$ when S confesses or is caught in a lie for simplicity. R now prefers to prosecute S if and only if $(1-\mu)\alpha\beta + \mu(1-\alpha)\gamma \leq \alpha(1-\mu)$. In equilibrium, R must again be indifferent in the lying region, so that her belief is now $\mu = \alpha' \equiv \frac{\alpha(1-\beta)}{(1-\alpha)\gamma+\alpha(1-\beta)} \in (0,1)$ and the equilibrium values of $\bar{y}$, y_c and $\boldsymbol{z}$ are determined by α' instead of α. As long as $\gamma \leq 1-\beta$, i.e. the court is more likely to convict a guilty than an innocent, $\alpha' > \alpha$. The incentives of a guilty type are again completely unaffected if, rather than according to $\boldsymbol{z}$ as determined by α', R now lets S go if and only if $z \geq z_\beta(m)$, where $z_\beta(m)$ solves $1 - G(z(m)) = \beta(G(z_\beta(m)) - G(m)) + 1 - G(z_\beta(m))$. Again, a solution $z_\beta(m) \in [z(m), 1)$ exists provided β is not too high. And as long as $\gamma \leq 1-\beta$, innocent types now have even stronger incentives to be honest since $(1-\gamma)(G(z_\beta(m)) - G(m)) + 1 - G(z_\beta(m)) \geq 1 - G(z(m))$.

In this flexible specification, Z_s represents the evidence strength standard required to prosecute a silent S.[23] For example, if $Z_s = H^{-1}\left(\frac{H(t)}{1-\alpha}\right) < 1$, R cannot use the informational content of silence and hence must make her decision as if she relied on the evidence alone (see equation (13)). Seidmann (2005) refers to this case as to the "American game" to contrast it to the "English game", in which an adverse inference is always allowed ($Z_s = 1$, corresponding to our baseline model). Provided the standard is met, so that prosecution is possible, parameter b_s measures S's eventual cost of reticence relative to confession, possibly lower than the cost of being caught in a lie.

The equilibrium of the baseline model easily adjusts for any Z_s and b_s. Sufficiently low guilty type, if any, confess, intermediate guilty types, if any, stay silent, and sufficiently high guilty type lie. Naturally, the interval of silent types enlarges as protection of silence gets stronger as measured by a lower Z_s or a lower b_s. Depending on parameters, this enlargement occurs at the expenses of the interval of confessors, of liars, or of both. This model can hence explain why confession and silence may coexist as optimal equilibrium strategies. Importantly, it also demonstrates that R may benefit from stronger protection of silence.

Proposition 5 (Protection of silence)**.** *Fix $f(y,z) = 2$ and $b \leq \frac{1-t-\alpha}{t}$. There exists a $Z_s < 1$ such that R's equilibrium payoff is strictly higher than in the baseline model if and only if $b_s < b$.*

A level of protection that induces some guilty types to remain silent may be beneficial for R because, if on the one hand it entails a type II error upon silence when the evidence is weak, on the other hand it reduces the fraction of liars and hence the pooling of innocents and guilties. Seidmann (2005) shows that S prefers the "American game" to the "English game", which is also the case in our setting, but R never does so. In spite of important differences between the two settings, proposition 5 also suggests how to reconcile these findings.[24] Indeed, for R to benefit from protection of silence it must necessarily be the case that, when prosecuted, a silent type obtains a higher payoff than a liar caught ($b_s < b$), so that a sufficiently weak level of protection suffices to incentivize silence while limiting the associated type II errors. The incentive structure of Seidmann (2005) does not allow this possibility. For simplicity, the proposition focuses on the case in which there are no confessors in the baseline model. When there are confessors in the baseline model, there are still parameter configurations for which R's payoff improves thanks to

[23] Equivalently, the standard could apply to any message other than confession. Indeed, in equilibrium, even an S who denies is always let go whenever the standard is not met.

[24] Differences concern both the information structure, in particular in our setting R's evidence cannot prove S's innocence but can prove his guilt and guilty types are heterogeneous in the strength of the evidence they expect, and the incentive structure, in particular Seidmann (2005) considers the leniency setup for S (see section 4.1.2).

protection of silence and these still require $b_s < b$. However, this condition is no longer sufficient since stronger protection of silence entails a first-order informational loss whenever it deters confession.

4.2.2 Standard for interrogating

As in the case of other restraints of individual freedom such as searches and arrests, law enforcers may be required to hold sufficiently strong evidence to interrogate the suspect in the first place. To analyze the effect of such evidence strength standard, consider the baseline model but suppose R can only interrogate if $z \leq Z_i$, where $Z_i \in (t, 1]$ is a commonly known parameter. Therefore, when S is interrogated, he knows R's evidence meets the standard. For simplicity, suppose also that when $z > Z_i$ then R must necessarily let S go.

The equilibrium analysis of our baseline model, which corresponds to $Z_i = 1$, easily generalizes. 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. R gives up the chance to interrogate S upon weak evidence but can conduct more informative interrogations upon strong evidence. The positive effect sometimes dominates.

Proposition 6 (Standard for interrogating). *Fix $f(y, z) = 2$ and $b \leq \frac{1-t-\alpha}{t}$. There always exists a $Z_i < 1$ such that R's equilibrium payoff is strictly higher than in the baseline model.*

For simplicity, the proposition again focuses on the case in which there are no confessors in the baseline model. Since the interrogation would otherwise be uninformative (see remark 1), R always strictly benefits from an appropriately chosen standard for interrogating. When there are confessors in the baseline model, there still exist parameter configurations for which R benefits from the standard. Also, one can show that in both cases, provided R sets the standard optimally, she always finds it optimal to indeed let S go when her evidence does not meet the standard, as we assumed at the outset.[25]

4.3 Empirical evidence and predictions

Empirical studies on interrogations indicate that the strength of the evidence as perceived by the suspect and law enforcers is a key predictor of the outcome of the interrogation and, in particular, of confession (Gudjonsson and Petursson, 1991; Moston et al., 1992; Stephenson

[25] In fact, the outcome under the optimal standard for interrogating also obtains in equilibrium of the game in which R reveals information about the evidence to S before he makes a claim, as in section 3.2, but by means of strategic disclosure without commitment as in section 3.3 (for details, see Ispano and Vida (2021)).

and Moston, 1994; Redlich et al., 2018). The following quote from Moston et al. (1992) best illustrates the foundations of our model.

> The strength of evidence against a suspect is likely to be a major determinant of both suspect behaviour and interviewing style.[...] When there is weak evidence, there is a limited range of interviewing strategies available, but with stronger evidence a greater range of possibilities emerges. Suspects' knowledge of the evidence against them is also likely to be a key predictor of how they will respond to an allegation. What is important here is the extent to which the suspect knows of the police evidence [...]. The police may well have strong evidence, such as witnesses and fingerprints, but if the suspect is unaware of this [...] the suspect may well begin an interview by denying [...] but later decide to confess as the police points out the evidence. Strength of evidence as perceived by both police and suspect is central to the process of interrogation. The interviewer manipulates the suspect's decision-making by using the available evidence as a persuasive technique. [...] The suspect's initial response is unlikely to bring an immediate end to the interview, particularly if it is a denial [...]. The initial response may prompt the interviewer to adopt a different questioning strategy, for example, using techniques to persuade the suspect that confessing may have its advantages.

Interestingly, early studies document rather poor interviewing techniques whereby suspects are immediately confronted with the evidence and those who do not confess are rarely induced to revise their initial claims (Moston et al., 1992; Baldwin, 1993; Stephenson and Moston, 1994). For example, Stephenson and Moston (1994) report that the accusatorial approach dominates over the information-gathering one, where

> The principal difference between the accusatorial and information-gathering strategies lies in the timing and context of the officer's "upgrading" of questioning by the introduction of whatever evidence is at his or her disposal. The standard line of questioning in the accusatorial style goes from an opening accusation by the interrogator followed by silence or a swift denial by the suspect, to "upgrading" by the interrogator which may be more or less effective in inducing an admission or damaging statement [...] By adopting the information-gathering strategy the interviewer increases the probability of eliciting an account from the suspect of what had occurred. The interviewer can then introduce evidence into questioning which contradicts this account, either in part or in a whole. In a successful interrogation

> employing this strategy the suspect's account may be incrementally modified such that eventually an admission of guilt is finally elicited.

Instead, more recent works report the use of more sophisticated tactics based on evidence revelation in reaction to the suspect's account, also as a result of training programs and reforms promoting a less accusatorial approach in favor of objective information gathering (Kassin et al., 2007; Soukara et al., 2009; Bull and Soukara, 2010; Walsh and Bull, 2010; Kelly et al., 2016; Leahy-Harland and Bull, 2017).

Our results speak to the merits of these more sophisticated tactics in that the back and forth game at section 3.3 can be maximally effective while, as discussed in section 3.2, immediate full evidence revelation makes the interrogation uninformative. Our results also highlight some empirical challenges from a rigorous test of this prediction, which may also contribute to explain why the common approach of measuring how the use of different tactics over the course of the interrogation correlates with confession rates typically yields mixed or only suggestive evidence. For a start, the game predicts that, as also noted in Soukara et al. (2009), these more sophisticated tactics are used when the suspect is not already voluntary confessing, which are by definition tougher cases. Also, the game summarizes in only few stages what in reality is often a longer, gradual, dynamic process in which the timing of confronting the suspect with the evidence and the exact content of disclosure are also key (see Kelly et al. (2016)). Importantly, the game demonstrates how only looking at confession rates as outcome variable may be reductive since in equilibrium it is also the case that claims of denial become more credible. Finally, an important aspect of the game is that withdrawn lies are forgiven in equilibrium. This information is typically unavailable in the data and hard to detect even by indirect measures, also considering that a confession obtained by an explicit promise or threat raises validity concerns in most legal systems. While our theory is agnostic about how equilibrium play is reached, it is consistent with the fact that an appeal to the suspect's self-interest from confessing is also typically part of these tactics (see for instance Kassin et al. (2007)).

Our static baseline model and its extensions also generate clear testable predictions that we discussed in section 2.2, 4.2.1 and 4.2.2. In particular, these are consistent with the observation above that the strength of the evidence, and the strength *perceived* by the suspect, are a key predictor of the suspect's strategy and confession. Also, our model at section 4.2.1, like the one of Seidmann (2005), can account for the stylized fact that a change in the level of protection of silence may not affect confession rates but simply the fraction of silent types and liars. Additionally, it can explain why confessing, staying silent and denying may be optimal for different guilty types, so that suspects may use all of these strategies even without any change in the observable characteristics of a case or the institutional framework.

4.4 Concluding remarks and avenues for future research

We provided a tractable framework to analyze interrogations and derived several implications for their design. In particular, we identified the commitment problems intrinsic to interrogations and solutions to alleviate or solve them. While our main objective of interest has been the accuracy of law enforcers' decisions, i.e. minimizing errors, all solutions discussed in this paper are not detrimental to the suspect's welfare and hence represent Pareto improvements. Also, while we did not consider the suspect's choice to engage in unlawful behavior, nor his care in avoiding to generate incriminating evidence, it seems plausible that, all else being equal, more accurate interrogations will also serve deterrent purposes.

In deriving these results, we maintained that all aspects of the strategic environment other than players' private information are common knowledge. However, law enforcers' power and arbitrariness are a major cause of criticism and an important reason behind the general movement towards the mandatory recording of interrogations (see for instance Sullivan (2005)). For example, there is evidence that suspects who are more fragile, less familiar with the legal system or not advised by a lawyer are also more prone to confess (see for instance Gudjonsson and Petursson (1991) and Moston et al. (1992)). Our model directly allows to identify the direction of the misleading efforts law enforcers would want to engage in if these are tolerated by law or go undetected and the suspect is prone to deception. Predictions agree with the logic behind common deceptive interrogations tactics (see for instance Kassin and McNall (1991)).[26] While surely objectionable on other grounds, if successful, these deceptive tactics improve information elicitation. Also, this improvement needs not come at the cost of extorting false confessions since innocent would still have no incentives to depart from honesty. As a next step, one could investigate if these deceptive tactics would remain effective in a framework in which the suspect is rational but uninformed about some institutional aspects (see Ispano and Vida (2022) for uncertainty about the interrogator's preferences). Likewise, since false confessions occur (see for instance Leo and Ofshe (1998)), it would be important to know how our insights modify when also innocents can have fundamental reasons not to be honest, for instance because they may expect even stronger evidence than some guilty types.

[26] Supposing S plays according to what he perceives as equilibrium behavior while R best responds given the true environment, we can easily calculate how R would want to mislead the suspect about several parameters of interest. R would always want to overstate the cost of reticence or the benefits from confessing (increase S's perception of b), exaggerate the strength of the incriminating evidence (decrease S's perception of ζ_m and Z_i as defined respectively in section 3.3 and 4.2.2) and misrepresent her true preferences over type I and type II errors (increase or decrease S's perception of α).

Besides, we did not consider laws that govern communication about the evidence to the suspect and we maintained that law enforcers' statements are voluntary but must be truthful. If law enforcers can make false claims, instead, new interesting strategic considerations arise due to the possibility that the suspect may in turn catch law enforcers in a lie, e.g. know that they are exaggerating the strength of the evidence. Regulation might also affect the law enforcers' strategic choice to interrogate the suspect in the first place and whether by means of a casual conversation or a formal interrogation. For example, by officially marking the start of a formal interrogation, the legal requirement that the suspect is notified of his right to silence may implicitly convey information about the presence of incriminating evidence.

References

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

Aumann, Robert J. and Sergiu Hart, "Long Cheap Talk," *Econometrica*, 2003, *71* (6), pp. 1619–1660.

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.

Baldwin, John, "Police interview techniques: Establishing truth or proof?," *The British Journal of Criminology*, 1993, *33* (3), 325–352.

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.

Bull, Jesse, "Interrogation and Disclosure of Evidence," *Working paper*, 2022.

Bull, Ray and Stavroula Soukara, "Four studies of what really happens in police interviews," in G. D. Lassiter and C. A. Meissner, eds., *Police interrogations and false confessions: Current research, practice, and policy recommendations*, American Psychological Association, 2010, pp. 81–95.

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.

Cuellar, Pablo, "Voluntary Disclosure of Evidence in Plea Bargaining," *Working paper*, 2020.

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.

Elliott, Matthew, Andrea Galeotti, Andrew Koh, and Wenhao Li, "Market segmentation through information," *Working paper*, 2021.

Forges, Francoise, "An approach to communication equilibria," *Econometrica*, 1986, *54* (6), 1375–1385.

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

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani, "Mediation, arbitration and negotiation," *Journal of Economic Theory*, 2009, *144* (4), 1397–1420.

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

Gudjonsson, Gisli H and Hannes Petursson, "Custodial interrogation: Why do suspects confess and how does it relate to their crime, attitude and personality?," *Personality and Individual Differences*, 1991, *12* (3), 295–306.

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," *American Law and Economics Review*, 07 2022.

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.

_ **and Péter Vida**, "Designing Interrogations," *Working paper*, 2021.

_ **and Péter Vida**, "Good cop-bad cop: delegating interrogations," *Working paper*, 2022.

Jehiel, Philippe, "Communication with forgetful liars," *Theoretical Economics*, 2021, *16* (2), 605–638.

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.

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

_ **, Richard A Leo, Christian A Meissner, Kimberly D Richman, Lori H Colwell, Amy-May Leach, and Dana La Fon**, "Police interviewing and interrogation: A self-report survey of police practices and beliefs," *Law and Human Behavior*, 2007, *31* (4), 381–400.

Kelley, Walter G and Allan C Peterson, *The theory of differential equations: classical and qualitative*, Springer Science & Business Media, 2010.

Kelly, Christopher E, Jeaneé C Miller, and Allison D Redlich, "The dynamic nature of interrogation.," *Law and human behavior*, 2016, *40* (3), 295.

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li, "Persuasion of a privately informed receiver," *Econometrica*, 2017, *85* (6), 1949–1964.

Krishna, Vijay and John Morgan, "The art of conversation: eliciting information from experts through multi-stage communication," *Journal of Economic Theory*, August 2004, *117* (2), 147–179.

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

Leahy-Harland, Samantha and Ray Bull, "Police strategies and suspect responses in real-life serious crime interviews," *Journal of police and criminal psychology*, 2017, *32* (2), 138–151.

Leo, Richard A and Richard J Ofshe, "The consequences of false confessions: deprivations of liberty and miscarriages of justice in the age of psychological interrogation," *Journal of Criminal Law and Criminology*, 1998, *88* (2), 429–496.

Leshem, Shmuel, "The Benefits of a Right to Silence for the Innocent," *The RAND Journal of Economics*, 2010, *41* (2), 398–416.

Mialon, Hugo M, "An economic theory of the fifth amendment," *Rand Journal of Economics*, 2005, pp. 833–848.

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

Moston, Stephen, Geoffrey M Stephenson, and Thomas M Williamson, "The effects of case characteristics on suspect behaviour during police questioning," *The British Journal of Criminology*, 1992, *32* (1), 23–40.

Myerson, Roger B, "Multistage Games with Communication," *Econometrica*, 1986, *54* (2), 323–358.

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.

Pearse, John and Gisli H Gudjonsson, "Measuring influential police interviewing tactics: A factor analytic approach," *Legal and Criminological Psychology*, 1999, *4* (2), 221–238.

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

Perez-Richet, Eduardo and Vasiliki Skreta, "Test design under falsification," *Econometrica*, 2022, *90* (3), 1109–1142.

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.

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

_ **and Alex Stein**, "The right to silence helps the innocent: A game-theoretic analysis of the Fifth Amendment privilege," *Harvard Law Review*, 2000, pp. 430–510.

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, "The Economic Case for Probablity-Based Sentencing," *Working paper*, 2019.

_ **and** _ , "Judicial Mechanism Design," *American Economic Journal: Microeconomics*, forthcoming.

Sobel, Joel, "Lying and deception in games," *Journal of Political Economy*, 2020, *128* (3), 907–947.

Soukara, Stavroula, Ray Bull, Aldert Vrij, Mark Turner, and Julie Cherryman, "What really happens in police interviews of suspects? Tactics and confessions," *Psychology, Crime and Law*, 2009, *15* (6), 493–506.

Stephenson, Geoffrey M and Stephen J Moston, "Police interrogation," *Psychology, Crime and Law*, 1994, *1* (2), 151–157.

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

Walsh, Dave and Ray Bull, "What really is effective in interviews with suspects? A study comparing interviewing skills against interviewing outcomes," *Legal and criminological psychology*, 2010, *15* (2), 305–321.

Appendix

A Proofs

A.1 Proof of proposition 1

Suppose that such a $\langle \boldsymbol{\ell}^*, \boldsymbol{z}^* \rangle$ exists. We already established sequential rationality of R's strategy. Given S's indifference condition, sequential rationality of S's strategy is straightforward. Since S's payoff is increasing in y, the strategy of confessors, if any, is optimal and no other type prefers to confess. The strategy of a liar is also optimal, since he is indifferent between sending any lie $m \in [y, \bar{y}]$ while any lie $m > \bar{y}$ is strictly dominated. For the same reasons, an innocent type $y \in [t, \bar{y}]$ is indifferent between being honest and sending any lie $m \in [y, \bar{y}]$, while he strictly prefers to be honest than to send any lie $m > \bar{y}$ or $m \in [t, y)$. Finally, by being honest an innocent type $y \geq \bar{y}$ earns the maximum attainable payoff.

We are left to show that $\langle \boldsymbol{\ell}^*, \boldsymbol{z}^* \rangle$ exists. We can construct an equilibrium of the baseline model as follows. Let $y_c, \bar{y}$, and R's decision rule be determined by the optimal mechanism (we suppress superscripts * everywhere when this does not cause confusion) as in the proof of proposition 2. By the Picard-Lindelöf theorem, we can calculate $\boldsymbol{\ell}$ using equation (7), which must hold for each $y \in [y_c, t)$ with initial condition $\ell(y_c) = t$. Notice that for the solution we must have $\lim_{y \to t} \ell(y) = \bar{y}$. This will hold because if R is indifferent after every message $\ell(y)$ then R is indifferent also on average (without knowing the message), i.e. equation (8) must hold and we know by lemma 2 (see again the proof of proposition 2) that the optimal mechanism satisfies this property, i.e. equation (8) indeed holds.

To see by direct calculation that $\lim_{y \to t} \ell(y) = \bar{y}$ holds, let $\ell(t) = \lim_{y \to t} \ell(y)$. Then

$$\int_{\ell(y_c)=t}^{\ell(t)} h(y')\mathrm{d}y' = \int_{y_c}^{t} h(\ell(y))\ell'(y)\mathrm{d}y = \frac{\alpha}{1-\alpha} \int_{y_c}^{t} h(y)\mathrm{d}y.$$

By substituting $y' = \ell(y)$, using equation (7) for $\ell'(y)$, and finally using the average indifference condition (equation (8)), which holds by lemma 2, we have that $\ell(t) = \bar{y}$.

A.2 Proof of corollary 1

Given some strategy of R, if a type y of S finds it optimal to lie then all types above y find it optimal to lie and obtain a strictly higher payoff than y. It is also clear that some guilty type must lie. Indeed, if no guilty types lie then R's action after message t should be 1 but then a guilty type sufficiently close to t would strictly prefer to lie. Hence, we can assume that S's equilibrium strategy is always some measurable function $\boldsymbol{\ell} : [y_c, t) \to [t, \bar{y})$, where $\bar{y}$ is the supremum of lies.[27] We establish now that the indifference condition for S has to hold as well in equilibrium and that the range of $\boldsymbol{\ell}$ is an interval. If this was not the case (say some y strictly prefers m to $m' < \bar{y}$) than all guilty types would strictly prefer m to m' and hence the range of $\boldsymbol{\ell}$ would not be an interval. In that case R's action after message m' should then be always 1 because no guilty type would send m'. As a consequence, no guilty type would ever send a message higher than m' because by sending m' they could obtain action 1, which would contradict that $\bar{y}$ is the supremum of the lies. For the same reasons, it follows that in an equilibrium in which R uses a cut-off strategy it must be that $z(m) > m$ for $m < \bar{y}$ and that $z(\bar{y}) = \bar{y}$. It follows that R's expected action must make S indifferent, i.e. for $a(m, z)$ we must have $\frac{\int_m^1 a(m,z)g(z)\mathrm{d}z}{1-G(m)} = \frac{1-G(z(m))}{1-G(m)}$ as the expected payoff of innocent type m fixed, where $\boldsymbol{z}$ is the equilibrium strategy from proposition 1.

It is also clear that the indifference condition for R must hold as well. After message t her action cannot be 1 because then a mass of guilty types would like to lie at t which would then change the optimal action of R to 0. After message t her action cannot be 0 because then no guilty type would lie to t which would change R's optimal action to 1. The indifference of R can be achieved by a lying function inducing a belief of R about the innocence of S equal to α because her belief cannot depend on z (see section (B.1) in the online appendix). Finally, given that R's indifference holds after any denying message it must be that it also holds on average. It follows that in any equilibrium equation (8) must also hold. Together with the possibly binding indifference condition for y_c between confessing and lying (otherwise $y_c = 0$ binds), the equilibrium values are pinned down just as in proposition 1. It follows that the payoff equivalence holds in any equilibrium for both S and R, both ex-ante and ex-post.

A.3 Proof of remark 1

When R does not interrogate, as $\mathbb{P}(y \geq t|z) = (H(z) - H(t))/H(z)$, by equation (1) it follows that R finds it strictly optimal to prosecute if and only if

$$z < H^{-1}\left(\frac{H(t)}{1-\alpha}\right). \tag{13}$$

[27] Throughout, for simplicity, we adopt the convention that the intervals of confessors, if any, and of lies sent, are right-open. There may also exist equilibria in which these intervals are right-closed.

When there are no confessors in equilibrium, i.e. $y_c^* = 0$, by equation (8) $\bar{y}^*$ is precisely equal to the RHS of equation (13). Since when $z < \bar{y}^*$ R always prosecutes S, while when $z \geq \bar{y}^*$ R could equivalently always let R go by the same argument that yields to equation (4), her payoff is then the same as when not interrogating. Conversely, when there are confessors in equilibrium, R is strictly better off because when she does not interrogate she necessarily either sometimes lets go some guilties who would confess or prosecutes some innocents who would separate.

A.4 Proof of lemma 1

Consider an arbitrary mechanism in the form of a measurable message space $\tilde{\mathcal{M}}$ and of a measurable function $\tilde{\mathcal{M}} \times [0,1] \to \Delta(\{-b,0,1\})$ which associates a random action to any pair (m,z), where m is S's message, z is R's actual true evidence and $\Delta(\{-b,0,1\})$ is the set of all probability distributions. The only constraint is that $q(y) \geq 0$ for each $y \in [0,1]$, where $q(y)$ denotes the expected payoff of type y when playing in the mechanism. Fix the resulting lowest expected loss of R when each type sends only messages which are optimal for him given the mechanism.

Consider now the deterministic cut-off direct mechanism $\boldsymbol{z}$ that, in expectation with respect to z, gives each type y exactly $q(y)$ when y reports that his type is y. This direct mechanism satisfies the truth-telling constraint (equation (10)). Indeed,

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

where the first inequality follows from the fact that, conditional on $z \geq y'$, it must be that y' expects $q(y')$ 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 may now prefer to report that he is type $y' < y$. Thus, all we can deduce is that the obtained direct mechanism is weakly better for R than the original mechanism in an environment where downward deviations are not possible. However, our optimal direct mechanism $\boldsymbol{z^*}$ is also optimal in the environment where downward deviations are not possible. Moreover, of course, $\boldsymbol{z^*}$ is also immune to such deviations. Therefore, focusing on direct deterministic cut-off mechanisms satisfying the truth-telling constraint is without loss of generality.

A.5 Proof of proposition 2

After providing an intuition for the relation between the optimal mechanism and the equilibrium (section A.5.1), we formally prove its optimality (section A.5.2).

A.5.1 Intuition

As pointed out in the body of the paper and explained in figure 3, in the optimal mechanism the truth-telling constraint binds for types sufficiently close to t. The exact counterpart of the truth-telling constraint in the equilibrium of the baseline model is that pooling types are indifferent among any lie. The optimal choice of $z(t)$ is then determined by the fact that R is trading off type I and type II errors. Suppose we increase the value of $z(t)$. At the 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 α. In the uniform case, these are measured by (or are proportional to) the appropriately weighted lengths of the $z(y)$ line from $z(t)$ respectively to the right of t (till the diagonal $z = y$) and to the left of t (till the line $z = 1$). The exact equilibrium counterpart of this optimality condition is the required average indifference of R described by equation (8), relating the measures of liars $(t - y_c)$ and of the lying region $(\bar{y} - t)$. Thus, when projecting the graph of the optimal $\boldsymbol{z}^*$ 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 equation (8). Somewhat surprisingly, the same argument goes through for the non-uniform case.

A.5.2 Proof

By lemma 1, focusing on direct deterministic cut-off mechanisms satisfying the truth-telling constraint is without loss of generality. Consider hence any decreasing function $\boldsymbol{z}$ hitting the vertical axis at some p above the horizontal axis, i.e. $p = z(t)$ and $p \in [t, 1]$. There are corresponding values of $\bar{y}(p) \in [t, 1]$ and $y_c(p) \in [0, t]$ such that $z(\bar{y}(p)) = \bar{y}(p)$ and either $z(y_c(p)) = 1$ or there is a function $k(p) \in [t, 1]$ such that $z(0) = k(p)$, in which case we set $y_c(p) = 0$.

Lemma 2. *There is a unique optimal direct cut-off mechanism. Moreover the first order condition is binding and it simplifies to R's indifference condition, i.e. to equation (8):*

$$\alpha \int_{y_c(p)}^{t} h(y)\mathrm{d}y = (1-\alpha) \int_{t}^{\bar{y}(p)} h(y)\mathrm{d}y. \tag{14}$$

Proof. As explained in the body of the paper, by the indifference condition of S, $\boldsymbol{z}$ must satisfy differential equation (12) with initial condition $p = z(t)$. The solution uniquely exists by the Picard-Lindelöf theorem (by gluing the local solutions together). Let us denote this solution by $z(.,p)$, which is differentiable with respect to p (see for instance Kelley and Peterson (2010)). Given such a $z(.,p)$ and the corresponding values of $\bar{y}(p)$ and $y_c(p)$ (which are also differentiable with respect to p), using equation (11), the optimal mechanism must minimize

$$\alpha \int_{y_c(p)}^{t} \int_{z(y,p)}^{1} h(y)g(z)\mathrm{d}z\mathrm{d}y + (1-\alpha)\int_{t}^{\bar{y}(p)} \int_{y}^{z(y,p)} h(y)g(z)\mathrm{d}z\mathrm{d}y.$$

Differentiating with respect to p, using Leibniz integral rule and that $z(y_c(p),p) = 1$ or $y_c'(p) = 0$ and $z(\bar{y}(p),p) = \bar{y}(p)$ for every p, the first order condition simplifies to

$$\alpha \int_{y_c(p)}^{t} h(y)g(z(y,p))\frac{\partial z(y,p)}{\partial p}\mathrm{d}y = (1-\alpha)\int_{t}^{\bar{y}(p)} h(y)g(z(y,p))\frac{\partial z(y,p)}{\partial p}\mathrm{d}y. \tag{15}$$

Using now the indifference condition between honesty and lying up to y of type $y_c(p)$, i.e.

$$1 - G(z(y_c(p),p)) = 1 - G(z(y,p)) - b(G(y) - G(y_c(p))),$$

and differentiating with respect to p yields

$$g(z(y,p))\frac{\partial z(y,p)}{\partial p} = D_pG(z(y,p)) = D_p(G(z(y_c(p),p)) + bG(y_c(p))) = K(p),$$

where $K(p) \neq 0$ is some function constant in y (in fact it is $g(p)$). Thus, the first order condition at equation (15) simplifies to

$$\alpha K(p)\int_{y_c(p)}^{t} h(y)\mathrm{d}y = (1-\alpha)K(p)\int_{t}^{\bar{y}(p)} h(y)\mathrm{d}y,$$

i.e. equation (14), where the LHS is the marginal decrement of type II errors and the RHS is the marginal increment of type I errors as p increases. Finally, it is easy to see that the optimum is interior, i.e. $p \in (t,1)$ and hence the first order condition is binding. When $p = t$, we have that $\bar{y}(p) = t, y_c(p) < t$, the RHS is 0, and the LHS is positive. When $p = 1$, we have that $y_c(p) = t, \bar{y}(p) > t$, the LHS is 0, and the RHS is positive. It then simply follows from the fact that $\boldsymbol{h} > 0$ and that $\bar{y}(p)$ and $y_c(p)$ are increasing in p that the optimum is unique. In fact, by differentiating the objective function again with respect to p one gets that it is strictly convex in p. It follows that the optimal mechanism coincides with the equilibrium strategy of R because both are determined by exactly the same conditions. □

A.6 Proof of remark 2

Consider the following variant of the optimal mechanism $\boldsymbol{z^*}$ parametrized by $Z \in [t, \bar{y}^*]$. If $z \leq Z$ then $a(y, z) = 0$ while if $z > Z$ then $a(y, z) = 1$ with probability $(1 - G(z^*(y)))/(1 - G(\max\{Z, y\}))$ and $a(y, z) = 0$ with complementary probability. This mechanism gives S the same expected payoff for every type y as $\boldsymbol{z^*}$. To investigate whether R finds it sequentially rational to report z truthfully we only have to check whether $z < Z$ wants to report $z' > Z$ and whether $z > Z$ wants to report $z' < Z$. If $Z = \bar{y}^*$, then clearly type $z = \bar{y}^* - \epsilon$ strictly prefers to report $z' > Z$ for all $\epsilon > 0$ sufficiently small. To see this, suppose that $Z = \bar{y}^*$. Then, by inducing the random action, $z = Z$ gets

$$(1-\alpha)\int_{y_c^*}^{t} \frac{1 - G(z^*(y))}{1 - G(Z)} \frac{h(y)}{H(Z)} \mathrm{d}y + \alpha \int_{t}^{\bar{y}^*} \frac{G(z^*(y)) - G(Z)}{1 - G(Z)} \frac{h(y)}{H(Z)} \mathrm{d}y <$$

$$(1-\alpha)\int_{y_c^*}^{t} \frac{1 - G(z^*(t))}{1 - G(Z)} \frac{h(y)}{H(Z)} \mathrm{d}y + \alpha \int_{t}^{\bar{y}^*} \frac{G(z^*(t)) - G(Z)}{1 - G(Z)} \frac{h(y)}{H(Z)} \mathrm{d}y = \int_{t}^{\bar{y}^*} \frac{h(y)}{H(Z)} \mathrm{d}y,$$

where the last equality follows from equation (8) and the RHS is the loss by inducing the action 0 (i.e. reporting some $z < Z$). If $Z = t$, then clearly type $z = t + \epsilon$ strictly prefers to report $z' < Z$ for all $\epsilon > 0$ sufficiently small. Hence there must be a $Z \in [t, \bar{y}^*]$ such that type $z = Z$ is just indifferent and stronger types prefer to induce action 0 while weaker types prefer to induce the random action. Notice that, when the random action is induced, R has no influence over the randomization, it only depends on y.

A.7 Proof of proposition 3

As described in the proposition, the persuasion rule matches each d with $z^*(d)$ for $d \in [y_c^*, \bar{y}^*]$ (hence type $\bar{y}^*$ to himself) while types $z \in [0, y_c^*) \cup (z^*(y_c^*), 1]$ do not send any signal. Notice that either $[0, y_c^*)$ or $(z^*(y_c^*), 1]$ are empty, or both. In the first case, when there is no signal guilties lie according to the equilibrium lying function and are let go. In the second case, when there is no signal guilties confess. When there is a signal $d \leq t$ types below d confess and are prosecuted while types above d lie covering the interval $[t, z^*(d))$ and are let go. By "covering the interval" we mean that the lying function induces R's belief to be constant over the interval. For signals $d \in (t, \bar{y}^*]$ guilties lie covering the interval $[t, d)$ and are prosecuted. R's actions are always sequentially rational and payoffs are exactly as in the optimal mechanism, in particular no lies are ever caught. Finally, we show that S's strategy above is optimal after any signal d. After signal $d \neq \bar{y}^*$, S's belief that $z = z(d)$ is just

$$\frac{g(z^*(d))}{g(z^*(d)) + \frac{g(d)}{z^{*\prime}(d)}},$$

while for $d = \bar{y}^*$ S believes that $z = \bar{y}^*$ with probability one. Hence, by lying above d his expected payoff is just 0 because $\boldsymbol{z^*}$ satisfies $g(z^*(d))z^{*\prime}(d) = -bg(d)$.

A.8 Proof of remark 3

Given the proof of lemma 1 and proposition 3, in any optimal persuasion rule no signal can induce a lie of S which is detected (more precisely, this has to be the case with probability one). Consider an optimal, possibly random persuasion rule $\boldsymbol{P} : [0,1] \to \Delta D$, where D is a Borel space and $\Delta(\cdot)$ is the set of all Borel probability measures, such that after any signal $d \in D$ in the continuation equilibrium innocents are always honest. Assume by contradiction that this rule satisfies monotonicity, i.e. no type of R has an incentive to deviate to a signal sent by a weaker type.

Consider any signal d and denote by y_d and y^d the supremum of the confessors and the infimum of the innocent types who separate in the continuation equilibrium after signal d respectively (with the convention that $y_d = 0$ if there are no confessors). It must be that y^d is also the supremum of the lies and y_d is the infimum of the liars. This is because innocents are honest, no lies can be detected hence R's action must be constant in any continuation equilibrium after any $m \in [t, y_d)$. We further assume that lies cover $[t, y^d)$ and type y^d sends the signal d (hence y^d is also the minimum of $z \geq t$ who send the signal d). Notice that y_d may or may not send the signal d. In case of R's action is 0 we allow y_d to vary if y_d does not send the signal d, otherwise y_d is the maximum of $z \leq t$ who send the signal d and the minimum of the liars. Notice that S's type y_d knows that the signal was not sent by R's type y_d. Clearly, no $z \in (y_d, y^d)$ ever sends the signal d. For $y \in [0, t)$ let $i(y) \in [t, 1]$ be given by $\alpha(t - y) = (1 - \alpha)(i(y) - t)$. Note that $i(y_c^*) = \bar{y}^*$ and that if $y^d > i(y_d)$ then R must play action 1 after any equilibrium message $m \geq t$ and if $y^d < i(y_d)$ then R must play 0 after any equilibrium message $m < y^d$ and action 1 after any equilibrium message $m > y^d$. In case $y^d = i(y_d)$ then R is indifferent between 0 and 1. We assume that the set of pairs (y_d, y^d) induced by the persuasion rule is a Borel set. We proceed in 4 steps to reach a contradiction. Points (1) and (2) below prove that a monotonic optimal P must induce a deterministic matching (pooling) of types of R, i.e. types $z \in [y_c^*, t)$ and type $i(z) \in (t, \bar{y}^*]$ send the same signal d with probability one and induce a continuation equilibrium characterized by $y_d = z$ and $y^d = i(z)$ (type t can separate). Points (3) and (4) below prove that a positive measure of types $z > \bar{y}^*$ must pool on some signals d with types in $z' \in [t, \bar{y}^*]$ which then contradicts monotonicity because then types in $z'' \in (z', \bar{y}^*]$ would like to send the signal d sent by the weaker pooling type $z > \bar{y}^*$ (and also by the stronger z') and induce a measure of confessors $i^{-1}(z') > i^{-1}(z'')$.

(1) It must be that $y^d \leq i(y_d)$. To see this suppose that there is an d for which $y^d > i(y_d)$. We claim that innocent types in $(i(y_d), y^d)$ obtain the same measure of 0-s and hence this contradicts optimality. Different innocent types $y', y'' \in (i(y_d), y^d)$ may get different measure of 0-s only if there is a signal d' for which $y^{d'} \in (y', y'')$ and $y^{d'} \leq i(y_{d'})$. Namely a signal after which y' can receive 0 and y'' gets 1. But by monotonicity, there is no s' such that $y^{s'} \leq y^d$ and $y_{s'} < y_d$ as otherwise the type inducing the signal s', close to $y^{s'}$ would like to send the signal d induced by the weaker type close to y^d, induce more confession and catch some lies. Hence, S's types in $(i(y_d), y^d)$ can receive 0 only if $y^{d'} > y^d$ and $y^{d'} \leq i(y_{d'})$ and then they all receive 0 with the same probability (possibly depending only on $y_{d'}, y^{d'}$).

(2) Conditional on that $y_d \in [y_c^*, t)$ we have that $y^d = i(y_d)$ with probability one. Given that the induced set of pairs (y_d, y^d) is a Borel set we only have to prove that there can be no $y', y'' \in [y_c^*, t)$ such that whenever $y_d \in (y', y'')$ we have that $y^d \neq i(y_d)$. Suppose by contradiction that such y' and y'' exist. We claim that all types of S in (y', y'') receive the same measure of 1-s which contradicts optimality. Two different types $\tilde{y}, \hat{y} \in (y', y'')$ may get different measure of 1-s only if there is a signal d such that $y_d \in (\tilde{y}, \hat{y})$ and $y^d \geq i(y_d)$. Namely a signal after which $\tilde{y}$ confesses and $\hat{y}$ can get 1 from R. But we know from (1) that if $y_d \in (y', y'')$ we have that $y^d < i(y_d)$ and hence these types can get 1 only if $y_d < y'$ and $y^d = i(y_d)$ in which case all these types get 1 with the same probability (possibly depending only on y_d, y^d).

(3) The following event must have positive probability: (i) S faces signals such that S does not know whether some signal d came from some type $z > \bar{y}^*$ or it came from some type $z' \in [t, \bar{y}^*]$ (in this case we say that z, z' pools on signal d) and (ii) R plays action 0. This is because if with probability one 0 is played only by types of R in $[t, \bar{y}^*]$ then the measure of 0-s allocated to innocent types would be strictly less then it should be in the optimal mechanism. This is because those types $z > \bar{y}^*$ who do not pool can only allocate action 1 with positive probability to innocents as otherwise type $\bar{y}^*$ of S would obtain a payoff of less than 1 which contradicts optimality. Hence, there must be a positive measure of types above $\bar{y}^*$ who pool with types in $[t, \bar{y}^*]$ on some signals d and R plays 0 after denials.

(4) It cannot be the case that this positive measure of types (identified in (3)(i)) all pool with $\bar{y}^*$ because then guilty types who do not confess would have an incentive to lie up to $\bar{y}^*$

and get 1 if not caught (which then would be infinitely more likely than getting caught) and avoid the 0-s. Hence some type $z > \bar{y}^*$ must pool with some type $z' < \bar{y}^*$ on some signal d.

Finally, by (2) it follows that types in $z'' \in (z', \bar{y}^*]$ would like to deviate and send the signal d sent by the weaker type z and induce $y_d = i^{-1}(z') > 0$ confessors ($i(y_d) = z'$) which is strictly more than what they induce in equilibrium as $i(.)$ is strictly decreasing. This contradicts our assumption that the persuasion rule is monotonic.

A.9 Proof of proposition 4

First, we complete the description of the equilibrium. S sends the message m as in the equilibrium at proposition 1 (we suppress superscripts * whenever this does not cause confusion). When S confesses then the game is over, with action $a = 0$ and S gets 0 (there is no need to punish detected false confessions). When $m \geq \bar{y}$, if S is not caught in a lie R lets him go believing that he is surely innocent, while if S is caught in a lie R immediately prosecutes him and punishes him at the level of $-b$ believing that he is surely guilty. Consider now some message $m < \bar{y}$ and the corresponding guilty type $y = \ell^{-1}(m)$. When $z \leq y$ then R immediately prosecutes him and punishes him at the level of $-b$ believing that he is surely guilty.[28] When z is such that $y < z \leq \zeta_m$ then R proves to S that her $z \leq \zeta_m$ and the game proceeds to stage three, in which case R knows that S is guilty if he was caught in a lie and otherwise believes that S is innocent with probability α. Finally, when $\zeta_m < z \leq z^*(y)$ then R prosecutes S and when $z \geq z^*(y)$ then R lets S go. In both cases R believes that S is innocent with probability α. The values of ζ_m are chosen to satisfy $G(m) - G(z^*(m)) = G(\zeta_m) - G(z^*(y))$ and to be in $[m, z^*(y)]$. It is easy to check that the so defined values of ζ_m will indeed fall in the right interval.

Suppose the game proceeds to stage four. When $m' = m$, then R lets S go if he was not caught in a lie, in which case she believes that S is surely innocent, while if S was caught in a lie R prosecutes him and punish him at $-b$. If $m' = 0$ then the game is over, with action $a = 0$ and S gets 0. Finally, in stage three all guilty types send the message 0 and all innocent types send the message m.

Payoffs are exactly as in the optimal mechanism: $m \in [t, \bar{y})$ gets $(1 - G(z^*(y)) + G(\zeta_m) - G(m))/(1 - G(m)) = (1 - G(z^*(m)))/(1 - G(m))$, and $y \in [y_c, t)$ gets $(1 - G(z^*(y)))/(1 - G(y))$.

Given S's strategy, it is clear that R's strategy is optimal. In particular, consider R's disclosure behavior. Strong types ($z \leq \zeta_m$) have no incentives to deviate since they are able to perfectly set guilties and innocents apart by sending ζ_m. Any deviation of weak types (sending

[28] An alternative and equivalent solution would be to punish type $y' < y$ when he is asked to confess in stage three but he confesses that he is some type different from y or he claims that he is y but this lie is detected.

some $\zeta' > \zeta$) is discouraged by S's skeptical, optimistic, belief that the evidence is as weak as possible consistent with the received message (i.e. ζ'_m), so that the continuation of the interrogation would be uninformative anyway. Finally, weak types ($z > \zeta_m$) of course would like to send the signal ζ_m but they cannot.

As for S, given R's strategy, letting $\ell(y) = m$ and $\ell(y') = m'$, we have to consider eight types of possible deviations: (1) a guilty type y behaves as a guilty type (a) $y' < y$ or (b) $y' > y$; (2) a guilty type y behaves as an innocent type (a) $m' < m$ or (b) $m' > m$; (3) an innocent type m behaves as an innocent type (a) $m' < m$ or (b) $m' > m$; (4) an innocent type m behaves as a guilty type (a) $y' < y$ or (b) $y' > y$. All these deviations result in a payoff as if the deviator was claiming to be a different type in the optimal mechanism and hence cannot be profitable.

A.10 Proof of proposition 5

In the uniform case, evaluating the LHS of equation (6) at $m = \bar{y}$ using that $z(\bar{y}) = \bar{y}$ yields that the expected payoff (multiplied by $1 - y$) for a guilty type y from lying is $1 - \bar{y} - (\bar{y} - y)b$, while the one from remaining silent is $1 - Z_s - (Z_s - y)b_s$. A comparison of the two and the fact that $b_s \leq b$ demonstrate that a guilty type, if any, strictly prefers to stay silent if and only if $y < y_s$, where

$$y_s \equiv \frac{(b+1)\bar{y} - (b_s + 1)Z_s}{b - b_s}.^{29} \tag{16}$$

Since there are no confessors in the baseline model, clearly there will not be any confessors for any $Z_s < 1$ either. Thus, either no guilty types are silent or the set of silent types and liars are respectively $Y_s = [0, y_s)$ and $Y_\ell = [y_s, t)$, in which case y_s must be strictly positive. Evaluating equation (16) for the equilibrium value of $\bar{y}$ in example 1 yields that no types are silent if $Z_s \geq \bar{Z}_s \equiv \frac{(1+b)t}{1+b_s(1-\alpha)}$, so that in this case the equilibrium is as in the baseline model. Conversely, if $Z_s < \bar{Z}_s$, y_s and $\bar{y}$ must solve equation (16) and (8) with $y_c = y_s$, yielding $\bar{y}^* = \frac{t(b-b_s)+(1+b_s)Z_s\alpha}{b-b_s(1-\alpha)+\alpha}$ and $y_s^* = \frac{(1+b)t-(1+b_s)Z_s(1-\alpha)}{b-b_s(1-\alpha)+\alpha}$.

Therefore, when $Z_s < \bar{Z}_s$, R's expected loss is proportional to

$$\begin{aligned}E(Z_s) &\equiv (1-\alpha)\int_t^{\bar{y}^*} (1-y)\,\mathrm{d}y + \alpha(1-Z_s)y_s^* \\ &= \frac{(Z_s - t)(1-\alpha)\alpha\,(1+b_s)\,(2(1-t)\,(b-b_s) + (2-t-Z_s)\alpha\,(1+b_s))}{2\,(b+\alpha-(1-\alpha)b_s)^{\,2}} \\ &\quad + \alpha(1-Z_s)\frac{(1+b)t - (1+b_s)Z_s(1-\alpha)}{b - b_s(1-\alpha)+\alpha},\end{aligned}$$

[29] For the sake of precision, if $b_s = b$, then the identity of liars and silent types will not be uniquely pinned down in equilibrium, only their measure. However, payoffs in any other equilibrium are the same as in the one we are selecting.

where the second term represent type II errors on silent types. Since the expression is convex in Z_s with $E'(Z_s)|_{Z_s=t} = -t\alpha < 0$, there exists a $Z_s < \bar{Z}_s$ such that $E(Z_s)$ is strictly lower than for $Z_s \geq \bar{Z}_s$ (i.e. the baseline model) if and only if $E'(Z_s)|_{Z_s=\bar{Z}_s} = \frac{(b-b_s)t\alpha}{b-b_s(1-\alpha)+\alpha} > 0$, which holds if and only if $b > b_s$. Finally, in such case $E(Z_s)|_{Z_s=\bar{Z}_s} - E(Z_s)|_{Z_s=t/(1-\alpha)} = \frac{(b-b_s)^2t^2\alpha^2}{2(1-\alpha)(b-b_s(1-\alpha)+\alpha)^2} > 0$, meaning R prefers the American game to the English game.

A.11 Proof of proposition 6

For any given standard for interrogating $Z_i \in (t, 1]$, in the uniform case, evaluating the LHS of equation (6) at $m = \bar{y}$ using that $z(\bar{y}) = \bar{y}$ yields that the expected payoff (multiplied by $Z_i - y$) for a guilty type y from lying is $(Z_i - \bar{y}) - (\bar{y} - y)b$. Thus, denoting again by y_c the smallest guilty type to lie, we have that $y_c > 0$, i.e. some types confess, if and only if

$$y_c = \frac{\bar{y}(1+b) - Z_i}{b} > 0. \tag{17}$$

Evaluating equation (17) for the equilibrium value of $\bar{y}$ in example 1 when there are no confessors yields that no types confess if $Z_i \geq \bar{Z}_i \equiv \frac{(1+b)t}{1-\alpha} \in (t, 1)$, in which case the equilibrium of the interrogation is as in the baseline model. Conversely, if $Z_i < \bar{Z}_i$, y_c and $\bar{y}$ must solve equation (17) and (8), yielding $\bar{y}^* = \frac{bt+Z_i\alpha}{b+\alpha}$ and $y_c^* = \frac{t+bt-(1-\alpha)Z_i}{b+\alpha}$.

Therefore, when $Z_i < \bar{Z}_i$, R's expected loss is proportional to

$$\begin{aligned}E(Z_i) \equiv& (1-\alpha)\int_t^{\bar{y}^*} (1-y)\,\mathrm{d}y + \alpha t(1-Z_i)\\ =& \frac{(t-Z_i)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2} + \alpha t(1-Z_i),\end{aligned}$$

where the second term is due to the fact that when $z > Z_i$ now R lets S go and hence makes a type II error when facing a guilty type. Since $E'(Z_i)|_{Z_i=\bar{Z}_i} = \frac{bt\alpha}{b+\alpha} > 0$, there always exists a $Z_i < \bar{Z}_i$ such that $E(Z_i)$ is strictly lower than for $Z_i = \bar{Z}_i$. Finally, note that if $Z_i \geq \bar{Z}_i$, then R's overall payoff is the same as in the equilibrium of the baseline model with no confessors or, equivalently, as if she does not interrogate. Indeed, in this case $\bar{y}^* = t/(1-\alpha) < Z_i$, so that when $z < \bar{y}^*$ then R always prosecutes while when $z \geq \bar{y}^*$ she could just always let R go (see the proof of remark 1).

B Online appendix

B.1 Lying and equilibrium updating

In this section, we show how to compute R's belief upon a message in the lying region for an arbitrary measurable lying function $\boldsymbol{\ell}$ and, in particular, that, as in the case in which $\boldsymbol{\ell}$ is strictly increasing and differentiable, this belief is independent from z and hence must be equal to α in equilibrium, i.e. R's indifference condition must hold. We show this result without any reference to belief restrictions, using only the concept of Nash equilibrium. Let $\frac{\mathrm{d}(\boldsymbol{\lambda}\circ\boldsymbol{\ell}^{-1})}{\mathrm{d}\boldsymbol{\lambda}}$ denote the Radon-Nikodym derivative, where $\boldsymbol{\lambda}$ is the adequate dimensional Lebesgue measure (here one-dimensional). Also, define $\boldsymbol{\beta}$ as $\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}\boldsymbol{\lambda}} = \boldsymbol{f}$ (here $\boldsymbol{\lambda}$ is two-dimensional) and let $\boldsymbol{id}$ denote the identity function. Then, for any two-dimensional Lebesgue measurable set A with positive measure, if R is indifferent at the elements in A we must have by Nash equilibrium that the payoffs (integrated over A) from letting S go and to prosecute him are the same, namely

$$(1-\alpha)\int_A \mathrm{d}\boldsymbol{\beta} = \alpha\int_A \mathrm{d}(\boldsymbol{\beta}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id})) = \alpha\int_A \frac{\mathrm{d}(\boldsymbol{\beta}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id}))}{\mathrm{d}\boldsymbol{\beta}}\mathrm{d}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id})$ is the pushforward measure and $\boldsymbol{\ell}^{-1}$ is the inverse correspondence of $\boldsymbol{\ell}$. Thus, we must have ($\boldsymbol{\beta}$−almost surely) that

$$(1-\alpha) = \alpha\frac{\mathrm{d}(\boldsymbol{\beta}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id}))}{\mathrm{d}\boldsymbol{\beta}}.$$

Manipulating the RHS,

$$\frac{\mathrm{d}(\boldsymbol{\beta}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id}))}{\mathrm{d}\boldsymbol{\beta}} = \boldsymbol{f}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id})\frac{\mathrm{d}(\boldsymbol{\lambda}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id}))}{\mathrm{d}\boldsymbol{\beta}} = \frac{\boldsymbol{f}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id})}{\boldsymbol{f}}\frac{\mathrm{d}(\boldsymbol{\lambda}\circ(\boldsymbol{\ell}^{-1},\boldsymbol{id}))}{\mathrm{d}\boldsymbol{\lambda}}$$

which, using our assumptions on $\boldsymbol{f}$ (see the more general sufficient condition at equation (20) in section B.2.1), simplifies to $\frac{\boldsymbol{h}(\boldsymbol{\ell}^{-1})}{\boldsymbol{h}}\frac{\mathrm{d}(\boldsymbol{\lambda}\circ\boldsymbol{\ell}^{-1})}{\mathrm{d}\boldsymbol{\lambda}}$, so that

$$1-\alpha = \alpha\frac{\boldsymbol{h}(\boldsymbol{\ell}^{-1})}{\boldsymbol{h}}\frac{\mathrm{d}(\boldsymbol{\lambda}\circ\boldsymbol{\ell}^{-1})}{\mathrm{d}\boldsymbol{\lambda}}. \tag{18}$$

In particular, when $\boldsymbol{\ell}$ is strictly increasing and differentiable, $\frac{\mathrm{d}(\boldsymbol{\lambda}\circ\boldsymbol{\ell}^{-1})}{\mathrm{d}\boldsymbol{\lambda}} = \frac{1}{\boldsymbol{\ell}'(\boldsymbol{\ell}^{-1})}$, so that equation (7) obtains.

B.2 More general distributions

B.2.1 Sufficient conditions on the prior

Let $\boldsymbol{f}$ be the joint density of y and z over the unit square and let it be 0 if and only if $z \leq y$. To ensure tractability of the model, $\boldsymbol{f}$ must satisfy the two following properties:

(i) there is a function $\boldsymbol{g}$ such that for every y

$$\frac{f(y,z)}{\int_y^1 f(y,z)\mathrm{d}z} = \frac{g(z)}{\int_y^1 g(z)\mathrm{d}z}; \tag{19}$$

(ii) there is a function $\boldsymbol{h}$ such that for every z, $y < t$ and $y' > t$

$$\frac{f(y,z)}{f(y',z)} = \frac{h(y)}{h(y')}. \tag{20}$$

Property (19) ensures that we can calculate R's strategy $\boldsymbol{z}$ so that it satisfies S' indifference condition. Property (20) ensures that R's belief about S's innocence in the lying region does not depend on z (i.e. that equation (18) obtains). Both properties are straightforwardly satisfied when, as assumed in the paper, $f(y,z) = kh(y)g(z)$.

B.2.2 A nonlinear example

As an addition to example 1, we show how to calculate the equilibrium and the optimal mechanism for the case when $h(y) = 1$, i.e. y is still uniform, and $g(z) = 2z$. We fix $\alpha = t = 1/2$ and $b = 1$. The baseline equilibrium values (we suppress the superscript *) of y_c and $\bar{y}$ are given by R's indifference condition $\bar{y} - 1/2 = 1/2 - y_c$ (because $\alpha = t = 1/2$) and by type y_c's indifference condition

$$\int_{\bar{y}}^1 2z\mathrm{d}z - \int_{y_c}^{\bar{y}} 2z\mathrm{d}z = 0,$$

from which $y_c = 2 - \sqrt{3}$ and $\bar{y} = \sqrt{3} - 1$ (alternatively, one solves first for the differential equation (see below) and determines $p, y_c, \bar{y}$ from the three equations (initial condition, terminal condition and R's indifference condition)). By the indifference condition for S for types between y_c and $\bar{y}$

$$\frac{z_p(y)}{y} z_p'(y) = -1,$$

from which $z_p(y) = \sqrt{p - y^2}$. Then, p is determined by any of the following conditions: $z_p(\bar{y}) = \bar{y}$ or $z_p(y_c) = 1$, yielding $p = 8 - 4\sqrt{3}$. The equilibrium lying function is linear (with slope one) and covers the interval $[t, \bar{y})$.

We now show directly that proposition 2 holds. R's loss in proportional to

$$\int_{y_c(p)}^{1/2}\int_{\sqrt{p-y^2}}^{1} 2z\mathrm{d}z\mathrm{d}y + \int_{1/2}^{\bar{y}(p)}\int_{y}^{\sqrt{p-y^2}} 2z\mathrm{d}z\mathrm{d}y,$$

where $y_c(p), \bar{y}(p)$ are determined by $z_p(\bar{y}(p)) = \bar{y}(p)$ and $z_p(y_c(p)) = 1$. The expression is minimized exactly for $p = 8 - 4\sqrt{3}$.

B.2.3 Biased type I errors in equilibrium

We now consider an example in which $f(y,z) \neq kh(y)g(z)$ but the two properties at section B.2.1 still hold. We calculate the optimal mechanism now for the case when y is drawn uniformly and then z is drawn from $(y,1]$ according to the density $g(z) = \frac{2z}{1-y^2}$ and show that the optimal mechanism is different from R's equilibrium decision rule. Notice that the equilibrium is the same as at section B.2.2 because R's indifference condition does not change and the indifference condition of type y of S's is just scaled by the factor $\frac{1}{1-y^2}$. However, the optimal mechanism is different and the equilibrium is biased in the sense that there R makes too many type I errors. We are looking for the p now which minimizes

$$\int_{y_c(p)}^{1/2}\int_{\sqrt{p-y^2}}^{1} \frac{2z}{1-y^2}\mathrm{d}z\mathrm{d}y + \int_{1/2}^{\bar{y}(p)}\int_{y}^{\sqrt{p-y^2}} \frac{2z}{1-y^2}\mathrm{d}z\mathrm{d}y,$$

which happens for $p \approx 1.04 < 8 - 4\sqrt{3} \approx 1.0718$. Clearly, the optimal mechanism shifts downward and to the left relative to the equilibrium because as y increases the density of z increases, i.e. the evidence gets weaker. Hence, under the equilibrium decision rule R makes too many type I errors. It follows that, while the equilibrium of the back and forth game at proposition 4 improves on the baseline model because withdrawn lies are forgiven, it does not implement the optimal mechanism. The same situation would arise if z was drawn uniformly from $(y,1)$ according to $g(z) = 1/(1-y)$ once y has been drawn uniformly.

B.2.4 Implementing the optimal mechanism with upwardly biased type I errors

Continuing the example at section B.2.3, we now show that as long as the optimal mechanism is not too far off from the equilibrium we still can implement the optimal mechanism by the same game described in section 3.3 but by constructing a different equilibrium. The idea is as follows. To reach the optimum, we must produce less type I errors and more guilty types should expect a positive payoff. Hence more guilties should lie. To keep R indifferent it means that these guilties should pool with more innocent types, some of which however, should expect a payoff of 1. We can achieve this by never prosecuting these innocent types but letting them go (together with the associated guilty types with whom they pool) and otherwise R proves that the evidence is sufficiently strong.

Formally, let us denote by $y_c(p) \approx 0.2 < y_c = 2 - \sqrt{3}$ and $\bar{y}(p) \approx 0.72 < \bar{y} = \sqrt{3} - 1$ the values for which $z_p(y_c(p)) = 1$ and $z_p(\bar{y}(p)) = \bar{y}(p)$ for the optimal $p \approx 1.04$. Let S lie above $y_c(p)$ according to a linear lying function $\boldsymbol{j}$ with slope one covering the interval $[1/2, \tilde{y})$, where $\tilde{y} - 1/2 = 1/2 - y_c(p)$ so that R is indifferent (recall that $t = \alpha = 1/2$). Notice that unlike in the equilibrium of proposition 4, where types who get payoff 1 separate immediately, here, to make R indifferent some of these types, namely those in $[\bar{y}(p), \tilde{y})$ will pool first and will separate only later with probability one (or they are let go) because after messages $m \in [\bar{y}(p), \tilde{y})$ R always proves that her evidence is stronger than $z_p(j^{-1}(m))$ and then guilty types confess (see the definition of ζ_m in what follows).

Again, large lies are immediately punished at $-b$. R lets S go when S separates (i.e. after $m > \tilde{y}$) and after message $m \in [1/2, \tilde{y})$ if $z > z_p(j^{-1}(m))$. Note that it should be the case that $z_p(j^{-1}(m)) \geq m$ which is indeed the case as $z_p(j^{-1}(\tilde{y})) = z_p(1/2) \approx 0.8888 > 0.8 \approx \tilde{y}$. R prosecutes S if $z \in [\zeta_m, z_p(j^{-1}(m))$ and proves that her evidence is stronger than ζ_m otherwise. ζ_m is chosen to satisfy the followings: $G(z_p(m)) - G(m) = G(z_p(j^{-1}(m))) - G(\zeta_m)$ for $m \in [1/2, \bar{y}(p))$ and for $m \in [\bar{y}(p), \tilde{y})$ we have that $z_p(j^{-1}(m)) = \zeta_m$. One can easily check (similarly to the proof of proposition 4) that the equilibrium conditions hold. This is because the payoffs are as in the optimal mechanism and any deviation can be identified with a deviation in the optimal mechanism.

To see by direct computation that guilty type $j^{-1}(m)$ has incentive to confess and withdraw his lie we must have that for $m < \bar{y}(p)$

$$G(\zeta_m) - G(m) \leq b(G(m) - G(j^{-1}(m))).$$

This holds because we know that $G(\zeta_m) - G(m) = G(z_p(j^{-1}(m))) - G(z_p(m))$ by construction and $z_p(m)$ is such that

$$1 - G(z_p(m)) - b(G(m) - G(y)) = 1 - G(z_p(y))$$

for any $y \in [y_c(p), \bar{y}(p))$ and $m \in [y, \bar{y}(p))$ and hence for $y = j^{-1}(m)$ as well. For $m \in [\bar{y}(p), \tilde{y})$ we must have that

$$G(z_p(j^{-1}(m))) - G(m) \leq b(G(m) - G(j^{-1}(m)))$$

because by construction $\zeta_m = z_p(j^{-1}(m))$ for such m-s. But this condition holds because we know that

$$1 - G(m) - b(G(m) - G(y)) < 1 - G(z_p(y))$$

for any $y \in [y_c(p), \bar{y}(p))$ for $m > \bar{y}(p)$.

In general, when in the optimal mechanism there are less type I errors than in equilibrium, the above construction works as long as $\tilde{y}$ can be chosen to make R indifferent and so that $z_p(j^{-1}(\tilde{y})) > \tilde{y}$. When instead in the optimal mechanism there are more type I errors than in equilibrium, a similar construction works as long as now one can include sufficiently many guilty types to lie in the first stage (among those who get payoff 0 in the optimal mechanism so that they will not separate and confess immediately in the first stage) - see the demonstration of the idea in section B.3.1. These conditions are satisfied as long as the optimal mechanism is not too far from the equilibrium decision rule. We were unable to find an example where these conditions are not met. But even if one can find such an example, it is still unclear whether these conditions are necessary to implement the corresponding optimal mechanism with the back and forth game that we consider at section 3.3.

B.3 Leniency for confession instead of punishment of lies

In this section, we suppose that S gets $-b$ whenever he is prosecuted and did not confess while he gets 0 if he confesses. R's loss is $(1+b)$ when $-b$ is given to an innocent and it is 0 when $-b$ is given to a guilty. Note that all the statements and constructions of this section directly translate to the "leniency setup" described in the body of the paper, i.e. when an S who is prosecuted gets 0 unless he confesses, in which case he gets $u = b/(1+b)$, and R's loss is 0 when giving u to a confessor and $1-u$ when giving u to an innocent.[30] Importantly, differently from our main model, R faces the constraint that in equilibrium she cannot give leniency (0 instead of $-b$, or u instead of 0 in the "leniency setup") to S unless he confesses. Absent this constraint, the two models become *equivalent* and the same equilibrium construction of the back and forth game at section 3.3 always implements the optimum. After highlighting the main differences with respect our baseline setup, we explain how, despite such constraint, R can improve her payoff using a slight variant of the equilibrium construction of the back and forth game used for proposition 4 (section B.3.1) and even attain the optimum if additionally innocents depart

[30] The only difference is that in such leniency setup the equilibrium cut-offs remain the same as in our baseline equilibrium while the optimal mechanism shifts away from the equilibrium and hence from the optimal mechanism at section 3.1. It is because in the optimal mechanism R uses only actions 1 and u instead of 1 and 0 and, differently from our baseline setup, R's optimality condition does not translate into R's average indifference condition in equilibrium.

from honesty (section B.3.2).

Even though lemma 1 holds under such payoffs, proposition 2 does not hold because: (1) R's indifferent condition in equilibrium now changes to

$$(1-\alpha)\mu(1+b) = \alpha(1-\mu),$$

where μ is her belief that S is innocent; (2) the equilibrium cut-off decision rule of R must be constant at $z(m) = \bar{y}$ as lies are not punished relative to prosecution; (3) the optimal mechanism remains unchanged. Fixing the parameters to $t = \alpha = 1/2$ and $b = 1$, in the uniform case in equilibrium now $y_c = 1/4$, $\bar{y} = 5/8$, $\boldsymbol{\ell}' = 1/2$ and $\boldsymbol{z} = 5/8$, with $\mu = 1/3 < \alpha$. In equilibrium R makes 1/64 type I errors and 3/32 type II errors while in our baseline equilibrium these are 1/36 and 1/24, respectively. In both cases, in the optimal mechanism these are 1/36 and 1/72, respectively. Henceforth, we report all losses without multiplying them by α and by 2 (the density).

B.3.1 Further improvement on payoffs with downwardly biased type I errors

Consider the following equilibrium of the multistage game. First, S communicates as in equilibrium. After separating messages R immediately makes a correct decision, i.e. gives 0 to confessors and 1 to innocents. Also R gives $-b$ immediately after lies which should not happen in equilibrium, lets S go when $z > \tilde{z}(m)$, prosecutes S if $z \in [\zeta_m, \tilde{z}(m)]$ and S gets $-b$ (notice that R is indeed indifferent because her belief about S's innocence is exactly $\mu = 1/3$), and proves that his evidence is stronger than ζ_m otherwise.

Now ζ_m and $\tilde{z}(m)$ are chosen to satisfy (1) $\zeta_m - m = m - \ell^{-1}(m)$ and (2) $\tilde{z}(m) - \zeta_m = 5/8 - m$. These choices of cut-offs ensure that all types get the same payoff as in equilibrium and that guilty types confess after learning that the evidence is stronger than ζ_m (in fact they are just indifferent by (1)). As a consequence, the strategies above indeed constitute an equilibrium because any deviation is equivalent to a deviation in the equilibrium of the one-shot game which is not profitable. The resulting type I and type II errors are 1/64 and 3/64, respectively.

In equilibrium some guilty types still get $-b$ which is inefficient, moreover, there are still too few type I errors and too many type II errors relative to the optimal mechanism. We now show how to reach the optimal level of type I errors and hence the only source of inefficiency which will remain is that still some guilty types get $-b$. The reason why guilty types must get $-b$ is that innocents cannot ever get 0 but must get $-b$ when they are prosecuted when pooled with a guilty, and hence guilty types must also get $-b$. The fact that innocent types may get $-b$

(resulting in a loss of 2) causes no inefficiencies in itself because this loss is exactly compensated by the smaller probability of making these mistakes in the following equilibrium.

One could further improve R's payoff by playing an equilibrium of the multistage game similar to the one described in subsection B.2.4. However, the situation is the reverse now. In section B.2.4, the optimal mechanism shifted downward and to the left relative to the equilibrium and hence when implementing the optimum we needed to involve more pooling innocent types (some of those who get 1 under the optimal mechanism) to maintain the indifference of R. Now we would like to produce more type I errors (as compared to the one-shot equilibrium or to the one above) and hence we need more guilty types (some of those who get 0 in the optimal mechanism) to lie so as to keep R indifferent. The same technique is much simpler when 0 could be given to someone who is prosecuted. The exact construction is as follows.

Recall that in the optimal mechanism types below 1/3 get 0 and types above 2/3 get 1. Hence, let now S types in $[1/6, 1/2)$ lie and cover the interval $[1/2, 2/3)$ according to a lying function with slope 1/2. This ensures the indifference of R whenever she chooses prosecution $(-b)$ or lets S go. Confessors get 0, separating innocents are let go and non-equilibrium liars get prosecuted and get $-b$ immediately when caught. R lets S go when $z \in [\tilde{z}(m), 1]$ prosecutes when $z \in [\zeta_m, \tilde{z}(m))$ and reveals that her evidence is stronger than ζ_m otherwise. For messages $m \in [1/2, 7/12)$ (which are sent by types in $[1/6, 1/3)$ who all should expect 0) $\tilde{z}(m) = 1/3 + m$ and $\zeta_m = 2m - 1/3$. For messages $m \in [7/12, 2/3)$ we set $\tilde{z}(m) = 3/2 - m$ and $\zeta_m = 5/6$. One can check that all types of S gets the same payoff as in the optimal mechanism and that S's type sending message $m \geq 7/12$, observing that R's evidence is stronger that ζ_m is just indifferent between confessing or sticking to the story m (types sending $m < 7/12$ strictly prefer to confess after observing ζ_m). This proves that the construction is indeed an equilibrium resulting in 1/36 of type I error (as in the optimal mechanism) and 1/24 type II errors. These errors are exactly as in the one-shot equilibrium of the baseline model where lies are punished relative to prosecution. Type II errors cannot be further decreased without further increasing type I errors, which in principle still could be an improvement. However, the optimal payoffs cannot be reached. Therefore, under these payoff specifications and the constraint that R cannot offer leniency unless S confesses the optimal mechanism cannot be implemented with our game in section 3.3, at least under the restriction that innocents are honest.

B.3.2 Implementation of the optimal mechanism with confessions of innocents

Assuming now that innocents are not always honest, we explain how to implement the optimum. Consider the equilibrium construction described in the proof of proposition 4 with the difference that when $z > \zeta_m$ the game continues and R invites each type of S to confess or stick to his story. Now if S confesses R plays 0 or 1 according to the cut-off described in the proof and if (off the equilibrium path) S sticks to his story R gives S a lower payoff (together with the appropriate belief that makes this behavior sequentially rational). Then, the optimal mechanism is implemented.

Designing Interrogations

Alessandro Ispano   Péter Vida

December 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 the law enforcer's evidence, which is informative about the suspect's status and may also disprove lies. We study the evidence strength standards for interrogating and drawing adverse inferences from silence that minimize prosecution errors. We consider the law enforcer's incentives to confront the suspect with the evidence, both at once and gradually. We describe the optimal mechanism under full commitment and a "back and forth" interrogation with discretionary punishment of lies that implements the optimum in equilibrium.

Keywords: lie, evidence, questioning, confession, law, prosecution, disclosure
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, Sidartha Gordon, Jeanne Hagenbach, Frédéric Koessler, Laurent Linnemer, Lucie Menager, Marco Ottaviani, Marcus Pivato, Régis Renault, Karl Schlag, Johannes Schneider, Joel Sobel, Egor Starkov as well as seminar participants at the THEMA theory meetings, Workshop on Signaling in Markets Auctions and Games, University of Copenhagen, Paris Applied Theory Day, Paris II University, Paris School of Economics, Annual Conference of the European Association of Law and Economics in Barcelona, European Summer Meeting of the Econometric Society in Copenhagen, and Hungarian Academy of Science for useful comments. Financial support from Labex MME-DII and ANR grant 19-CE26-0010-03 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. Comparable situations arise in the assessment of an employee's misconduct, the investigation of a student's fraud, the determination of a spouse's betrayal, etc. With law enforcement as leading application, this paper develops a theoretical framework that describes how interrogations unfold. It then explores several questions on the design, conduct and regulation of interrogations to determine which institutions enhance information revelation and yield to more accurate 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 is privately informed about his status as guilty or innocent and the likely strength of the evidence against him. The law enforcer is privately informed about the actual evidence.

In spite of the potential complexity resulting from the correlation between the private information of the two parties, we provide a handy information structure.[2] To fix ideas, consider the following stories, in which y and z represent the private information, or "type", of the suspect and the law enforcer, respectively, and the suspect is guilty when $y < t$, where t is known:

- Due to a restraining order, a husband's distance y to the house of his wife cannot be less than t. A caring neighbor living at distance z spotted the husband from the window;
- A jeweler is selling gold rings whose purity y allegedly falls short of the declared purity t. A forensic test provides an upper bound z on the purity level;
- A telephone operator left work at time y, allegedly before the time of the end of his shift t. An unanswered call occurred at time z, proving he had left by then.

[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. From a normative perspective, it can be thought of as reflecting the social costs of each error. Throughout, we abstract from details about the separation of roles in the legal system and assume the law enforcer directly takes prosecution decisions. The model can equivalently apply to any other decision that the law enforcer would want to base on the suspect's guilt, e.g. an arrest, and generates disutility to the suspect irrespectively.

[2] For instance, in the related context of plea bargaining, Reinganum (1988) writes:

> A more difficult task is to incorporate the discovery process. One way to do this is to assume that the defendant receives a signal which is (imperfectly) correlated with the strength of the case. If the prosecutor also observes this signal, then this is basically an exercise in updating priors. [...] If the signal is private information for the defendant, matters could become considerably more complicated.

In our baseline model, the suspect sends a single message (m), which must be interpreted as a claim about his type, if any, in reply to the law enforcer's inquiry, and the law enforcer makes a decision. The suspect additionally incurs some costs if he is caught in a lie ($m > z$), or if he stays silent and the evidence is unambiguously incriminating ($z < t$). Under an equilibrium refinement that gives prominence to honesty adapted from Hart et al. (2017), innocent types and confessors (types whose $m < t$), if any, claim the truth (proposition 0). Conversely, some sufficiently high, unsuspicious, guilty types necessarily lie and mimic low innocent types. Sufficiently high innocent types separate, instead, so that high, strong, claims of denial are credible. Clear predictions then obtain on players' equilibrium strategies (proposition 1) and payoffs (corollary 1).

This baseline model already sheds light on some fundamental aspects of interrogations, starting from their purpose in relation to the right to silence and the elicitation of a confession. 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, there is a longstanding debate on whether an adverse inference, i.e. a conclusion pointing at the suspect's guilt, can be drawn (see for instance Seidmann and Stein (2000), Seidmann (2005), and the discussion between O'Reilly (1994) and Ingraham (1995)). In our model, these differences are parametrized by the evidence strength standard required to prosecute the suspect when he stays silent, which in equilibrium is equivalent to a standard for prosecuting in the absence of a confession (remark 1). A basic purpose of interrogating is hence the possibility to elicit a confession from a suspect who otherwise would sometimes necessarily be let go. And conversely, when prosecution is always possible, interrogating is useful if and only if some types indeed confess (remark 2). Still, the optimal standard for prosecuting (proposition 2) is not always minimal. Indeed, anticipating that by staying silent he will sometimes be let go, the suspect has less need to lie, which enhances information transmission. As a result, not only guilties and innocents alike benefit, as Seidmann (2005) points out, but also the accuracy of the law enforcer's decisions sometimes increases. Thus, while in accordance with the logic behind adverse inferences a silent suspect is necessarily guilty in our setting, our results provide a purely information based justification for adverse inferences being insufficient to trigger prosecution without additional supporting evidence.[3]

As in the case of other restraints of individual freedom such as searches and arrests, the

[3] 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."

law enforcer may be required to hold sufficiently strong evidence to interrogate the suspect in the first place. When the suspect is interrogated, he then knows that the evidence must be sufficiently strong. Under a more stringent standard for interrogating, a guilty suspect is less inclined to lie and more inclined to confess, so that the interrogation becomes more informative. Therefore, even assuming that the suspect must be let go when the standard is not met, the optimal standard for interrogating (proposition 3) is often not minimal. Moreover, it is never optimal to set the standard for interrogating so that the suspect is ever prosecuted without being interrogated (remark 3).

We then enrich the baseline model by supposing that the law enforcer can directly confront the suspect with the evidence before he makes a claim. Uncertainty about what the law enforcer knows is crucial. Indeed, if the suspect perfectly observes the evidence the interrogation becomes uninformative (proposition 4). Therefore, mandatory disclosure requirements about the evidence, which are typically in place further down the prosecution process, would be detrimental if extended to interrogations. As for voluntary disclosure, when the law enforcer can only reveal the exact evidence or nothing there are both equilibria with no evidence revelation, so that the interrogation unfolds as in the baseline model, and equilibria with full evidence revelation, so that the interrogation is uninformative (corollary 2). The latter are sustained by the suspect's skeptical, optimistic, belief that undisclosed evidence is the weakest, i.e. the notorious "You have nothing on me". When partial disclosure is also possible, instead, and the law enforcer can understate the strength of her evidence, there exist partially revealing equilibria in which types with sufficiently strong evidence pool (corollary 3). This type of equilibrium can replicate the outcome under the optimal standard for interrogating, making such standard redundant and sometimes improving the informativeness of the interrogation relative to the baseline model.

The law enforcer can do even better by means of gradual evidence revelation, so as to intertemporally screen suspects and benefit from each and every piece of her information. We present a dynamic interrogation policy in which, by revealing that the evidence is stronger and stronger as time passes, the law enforcer is able to invite more and more confessions from guilties and credible denials from innocents. We calculate the optimal interrogation policy of given length and then show that the law enforcer always prefers longer interrogations (proposition 5).

Overall, the design instruments we considered can be effective because they serve as partial commitment devices for the law enforcer's prosecution decisions, which must always be sequentially rational in equilibrium. 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.[4] A comparison of the optimal mechanism (proposition 6) and the equilibrium of the baseline model clarifies the exact nature of the law enforcer's commitment problem. The law enforcer would benefit from committing to sometimes prosecute some innocents and to let some guilties go. And while the possibility of catching and punishing lies makes interrogations informative in the first place, actually punishing lies is inefficient. Also, as the dynamic interrogation policy described above gets infinitely long, payoffs converge to those under the optimal mechanism.

The optimal mechanism can be implemented in equilibrium of a two-round variation of the baseline model (proposition 7). The law enforcer must be able to reveal information about her evidence in reaction to the suspect's claim, who can then reply back, and she must have discretion on whether to punish lies. When the evidence is sufficiently strong relative to the suspect's claim in the first round, the law enforcer proves it rather than immediately taking a decision. In the second round, a guilty suspect will step back on his lie, which will be forgiven, and an innocent type will stick to his story. The equilibrium behavior 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. However, while those models require the court to sometimes convict a suspect known to be surely innocent, in our game the law enforcer's decisions are sequentially rational at each information set. In particular, in the second round an innocent suspect is always let go. Since the equilibrium exhibits an implicit promise of leniency for confession, our results also support the view that these kinds of agreements, on which the law is often blurry or controversial, should not be disallowed.[5]

The paper is structured as follows. After a discussion of the related literature here below, section 2 presents the baseline model and its equilibria. Section 3 explores its implications. Section 4 considers evidence revelation. Section 5 presents the optimal mechanism and its game implementation. Section 6 concludes by considering extensions and additional questions, namely properties and interpretations of the information structure, alternative moves and payoffs (more severe punishment for perjury, no punishment but leniency for confession, continuous law enforcer's actions), asymmetric information about the law and deceptive tactics, and regulation of evidence revelation. The appendix contains additional material (section A and C) and all proofs (section B).

[4] See 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.

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

Relation to the literature. While the entire judicial process is a prominent field of application of information economics, suspects' interrogations have received only limited attention. A notable exception is Baliga and Ely (2016), who study the interrogator's commitment problems inherent to torture. Most closely related is Seidmann (2005), who considers how adverse inferences affect the suspect's strategy in the interrogation.[6] We offer new results on the issue, which we compare to those of Seidmann (2005) in detail in section 6.2, and we consider several other design aspects. Otherwise, the law and economics literature generally studies 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. Hence, capturing essential features of the judicial process only in stylized form, our framework accounts for important specificities that arise in interrogations due to the information structure and the nature of communication.[7] 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.[8] Thus, taking two-sided asymmetric information between the suspect and the law enforcer a step further than previous work, our model allows for heterogeneity not only between a guilty and an innocent, 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. Besides, 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.[9] 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 theoretical models of lying (e.g. Kartik (2009), Dziuda and Salas (2018), Balbuzanov (2019) and

[6] Leshem (2010) extends the analysis to a setting in which even innocent suspects may prefer to exercise the right to silent as their honest claims may be disproved.

[7] In particular, we take as given that guilt and reticence entail some punishment without considering the complex determinants of plea bargaining (Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezzetti, 2001) 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.

[8] 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). Cuellar (2020) studies plea bargaining outcomes when the prosecutor can acquire and disclose evidence over time.

[9] See for instance Shin (1994), Mialon (2005), Bhattacharya and Mukherjee (2013) and Hart et al. (2017).

Jehiel (2021)), the detectability of a lie and its cost for the suspect derive explicitly from the law enforcer's private information, which in particular naturally implies that the detectability of a lie increases with its size.[10] Moreover, our framework therefore allows studying the effects of revelation of the law enforcer's private information to the suspect. 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.[11] 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 disproved by the law enforcer's evidence, the interrogation would be completely uninformative.

2 A simple model of interrogations

2.1 Model

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 and R privately observes her evidence z, drawn uniformly from $\left\{(y,z) \in [0,1]^2 : y < z\right\}$. S is **guilty** when $y < t$ and **innocent** when $y \geq t$, where $t \in (0,1)$ is a commonly known parameter.

R's evidence is a signal about S's type proving that $y < z$, the lower the z the stronger the evidence since S's guilt becomes more likely. In particular, when $z \leq t$ we say that the evidence is **conclusive** since it proves that S is surely guilty (see figure 1). Likewise, in addition to his status as guilty or innocent, S's type determines the strength of the evidence he expects R to possess, which is stronger the lower the y as he knows that $z > y$. While ensuring tractability, the information structure respects natural properties, which we discuss in section 6.1.

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

[10] 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. Jehiel (2021) considers a repeated communication setting in which a sender who lies then forgets his message and may hence later contradict himself. Relatedly, in Ioannidis et al. (2020) the message of the sender determines the receiver's costly investigation technology. Perez-Richet and Skreta (2020) consider general cost functions for the sender from manipulating a test that the receiver designs. Also, see Sobel (2020) for a general definition of lying.

[11] 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.

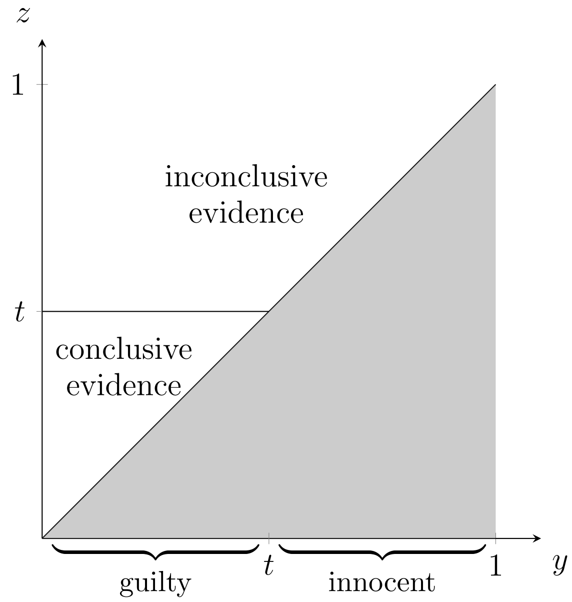


Figure 1 The sample space, the suspect's status and the evidence

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 that 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 be prosecuted, i.e. $a = 0$, or let go freely without charges, i.e. $a = 1$.

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

$$\alpha\, a \mathbb{1}_{y<t} + (1-\alpha)\,(1-a)\,\mathbb{1}_{y\geq t}, \tag{1}$$

where $\mathbb{1}_{y<t}$ and $\mathbb{1}_{y\geq t}$ are indicator functions for S's status as guilty and innocent, respectively, and $\alpha \in (0,1)$ a commonly known parameter. S's payoff is

$$a - b\mathbb{1}_{m\geq z \,\text{or}\, (m=s \,\text{and}\, z\leq t)}, \tag{2}$$

where $b > 0$ is a commonly known parameter and $\mathbb{1}_{m\geq z \,\text{or}\, (m=s \,\text{and}\, z\leq t)}$ the indicator function for when S is caught in a lie or when he is silent but the evidence is conclusive.

R aims at prosecuting a guilty and letting an innocent go and α measures the relative importance of a type II error over a type I error. S aims at being let go but also incurs some cost b for being caught in a lie, or for refusing to answer R's question when the evidence is already unambiguously incriminating. This **punishment** for **reticence** may be direct and explicit, e.g. a penalty for lying, or be due to R taking a more unfavorable action when S is proving uncooperative. Section C.1 in the appendix formalizes the latter interpretation. In section 6.2, we allow S's cost when he is silent but the evidence is conclusive to be lower (possibly zero) than when he is caught in a lie. We also consider the alternative but closely related incentive structure in which S bears no costs but enjoys some leniency for confessing.

Equilibrium selection. Throughout, we restrict our attention to equilibria in which innocent types and (guilty) confessors are honest. In section C of the appendix we derive this restriction as a result in a more general game using an extremely weak, possibly-mixed, equilibrium concept and requiring only a very weak form of truth-leaning adapted from Hart et al. (2017) (proposition C.0 in the appendix contains a more formal and complete statement).

Proposition 0 (Honesty of innocents and confessors)**.** *Under a truth-leaning refinement, in equilibrium innocent types and confessor are honest.*

Thus, R always finds it optimal to prosecute when S confesses and/or S is caught in a lie (and/or the evidence is conclusive). Likewise, since a silent type is necessarily guilty, R always finds it optimal to prosecute upon silence. And anticipating R's action, no guilty type finds it optimal to stay silent (even if there was no punishment for doing so). However, we consider a *more general* version of the model in which when S is silent, R must necessarily let S go if $z > Z_s$, where $Z_s \in (t, 1]$ is a commonly known parameter. Again by proposition 0, which holds for any arbitrary action policy of R upon silence, R therefore chooses

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

where $a(s, z)$ denote R's action upon $m = s$ and evidence z. Z_s can be interpreted as the evidence strength standard required to prosecute S, inversely related to the level of **protection of silence**. If $Z_s = 1$, prosecution is always possible, while in the limit case $Z_s = t$ prosecution is possible only with conclusive evidence. If $Z_s = t/(1-\alpha) < 1$, R cannot use the informational content of silence and hence takes the *optimal* decision based on the evidence alone, a situation Seidmann (2005) refers to as the "American game" to contrast it to the "English game" ($Z_s = 1$).

2.2 Equilibrium

Because of proposition 0, a pure strategy of S is fully described by a partition of the set of guilty types into the sets Y_c, Y_s and Y_ℓ, 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. Likewise, we can concentrate on R's actions at information sets in which S denies and he is not caught in a lie, i.e. in the set $\{(m, z) | m \in [t, 1], z > m\}$, since in all other instances R's sequentially rational behavior follows straightforwardly from proposition 0. A pure strategy $\boldsymbol{a}$ of R 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 $\boldsymbol{\mu}$ specifies a probability $\mu(m, z) = \mathbb{P}(y \geq t | m, z)$ that S is innocent. Without loss of generality, to ease exposition we focus on pure strategies only.[12] 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_ℓ 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.[13]

The following proposition describes an equilibrium and some properties of all equilibria (see proposition B.1 in the appendix for a more formal and complete version of the statement).

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

(i) low guilty types confess, middle guilty types are silent and high guilty types lie (only the interval of liars is always non-empty);

(ii) liars lie according to an increasing lying function and pool low innocent types, while sufficiently high innocent types separate;

(iii) upon a pooling message and not catching S in a lie R is indifferent and lets S go if and only if the evidence is sufficiently weak, where higher messages require weaker evidence;

(iv) all guilty types who do not confess are indifferent with respect to any of their equilibrium messages.

Moreover, any other equilibrium has the same interval of confessors and lies sent, measure of liars and silent types, and expected action $A(m)$ of R upon each message.

[12] All of our results hold after appropriately accommodating for mixed strategies.

[13] 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.

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. In the equilibrium singled out in the proposition the set of liars is the interval containing the highest guilties, the lying function is increasing, and R's action policy takes a 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 the punishment for reticence b is higher, when R is tougher as measured by a higher weight α she attaches to type II errors, when the prior likelihood of innocence is lower as measured by a higher t 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. A higher punishment 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

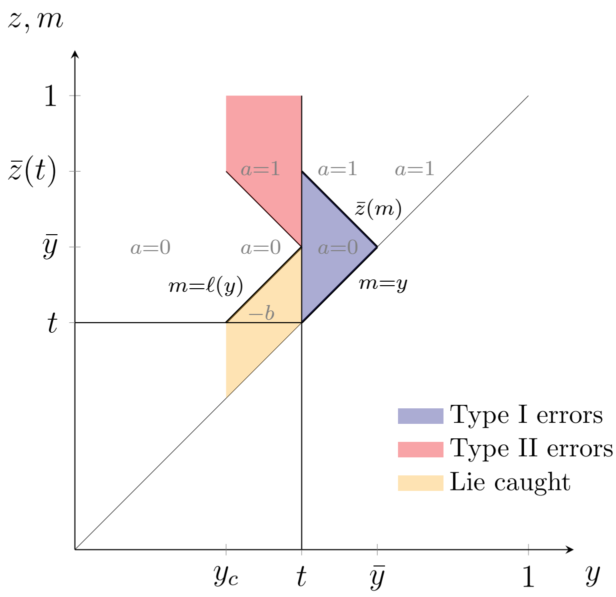


Figure 2 Equilibrium payoffs
($t = 1/2$, $b = 1$, $\alpha = 1/2$, $Z_s \geq 5/6$)

the same in every equilibrium. In particular, R's ex-ante expected loss is[14]

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

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.*

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 parameter configuration chosen. Separating guilty types, i.e. types below y_c, get $a = 0$ and separating innocent types, i.e. types above $\bar{y}$, 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.

[14] For the sake of precision, the expression should be multiplied by two since the joint density of y and z is $f(y, z) = 2$. Throughout, we ignore this scaling factor when computing expected payoffs.

3 Implications

The model directly yields two implications for the purpose of interrogations and, in particular, the idea that the possibility of eliciting a confession plays a central role.

Remark 1 (Protection of silence as a standard for prosecuting)**.** *In the equilibrium at proposition 1, if S does not confess he is always let go whenever the evidence is weaker than the standard Z_s required to prosecute a silent type. Therefore, Z_s can be equivalently thought of as the standard required to prosecute S unless he confesses.*

The intuition behind this remark is that R's action policy upon a denying message not contradicted by the evidence is always weakly more lenient than upon silence, i.e. $\bar{z}(m) \leq Z_s$. Also, no guilty type ever sends a lie that would be caught by evidence which is weaker than the standard, i.e. it must be that $\bar{y} \leq Z_s$. Hence, by interrogating, R has a chance to obtain a confession and prosecute some guilty types that would otherwise necessarily be let go when the evidence is too weak ($z > Z_s$). Additionally, when the evidence is sufficiently strong to prosecute but inconclusive, R may be able to tell a guilty and an innocent apart, e.g. because S is caught in a lie. However, if prosecution is always possible ($Z_s = 1$), so that no type elects to stay silent, this informational benefit is still indissolubly related to the possibility of obtaining a confession.

Remark 2 (Informational benefit of confession)**.** *Suppose that prosecution is always possible ($Z_s = 1$). R's expected payoff from interrogating is strictly higher than when she relies only on the evidence to take a decision if and only if some types confess.*

Somehow more surprisingly, a more stringent standard for prosecuting, or equivalently a higher level of protection of silence, sometimes increases the accuracy of prosecution decisions.

Proposition 2 (Optimal standard for prosecuting)**.** *Let $Z_s^\star$ denote R's optimal standard for prosecuting (i.e. unless S confesses, R can prosecute if and only if $z \leq Z_s^\star$).*

- *If no types confess when prosecution is always possible ($Z_s = 1$), then the optimal standard for prosecuting is more stringent ($Z_s^\star < 1$) and such that some types stay silent.*
- *If instead some types confess when prosecution is always possible ($Z_s = 1$):*
  - *having no standard for prosecution is optimal ($Z_s^\star$=1) if t, α and b are large;*[15]

[15] 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 α.

- *if the optimal standard for prosecuting is more stringent ($Z_s^\star < 1$) and such that some types stay silent, then it is also such that no types confess.*

A binding standard for prosecuting, 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 without any standard all guilty types lie, the interrogation is uninformative (see remark 2). A binding standard is then always optimal since the loss introduced on silent types is initially negligible relative to the benefits of increased separation. Instead, a more stringent standard has less clear benefits for R when it also discourages some guilties 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 standard is large and type II errors are rather costly for R, i.e. if t, α and b are large, R's expected loss is always increasing as the standard gets more stringent. Otherwise, R's expected loss is non-monotone and a binding standard can be optimal if it is stringent enough so that only liars and silent types, but no confessors, remain. While the proposition is expressed in terms of R's welfare, it also implies that imposing a binding standard is sometimes unambiguously Pareto improving since S always favors a lower Z_s.

Finally, we consider the effects of a standard for *interrogating*, denoted by Z_i, i.e. R can only interrogate if $z \leq Z_i$. Since Z_i is observable by law, when S is interrogated he knows R's evidence meets the standard. The equilibrium analysis of our baseline model, which corresponds to $Z_i = 1$, easily generalizes to describe the outcome of the interrogation under any standard $Z_i \in (t, 1]$.[16] For the moment, we suppose the standard for prosecuting (Z_s) is also fixed at Z_i, so that when $z > Z_i$ then R must let S go. A more stringent standard for interrogating 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 for interrogating)**.** *Let $Z_i^\star$ denote R's optimal standard for interrogating (i.e. she interrogates if and only if $z \leq Z_i^\star$ and otherwise she lets S go).*

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

- *If no types confess when interrogating is always possible ($Z_i = 1$), then the optimal standard for interrogating is more stringent ($Z_i^\star < 1$).*
- *If instead some types confess when interrogating is always possible ($Z_i = 1$), then:*
  - *the optimal standard for interrogating is more stringent ($Z_i^\star < 1$) if and only if t, b and α are small;*[17]
  - *in particular, having no standard for interrogating is optimal ($Z_i^\star = 1$) if some types always confess for any b, i.e. if $\alpha \geq 1 - t$.*

The general message of the proposition is that a stringent standard for interrogating 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 low weight α R attaches to type II errors and a lower prior that S is guilty (a lower t).

While we have assumed that $Z_s = Z_i$, i.e. that the standard for prosecuting coincides with the standard for interrogating, there is no gain for R from setting a weaker standard for prosecuting than for interrogating, i.e. to sometimes prosecute without interrogation.

Remark 3 (No prosecution without interrogation)**.** *It is never optimal for R to set the standards for prosecuting and interrogating in such a way that she sometimes prosecutes without interrogating (i.e. such that $Z_i < Z_s$). In particular, under the optimal standard for interrogating, whenever R does not interrogate she indeed finds it sequentially rational not to prosecute S.*

The ultimate difference between a standard for prosecuting and a standard for interrogating is that under the latter S learns that the standard is met. In the next section, we study what happens when there is no standard in place but R can directly disclose information about the evidence to S. From now on, unless stated otherwise, we assume that $Z_s = 1$ (and also $Z_i = 1$), i.e. silence is given no protection, or equivalently prosecution is always possible, so that no type is silent in equilibrium.

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

Assumption (NS)**.** *Henceforth, silence is given no protection or, equivalently, prosecution is always possible ($Z_s = 1$), so that without loss of generality we can restrict S's message space to $\mathcal{M} = [0, 1]$.*

4 Confronting the suspect with the evidence

4.1 One-shot revelation

Common intuition suggests that R should not give away to S what she knows, at least not exactly and not immediately. The next proposition formalizes this argument.

Proposition 4 (Public evidence)**.** *Consider the baseline model but suppose R's evidence z is observed by S as well. Then, an equilibrium exists and any equilibrium is uninformative, i.e. the expected payoff of each type z of R is the same as when R makes decisions relying on z only.*

Once S knows z, unless R holds conclusive evidence, in which case she always prosecutes S regardless of the game being played, innocents cannot be set apart from guilties since these know how to tailor their lies to ensure they are not caught.

We now consider R's equilibrium incentives to voluntarily reveal her information to S, depending also on her available messages. Throughout, we maintain that R cannot make false claims, due to the risk of legal action or the inadmissibility of the interrogation in court. Equivalently, S only believes claims backed up by physical proof.

Consider the baseline game and assume that, before S sends his message, R can choose whether to disclose her exact type z or not to do so.

Corollary 2 (Exact disclosure)**.** *When R can disclose z or nothing, there is an equilibrium in which she never discloses and the continuation is as in the baseline model. There is another equilibrium in which she always discloses and the continuation is as at proposition 4.*

The first part of the corollary is a direct consequence of proposition 4 since, provided S expects no disclosure, each type of R holding inconclusive evidence is weakly harmed by deviating and revealing z. The second part is reminiscent of the familiar unraveling argument in persuasion games (Milgrom, 1981). R's skeptical off the equilibrium path belief that undisclosed evidence is the weakest, i.e. $z = 1$, makes disclosure weakly optimal for each type.

While these observations may suggest that the possibility to reveal information to S is unhelpful or detrimental for R, this need not be the case when partial revelation is possible. Suppose now R can disclose garbled information about her type z in the sense that she can prove that her evidence is stronger than ζ for any $\zeta \geq z$. Alternatively, z can disclose any degraded evidence $\zeta \geq z$. Technically, we assume that any type z of R sends a signal $\zeta \in [z, 1]$ to S, where signal $z = 1$ is equivalent to nondisclosure.

Corollary 3 (Partial disclosure)**.** *When R can disclose that her evidence is stronger than ζ for any $\zeta \geq z$, for any $Z > t$ there is an equilibrium in which types $z \leq Z$ pool on the signal Z and each type $z > Z$ separates and discloses her type by sending the signal z.*

The pool does not unravel because even a type with inconclusive evidence stronger than Z has no incentives to reveal herself. Upon a deviation to some $\zeta < Z$, S would again skeptically think R's type is the weakest who could have possibly sent ζ, i.e. type ζ. Combining corollary 3 with proposition 3 and remark 3, it follows that R can replicate the effect of a standard for interrogating and sometimes strictly improve her payoff relative to the baseline model.

4.2 Gradual revelation

The main drawback for R of revealing in one-shot that her evidence is sufficiently strong is that when her evidence is weak she obtains no information from S. Revealing less, i.e. that her evidence is stronger than a weaker level, would still suffice to set some types apart. Indeed, some low guilties may still be willing to confess and some sufficiently high innocents may still credibly separate by means of strong claims of denial that no guilty is willing to make. We now describe a dynamic interrogation in which R attenuates this problem thanks to gradual information revelation and inter-temporal screening. As time passes, R proves that her evidence is stronger and stronger provided she can, and otherwise she stops the interrogation and makes a decision. The more information R discloses, the higher the guilty type she can successfully invite to confess and the lower the innocent type she can invite to credibly deny. Once R has reached the end of the screening phase and revealed the maximal desired amount of information but S has not identified himself with any of the invited types, then S must make a final claim about his type and R makes a decision.

Formally, R commits to an interrogation policy $(\zeta_\tau, y^g_\tau, y^i_\tau)_{0 \leq \tau \leq T}$ where, as τ increases from 0 to T, ζ_τ is a continuously decreasing (hence, stronger) evidence strength level to be proven, y^g_τ is a continuously increasing guilty type invited to confess and y^i_τ is a continuously decreasing innocent type invited to deny.[18] Whenever R takes an action, i.e. a prosecution decision, the game is over. The timing and the terminal payoffs are as follows. If R cannot prove at time 0 that her evidence is weakly stronger than ζ_0, then she must take an action. At time τ, if

[18] At time 0, R invites all types below y^g_0 to confess and all types above y^i_0 to deny.

reached, R must prove that her evidence is weakly stronger than ζ_τ and ask the following two questions:"Are you type y^g_τ?" and "Are you type y^i_τ?". R's answer to each of the two question must be "Yes" or "No" but only one can be answered with "Yes". If one answer is "Yes", then R must take an action. When R detects a lie, i.e. $z \leq y^g_\tau$ in case of confession or $z \leq y^i_\tau$ in case of denial, then she prosecutes S who gets a payoff of $-b$. When she does not detect a lie, she prosecutes a type who confesses and she lets go a type who denies. If both answers are "No" and $z = \zeta_\tau$, then R must take an action. If time T has passed without R taking any action, i.e. the screening phase is over, then in the last phase of the interrogation S must send a message $m \in [y^g_T, y^i_T]$, after which R must take an action and payoffs realize as in the baseline model.

S's incentives are best summarized by the fictitious speech R could give at time $\tau > 0$ after she has proven that her evidence is stronger than ζ_τ:

> "You see, my evidence is stronger than ζ_τ so if you are guilty type y^g_τ you should tell me now and get a payoff of 0. You can wait and try to confess later at some time $\tau' > \tau$ but then I will accept only the confession of some specific type $y^g_{\tau'} > y^g_\tau$ and you will get punished at $-b$ if I catch you in a lie. Of course, you are trading this risk off with the chance that I decide to let you go because my evidence is not stronger than $\zeta_{\tau'}$. You can also decide not to confess at all and wait until time T to see whether my evidence is stronger than ζ_T. However, by then I will only accept confessions larger than y^g_T because you have answered with 'No' to all of my question and you will again get punished at $-b$ if caught in a lie. Of course, you can also try and claim any time τ' that you are $y^i_{\tau'}$ but there is a chance that I catch you.
>
> If you are innocent type y^i_τ then please tell me and I let you go. If you are a lower innocent type you can try to convince me, but I might catch you in a lie.
>
> If we reach time T and you have not yet confessed or convinced me about your innocence, you will have to make a claim about your type and then I will not be able to tell the different types apart."

We concentrate on interrogation policies of R that induce no lying until T in equilibrium and upon which R cannot improve given ζ_T.[19] It follows that, for any given ζ_T, after time T the continuation is as in the equilibrium of the game with one-shot revelation described in corollary 3 after the pooling signal ζ_T, with the difference that guilties who confess and innocents who separate in the one-shot game have already done so at some time $\tau \leq T$. This equilibrium condition and no lying pin down S's behavior all along the game and R's behavior in the last

[19] Lying necessarily happens in the last phase. This last phase must be part of the interrogation to maintain the equilibrium because since $\zeta_T > t$ there is an interval of types left, including t, after the screening phase.

stage. Solving backward, the equilibrium conditions for S and R, whose prosecution decisions must always be sequentially rational, and the optimality of the interrogation policy given ζ_T pin down the whole interrogation.

Proposition 5 (Dynamic policy)**.** *For any given $\zeta_T > t$ there exists a unique interrogation policy $(\zeta_\tau, y^g_\tau, y^i_\tau)_{0\leq\tau\leq T}$, and a corresponding equilibrium with no lying before time T, that maximizes R's payoff among these policies. Furthermore, all types of R prefer a lower ζ_T.*

In this equilibrium, each type $z \geq \zeta_T$ of R is satisfied in the sense that she identifies exactly the same confessors and credible deniers as in the one-shot revelation equilibrium with pooling signal $Z = z$ (hence, she obtains the same payoff), while when she has to make a decision only knowing that S's type does not belong to either of these sets she is again indifferent between prosecuting and letting S go. R's indifference is crucial for her to be able to give S the right incentives during the entire screening phase so as to receive from him the right, optimal amount of information, which in turn makes her indifferent. Each type of R higher than t can be satisfied in this sense by choosing ζ_T above but arbitrarily close to t. And in fact, as ζ_T approaches t, payoffs get arbitrarily close to those under the optimal mechanism described at the next section (see the proof of proposition 5 and 6). However, the equilibrium ceases to exist when $\zeta_T = t$ because type t cannot reveal himself, so that R's optimum could only be attained by means of an endless interrogation policy.[20]

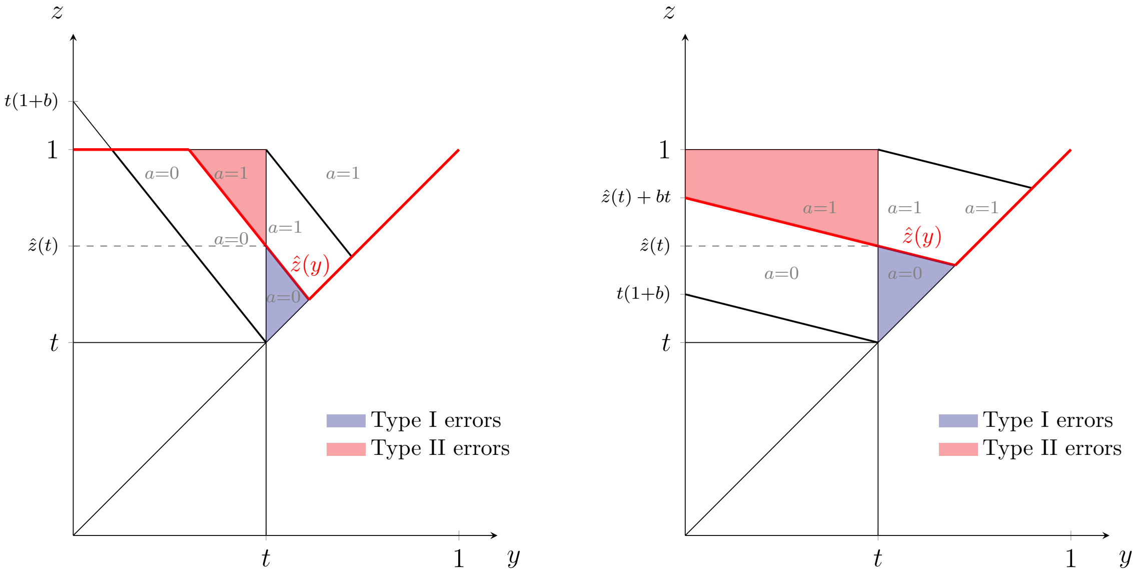
5 The optimal interrogation

5.1 Optimal mechanism under full commitment

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[21] 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 $z \geq \hat{z}(y)$.

[20] Type t would never have an opportunity to reveal himself since if $z = t$, then $y < t$. Still, the nonexistence problem of the equilibrium for $\zeta_T = t$ is deeper. For instance, even if type t of S could reveal himself, say because R's evidence z proved that $y \leq z$, R would then necessarily let him go with probability one, and some other types would find it profitable to wait until t. R's best interrogation policy is endless in that R must ask questions for each $y \in (t, 1]$, which is left-open, even though the probability that the interrogation actually never stops is zero since so is the probability that S's type is exactly t.

[21] As shown in section B.6.3 in the appendix, when looking for the optimal mechanism all the imposed restrictions on the class of mechanisms are without loss of generality.



(a) Non binding constraint for low guilties

(b) Binding constraint for all guilties

Figure 3 Determination of the optimal mechanism

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 and not lie upward, i.e. for every $y, y' \in [0,1]$ such that $y < y'$

$$1 - \hat{z}(y) \geq 1 - \hat{z}(y') - b\,(y' - y)\,.^{22} \tag{6}$$

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 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 3 demonstrates. We distinguished two cases depending on whether only sufficiently high types obtain a positive expected payoff (figure 3a) or all types do so (figure 3b).

The optimal mechanism, denoted by $\hat{\boldsymbol{z}}^{\star}$, has a close relation with the equilibrium of the baseline model (section B.6.1 in the appendix provides detailed intuitions). Let us use the nota-

[22] 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.

tion introduced at section 2.2 and take the highest confessor y_c, the lowest separating innocent $\bar{y}$, and R's cutoff action policy $\bar{z}(m)$ at their equilibrium levels (with $y_c = 0$ by convention if $\lambda(Y_c) = 0$). For types who pool in the equilibrium of the baseline model, i.e. for $y \in [y_c, \bar{y})$, R still uses the same cutoff strategy as in equilibrium and extends it to honest confessions of types who lied. That is, defining $m(y) = \ell(y)$ for a guilty type and $m(y) = y$ for an innocent type, let $\hat{z}^{\star}(y)$ be such that $1 - \hat{z}^{\star}(y) = 1 - \bar{z}(m(y)) - (m(y) - y)\, b = 1 - \bar{z}(y)$. Besides, guilty types and innocent types who separate in the equilibrium of the baseline model still get always 0 and 1, respectively, i.e. $\hat{z}^{\star}(y) = 1$ for $y < y_c$ and $\hat{z}^{\star}(y) = y$ for $y \geq \bar{y}$.

Proposition 6 (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.*

A comparison of R's decision rule in the optimal mechanism (figure 3a) and in equilibrium (figure 2) clarifies the effects of R's lack of commitment. R would benefit from committing to sometimes prosecute some innocents who deny and to sometimes let go some guilties who confess. However, anticipating they will be prosecuted, in equilibrium those guilty types elect to lie instead. As a result, R's optimal decision rule for innocent types becomes sequentially rational, which explains why type I errors are the same. Instead, since liars are punished when caught, in order to preserve S's sequential rationality, R's decision rule when they are not caught must be more lenient than the optimal one, which explains why type II errors are higher.

5.2 Implementation without commitment

R's expected loss under the optimal mechanism can be replicated in equilibrium of a simple two-stage game built on the baseline model that combines "back and forth" information revelation and discretion for R on whether to punish detected lies. Consider the following game:

- **Stage 0** S and R observe their private information as in the baseline model;
- **Stage 1** S sends a message $m \in [0, 1]$. R can either immediately choose an action $a(m, z)$, so that the game is over and payoffs realize as in the baseline model, or disclose that her evidence is such that $z \leq \zeta_m$ (i.e. as in section 4.1, type z of R sends a signal $\zeta \in [z, 1]$) and continue to stage 2;

- **Stage 2** S sends a new message $m' \in \{0, m\}$. R chooses a new action $a'(m, m', z)$ and payoffs realize as in the baseline model with a' replacing a.

S's message in stage two should be interpreted as an answer to R's question "Are you guilty?" and message 0 as a confession. Besides, when in the first stage S denies and is caught in lie it is now R's discretion to decide whether she stops the interrogation, in which case S immediately gets $-b$, or continue to the second stage. In the latter case, if S's new message is m, i.e. he insists that he is innocent, then he gets $-b$.

Proposition 7 (Implementation without commitment)**.** *There is an equilibrium of this two-stage game with discretionary punishment of lies in which expected payoffs are as in the optimal mechanism and S's behavioral strategy in stage 1 is as in the equilibrium at proposition 1.*

The structure of the equilibrium is intuitive. R immediately takes the correct action for separating types. Instead, for pooling types, R's continues the interrogation provided the evidence is sufficiently strong relative to S's claim so that by confronting him with the evidence she will be able to set a guilty and an innocent apart. More in detail, when $z > \zeta_m$, where ζ_m is an evidence strength level contingent on m, she immediately makes a decision. As usual, she lets S go if the evidence is sufficiently weak and prosecutes S if the evidence is strong (but still above ζ_m). Instead, when $z \leq \zeta_m$, she discloses this information to S and offers him a second opportunity to confess. The appropriate choice of ζ_m now makes type $\ell^{-1}(m)$ barely willing to do so. Instead, innocent type m sticks to his stage one story anticipating he will be let go. This game entails a combination of "carrot and stick" in the treatment of lies. Small lies, i.e. equilibrium lies in stage one are forgiven. Conversely, big lies, i.e. off the equilibrium path lies of a guilty type who mimics an innocent type higher than he is supposed to, would still be punished immediately in stage one if caught ($z \leq \ell^{-1}(m)$).

6 Discussion

We provided a tractable framework to analyze interrogations and derived several implications for their design. This section covers some extensions and additional questions.

6.1 Properties and interpretations of the information structure

Consider first the joint distribution of S's status as guilty or innocent and the evidence. Seeing z as a signal about S's status, the monotone likelihood ratio property holds since the probability that S is guilty $\mathbb{P}(y < t|z) = min\{t/z, 1\}$ is decreasing, which clarifies in which sense a lower z represents stronger evidence. Likewise, the distribution of z conditional on S's guilt is dominated in the sense of first order stochastic dominance by the corresponding one conditional on S's innocence. Our information structure allows for heterogeneity within guilty and innocent suspects and ensures these natural properties extend to the joint distribution of S's type and the evidence.[23] The fact that the evidence excludes some of S's types - technically the conditional distribution of y given z does not have full support - allows for a natural definition of lie catching. The fact that excluded types are the high ones, e.g. R's evidence can never prove S's innocence, captures the idea that whistleblowing, anonymous tips and voluntary reports to law enforcement authorities are inherently incriminating. The specific joint distribution of y and z we consider has the convenient property that conditional distributions are uniform.

The information structure can be equivalently thought of as arising within an isomorphic model in which y and z are drawn independently and uniformly from $[0, 1]$ but S is only interrogated when $z > y$.[24] A first interpretation for this simplifying assumption is then that in the case $z \leq y$ R does not observe any evidence and is not even aware of S. Alternatively, as S's guilt becomes less likely than under the prior, R does not find it worth it or has no legal ground to go after S. Ultimately, in this isomorphic model the interrogation process is not random. The prior probability that S is guilty is t but becomes $(2 - t)t$ when S is interrogated, which is higher (and increasing in t).

6.2 Alternative moves and payoffs

In this section, we discuss how the equilibrium of the baseline model modifies under some natural variations.

Lower punishment for silence. Suppose S's punishment for staying silent when evidence is conclusive is lower than when he is caught in a lie, i.e. equation (2) becomes $a - b_\ell \mathbb{1}_{m \geq z} - b_s \mathbb{1}_{m=s \text{ and } z \leq t}$ with $b_\ell > b_s \geq 0$. Then, in any equilibrium the set of silent types and liars 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_ℓ replacing b and the difference that, provided some types indeed stay silent, the highest confessor, the lowest liar, the lowest separating innocent type and R's action policy will adjust based on b_s. In particular, the lowest liar y_ℓ will be indifferent

[23] Related to assumption 2 in Reinganum (1988), for any interval $\delta \subseteq [0, 1]$, $\mathbb{E}[z|z \in \delta, y < t] \leq \mathbb{E}[z|z \in \delta, y \geq t]$ (provided that the evidence is not conclusive, so that these expectations are well defined). Additionally, in our setting, for any type y and y' such that $y < y'$ (which are not excluded by δ, so that again expectations are well defined), $\mathbb{E}_y[z|z \in \delta] \leq \mathbb{E}_{y'}[z|z \in \delta]$, where $\mathbb{E}_y$ and $\mathbb{E}_{y'}$ denote the expectations respectively of type y and y'.

[24] This model can also accommodate the alternative assumption that, when $z \leq y$, R can still interrogate S. In this case, R's private information is z and whether $z > y$ or $z \leq y$, while S's private information is y.

between lying, say 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 $A(s) = \int_{[t,1]} a(s,z)/(1-t)\mathrm{d}\lambda$.

No punishment. If S incurs no costs, i.e. $b = 0$, then proposition 0 still holds, i.e. innocents and confessors are honest in any truth-leaning equilibrium, and proposition 1 also holds whenever $t/(1-\alpha) < 1$.[25] In particular, $\bar{y} = t/(1-\alpha) = \bar{z}(m)$ for all $m \in L$ as long as $Z_s \geq t/(1-\alpha)$, all guilty types lie and the equilibrium is uninformative. When $Z_s < t/(1-\alpha)$ then $\bar{y} = Z_s = \bar{z}(m)$ for all $m \in L$ and there are also some silent types but obviously no confessors. In this case, since some guilties remain silent, the equilibrium is informative but R's payoff is even worse than when she relies only on the evidence z to make a decision because protection of silence is too strong, i.e. Z_s is too low.

No punishment and leniency for confession. Suppose again that S incurs no costs, i.e. $b = 0$ as above, but that by confessing he obtains a premium $U \in (0,1)$ in addition to R's action. The incentive structure is then as in Seidmann (2005), who considers the two levels of protection of silence $Z_s = 1$ (the "English game") and $Z_s = \min\{1, t/(1-\alpha)\}$ (the "American game"). We highlight four main points from a comparison of this leniency setup with the baseline model:

(i) in the leniency setup in both the American and the English games no (positive measure of) types are silent, hence the American level of protection of silence can never hurt R, as opposed to Seidmann (2005);

(ii) after an appropriate rescaling, equilibrium payoffs in these games are the same as in our baseline model (as long as no types are silent also in the baseline model);

(iii) in the leniency setup, R cannot benefit from protection of silence no matter its level, as in Seidmann (2005) for the American level, and she is hurt when protection of silence is stronger than the American level;

(iv) instead, in the baseline model, R sometimes benefits and sometimes she is hurt from the American level of protection of silence.

For points (i) and (ii), given that innocents and confessors are honest, the equilibria in both games are described in the previous paragraph, with the difference that now there might be some types who confess and there are no silent types. Since there are no silent types, the informativeness of the equilibria is as described by remark 2 and payoffs are exactly as in our

[25] When $t/(1-\alpha) \geq 1$ then all innocents pool, i.e. $\bar{y} = 1$ as opposed to $\bar{y} < 1$ as described in the proposition.

baseline model with $b = U/(1-U)$ and, say, $Z_s = 1$.[26] Point (iii) holds as in the leniency setup it turns out that R is harmed from choosing any $Z_s < t/(1-\alpha)$ which induces a positive measure of silent types. It demonstrates that the finding of Seidmann (2005) is robust to our information structure, which is remarkable given that there are important differences, such as the fact that R's evidence cannot prove S's innocence and can prove his guilt and guilty types are heterogeneous in the strength of the evidence they expect. Point (iv), which simply follows from proposition 2, stands instead in stark contrast with Seidmann (2005) as it implies that protection of silence can even result in a Pareto improvement. This difference holds because in our baseline model there are levels of protection of silence which can induce some types to stay silent while under the same levels there is no silent type in the leniency setup. Beyond equilibrium forces, the reason is that, differently from the leniency setup, in our baseline model silent types never get punished when the evidence is inconclusive, i.e. R is sometimes lenient even with silent types (see footnote 26).

Continuous actions. The equilibrium of the baseline model remains unaffected if R chooses her action continuously from $[0,1]$, e.g. if she must decide at which intensity to prosecute S. Indeed, for any pooling message, upon not catching S in a lie R would again be indifferent among actions. If simultaneously R's loss becomes quadratic, e.g. $a^2 \mathbb{1}_{y<t} + (1-a)^2 \mathbb{1}_{y \geq t}$ 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 incentives to do so 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 as well as the measure of liars and the measure of the lying region 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.

[26] Indeed, after appropriately rescaling S's payoff (i.e. $-b$ to 0, 0 to U and 1 to 1 or the other way around) the equilibrium measure of confessors and the lying region L are determined exactly by the same conditions in both models, the only difference being that in the leniency setup $\bar{z}(m) = \bar{y}$ while in the baseline model $\bar{z}(m) = \bar{y} + b(\bar{y} - m)$ for all $m \in L$. Yet, given that in the baseline model when $z > \bar{y}$ lies are never detected and R is indifferent when $m \in L$, R could just as well let everyone go whenever $z > \bar{y}$, which is exactly what happens in the equilibrium of the leniency setup. Finally, note that even after rescaling payoffs, the two setups are different because while in our setup liars may get $-b, 0$ or 1 (or $0, U$ or 1) in the leniency setup liars always get $-b$ or 1 but never 0 (or equivalently, 0 or 1 but never U). The same difference holds for silent types.

6.3 Asymmetric information about the law and deceptive tactics

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. However, law enforcers' arbitrariness is a major cause of criticism and an important reason behind the general movement towards the mandatory recording of interrogations.[27] Our model directly allows to identify the direction of the misleading efforts R would want to engage in if these are tolerated by law or go undetected and S is prone to deception. Indeed, supposing S plays according to what he perceives as equilibrium behavior while R best responds given the true environment, we can easily calculate how R would want to mislead the suspect about several parameters of interest. In accordance with the logic behind common interrogations tactics,[28] R would always want to overstate the punishment for reticence (increase S's perception of b), exaggerate the strength of the incriminating evidence (decrease S's perception of Z_i and Z as defined in section 3 and 4) and misrepresent her true preferences over type I and type II errors (increase or decrease S's perception of α). Besides, R would sometimes benefit from concealing the right to silence to an otherwise unaware S. While surely objectionable on other grounds, if successful these deceptive tactics improve information elicitation. Also, this improvement does not come at the cost of extorting false confessions since the only possibility for these tactics to induce innocent types to depart from honesty is if they generate a shift away from truth-leaning (see section C in the appendix). As a next step, one could investigate if these deceptive tactics would remain effective in a persuasion framework where S is rational but uninformed.

Parameters b and Z_s can be directly thought of as expectations of S about random variables whose realizations are private information of R. We now explain how also the parameter α can be made R's private information. In the equilibrium of the baseline model, R 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 $\frac{1}{1+\ell^{-1'}(m)} = \bar{\alpha}(m)$. This differential equation pins down the lying region and the set of liars.

[27] See for instance Sullivan (2005).

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

6.4 Regulating evidence revelation

We did not consider comprehensively laws that govern communication about the evidence to the suspect. In particular, we maintained that the law enforcer's statements must be truthful. If the law enforcer can make false claims, instead, new interesting strategic considerations arise due to the possibility that the suspect may in turn catch the law enforcer in a lie, e.g. know that she is exaggerating the strength of the evidence. Importantly, regulation might also affect the law enforcer's choice to interrogate the suspect in the first place, and whether by means of a casual conversation or a formal interrogation. For example, by officially marking the start of a formal interrogation, the legal requirement that the suspect is explicitly notified of his right to silence may implicitly convey additional information.[29]

References

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.

Cuellar, Pablo, "Voluntary Disclosure of Evidence in Plea Bargaining," *Working paper*, 2020.

[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."

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.

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.

Jehiel, Philippe, "Communication with forgetful liars," *Theoretical Economics*, 2021, *16* (2), 605–638.

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.

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

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

Leshem, Shmuel, "The Benefits of a Right to Silence for the Innocent," *The RAND Journal of Economics*, 2010, *41* (2), 398–416.

Mialon, Hugo M, "An economic theory of the fifth amendment," *Rand Journal of Economics*, 2005, pp. 833–848.

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

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.

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

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

_ **and Alex Stein**, "The right to silence helps the innocent: A game-theoretic analysis of the Fifth Amendment privilege," *Harvard Law Review*, 2000, pp. 430–510.

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.

Sobel, Joel, "Lying and deception in games," *Journal of Political Economy*, 2020, *128* (3), 907–947.

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_ℓ 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 $\mu(m, z)$ 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, the total pushforward measure of the Lebesgue measure λ (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 Proofs

B.1 Proof of proposition 1

We will prove this formal version of proposition 1.

Proposition B.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) *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)*:*

$$a(m, z) = \begin{cases} 0 & \text{if } z \leq \bar{z}(m) \\ 1 & \text{if } z > \bar{z}(m) \end{cases}, \tag{8}$$

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.*

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 B.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

$$\pi_\ell(m; y) = \underbrace{(1 - m)A(m)}_{\text{lie not detected}} - \underbrace{(m - y)b}_{\text{lie detected}}. \tag{9}$$

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

$$\pi_{s,g}(y) = \underbrace{1 - Z_s}_{\text{inconclusive evidence}} - \underbrace{(t - y)b}_{\text{conclusive evidence}} \tag{10}$$

and when he is innocent is

$$\pi_{s,i}(y) = 1 - Z_s \mathbb{1}_{y \leq Z_s}. \tag{11}$$

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

$$\pi_{d\ell,i}(m; y) = \int_y^1 a(m, z)\mathrm{d}z. \tag{12}$$

Finally, the expected payoff of innocent type y from being honest is simply equation (9), or equivalently equation (12), evaluated in $m = y$

$$\pi_{h,i}(y) = (1 - y)A(y). \tag{13}$$

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;

[30] Formally, the expected payoff of each type y is defined conditional on $z > y$. 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.

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

$$A(m) = \frac{1 - \bar{y} - b\,(\bar{y} - m)}{1 - m}. \tag{14}$$

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_ϵ who also sends this message. Using equation (9), y_ϵ earns $\epsilon A(m_\epsilon) - (1 - \epsilon - y_\epsilon)b$, which letting ϵ go to 0 converges to $-(1 - y_\epsilon)b$. There is therefore an arbitrary small ϵ such that type y_ϵ 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 (9) 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 (9) it is apparent that the expected payoff of type t_ϵ^- 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 (10). The expected payoff of type t_ϵ^- from deviating to $m = t$ is $1 - t - b\epsilon$, which for ϵ 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 (9) and (10) clarifies that each guilty type sending m could profitably deviate to staying silent or confessing honestly.

Point (iii): From equation (9) 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 (9) 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 (9) 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

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

Solving this differential equation with terminal condition $A(\bar{y}) = 1$ yields equation (14). □

Lemma B.2. *In any equilibrium:*

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

$$\mu(m, z) = \alpha; \tag{15}$$

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

$$\frac{\lambda\,(Y_\ell)}{\lambda\,(L)} = \frac{1-\alpha}{\alpha}. \tag{16}$$

Proof. Point (i) follows directly from equation (1) 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 (16) follows from choosing $A = L$. □

Evaluating equation (14) in t shows that there must be a one to one relationship between $A(t)$ and $\bar{y}$, i.e.

$$A(t) = \frac{1 - \bar{y} - b\,(\bar{y} - t)}{1 - t} \qquad \text{or, equivalently,} \tag{17}$$

$$\bar{y} = \frac{1 - (1 - t)A(t) + bt}{1 + b}. \tag{18}$$

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 (9) in t yields the expected payoff from lying for type y

$$\pi_\ell(y) = (1-t)A(t) - (t-y)b. \tag{19}$$

Let us denote by v the expected payoff for a guilty type from remaining silent conditional on evidence being inconclusive, i.e., from equation (3),

$$v \equiv \mathbb{P}\left(z > Z_s \,|\, \text{inconclusive evidence}, y < t\right) = \frac{1-Z_s}{1-t}. \tag{20}$$

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

$$A(t) \geq v, \tag{21}$$

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

Proof. Replacing equation (20) in (10) and subtracting from equation (19) yields $\pi_\ell(y) - \pi_{s,g}(y) = A(t) - v$. If inequality (21) 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*

$$y_c \equiv \frac{bt - (1-t)A(t)}{b}, \tag{22}$$

and Y_c has positive measure if and only if

$$b > \frac{1-t}{t}A(t). \tag{23}$$

Proof. By lemma B.4, $\pi_\ell(y)$, which is defined in equation (19), 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 (22) is the unique solution to $\pi_\ell(t; y) = 0$. Also, $y_c > 0$ if and only if equation (23) holds. □

B.1.2 Possible cases

Throughout superscripts index the respective cases. Also, we innocuously ignore sub-cases 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 (22) and $\lambda\left(Y_\ell^I\right) = t - y_c$ so that, by equation (16),

(17) and (18), $\bar{y}^I = \frac{\alpha+bt}{\alpha+b}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I = \frac{(1+b)t-(1-\alpha)}{b+\alpha}$. By equation (23), 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 (16), (17) and (18) 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 (23) 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 (21), so that, by equation (18), $\bar{y}^{III} = \frac{1-(1-t)v+bt}{1+b}$ (where, using equation (20), indeed $\bar{y}^{III} < Z_s$). By equation (16), $\lambda\left(Y_\ell^{III}\right) = \frac{(1-\alpha)(1-t)(1-v)}{\alpha(b+1)}$. 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 (23), $\lambda\left(Y_c\right) > 0$ if and only if

$$v < v_{IIIa} \equiv \frac{tb}{(1-t)}. \tag{24}$$

- **Case IIIa: some types confess**. When $v < v_{IIIa}$, from equation (22), $\lambda\left(Y_c^{III}\right) = y_c^{III} = \frac{bt-(1-t)v}{b} > 0$, 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*
$$\frac{1-Z_s}{1-t} > max\left\{\frac{(1-\alpha)b}{b+\alpha}, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\}; \tag{25}$$

- $\lambda\left(Y_c\right) > 0$ *iff*
$$b > \frac{1-t}{t} max\left\{\frac{1-Z_s}{1-t}, \frac{(1-\alpha)b}{b+\alpha}\right\}, \tag{26}$$

  which if $\lambda\left(Y_s\right) = 0$ simplifies to
$$b > \frac{1-t-\alpha}{t}. \tag{27}$$

Proof. In each of the three cases described at section B.1.2, $\lambda\left(Y_c\right)$, $\lambda\left(Y_s\right)$ and $\lambda\left(Y_\ell\right)$ 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 (22), 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 $\lambda\left(Y_\ell\right)$ by lemma B.2, it follows that $\bar{y}^I > \bar{y}^{III}$ and, from equation (17), 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 α:

- 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

$$A(t) = \begin{cases} max\left\{v, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\} & \text{if}\, b \leq \frac{1-t-\alpha}{t} \\ max\left\{v, \frac{(1-\alpha)b}{b+\alpha}\right\} & \text{if}\, b > \frac{1-t-\alpha}{t}. \end{cases} \tag{28}$$

Using the definition of v (equation (20)), equation (25) 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 (26) obtains by evaluating equation (23) using equation (28) and noting that $\lambda\left(Y_c\right) = 0$ by construction whenever case II occurs. Finally, equation (27) 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 (26), which by the previous observations must necessarily be the case when $\lambda\left(Y_s\right) = 0$. □

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 (9) 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 (12) and (13) 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 (11) and (13) 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 (16) 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 (15). 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 (14) 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

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

$$=(1-\alpha)\int_L (1-y)\,\mathrm{d}\lambda + \alpha\lambda\,(Y_s)\,(1-Z_s) \tag{30}$$

$$=\alpha\lambda\,(Y_\ell)\left(1-t-\frac{\alpha\lambda\,(Y_\ell)}{2(1-\alpha)}\right) + \alpha\lambda\,(Y_s)\,(1-Z_s). \tag{31}$$

Equation (29) obtains by a change of variables under pushforward integrability where, by lemma B.2, $(1-\alpha)\lambda = \alpha\,(\lambda \circ \boldsymbol{g})$. Equation (30) obtains because the first term of equation (29) 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 (31) 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 (16).

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

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

which is again identical across equilibria.

B.3 Proofs for section 3

B.3.1 Proof of remark 1

The statement follows directly from the observation that R's equilibrium action policy upon a pooling message is such that $\bar{y} \leq \bar{z}(m) \leq Z_s$.

B.3.2 Proof of remark 2

Suppose that R must rely on z only to make a decision. Obviously, R finds it optimal to prosecutes if $z \leq t$, while for $z > t$, she prosecutes if and only if $\alpha t/z \geq (1-\alpha)(1-t/z)$, i.e. if

$$z \leq \frac{t}{1-\alpha}. \tag{32}$$

Consider now R's decisions in the baseline model when $Z_s = 1$ for each realization of R's type z. When $z \leq t$ again R always prosecutes. If $z \in (t, \bar{y})$, then equation (32) holds and R always prosecutes even when she does not catch S in a lie (because $\bar{z}(m) \geq \bar{y}$). If $z \geq \bar{y}$ and there are no confessors in the baseline equilibrium then we also have $z \geq \frac{t}{1-\alpha}$. Again, since z is indifferent in the baseline equilibrium she could just as well let everyone go while keeping the same payoff as when she does not interrogate because she never catches a lie. Instead, if $z > \bar{y}$ and there are confessors then R is strictly better off in the baseline equilibrium because when she does not interrogate she either lets go guilties who confess in the baseline equilibrium or prosecutes innocents between $\bar{y}$ and z who separate in the baseline equilibrium.

B.3.3 Proof of proposition 2

In light of remark 1, we prove the proposition with reference to the level of protection of silence, defined in terms of $v = \frac{1-Z_s}{1-t}$ rather than Z_s as per equation (20). 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 to condition $b \leq \frac{1-t-\alpha}{t}$ by inequality (27), since inequality (25) for having $\lambda(Y_s) > 0$ 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 $\lambda(Y_\ell)$ and $\lambda(Y_s)$ in equation (31), R's expected loss is

$$\begin{aligned} E(v) =&(1-\alpha)\frac{(1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)}{1+b} \\ &+\alpha(1-t)v\left(t-\frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \end{aligned} \tag{33}$$

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

$$\tilde{v} = \frac{1-t-\alpha-b^2t\alpha+2b(1-t-\alpha)}{(1+2b)(1-t)(1-\alpha)} \in \left(A^{II}(t), 1\right). \tag{34}$$

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

Suppose now that $b > \frac{1-t-\alpha}{t}$, instead, so that the value of v can affect whether $\lambda(Y_c) > 0$. 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 (24) represents the level of protection of silence above which $\lambda(Y_c) = 0$,
  - 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 $\lambda\left(Y_\ell\right)$ and $\lambda(Y_s)$ in equation (31), R's expected loss is

$$\begin{aligned} E(v) =&(1-\alpha)\frac{(1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)}{1+b} \\ &+\alpha(1-t)v\left(t-\frac{bt-(1-t)v}{b}-\frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \end{aligned} \tag{35}$$

Since $E(v)$ is convex[31] 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 $\lambda\left(Y_s\right) = 0$, yielding

$$E(v)\left.\right|_{v=A^I(t)} = \frac{(1-t)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2}. \tag{36}$$

Consider now the region $v \geq v_{IIIa}$. R's expected loss in this region is again given by equation (33) 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 (34). 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,[32] $E'(v)\left.\right|_{v=v_{IIIa}} \geq 0$ and R's global optimal level of protection is $v = A^I(t)$, i.e. it is such that $\lambda\left(Y_s\right) = 0$. Indeed, R's expected loss is always continuous in v (i.e. equation (33) and (35) 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 $E'(v)\left.\right|_{v=v_{IIIa}} < 0$, R's optimal level of protection is either $v = A^I(t)$, i.e. such that $\lambda\left(Y_s\right) = 0$, or it is such that $\lambda\left(Y_s\right) > 0$ and equal to $\tilde{v}$, depending on whether equation (33) evaluated at $\tilde{v}$ is greater or lower than equation (36) (and there exist parameter combinations for which the optimal level is such that $\lambda\left(Y_s\right) > 0$, as for instance $t = 17/64$, $\alpha = 1/2$ and $b = 1$).

B.3.4 Proof of proposition 3

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

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

[31]
$$E''(v) = \frac{(1-t)^2\left(b+2b^2+2\alpha+3b\alpha\right)}{b(1+b)^2} > 0.$$

[32]
$$\frac{\partial\hat{t}\,(b,\alpha)}{\partial b} = -\frac{(1-\alpha)((1-\alpha)+2b(1+b)(2-\alpha))}{(1+b)^2(1-b(\alpha-2))^2} < 0 \quad \frac{\partial\hat{t}\,(b,\alpha)}{\partial\alpha} = -\frac{1+2b}{(1+b(2-\alpha))^2} < 0.$$

For any given standard $Z_i \in (t, 1]$, 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 $\bar{y} < Z_i$ replacing $\bar{y} < 1$, where $\bar{y}$ hence the lying region L may now depend on Z_i, and

$$A(m) = \frac{Z_i - \bar{y}(Z_i) - b\left(\bar{y}(Z_i) - m\right)}{Z_i - m} \tag{37}$$

replacing equation (14). Equation (17) and (18) become respectively

$$A(t) = \frac{Z_i - \bar{y}(Z_i) - b\left(\bar{y}(Z_i) - t\right)}{Z_i - t} \quad \text{and} \tag{38}$$

$$\bar{y}(Z_i) = \frac{Z_i - (Z_i - t)A(t) + bt}{1 + b}. \tag{39}$$

Also, equation (19), i.e. the expected payoff from lying in the lying region for a guilty type, is

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

so that equation (22), i.e. the highest confessor (if any), becomes

$$y_c(Z_i) \equiv \frac{bt - (Z_i - t)A(t)}{b} \tag{40}$$

and the necessary and sufficient condition (23) to have that $\lambda\left(Y_c\right) > 0$ becomes

$$b > \frac{Z_i - t}{t} A(t). \tag{41}$$

Since $Z_i = Z_s$ and S knows that $z \leq Z_i$, 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 $\lambda\left(Y_c\right) > 0$, $\lambda\left(Y_\ell^I\right) = t - y_c(Z_i)$, where $y_c(Z_i)$ is defined in equation (40). Using equation (16) and (38), $\bar{y}^I(Z_i) = \frac{bt + Z_i\alpha}{b+\alpha}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I(Z_i) = \frac{t + bt - (1-\alpha)Z_i}{b+\alpha}$. This case can only occur if $b > \frac{(1-\alpha)Z_i - t}{t}$, i.e. if $Z_i < \frac{(1+b)t}{1-\alpha}$ (which is satisfied for any b and Z_i if $t \geq 1-\alpha$), so that condition (41) holds (this then automatically implies that $\bar{y}^I < Z_i$).

Case II: all guilty types lie. If $\lambda\left(Y_c\right) = 0$, the measure of liars Y_ℓ^{II} is then $\lambda\left(Y_\ell^{II}\right) = t$. Using equation (16) and (38), it then follows that $\bar{y}^{II} = \frac{t}{1-\alpha}$ and $A^{II}(t) = \frac{(1-\alpha)Z_i - t(1+b\alpha)}{(Z_i - t)(1-\alpha)}$. By equation (41), this case can only occur if $b \leq \frac{(1-\alpha)Z_i - t}{t}$ (this then automatically implies that $\bar{y}^{II} < Z_i$ and $A^{II}(t) > 0$).

For any Z_i 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 follows analogous steps as section B.1.4 and B.2. R's expected loss under interrogation standard Z_i is then

$$E(Z_i) = (1-\alpha)\int_{L(Z_i)} (Z_i - y)\mathrm{d}\lambda + \alpha\, t(1-Z_i) \tag{42}$$

$$= \alpha\lambda\left(Y_\ell\left(Z_i\right)\right)\left(Z_i - t - \frac{\alpha\lambda\left(Y_\ell\left(Z_i\right)\right)}{2(1-\alpha)}\right) + \alpha\, t(1-Z_i) \tag{43}$$

The first term in equation (42) obtains by analogous simplifications as at equation (30), while the second term is due to the fact that when $z > Z_i$ now R takes action $a = 1$ and hence makes a type II error when facing a guilty type. Equation (43) follows from some rearranging using equation (16).

Assume for the moment that given Z_i case I above obtains, i.e. $\lambda\left(Y_c\right) > 0$. Then, using that $\lambda\left(Y_\ell\right) = \frac{(1-\alpha)(Z_i-t)}{\alpha+b}$, equation (43) becomes

$$E(Z_i) = \frac{(t-Z_i)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2} + t(1-Z_i)\alpha. \tag{44}$$

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

$$t < \bar{t}\left(b,\alpha\right) \equiv \frac{(1-\alpha)(2b+\alpha)}{2b+\alpha+b^2}, \tag{45}$$

where $\bar{t}\left(b,\alpha\right)$ is strictly decreasing in its arguments.[33] In such a case, the FOC gives

$$\tilde{Z}_i = \frac{t(b(2+b)+\alpha)}{(1-\alpha)(2b+\alpha)}. \tag{46}$$

Now, if case I above obtains even at $Z_i = 1$, condition (45) is necessary and sufficient for an interior standard to be optimal. If $\lambda\left(Y_c\right) = 0$ for $Z_i = 1$, instead, equation (44) on which the minimization was taken over represents R's expected loss only in the region $Z_i < \hat{Z}_i \equiv \frac{(1+b)t}{1-\alpha}$. However, in such case the optimal standard is always interior and hence given by equation (46). Indeed, when $Z_i \geq \hat{Z}_i$, by remark 2 the interrogation is uninformative and, since by equation (32) R would choose anyway $a = 1$ whenever $Z_i \geq \hat{Z}_i$ if allowed to, R's payoff is as in an uninformative equilibrium. Therefore, a Z_i that induces $\lambda\left(Y_c\right) > 0$ is strictly optimal for R.

[33]
$$\frac{\partial\bar{t}\left(b,\alpha\right)}{\partial b} = -\frac{2b(1-\alpha)(b+\alpha)}{(b(2+b)+\alpha)^2} < 0 \quad \frac{\partial\bar{t}\left(b,\alpha\right)}{\partial\alpha} = -\frac{(b+\alpha)(b(3+2b)+\alpha)}{(b(2+b)+\alpha)^2} < 0.$$

B.3.5 Proof of remark 3

We prove the statement by showing that whenever R would find it optimal to prosecute for some $z > Z_i$, where Z_i is the standard for interrogating, then R's expected loss $E(Z_i)$ is decreasing in Z_i. For the moment, fix the standard for prosecuting, which by assumption is such that $Z_i \leq Z_s$, at $Z_s = 1$, so that R's action when S is not interrogated is unconstrained. Suppose first that equation (32) holds, i.e. $t/(1-\alpha) > 1$, so that, for any $Z_i < 1$, R always finds it optimal to prosecute whenever $z > Z_i$. Then, $E(Z_i)$ is given by the sum of the first term of equation (44), i.e. R's loss when she interrogates, and $(1-\alpha)(1-2t+Z_i)/2$, i.e. type I errors on innocents when she does not interrogate. Indeed $E'(Z_i) = -\frac{(1-\alpha)b^2(Z_i-t)}{(\alpha+b)^2} < 0$, so that R prefers $Z_i = 1$. Next, suppose that $Z_i \leq t/(1-\alpha) \leq 1$, so that, when $z > Z_i$, R finds it optimal to prosecute if and only if $z \leq t/(1-\alpha)$. Then, $E(Z_i)$ is given by the sum of three terms: the first term of equation (44), i.e. R's loss when she interrogates, $(1-\alpha)(t/(1-\alpha)-2t+Z_i)(t/(1-\alpha)-Z_i)/2$, i.e. type I errors on innocents when she does not interrogate and prosecutes, and $\alpha t(1-t/(1-\alpha)$, i.e. type II errors on guilties when she does not interrogate and lets S go. Again, $E'(Z_i) = -\frac{(1-\alpha)b^2(Z_i-t)}{(\alpha+b)^2} < 0$, so that R prefers $Z_i = t/(1-\alpha)$. Since in both cases R would like to increase Z_i as much as possible, this must be even more the case when $Z_s < 1$, since then when $z > Z_i$ she is also sometimes constrained in her action. The second part of the remark is a direct implication of the first. It can also be seen directly from the fact that whenever $Z_i^* < 1$, so that Z_i^* is as in equation (46), $Z_i^* > \frac{t}{1-\alpha}$, which by equation (32) is the cutoff for z above which R finds it optimal to let S go.

B.4 Proofs for section 4.1

B.4.1 Proof of proposition 4

Since S knows z, he will never send a message $m \geq z$, which will be surely caught in a lie. It follows that upon any on the equilibrium path message m and m', $a(m,z) = a(m',z)$, since otherwise S would always choose the message upon which $a = 1$. Likewise, S will never confess if $a(m,z) = 1$ for some $m > t$ and $z > t$, which means that whenever S confesses R is always choosing $a = 0$ anyway. It follows that any equilibrium is necessarily uninformative. The following lemma describes one equilibrium, for which checking sequential rationality of both S and R is straightforward.

Lemma B.6 (An equilibrium when S knows z)**.** *The following is an equilibrium. If $z \leq t$ then all types of S confess. If $z > t$ then S uses the following strategy, where $y_z = max(0, t-(z-t)\frac{1-\alpha}{\alpha})$: $y < y_z$ confesses, $y \in [y_z, t)$ lies according to $\ell_z(y) = t + (y-y_z)\frac{\alpha}{1-\alpha}$. R then lets S go after* *any message $t \leq m < z$ if $z > \frac{t}{1-\alpha}$ and prosecutes S if $z \leq \frac{t}{1-\alpha}$ (for the sake of precision if $z = \frac{t}{1-\alpha}$ then both type of equilibria exist). In both cases, R's belief is α after any pooling message. Hence, the strategy of S is the same for all $z \geq \frac{t}{1-\alpha}$ because then $t - (z-t)\frac{1-\alpha}{\alpha} \leq 0$ and $\ell_z(t) \leq z$. Notice that for $z = \bar{y}$ we have that $\boldsymbol{\ell}_{\bar{y}} = \boldsymbol{\ell}$.*

B.4.2 Proof of corollary 2

The first statement is obvious once one defines the out of equilibrium path behavior of S and R as in one of the equilibria of proposition 4 (see for example lemma B.6). For the second statement, we define the out of equilibrium path strategies and beliefs after nondisclosure from R as follows. S believes that $z = 1$, and type $z = 1$ and S play as described in lemma B.6 for $z = 1$. All we have to check is that no type $z > t$ wants to deviate to nondisclosure. If $z \leq \frac{t}{1-\alpha}$ then z prosecutes everyone in equilibrium. Deviating to nondisclosure is not profitable because those ys who are now caught in a lie were prosecuted before as well, while when z has to move she is in the same situation as on the equilibrium path, namely she is indifferent given that her belief is still α. If $z > \frac{t}{1-\alpha}$ then z lets all types of S go on the equilibrium path but then deviating to nondisclosure does not change the strategy of S (the strategy of S is the same for all $z \geq \frac{t}{1-\alpha}$, see lemma B.6).

B.4.3 Proof of corollary 3

We start by completing the description of the equilibrium. After signal Z, S and R play as in the equilibrium of the game described at section B.3.4 in which the standard for interrogation Z_i is Z. After any signal $\zeta > Z$, type ζ and S play according to the strategy described in lemma B.6 for $z = \zeta$. After any out of equilibrium signal $\zeta < Z$, S believes that $z = \zeta$, and S and type ζ of R play as described in lemma B.6 for $z = \zeta$.

First, we check that no type $z > Z$ wants to deviate to a signal $\zeta > z$. If $\frac{t}{1-\alpha} \leq Z$ then each $z > Z$ lets all types of S go on the equilibrium path and deviating to some $\zeta > z$ does not change the strategy of S (the strategy of S is the same for all $z \geq \frac{t}{1-\alpha}$, see lemma B.6). If $z \geq \frac{t}{1-\alpha} > Z$ the same argument applies. If $\frac{t}{1-\alpha} > Z$ but $z < \frac{t}{1-\alpha}$ then z prosecutes everyone in equilibrium. Deviating is not profitable because those ys who are now caught in a lie were prosecuted before as well and when z has to move she is in the same situation as on the equilibrium path, namely she is indifferent given that her belief is still α.

Next, we check that no type $z \leq Z$ wants to deviate to a signal $\zeta \in [z, 1]$. The arguments are very similar. Consider the usual equilibrium construction (on path), in which in the pooling region R's belief is α, induced by the strictly increasing lying function $\boldsymbol{\ell}_Z : [y_c(Z), t) \rightarrow [t, \bar{y}(Z))$

with slope $\frac{\alpha}{1-\alpha}$. Using the notation introduced in lemma B.6, notice that $\boldsymbol{\ell}_Z = \boldsymbol{\ell}_{\bar{y}(Z)}$ and $y_c(Z) = y_{\bar{y}(Z)}$, so if $\zeta = \bar{y}(Z)$ then S's behavior does not change. If $\zeta < \bar{y}(Z)$, then the set of confessors increases relative to the one on the equilibrium path by $y_c(Z) - y_\zeta$. But now z no longer catches in a lie exactly the same measure of guilty types. And whenever R has to move out of the equilibrium path, she is in the same situation as on path given that her belief is kept at α. If $\zeta > \bar{y}(Z)$, then the set of confessors decreases by $-(y_c(Z) - y_\zeta)$. But now z additionally catches in a lie the same measure of guilty types. Again, z cannot be better off by deviating because when she has to move out of the equilibrium path, her belief is just as on path.[34]

B.5 Proof of proposition 5

We use the notation and calculations developed in section B.3.4. To simplify notation, we drop the time index τ and write $\zeta_\tau = \zeta, y_\tau^{g,i} = y^{g,i}(\zeta)$, we fix $\zeta_T = Z$ and we keep ζ_0. We write a_z for the prosecution decision when $\zeta = z$ and there was neither confession nor denial. We give first the interrogation policy conjectured to be optimal given Z and the corresponding equilibrium. Then we argue that the policy is indeed optimal given Z. Finally, we show that R always prefers a lower Z.

If $\lambda(Y_c(Z)) = 0$ then the interrogation is trivial: $T = 0$, R lets S go when $z > Z$ and otherwise she proves that $z \leq Z$, $y^g(Z) = 0, y^i(Z) = \bar{y}(Z) = \bar{y}$ and the last phase strategies are as in the baseline model with the appropriate adjustment on $A(m)$. Hence, suppose that $\lambda(Y_c(Z)) > 0$. Let us define $\zeta_0 = \sup\{1 \geq \zeta | y_c(\zeta) > 0\}$, set

$$y^g(\zeta) = y_c(\zeta) = \frac{t + bt - (1-\alpha)\zeta}{\alpha + b}$$

and

$$y^i(\zeta) = \bar{y}(\zeta) = \frac{bt + \alpha\zeta}{\alpha + b}$$

for $\zeta \in [Z, \zeta_0]$ (so ζ is going from ζ_0 to Z). R lets S go when $z > \zeta_0$. For $\zeta \in [Z, \zeta_0]$, R's prosecution decision a_z is random and she lets S go with probability $\frac{b(1-\alpha)}{\alpha+b}$ (alternatively think of $a_z \in [0, 1]$, since as mentioned in section 6.2 all of our statements are still valid). In the last phase S and R plays as in the one-shot revelation game after the pooling signal Z (for simplicity we assume that type $y_c(Z)$ does not confess at time T but lies up to t so we have the increasing

[34] This is exactly the reason why we had to construct the out of equilibrium behavior of S carefully and keep the belief of R at α on an interval including $[t, \bar{y}(Z))$. Otherwise some type z close to t could "pretend" that her evidence is relatively weak by sending some signal ζ such that $\zeta >> z$. She could "convince" types of S who lie below z on the equilibrium path to lie above z. She can then catch these lies and when she has to move after messages in the interval $[t, z)$ let S go because she becomes more confident (in fact now sure) about his innocence.

lying function).

Checking equilibrium conditions for R is simple. Indeed, if $z > \zeta_0$ it is sequentially rational to let S go and by definition of the interrogation policy when the action a_z must be taken R knows that S's type is between $y_c(z)$ and $\bar{y}(z)$ so she is indeed indifferent between prosecuting and letting S go. In the last phase, by definition R plays sequentially rationally, where the same is true for S. What is left is to check is that the no lying conditions hold. First, direct calculations show that when $z \leq \zeta$ is revealed the expected payoff of each y still in the game (i.e. $y \in [y_c(z), \bar{y}(z)]$) is:

$$\frac{\frac{(\zeta(1-\alpha)-t-bt)b}{\alpha+b} + by}{\zeta - y}.$$

Notice that this expression is 0 for $y = y_c(\zeta)$ and 1 for $y = \bar{y}(\zeta)$, strictly decreases in ζ for all $y > t$, strictly increases for all $y < t$ and it is constant with value $\frac{b(1-\alpha)}{\alpha+b}$ for $y = t$ (remember that ζ decreases in equilibrium). Now any y who decides to claim that he is $y^i(\zeta) = \bar{y}(\zeta) > y$ or decides to play as if he was any $y' > y$ gets

$$\frac{\frac{\frac{(\zeta(1-\alpha)-t-bt)b}{\alpha+b}+by'}{\zeta-y'}(\zeta - y') - b(y' - y)}{\zeta - y},$$

which is exactly his equilibrium payoff, and in fact these are S's baseline equilibrium payoffs as well. What is left to check is that no guilty type y can profit from waiting with his confession, say until $z \leq \zeta$ is revealed, and then claiming that he is $\bar{y}(\zeta)$ if the interrogation continues until then. But this deviation is payoff equivalent to play as if he was type $y_c(\zeta)$, which as we checked already is not beneficial. In fact, all types of S are completely indifferent with respect to their strategy during the interrogation, as they always get their baseline equilibrium payoff.

We now show that the given policy is indeed optimal. R's payoff is completely determined by the payoffs the different types of S expect before R reveals any information. Given that R has no control over her payoff when S's type is in $[y^g(Z), y^i(Z)]$ the question is how much types of S outside this interval should expect to get to have R's payoff maximized. The binding no-lying condition, however, immediately determines these expectations once the expectation of an arbitrary type from this interval, say type t's, is pinned down. Direct calculation shows (exactly as in the proof of proposition 6, see also the intuition in section B.6.1) that R's payoff is maximal if type t's and hence all types' expected payoff is exactly as in the baseline equilibrium and that our policy provides exactly these expectations for S. It follows that R's optimal payoff is what R gets under the optimal mechanism (see proposition 6) minus α times the type II errors she makes due to compensating the punished caught lies when $z < y^i(Z)$. Hence, as Z

converges to t, this payoff converges to the one under the optimal mechanism.

We now argue that the optimal policy is uniquely pinned down. Given that R must randomize to provide these payoffs for S when $z = \zeta$ and R knows only that S's type is in $(y^g(\zeta), y^i(\zeta))$ it follows that R must be indifferent. Among the many possible policy in which R is indeed indifferent, the one that we have singled out gives the most information to R, because under more informative such policies $y^g(\zeta)$ could profitably claim that he is $y^i(\zeta)$ for any $\zeta \in [Z, \zeta_0]$.

We still have to show that R prefers a lower Z. This is also simple. Let us choose a $Z' \in (t, Z)$. We cannot say that all zs are strictly better off because for types $z \geq Z$ nothing changes, the screening phase is the same till $\zeta = Z$. Also types below $\bar{y}(Z')$ are indifferent between the two options because even though types $y \in [y_c(Z'), y_c(Z)]$ do not reveal themselves under Z but they do under Z', a type z below $\bar{y}(Z')$ will catch the same measure of guilties in a lie under Z. These types of R do not care about identifying "less" innocents because those are all below their own z. Types $z \in [\bar{y}(Z'), Z)$ however, strictly prefer Z' to Z because the pooling region shrinks and they can identify more types of S on both sides. Finally, notice that when $Z = t$, we have that $y^g(t) = y_c(t) = y^i(t) = \bar{y}(t) = t$, so t cannot reveal himself.

B.6 Proof of proposition 6

After providing an intuition for the relation between the optimal mechanism and the equilibrium (section B.6.1), we prove its optimality in the class of mechanisms considered in the main body of the paper (section B.6.2). We then show how arbitrary mechanisms offer no improvement (section B.6.3).

B.6.1 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 3a) or the vertical axis (the case of figure 3b). 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 α. 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, 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.

B.6.2 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 3a, 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

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

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

$$\hat{z}^{\star}_{y_c>0}(t) = \frac{\alpha + b(t + \alpha - t\alpha)}{b+\alpha}. \tag{47}$$

Suppose now that $(1 + b)t \leq 1$, instead, so that $y_c(t) = 0$, i.e the case of figure 3b 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

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

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 (47) 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.

B.6.3 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 type, R makes no error. However, when $-b$ is given to an innocent type, 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.

B.7 Proof of proposition 7

First we complete the description of the equilibrium. S sends the message m as in the baseline equilibrium using the strictly increasing lying function $\boldsymbol{\ell}$. When S confesses then the game is over, with automatic action $a = 0$ and S gets 0 (there is no need to punish detected false confessions). When $m \geq \bar{y}$, if S is not caught in a lie R lets him go believing that he is surely innocent, while if S is caught in a lie R immediately prosecutes him and punishes him at the level of $-b$ believing that he is surely guilty. Consider now some message $m < \bar{y}$ and the corresponding guilty type $y = \ell^{-1}(m)$. When $z \leq y$ then R immediately prosecutes him and punishes him at the level of $-b$ believing that he is surely guilty.[35] When z is such that $y < z \leq m + b(m - y) = \zeta_m$ then R proves to S that her $z \leq \zeta_m$ and the game proceeds to stage two, in which case R knows that S is guilty if he was caught in a lie and otherwise believes that S is innocent with probability α. Finally, when $\zeta_m < z \leq \hat{z}^*(y)$ then R prosecutes S and when $z \geq \hat{z}^*(y)$ then R lets S go. In both cases R believes that S is innocent with probability α. Suppose the game proceeds to stage two. When $m' = m$, then R lets S go if he was not caught in a lie, in which case she believes that S is surely innocent, while if S was caught in a lie R automatically prosecutes him and punish him at $-b$. If $m' = 0$ then the game is over, with automatic action $a = 0$ and S gets 0. Finally, in stage two all guilty types send the message 0 and all innocent types send the message m.

We now verify that the strategy of R is indeed a strategy. Namely, we have to show that

[35] An alternative and equivalent solution would be to punish type $y' < y$ when he is asked to confess in stage 2 but he confesses that he is some type different from y or he claims that he is y but this lie is detected.

$m + b(m - y) = \zeta_m \leq \hat{z}^*(y)$. This is clear from the fact that y expects a non-negative payoff in the baseline equilibrium just as in the optimal mechanism. Formally, we have that $\hat{z}^\star(y) = \bar{z}(m) + b(m - y)$ and $\bar{z}(m) \geq m$. Payoffs are exactly as in the optimal mechanism: $m \in [t, \bar{y})$ gets $b(m - y) + 1 - \hat{z}^\star(y) = 1 - \bar{z}(m) = 1 - \hat{z}^\star(m)$, and $y \in [y_c, t)$ gets $1 - \hat{z}^\star(y)$.

Given S's strategy it is clear that R's strategy is optimal (type $z > \zeta_m$ of R of course would like to send the signal ζ_m but she cannot). Given R's strategy we have to consider 8 types of possible deviations for S. Let $\ell(y) = m$ and $\ell(y') = m'$. (1) a guilty type y behaves as a guilty type (a) $y' < y$ or (b) $y' > y$. (2) a guilty type y behaves as an innocent type (a) $m' < m$ or (b) $m' > m$. (3) an innocent type m behaves as an innocent type (a) $m' < m$ or (b) $m' > m$. (4) an innocent type m behaves as a guilty type (a) $y' < y$ or (b) $y' > y$. 1(a), 3(a) and 4(a),(b) are clearly suboptimal for S. We show that in the rest of the cases S obtains exactly his equilibrium payoff. 1(b) gives $1 - \hat{z}^\star(y') - b(y' - y) = 1 - \hat{z}^\star(y)$. 2(a) gives $1 - \hat{z}^\star(y') + b(m' - y') - b(m' - y) = 1 - \hat{z}^\star(y)$. 2(b) gives $1 - \hat{z}^\star(y') - b(y' - y) - b(m' - y') + b(m' - y') = 1 - \hat{z}^\star(y)$. 3(b) gives $1 - \hat{z}^\star(y') + b(m' - y') - b(m' - m) = 1 - \hat{z}^\star(y') + b(m - y') = 1 - \hat{z}^\star(y) + b(y' - y) + b(m - y') = 1 - \hat{z}^\star(y) + b(m - y) = 1 - \hat{z}^\star(m)$.

C Truth-leaning and honesty

Consider the game described at section 2.1 with the following more general payoff of S:

$$a - b_\ell \mathbb{1}_{m \geq z} - b_s \mathbb{1}_{m = s \,\text{and}\, z \leq t}, \tag{48}$$

where $\mathbb{1}_{m \geq z}$ and $\mathbb{1}_{m = s \,\text{and}\, z \leq t}$ are indicator functions for when S is caught in a lie and for when he is silent but the evidence is conclusive, respectively, and b_ℓ and b_s are commonly known parameters such that $b_\ell \geq b_s \geq 0$.

We say that $\langle \boldsymbol{M}, \boldsymbol{a} \rangle$ is a **quasi-equilibrium** if $\boldsymbol{M} = (M_y)_{y \in [0,1]}$ and (1) 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 $\boldsymbol{a}$ and (2) if $m \in M_y$ for some y then $a(m, z) = 0$ if $m \geq z$ and there is no $y' \in [t, z)$ such that $m \in M_{y'}$. Notice that R's action can be arbitrary off the path (i.e. when $\nexists y : m \in M_y$) and most of the time on the path as well, the only restriction is that when R knows that an on path lie was sent by a guilty for sure then she must prosecute him as described by (2). 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 2.2 represents the expected payoff of type m from choosing message m determined by $\boldsymbol{a}$. For $m = s, A(s)$ can be arbitrary and 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 $m \neq s$ we write $A_y(m) = \int_{[y,1]} a(m,z)/(1-y)\mathrm{d}\lambda$ and write ${}_yA(m) = \int_{[y,m]} a(m,z)/(1-y)\mathrm{d}\lambda$ when $m > y$. Notice that $A_m(m) = A(m)$ and $A_y(m) = (1-m)A(m)/(1-y) + {}_yA(m)$ when $m > y$.

A **truth-leaning test sequence** of the baseline game with $(\varepsilon^n_y)_{y\in[0,1]} \to 0$, where $\varepsilon^n_y \geq \varepsilon^n > 0$ for all y and for all n, is a sequence of games such that each type $y \in [0,1]$ obtains an extra $\varepsilon^n_y > 0$ when choosing $m = y$ relative to its baseline game payoff and if for each $m \in [0,1]$ we have that: if $t \leq m \notin \cup_{y<t} M^n_y$ then $A^n(m) = 1$ and if $t > m \notin \cup_{y \geq t} M^n_y$ 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 $y \in [0,1], M_y \subseteq \limsup_{n\to\infty} M^n_y = \{m \in \mathcal{M} | \exists (n_k)_{k\in\mathbb{N}}, m_{n_k} \in M^{n_k}_y : \lim_{k\to\infty} m_{n_k} = m\}$ and for all $m : A^{n_k}(m) \to A(m)$.

Proposition C.0. *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 at 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 $\exists y \geq t, m \neq y : m \in M_y$. Then we must have $m^n \neq y$ in M^n_y for some n in the nearby equilibrium. Consider in this nearby equilibrium the smallest $y \geq t$ such that we have $m^n \neq y$ in M^n_y.[36] There must be some $y' < t : y \in M^n_{y'}$. 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

[36] If the minimum does not exist then the following argument goes through for some y with this property sufficiently close (less than $\varepsilon^n_y \geq \varepsilon^n$) to the infimum of these y-s.

saying that y' weakly prefers y to m^n, yielding respectively:

$$\begin{aligned}
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-m^n)A^n(m^n)+{}_yA^n(m^n)(1-y)-b_\ell(m^n-y)\\
(1-m^n)A^n(m^n)+A^n_{y'}(m^n)(1-y')-b_\ell(m^n-y') &\leq (1-y)A^n(y)+A^n_{y'}(y)(1-y')-b_\ell(y-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A^n_y(s)\\
(1-t)A^n(s)-b_s(t-y') &\leq (1-y)A^n(y)+A^n_{y'}(y)(1-y')-b_\ell(y-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A^n_y(m^n)=(1-m)A^n(m^n)\\
(1-m^n)A^n(m^n)+A^n_{y'}(m^n)(1-y')-b_\ell(m^n-y') &\leq (1-y)A^n(y)+A^n_{y'}(y)(1-y')-b_\ell(y-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A^n_y(m^n)=(1-m^n)A^n(m^n)\\
(1-y')A^n_{y'}(m^n)=(1-m^n)A^n(m^n) &\leq (1-y)A^n(y)+A^n_{y'}(y)(1-y')-b_\ell(y-y').
\end{aligned}$$

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 of the inequality for y is less than or equal to the LHS of the inequality for y' by noting that $A^n_{y'}(y)(1-y')=0$ (or arbitrarily close to it, i.e. less than ε_y^n - see footnote 36) by (2) from the definition of quasi-equilibrium because no innocent type is lying to y since y is the smallest such type and that clearly $A^n_y(m^n)(1-y)\leq A^n_{y'}(m^n)(1-y')$ holds. For the second pair of inequalities one should also use the facts that $b_\ell\geq b_s$ and $(1-t)A^n(s)\geq(1-y)A^n_y(s)$. In all the cases 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^n to y because y' does not obtain the extra ε_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 ε^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 exactly the same as the equilibrium in proposition 1, no matter whether there are silent types or not.

Consider hence 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 $[t,\bar{y})$. For simplicity we describe the proof for a fixed parameter combination, which can be then easily generalized. Consider the case with parameter values: $Z_s = 1, \alpha = t = 1/2, b_\ell = b_s = 1$. In this case $y_c = 1/6, \bar{y} = 5/6$, and $A(1/2) = 2/3$. 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 ε_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: $A^n(1/2) = 2/3 + \eta^n, y_c^n = y_c + \eta^n$, 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 y_c and 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.[37] □

C.1 The full-fledged game

Consider now the most general game where S's payoff depends directly only on the action of R, and only for simplicity and brevity take $b_\ell = b_s = b$. Thus, we allow R to take actions from the set $\{-b, 0, 1\}$ and S's payoff is equal to R's action. We tailor property (2) of the quasi-equilibrium defined above to the current setup as follows:

(2a) $a(s, z) \in \{0, 1\}$ if $z > t$ and $a(s, z) = -b$ if $z \leq t$;

(2b) suppose that $m \in M_y$ for some y. Then $a(m, z) \in \{0, 1\}$ if $m \geq z$ and there is a $y' \in [t, z) : m \in M_{y'}$ and $a(m, z) = -b$ if $m \geq z$ and there is no $y' \in [t, z) : m \in M_{y'}$.

Point (2a) is natural. As for point (2b), it requires on the one hand that if there is an innocent who could have sent the lie m then R will not choose action $-b$ because choosing action 0 is a better response. On the other hand, if the liar is surely guilty then R punishes him at the level of $-b$. Condition (2) is required to hold only after on the equilibrium path messages. Hence, the main difference to the previous game is that now an equilibrium lie of an innocent type will

[37] 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.

not be punished at the level of $-b$ and so sometimes the lies of guilties are covered by the lies of innocents and are not punished. Now we prove the corresponding version of proposition C.0. We only have to prove that in truth-leaning quasi-equilibria innocents are honest and the rest of the proposition trivially follows. The structure of the proof is the same as before but now we have to write the inequalities as follows:

$$\begin{aligned}
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-m^n)A^n(m^n)+A_y^n(m^n)(1-y)\\
(1-m^n)A^n(m^n)+{}_{y'}A^n(m^n)(1-y') &\leq (1-y)A^n(y)+A_{y'}^n(y)(1-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A_y^n(s)\\
(1-t)A^n(s)-b(t-y') &\leq (1-y)A^n(y)+A_{y'}^n(y)(1-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A_y^n(m^n)=(1-m)A^n(m^n)\\
(1-m^n)A^n(m^n)+A_{y'}^n(m^n)(1-y') &\leq (1-y)A^n(y)+A_{y'}^n(y)(1-y');\\
(1-y)A^n(y)+\varepsilon_y^n &\leq (1-y)A_y^n(m^n)=(1-m^n)A^n(m^n)\\
(1-y')A_{y'}^n(m^n)=(1-m^n)A^n(m^n) &\leq (1-y)A^n(y)+A_{y'}^n(y)(1-y').
\end{aligned}$$

We can reach a contradiction in all the four cases by adding $A_{y'}^n(y)(1-y')$ to both sides of the inequalities for y. For the first pair, observe that $A_{y'}^n(y)(1-y')+A_y^n(m_n)(1-y)=A_{y'}^n(m^n)(1-y')$ (or gets arbitrarily close to it) because, using (2b), no innocents smaller than y are lying up to y or to m_n (or those are arbitrarily close to y). Then subtracting the LHS of the inequality for y from the RHS of the inequality for y' and vice versa we get that $-\varepsilon_y^n$ is larger than a quantity which is arbitrarily close to 0. For the second pair of inequalities one observes that $A_{y'}^n(y)(1-y')\leq b(t-y')$ (again using (2b) and (2a) with equality for sure if $y=t$) and that $(1-y)A_y^n(s)\leq(1-t)A^n(s)$. For the third pair, $A_{y'}^n(y)(1-y')\leq A_{y'}^n(m^n)(1-y')$ because $y>m^n$ and y' gets punished more with such a lie using again (2b). For the fourth pair we have exact equality of the RHS of the inequality for y (after adding $A_{y'}^n(y)(1-y')$) and the LHS of the one for y'.