THEMA

théorie économique,
modélisation et applications

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

Designing Interrogations

Alessandro Ispano, Peter Vida

November 2023

CERGY PARIS UNIVERSITÉ

cnrs

Designing Interrogations

Alessandro Ispano Péter Vida

November 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 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 the editor, four anonymous referees, Helmuts Āzacis, Eric Danan, Harry Di Pei, Sidartha Gordon, Jeanne Hagenbach, Frédéric Koessler, Mark Le Quement, 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, University of East Anglia and University of Mannheim 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 better decisions.[2]

Interrogations exhibit two distinctive features that our model seeks 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 typically, while law enforcers privately know the actual evidence gathered, the suspect privately knows 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

[1] At the moment of writing, a search based on the keyword "interrogation" on the popular IMDB internet movie database yields 4477 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.

enforcer's private information can be thought of as 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 and whether to impose him some discretionary additional cost.[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. In equilibrium, innocents are honest. Possibly, some unconfident guilties are honest as well, i.e. they confess, to avoid the risk of being caught in a lie, which entails prosecution and the cost. Instead, sufficiently confident guilties 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) 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. However, doing so is clearly suboptimal ex-post once the suspect has revealed his status, as it occurs in the optimal mechanism. Besides, while the threat of catching and punishing lies makes eliciting information from the suspect possible in the first place, the lying that occurs in equilibrium harms the law enforcer and overall efficiency, since too many guilties are let go.

We then investigate the scope for richer communication protocols to compensate for the law enforcer's lack of commitment over decisions. We demonstrate how the optimal mechanism can be implemented without any commitment in a back-and-forth variation of the baseline model based on the idea of letting the suspect provide his own account and then challenging him with the evidence accordingly (proposition 3). 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. In equilibrium, the law enforcer will use her discretion on whether to punish or forgive

[4] The model equivalently applies to any decision other than prosecution that the law enforcer would want to base on the suspect's guilt, e.g. an arrest or a conviction, and generates disutility to the suspect irrespectively. Until section 4, we abstract from details about the separation of roles in the legal system. There, we also provide interpretations of the cost and explain how insights carry through if the law enforcer's discretion takes the form of leniency instead of punishment.

lies 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. 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 additional cost 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. This decision may also depend on the uncovering of new evidence. Next, we demonstrate how two recurrent legal institutions, namely protection of the suspect's right to silence (proposition 4) and evidence strength standards for interrogating (proposition 5), can alleviate the law enforcer's commitment problem identified above and improve decisions. We also show how the problem can be fully solved when the law enforcer can commit to any arbitrary revelation policy about the evidence before the suspect makes a claim (proposition 6), even though this solution has limited practical appeal because it is not robust to the law enforcer's incentives to understate the strength of the actual evidence. 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, we simultaneously confirm how his results extend to our setting and we offer 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

[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.

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 guilty suspects prefer different strategies.

Second, in contrast to the plea bargaining literature, in which the suspect must typically choose whether to accept the deal the law enforcer proposes, we explicitly model communication between the two parties about information relevant to assess the suspect's guilt. The equilibrium behavior in the second round of the back-and-forth interrogation is reminiscent of screening outcomes in plea bargaining (Grossman and Katz, 1983; Reinganum, 1988) and the judicial mechanism of Siegel and Strulovici (2023), 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.

Third, 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 cheap talk (Crawford and Sobel, 1982). Still, these might be disproved by the law enforcer's evidence. Thus, differently from theoretical models of lying (e.g. Kartik (2009),

[6] In particular, we take as given that guilt and reticence may 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, 2023). 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).

Dziuda and Salas (2018), Balbuzanov (2019) and Jehiel (2021)), the detectability of a lie of the suspect derives explicitly from the law enforcer's private information, which in particular implies that bigger lies are more easily detected.[9] Moreover, our framework allows studying communication protocols with two-sided information revelation and the effects of revelation of the law enforcer's evidence to the suspect.

Our model hence also joins the 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 be hurt by her information since as a result the sender may reveal less. In our setting, absent the possibility that the law enforcer's evidence may disprove the suspect, who may then be punished, 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 communication protocol 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, possibly mediated, communication (e.g. Aumann and Hart (2003), Krishna and Morgan (2004), Goltsman et al. (2009) and Gottardi and Mezzetti (2022)).

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 $y \in [0, 1)$ drawn according to density $f(y)$. S is guilty when $y < t$ and innocent when $y \geq t$, where $t \in (0, 1)$ is a commonly known parameter. Likewise, R privately observes her type, or evidence, $z \in (y, 1]$ drawn conditional on y according to density $f(z|y)$. Thus, the evidence is a signal about S's

[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. 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 Olszewski (2004), 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.

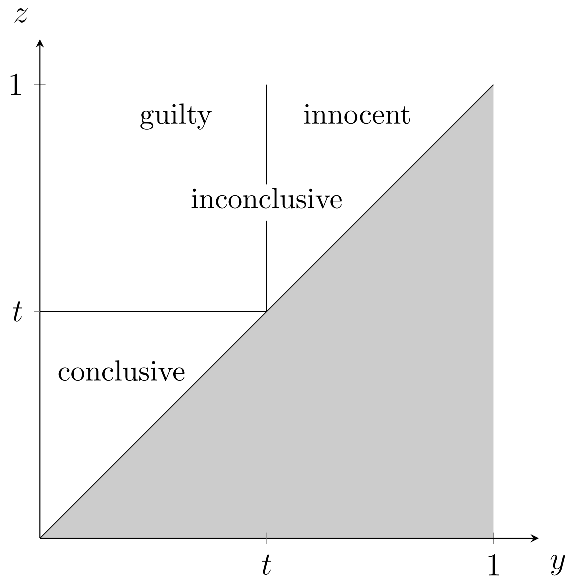

Figure 1 The sample space, the suspect's status and the evidence

type proving that $y < z$. When $z \leq t$ we say that the evidence is **conclusive** since R knows that S is surely guilty (see figure 1). This model can be motivated with very simple stories.[11] For tractability, we restrict our attention to the case in which the private information of each party does not contain information about the private information of the other beyond the fact that $z > y$. Technically, we assume that there is a density $\boldsymbol{g}$ over $[0, 1]$ such that for every y and z with $y < z$

$$f(z|y) = \frac{g(z)}{1 - G(y)}, \tag{1}$$

where $G(x) \equiv \int_0^x g(z)\mathrm{d}z$. Indeed, defining $h(y) \equiv f(y)/(1-G(y))$, the joint distribution of y and z, with support $\left\{(y, z) \in [0, 1]^2 : y < z\right\}$, can be written as $f(y, z) = h(y)g(z)$. The conditional density of y given z is $f(y|z) = \frac{h(y)}{H(z)}$, where $H(x) \equiv \int_0^x h(y)\mathrm{d}y$. Thus, the assumption also implies that the lower the z the lower the conditional expectation of S's type and hence the stronger the evidence in that S's guilt becomes more likely. Likewise, the lower S's type, the stronger the evidence he expects R to possess. Throughout, functions $\boldsymbol{h}$ and $\boldsymbol{g}$ additionally satisfy mild regularity conditions.[12]

[11] 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.

The fact that the evidence provides an upper bound on the suspect's lawfulness, so that for example it can never prove his innocence, captures the idea that whistleblowing, anonymous tips and voluntary reports to law enforcement authorities are inherently incriminating, which may also explain why the suspect qualifies as one and is interrogated in the first place.

[12] Namely, $\boldsymbol{h}$ and $\boldsymbol{g}$ are Lipschitz continuous, bounded from above and bounded away from 0 over the interval $[0, 1]$ and, additionally, $\boldsymbol{g}$ is differentiable with continuous derivative.

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.[13] R then takes an action $a \in \{-b, 0, 1\}$, where $b > 0$ is a commonly known parameter, and payoffs realize as described below.

S's message can be interpreted as 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, we say that he is **caught in a lie** when R's evidence contradicts his claim, i.e. when $m \geq z$. In section 4.2.2, we consider S's possibility to stay silent. R's action can be interpreted as a decision on whether S should be let go freely, i.e. $a = 1$, prosecuted, i.e. $a = 0$, or prosecuted and inflicted some additional cost, i.e. $a = -b$, which we refer to as the **punishment**.

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

$$\alpha \max(a, 0) \, \mathbb{1}_{y<t} + (1 - \alpha)(1 - a) \, \mathbb{1}_{y \geq t}, \tag{2}$$

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 equal to R's action a.

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 above which R finds it optimal to let him go. When S is prosecuted, if he is guilty R makes no loss regardless of whether he incurs the punishment (hence the $\max(a, 0)$ in equation (2)). Instead, R suffers extra disutility when an innocent incurs the punishment since in that case her loss is $1 + b$ instead of 1. S aims at being let go and avoiding the punishment regardless of whether he is innocent or guilty. In section 4.1, we discuss interpretations of the incentive structure and alternative specifications.

2.2 Equilibrium

Equilibrium concept. Throughout, we restrict our attention to Nash equilibria in pure strategies in which innocent types are honest.[14] By Nash equilibrium, R's behavior must be sequentially rational on the equilibrium path. Therefore, on the equilibrium path, after any message which is sent only by guilty types, after a detected lie, or when the evidence is con-

[13] Equivalently, R may observe z only after receiving m, e.g. as a result of an unmodeled verification stage of S's message, which can help explain why R interrogates S even when the evidence is conclusive.

[14] 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).

clusive, R must infer that S is surely guilty and prosecute him, i.e. choose action $a = 0$, or punish him, i.e. choose action $a = -b$. Likewise, upon any message only sent by an innocent type (which cannot be contradicted by the evidence as innocents are honest), R must infer that S is surely innocent and let him go, i.e. choose action $a = 1$. Instead, after messages which are sent by both guilties and innocents, R's action must be sequentially rational when beliefs are calculated according to a generalized version of Bayes rule (see footnote 17 and section B.1 in the online appendix for details).[15] Among pure Nash equilibria in which innocents are honest, we focus on those in which on the equilibrium path R always chooses action $a = -b$ when S is caught in a lie and action $a = 0$ when S confesses and, throughout, the term *equilibrium* denotes an equilibrium in this class.[16]

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 (see corollary 1). Accordingly, the behavior of S is described by a strictly increasing and everywhere differentiable **lying function** $\boldsymbol{\ell} : [y_c, t) \to [t, \bar{y})$ with range $[t, \bar{y})$ 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 understanding that types $y < y_c$ (if any, i.e. if $y_c > 0$) confess honestly. We refer to the range of $\boldsymbol{\ell}$ 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 over $(y_c, \bar{y})$, 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 $z \geq z(m)$ and $a(m, z) = 0$ if $z < z(m)$, i.e. the weakness of the evidence above which S is let go and below which S is prosecuted. 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. Such an equilibrium can hence by described by a pair $\langle \boldsymbol{\ell}, \boldsymbol{z} \rangle$ for which 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.

In particular, given S's strategy, to compute her expected payoff from prosecuting and from letting S go 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

$$\frac{f(\ell(y), z)}{f(\ell(y), z) + \frac{f(y,z)}{\ell'(y)}}.^{17} \tag{3}$$

[15] For the sake of precision, all these properties should hold only with probability one but we assume they hold exactly at all information sets on the equilibrium path to ease exposition.

[16] As shown in section B.2 of the online appendix, these are R's most preferred equilibria (among pure Nash equilibria in which innocents are honest).

Given the assumption at equation (1), so that $f(y,z) = h(y)g(z)$, this probability does not depend on z. Intuitively, given that $f(y,z)$ 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 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 α, so that she is indifferent between actions.

We now establish an indifference condition for S that R's equilibrium strategy in turn will have to satisfy. We say that S is indifferent among denying messages given R's strategy $\boldsymbol{z}$ 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'. It must hold because, again given the assumption at equation (1), if one such type strictly preferred one message to another, so would any other type and the equilibrium construction would break down. Differentiating with respect to m, the condition simplifies to

$$g(z(m))z'(m) = -bg(m). \tag{4}$$

It is then 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 and the consequent punishment that higher lies entail.

It will become handy to choose $\boldsymbol{z}$ so that this indifference condition and equation (4) 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.[18] 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

[17] Formally, by Nash equilibrium, R's indifference between prosecuting and letting S go in any rectangle is given by

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

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

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

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

[18] 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.

be written as

$$1 - G(z(m)) - b(G(m) - G(y)) = 1 - G(z(y)), \tag{5}$$

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 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 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:*

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

$$\frac{h(\ell(y))}{h(\ell(y)) + \frac{h(y)}{\ell'(y)}} = \alpha, \tag{6}$$

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

(ii) in the relevant region $\boldsymbol{z}^*$ is the solution of differential equation (4) with terminal condition $z(\bar{y}^*) = \bar{y}^*$ and $\min\{y_c^*, 1 - G(z(y_c^*))\} = 0$, i.e., more precisely

$$z^*(m) = \begin{cases} 1 & \textit{if } m < y_c^* \\ \textit{the solution of } g(z(m))z'(m) = -bg(m) & \textit{if } m \in [y_c^*, \bar{y}^*) \\ m & \textit{if } m \geq \bar{y}^* \end{cases} . \tag{7}$$

As an illustration, we report closed formed solutions for the equilibrium when the joint distribution of y and z is uniform over the triangle (see again figure 1).

Example 1 (Uniform case)**.** *Let $f(y, z) = 2$, so that $f(y) = 2(1 - y)$, $g(z) = 1$, $f(z|y) = 1/(1 - y)$, $h(y) = 2$, $H(y) = 2y$, $G(z) = z$ and $f(y|z) = 1/z$. Then, $\ell^*(y) = \frac{\alpha}{1-\alpha} y + \bar{y}^* - \frac{\alpha}{1-\alpha} t$ and, for $m \in [y_c^*, \bar{y}^*)$, $z^*(m) = \bar{y}^* + b(\bar{y}^* - m)$, 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^*,

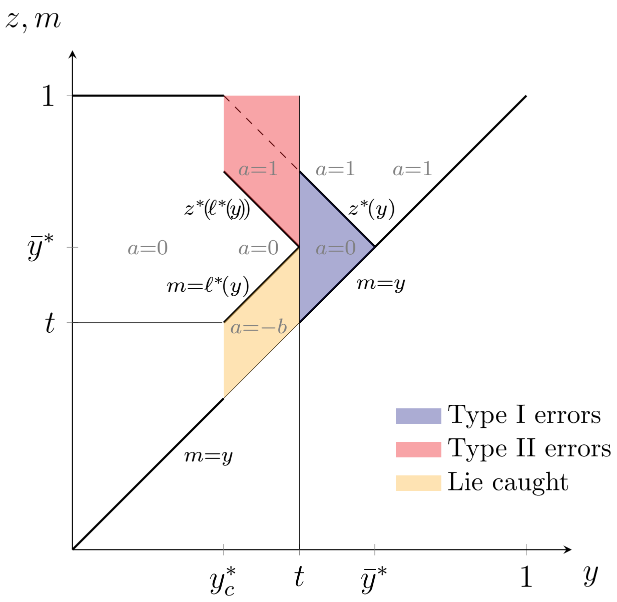

Figure 2 Equilibrium payoffs in the uniform case ($t = 1/2$, $b = 1$, $\alpha = 1/2$)

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.

get $a = 0$ and separating innocent types, i.e. types above $\bar{y}^*$, get $a = 1$, so that R makes no errors. A guilty type above y_c^* is caught in a lie when $z \leq \ell^*(y)$ and in this case he gets $a = -b$. 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.

Proposition 1 also implies that R's ex-ante expected loss in equilibrium is

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

$$= (1-\alpha)\int_t^{\bar{y}^*} (1 - G(y))h(y)\mathrm{d}y = (1-\alpha)\int_t^{\bar{y}^*} f(y)\mathrm{d}y. \tag{8}$$

Equation (8) 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 have focused on equilibrium 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, not only ex-ante, i.e. before S has observed y and R has observed z, but also ex-post.

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

Comparative statics. The model generates some intuitive comparative statics, which are direct consequence of the indifference of the lowest liar with respect to confessing whenever some types confess ($y_c^* > 0$) and the following equation that must hold in any equilibrium

$$\alpha \int_{y_c^*}^{t} h(y)\mathrm{d}y = (1-\alpha) \int_{t}^{\bar{y}^*} h(y)\mathrm{d}y. \tag{9}$$

Indeed, for any given $z \geq \bar{y}^*$, in the lying region R must be indifferent on average as well given that she is indifferent upon any $m \in [t, \bar{y}^*)$.

- When the punishment b is higher, more types confess and the lying region decreases. If $b \to \infty$ all types separate so that the equilibrium becomes fully informative, while if $b \to 0$ the equilibrium becomes uninformative in that R does not benefit from interrogating.
- When R is tougher as measured by a higher weight α she attaches to type II errors, more types confess. Less trivially, the lying region increases, which can be intuitively understood as that R requires more convincing to let S go.[19]
- When the prior probability of innocence is lower as measured by a higher t, more types confess.
- Some types confess, i.e. $y_c^* > 0$, if and only if $G^{-1}(\frac{1}{1+b}) < H^{-1}(\frac{H(t)}{1-\alpha})$, which is more easily satisfied for higher b, α and t.[20]

Comparative statics can also be performed with respect to features of the prior distribution. For example, upon a right shift of the prior distribution of innocent types in the first-order stochastic dominance sense, i.e. when innocents become more confident, more guilties confess and the lying region increases.

[19] Technically, as α increases, equation (9) 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^* strictly 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.

[20] The inequality obtains by calculating $\bar{y}$ assuming $y_c = 0$ in equation (9), which yields $\bar{y} = H^{-1}(\frac{H(t)}{1-\alpha})$ (if such value does not exist, we adopt the convention that it is equal to 1) and then imposing that the expected payoff of type $y = 0$ from lying at such $\bar{y}$, which can be calculated using the LHS of equation (5), is negative.

Finally, in the context of this baseline 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, as $\mathbb{P}(y \geq t|z) = (H(z) - H(t))/H(z)$, given her preferences at equation (2), R finds it optimal to prosecute S if

$$z < H^{-1}\left(\frac{H(t)}{1-\alpha}\right) \tag{10}$$

and to let him go otherwise. By equation (9) evaluated in $y_c^* = 0$, this cutoff 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 $G^{-1}(\frac{1}{1+b}) < H^{-1}(\frac{H(t)}{1-\alpha})$.*

Still, in what follows, we will show that there are ways to make interrogations always useful for R. These rely on richer communication protocols (section 3.2 and 4.2.4) or even simple instruments such as protection of the suspect's right to silence and evidence strength standards for interrogating (sections 4.2.2 and 4.2.3).

3 The optimal interrogation

In this section, we first consider the mechanism design problem of maximizing R's payoff with full commitment. We then show how to implement the optimum without any commitment with back-and-forth communication.

3.1 Optimal mechanism

In contrast with section 2.2, we here 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.

Formally, a mechanism consists of a measurable message space $\tilde{\mathcal{M}}$ and 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 evidence and $\Delta(\{-b, 0, 1\})$ is the set of all probability distributions over R's possible actions. While the space of such mechanisms is large, to determine R's optimum it suffices to consider a 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 $z \geq z(y)$, $a(y,z) = 0$ if $z \in (y, z(y))$ and $a(y,z) = -b$ if $z \leq y$, i.e. R lets S go if the evidence is sufficiently weak relative to his claim, prosecutes him otherwise, and additionally punishes him upon detecting a lie. Such a mechanism satisfies the **truth-telling constraint** if for every $y, y' \in [0,1]$ such that $y < y'$

$$1 - G(z(y)) \geq 1 - G(z(y')) - b\left(G(y') - G(y)\right). \tag{11}$$

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

$$\alpha \int_0^t (1 - G(z(y)))h(y)\mathrm{d}y + (1-\alpha)\int_t^1 \left(G(z(y)) - G(y)\right)h(y)\mathrm{d}y \tag{12}$$

subject to the truth-telling constraint (equation (11)). 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

$$g(z(y))z'(y) = -bg(y). \tag{13}$$

This constraint is just S's equilibrium indifference condition, i.e. equation (4). 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). Therefore, the optimal mechanism and the equilibrium only differ in S's behavior.

Proposition 2 (Optimal mechanism)**.** *R's equilibrium strategy $\boldsymbol{z}^*$ at proposition 1 (i.e. equation (7)) 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 being caught. It turns out that this is the only source of inefficiency. Indeed, in equilibrium 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 in equilibrium.

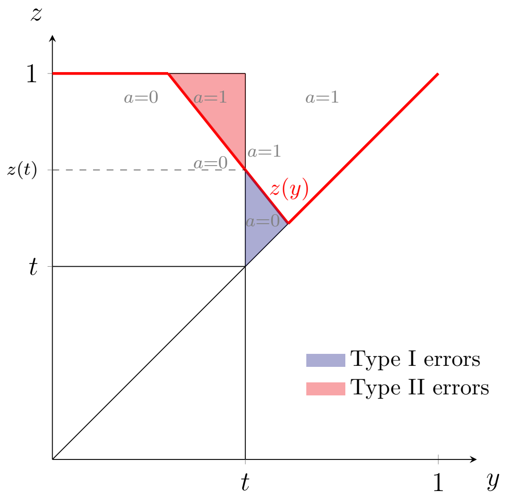

Figure 3 Determination of the optimal mechanism in the uniform case

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 (11) 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 (11) 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.

Thus, due to truth-telling, the optimal mechanism sometimes prescribes a wrong decision for a type whose status as innocent or guilty is certain, i.e. to let go a confessor surely known to be guilty and to prosecute an honest denier surely known to be innocent. This feature is likely to limit its practical appeal and help explain why richer interrogation protocols are typically used in practice.

3.2 Implementation without commitment

We now 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 features "back-and-forth" communication between S and R. In modeling R's communication about the evidence 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 that does not exceed the true one. Technically, any type z of R can send a signal $\zeta \in [z, 1]$ to S and message $\zeta = 1$ can be thought of as nondisclosure.

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 $a \in \{-b, 0, 1\}$, so that the game ends and payoffs realize, or sends a signal $\zeta \in [z, 1]$ and the game continues to stage 3;
- **stage 3** S sends a new message $m' \in \{0, m\}$;
- **stage 4** R chooses an action $a' \in \{-b, 0, 1\}$ and payoffs realize accordingly.

Proposition 3 (Implementation without commitment)**.** *There is an equilibrium of the back-and-forth game in which ex-post expected payoffs are as in the optimal mechanism.*

We provide here an intuitive description of the equilibrium while the formal and exact details are in section A.6 of the appendix. First, we describe the behavior on the equilibrium path. Then, we discuss the incentives which drive the equilibrium together with off the equilibrium path behavior.

S's behavioral strategy in stage one is as in the equilibrium at proposition 1. Innocents are honest, types below y_c^* confess and guilty types above y_c^* lie according to the lying function $\boldsymbol{\ell^*}$. R immediately takes the correct action for separating types, i.e. upon messages below y_c^* she prosecutes and upon messages above $\bar{y}^*$ she lets R go. Instead, after any message m in the lying region, i.e. after pooling messages, there are two possibilities. If the evidence is sufficiently strong relative to S's claim m, R continues the interrogation. Namely, there is a threshold ζ_m such that if $z \leq \zeta_m$ then R proves that her evidence is at least as strong as ζ_m by sending the signal ζ_m and offers S the possibility to withdraw his lie in stage three. 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 that you are type m?" and message 0 as a confession. A guilty withdraws his lie and an innocent sticks to his story and R again takes the correct action. If instead the evidence is not strong enough, i.e. when $z > \zeta_m$, R stops the interrogation and makes an immediate decision. In particular, R prosecutes S if the strength of the evidence is moderate, i.e. if $z \in (\zeta_m, z_m]$, and lets S go if her evidence is weak, i.e. if $z \in (z_m, 1]$, for some threshold value z_m which also depends on the message m.

Let us now consider the incentives of R. When the evidence is strong, i.e. when $z \leq \zeta_m$, R is completely happy to continue the interrogation and prove that her evidence is stronger than ζ_m because she knows she will make no mistake given S's strategy. This observation remains true even if R has caught S in an equilibrium lie in stage one. Instead, when the evidence is weak, i.e. when $z > \zeta_m$, R cannot continue the interrogation in such a way to provide a guilty S with the proper incentives to withdraw his lie. S would stick to his story in case he is confronted with a signal that does not prove that the evidence is at least as strong as ζ_m. It is because, given S's lying strategy in stage one, ζ_m is constructed in such a way that if $z > \zeta_m$ then no lies are ever detected. Hence, R would reveal that S was surely not caught in a lie. It follows that R is just indifferent between prosecuting and letting S go or continuing the interrogation with some weaker, unconvincing proof. R's information about S is, or would remain, just the same as in the equilibrium at proposition 1 when S is not caught in a lie.

Finally, we turn to the incentives of S. The values of z_m and ζ_m are constructed in such a way that S's payoff from following his equilibrium strategy is exactly the same as what he expects from telling the truth in the optimal mechanism. More precisely, an innocent type m expects to be prosecuted only when $z \in (\zeta_m, z_m]$ and a guilty type y is let go only when $z \in (z_{\ell^*(y)}, 1]$. Any possibly profitable deviation of some type y from the equilibrium strategy is identical to the equilibrium behavior of a type $y' > y$. Given that R knows S's communication strategy in stage one, these deviations are detected exactly when $z \in (y, y']$ and are punished with action $-b$ just as in the optimal mechanism. More precisely, off the equilibrium path lies in stage one are immediately punished with action $-b$ in stage two once the lie was identified as an off the equilibrium path one. A detected equilibrium lie which is not withdrawn in stage three is also punished with action $-b$ but only in stage four. Only withdrawn detected equilibrium lies are forgiven. It follows that any deviator of type y expects the same payoff as he was reporting type $y' > y$ in the optimal mechanism, which is clearly not profitable. In fact, on the equilibrium path a guilty type y who lies in stage one to $m = \ell^*(y)$ is just indifferent between withdrawing his lie or sticking to his story that he is type m once confronted with the signal ζ_m. It is because in the optimal mechanism type y is just indifferent between reporting y or m.

4 Discussion

4.1 The incentive structure

Interpretations of the punishment. In our model, R can take, or threaten to take, a more unfavorable action, i.e. inflict punishment b, when S is proving uncooperative. It has long been recognized that law enforcers may resort to such discretion even when they do not have formal authority, most notably in the case of police officers.[21] Such discretion can take many different

[21] 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).

forms, e.g. a flashier arrest or longer detainment in custody, and also be exercised indirectly by influencing the decisions of other parties along the prosecution process, e.g. a recommendation of higher bail or more severe charges. The punishment can also be interpreted as S's cost for being caught in a lie which, depending on the context, may be direct and explicit, e.g. if lying entails a penalty or constitutes an independent offense, indirect, e.g. if the jury is more severe with a reticent suspect, include reputational damages, e.g. if S's loss of credibility compromises his position in other investigations, and also incorporate psychological costs associated with S's dishonesty being exposed. Indeed, with small qualifications, the entire analysis carries trough in each of these alternative scenarios:

(i) S can incur the punishment even when he is let go, i.e. R can also choose action $a = 1 - b$ (see section B.2 of the online appendix);

(ii) the intensity of prosecution and punishment can vary, i.e. R can choose any action $a \in [-b, 1]$ (see again section B.2 of the online appendix);

(iii) S incurs cost $-b$ only if R chooses to expose that S has been caught in a lie;

(iv) R only decides whether to let go or prosecute S, i.e. $a \in \{0, 1\}$, with the objective to avoid errors, i.e. minimize

$$\alpha\, a \mathbb{1}_{y<t} + (1-\alpha)\,(1-a)\,\mathbb{1}_{y\geq t}, \tag{14}$$

and either:

(a) S automatically incurs cost $-b$ whenever he is caught in a lie and does not withdraw it, i.e. his payoff is $a - b\mathbb{1}_{m\geq z}$ in the baseline model and $a - b\mathbb{1}_{m\geq z \& m'=m}$ in the back-and-forth game;

(b) a third party (with preferences as at equation (2)) chooses whether to inflict S the additional punishment $-b$.

No punishment. Our results admit the possibility that there is no punishment or cost, i.e. $b = 0$, as limit case. However, 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.

Leniency instead of punishment. Suppose now that S incurs no punishment but, if he confesses (and only in such case), he obtains a premium $u \in (0,1)$, e.g. a plea bargaining deal. Mirroring R's preferences at equation (2), we specify that when S obtains u then 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 in the leniency setup is *the same* as in the punishment setup. This comparison illustrates in which sense parameter b can also be thought of 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. 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 should also be able to be lenient with denying types of S when the interrogation stops early and the evidence is weak.

Preferences of the law enforcer. We have assumed that R aims to minimize 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. And for normative statements, with scenario iv(a) above in mind, R's preferences at equation (2) can be thought of as reflecting those of society. Importantly, welfare results do not rely on our exact specification of preferences over S's punishment, i.e. an extra error of size b if S is innocent and no error if S is guilty. Insights are robust to other sensible alternatives, including some taste or distaste for punishing a guilty.[22] Finally, as formalized in section 4.2.1 below, R's preferences may concern other decisions beyond prosecution further down the prosecution process such as judicial errors.

4.2 Extensions

Throughout this section, which covers four extensions, we adopt the simplification discussed at scenario iv(a) above that R only makes a prosecution decision and S additionally incurs cost b whenever caught in a lie.

[22] Letting $(1-\alpha)(1+x)$ and αw denote the designer's loss when an innocent and a guilty get $-b$, respectively, with $x \geq 0$, a simple sufficient condition for the revelation principle to hold and the optimal mechanism to be unaffected relative to our model ($x = b$ and $w = 0$) is that $x \geq b$ and $w \geq -b$.

4.2.1 Continuation of the prosecution process

In this section, we show how our model 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. the court's decision c to convict ($c = 0$) or acquit ($c = 1$) replaces R's action a in equation (14). 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 let S go if and only if $(1-\mu)\alpha\sigma(z, m)+\mu(1-\alpha)(1-\sigma(z, m)) \geq \alpha(1 - \mu)$. The inequality simplifies to $\mu \geq \alpha$, i.e. the same threshold as in the baseline model, so that equation (14) still captures her incentives. Moreover, once S's incentives are also taken into account, the equilibrium construction of the baseline model easily adjusts.

Indeed, suppose S's payoff is given by the court's decision c instead of R's action a and, for simplicity, that conviction is certain, i.e. $\sigma(z, m) = 0$, 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

$$1 - G\left(z^*\left(m\right)\right) = \int_m^{z_\sigma(m)} \sigma(z, m)g(z)\mathrm{d}z + 1 - G\left(z_\sigma(m)\right).$$

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 4.2.2).

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 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 in that she will be less inclined to let S go. Indeed, 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. For simplicity, we take $\beta = 0$ and $\gamma = 1$ when S confesses or is caught in a lie and $\beta \in (0, 1)$ and $\gamma \in (0, 1)$ constant otherwise. R now prefers to let S go if and only if $(1 - \mu)\alpha\beta + \mu(1 - \alpha)\gamma \geq \alpha(1 - \mu)$. 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, $\alpha' \geq \alpha$. 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

$$1 - G\left(z\left(m\right)\right) = \beta\left(G\left(z_\beta(m)\right) - G\left(m\right)\right) + 1 - G\left(z_\beta(m)\right).$$

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 $(1-\gamma)\left(G\left(z_\beta(m)\right) - G\left(m\right)\right) + 1 - G\left(z_\beta(m)\right) \geq 1 - G\left(z\left(m\right)\right).$

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.

4.2.2 Protection of silence

Most legal systems recognize the suspect's right to refuse to answer law enforcers' questions. 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)).

Thus, 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 in that he may just as well confess. 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.

In this flexible specification, Z_s represents the evidence strength standard required to prosecute a silent S.[23] For example, if $Z_s = H^{-1}\left(\frac{H(t)}{1-\alpha}\right)$, it is exactly as if R could not use the informational content of silence and should make her decision relying on the evidence alone (see equation (10)). Seidmann (2005) refers to this case as to the "American game" to contrast it to

[23] Equivalently, Z_s could apply to any message $m \notin [0, t)$, i.e. represent the legal standard for prosecution in the absence of a confession. Indeed, given Z_s, in equilibrium even an S who denies will be always let go whenever $z > Z_s$.

the "English game", under which an adverse inference is always allowed, i.e. the case $Z_s = 1$, 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, possibly lower than the one for being caught in a lie.

The equilibrium of the baseline model easily adjusts for any Z_s and b_s. Sufficiently low guilty types, if any, confess, intermediate guilty types, if any, stay silent, and sufficiently high guilty types 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 expense 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. For simplicity, we focus on the case in which there are no confessors in the baseline model so that, without any protection of silence, the interrogation would be uninformative (see remark 1).

Proposition 4 (Protection of silence)**.** *Suppose $G^{-1}(\frac{1}{1+b}) \geq H^{-1}(\frac{H(t)}{1-\alpha})$. There exists a Z_s such that R's equilibrium payoff is strictly higher than in the baseline model if and only if $b_s < b$. In particular, one such Z_s is then $H^{-1}\left(\frac{H(t)}{1-\alpha}\right)$.*

A level of protection that induces some guilty types to remain silent can 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 proportion 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 4 also suggests how to reconcile these findings.[24] Indeed, for R to benefit from protection of silence it must be 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.

4.2.3 Standards 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

[24] Differences concern both the information structure, e.g. 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, e.g. Seidmann (2005) considers the leniency setup (see section 4.1).

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 incentivizes confession and discourages 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 may dominate. For an appropriately chosen standard, R's payoff always improves when there are no confessors in the baseline model, so that the interrogation would otherwise be uninformative (see remark 1).

Proposition 5 (Standard for interrogating)**.** *Suppose $G^{-1}(\frac{1}{1+b}) \geq H^{-1}(\frac{H(t)}{1-\alpha})$. There always exists a Z_i such that R's equilibrium payoff is strictly higher than in the baseline model. In particular, one such Z_i is $H^{-1}\left(\frac{H(t)}{1-\alpha}\right)$.*

4.2.4 Persuasion

In this section, 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. As such commitment power may partly derive from laws governing communication about the evidence, this approach allows first of all to identify the maximal effectiveness that such law can possibly have. Besides, it sheds lights on the extent to which back-and-forth communication is necessary, and not only sufficient, to reach the optimum when committing to wrong prosecution decisions is non-credible for R.

Accordingly, R commits to an experiment, or **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. The baseline model is then 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.

To determine R's optimum, we can concentrate on simple, deterministic, persuasion rules in which some types of R send a designated, "empty", signal, which we interpret as sending no signal whatsoever, and the rest are partitioned according to the (non-empty) signals they send. Namely, every such signal identifies a unique set of types of R who pool on that signal, i.e. an element of the partition. Furthermore, it is enough to consider persuasion rules for which any element of the partition contains at most two types. Thus, a type who sends a signal either reveals herself or pools with exactly one other type. Formally, such a persuasion rule, which we refer to as a **deterministic matching**, can be described by an appropriate signal realization space $D \subseteq [0,1]$ and a corresponding injective matching function $\boldsymbol{z} : D \to [0,1]$ with the interpretation that each signal $d \in D$ is sent by types d and $z(d)$, thereby identifying the signal with type d in the pool. Types outside $D \cup z(D)$ do not send any signal. Hence, when S observes the signal $d \in D$, he forms belief about R's type being d or $z(d)$.[25]

Proposition 6 (Optimal Persuasion)**.** *The deterministic matching with $D = [y_c^*, \bar{y}^*]$ and matching function $\boldsymbol{z}^*$ as at proposition 1 in equation (7) restricted to D 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, given that the equilibrium implements the optimal mechanism, no lies are ever detected. Thus, after signal t, which is sent by types t and $z^*(t)$ of R, all guilties confess. Hence, 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 more work, one can show that a similar issue would arise under any other optimal persuasion rule so that optimal persuasion cannot be made immune to such deviations. 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, doing so is impossible under the natural assumption at section 3.2 that R's private information is hard, i.e. it can be disclosed, possibly vaguely, or withheld but not fabricated. Under this assumption, these results imply that R cannot attain her optimal payoff in a "short" game in which first R, then S, reveal information.

[25] In fact, such persuasion rule is given by the graph of $\boldsymbol{z}$ (see for example the decreasing portion of the red line on figure 3). A point on the graph represents a pair of types who pool with each other by sending the same signal. Types who are matched with themselves are on the intersection of the graph with the diagonal and these types reveal themselves. All remaining types do not send any signal. We hasten to say that not all pairs $(D, \boldsymbol{z})$ specify a deterministic matching. Beyond the requirement that $\boldsymbol{z}$ is injective, a well-defined deterministic matching must satisfy that if $z(d) \in D$ then $z(d) = d$. For example, $D = [0, 1/2]$ and $z(d) = 1 - d$ describe the persuasion rule in which every type sends a signal, each pair of types $d \in [0, 1/2)$ and $1 - d \in (1/2, 1]$ pool by sending the signal d and separates from the other pairs, and type $1/2$ reveals herself. Similar, although stochastic, persuasion rules can be found in Elliott et al. (2021).

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 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 3.2 can be maximally effective while, as discussed in section 4.2.4, 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 voluntarily 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 observation 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, 4.2.2 and 4.2.3. 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 4.2.2, 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 proportions 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 commitment problems intrinsic to interrogations and remedies 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 to avoid generating incriminating evidence, it seems plausible that, all else being equal, more informative 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. Besides, this improvement need not come at the cost of extorting false confessions since innocents 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

[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 3.2 and 4.2.3) and misrepresent her true preferences over type I and type II errors (increase or decrease S's perception of α).

uncertainty about the interrogator's preferences). Likewise, since false confessions occur (see for instance Leo and Ofshe (1998)), it is important to study how our insights would modify when also some innocents can have fundamental reasons not to be honest, for instance because they expect even stronger evidence than guilty types.

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 of notifying the suspect 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.

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.

Gottardi, Piero and Claudio Mezzetti, "Shuttle diplomacy," *Working paper*, 2022.

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.

Olszewski, Wojciech, "Informal communication," *Journal of Economic Theory*, 2004, *117* (2), 180–200.

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*, August 2023, *15* (3), 243–70.

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 (6), 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 (9) 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 (9) 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 (6) for $\ell'(y)$, and finally using the average indifference condition (equation (9)), 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 (9) 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 payoff equivalence holds in any equilibrium for both S and R, both ex-ante and ex-post.

A.3 Proof of remark 1

When there are no confessors, i.e. $y_c^* = 0$, by equation (9) it follows that $\bar{y}^*$ is precisely equal to the RHS of equation (10), i.e. the evidence cutoff above which R would let S go when

[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.

not interrogating. But in equilibrium, when $z < \bar{y}^*$ R could equivalently always prosecute S, and when $z \geq \bar{y}^*$ R could equivalently always let S go, without changing her payoff. Indeed, R is indifferent between prosecuting and letting S go when S is not caught in a lie, which is always the case when $z \geq \bar{y}^*$. It follows that R's equilibrium payoff is the same as when not interrogating.

A.4 Proof of lemma 1

Consider an arbitrary mechanism and fix the resulting lowest expected loss of R when each type sends only messages which are optimal for him given the mechanism. For each $y \in [0,1]$ let $q(y) \geq 0$ denote the maximum of 0 and the expected payoff of type y when playing in 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 (11)). 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, in the original mechanism, conditional on $z \geq y'$, it must be that y' expects $q(y')$ (or something negative), 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 $y' \geq z \geq y$. Note, that changing the expectations from some negative number to 0 does not affect the argument because all negative expectations (and only those) were set to 0.

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 (9), 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 (9). 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 (9):*

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

Proof. As explained in the body of the paper, by the indifference condition of S, $\boldsymbol{z}$ must satisfy differential equation (13) 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 (12), 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

$$\alpha \int_{y_c(p)}^{t} h(y)g(z(y,p))\frac{\partial z(y,p)}{\partial p}\mathrm{d}y = (1-\alpha) \int_{t}^{\bar{y}(p)} h(y)g(z(y,p))\frac{\partial z(y,p)}{\partial p}\mathrm{d}y. \tag{16}$$

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

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

where $K(p) \neq 0$ is some function constant in y (in fact it is $g(p)$). Thus, the first order condition at equation (16) 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 (15), 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 the same conditions. □

A.6 Proof of proposition 3

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 $z > z^*(y)$ 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 punishes 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 can perfectly set guilties and innocents apart by sending ζ_m. Any deviation of weak types (sending 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

[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.

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 were claiming to be a different type in the optimal mechanism and hence cannot be profitable.

A.7 Proof of proposition 4 and 5

We recall that in the baseline model when there are no confessors $\bar{y}^* = H^{-1}(\frac{H(t)}{1-\alpha})$ and, by remark 1, R's loss is the same as without interrogation.

Proof of proposition 4. The equilibrium construction of the baseline model ($Z_s = 1$) easily generalizes to the case $Z_s < 1$. And as $b_s \leq b$, the smallest innocent type who separates in equilibrium, which we denote by $\bar{y}$, must be weakly lower than Z_s. Consider first the case $b_s = b$. Guilty types weakly prefer to lie, say at $\bar{y}$, than to stay silent. Hence the equilibrium is either equivalent to the baseline equilibrium (when $Z_s \geq \bar{y}^*$) or $\bar{y} = Z_s$ (when $Z_s < \bar{y}^*$) and some guilties stay silent because they are indifferent. In the latter case, R is strictly worse off than without interrogation. The reason is that, for $z < Z_s$, R could just as well always prosecute as in the case of no interrogation and, for $z \geq Z_s$, R could just always let S go even without listening to his message. Hence, the only payoff difference between the equilibrium and the case of no interrogation stems from the decisions of R when $z \in [Z_s, \bar{y}^*)$. In those cases, R makes a suboptimal decision on average because we know that for those z-s, when making a decision independently of what S's message is, i.e. when there is no interrogation, R is strictly better off by prosecuting. Consider now the case $b_s < b$. Guilty types strictly prefer to stay silent than to lie at Z_s and hence $\bar{y} < Z_s$. For $Z_s = \bar{y}^*$, this means that more innocents separate and there will be some silent types. Hence, relative to no interrogation, R makes strictly less type I and type II errors when $z \in (\bar{y}, \bar{y}^*)$ and otherwise makes exactly the same amount of errors.

Proof of proposition 5. The equilibrium construction of the baseline model ($Z_i = 1$) again easily generalizes to the case $Z_i < 1$. And the smallest innocent type who separates in equilibrium, which we denote by $\bar{y}$, must be strictly lower than Z_i. Indeed, a lie at Z_i would be caught with probability one. Thus, for $Z_i = \bar{y}^*$, again $\bar{y} < \bar{y}^*$. It means that more innocents separate and there will be confessors. R makes less errors relative to the case of no interrogation for exactly the same reasons as in the case of $b_s < b$ and $Z_s = \bar{y}^*$ above.

A.8 Proof of proposition 6

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 herself) 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 $\boldsymbol{z^*}$ satisfi

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

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

Letting $\boldsymbol{f}$ denote the joint distribution of y and z and manipulating the RHS,

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

Using our assumptions on $\boldsymbol{f}$, the expressions 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

$$1-\alpha = \alpha\frac{\boldsymbol{h}(\boldsymbol{\ell}^{-1})}{\boldsymbol{h}}\frac{\mathrm{d}(\boldsymbol{\lambda}\circ\boldsymbol{\ell}^{-1})}{\mathrm{d}\boldsymbol{\lambda}}. \tag{17}$$

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 (6) obtains.

B.2 Other equilibria are dominated

In this section, we show that the equilibria we focus on in section 2.2 are R's favorites among all Nash equilibria in pure strategies in which innocents are honest. We show this result holds even when R can choose her action continuously, i.e. $a \in [-b, 1]$ as opposed to $a \in \{-b, 0, 1\}$. We thereby also cover the case in which R may choose to punish S even when she lets him go, i.e. take action $1 - b$.

Lemma B.1. *In any other pure Nash equilibrium in which innocents are honest R's ex-ante expected payoff is weakly lower than in those considered in proposition 1 and corollary 1.*

Proof. In any equilibrium, there is a $\tilde{y} \in (t, 1]$ above which S's types separate while when there is a denial lower than $\tilde{y}$ and S is not caught in a lie, in which case we say R *has discretion*, R is indifferent among any action in $[0, 1]$ or strictly prefers action 0.[29] This is because when R's belief is β, then either 0 or 1 is the optimal action or every action is optimal (when $\beta = \alpha$). Hence R's loss is given by:

$$(1 - \alpha) \int_t^{\tilde{y}} f(y)dy.$$

Assume by contradiction that $\tilde{y} < \bar{y}^*$, so that R's loss is lower than in any equilibrium we consider in the main body of the paper. Then, R cannot ever strictly prefer action 0 when she has discretion. This is because then the innocent S who gets always 0 would have an incentive to deviate up to message $\tilde{y}$. It follows that there must be confessors and the highest confessor is strictly higher than y_c^* by R's indifference. This confessor will have strict incentives to deviate and lie to $\tilde{y}$ even if he gets 0 when confessing and $-b$ whenever he is caught in a lie. This is a contradiction. □

Finally, we note that the proof of lemma 1 remains unchanged in case under arbitrary mechanisms R can allocate to S values from the interval $[-b, 1]$.

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

[29] Note $\tilde{y}$ must be strictly larger than t otherwise there would be a mass of liars at t. Moreover, $\tilde{y}$ is the supremum of messages after which R plays 0 with positive probability in case she has discretion and hence she must also play 0 with positive probability for lower messages. Note also that S will never lie above $\tilde{y}$.

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 3.2 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 3 (section B.3.1) and even attain the optimum if additionally innocents depart 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 type I and type II errors without multiplying them by 2 (the prior 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.

[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.

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, and, 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 some guilty types still 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 another equilibrium of the multistage game. 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 exact construction of the equilibrium 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 3.2, at least under the restriction that innocents are honest. We conjecture that in this setup the longer the interrogation is the closer one can get to the optimal mechanism.

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 3 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 by (together with the appropriate belief that makes this behavior sequentially rational). Then, the optimal mechanism is implemented.