THEMA
théorie économique, modélisation et applications

THEMA Working Paper n°2020-05
CY Cergy Paris Université, France

Custodial Interrogations

Alessandro Ispano, Péter Vida

July 2020

Custodial Interrogations

Alessandro Ispano Péter Vida

July 2020

Abstract

We provide a model of custodial interrogations in which the suspect is privately informed about his guilt and the likely strength of incriminating evidence and law enforcers are privately informed about the actual evidence. The evidence is directly informative about the suspect's guilt and may also disprove his eventual lies. We study how communication in the interrogation and the accuracy of prosecution decisions vary with the scope of protection of the suspect's right to silence, the relative costs of type I and type II errors for law enforcers and the evidence strength standard for interrogating. We also evaluate the scope for deceptive interrogation tactics when the suspect is prone to deception. Finally, we describe the optimal mechanism under full commitment over law enforcers' decisions and a natural sequential game that implements it. Our results offer important insights for the design of the legal system.

Keywords: lie, evidence, leniency, questioning, confession, law, prosecution
JEL classifications: D82, D83, C72, K40

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

1 Introduction

In legal systems based on the rule of law, interrogation of suspects plays an important role in the investigation phase that may lead to prosecution.[1] Due process standards hence govern interrogations to ensure both the respect of the suspect's rights and the admissibility in court of the information law enforcers eventually obtain. This paper adopts the perspective of information economics, which studies the sorts of strategic considerations inherent to interrogations, and has both a positive and a normative objective. On a first level, it aims to provide a formal framework that describes how interrogations unfold based on essential features of the legal system. Besides, it aims to determine which institutions enhance information revelation from the suspect and yield to more accurate prosecution decisions.

Our model captures the conflict of interests between the suspect, who aims to convince law enforcers of his innocence and be let go, and law enforcers, who aim to obtain truthful information from the suspect to minimize type I (prosecuting an innocent) and type II (letting a guilty go) errors. The suspect is privately informed about his status as guilty or innocent as well as the likely strength of the incriminating evidence. Law enforcers are privately informed about the evidence actually gathered in the course of the investigation. First, the evidence is directly informative about the suspect's status. Besides, the evidence may disprove eventual false claims of the suspect during questioning. For his part, the suspect typically enjoys the right against self-incrimination, i.e. he is not obliged to make any claim, but also some leniency for confessing relative to when he is reticent, i.e. he denies and he is caught in a lie or he stays silent and the evidence is unambiguously incriminating.

Initially, we maintain that communication in the interrogation is one-shot and unidirectional from the suspect to law enforcers. The suspect's message must be interpreted as a reply to law enforcers' inquiry about his type, e.g. "what time did you leave that night?", and a sufficiently weak claim amounts to a confession. Under a refinement that gives

[1] For instance, Mueller (1961) refers to the phase of police action against or upon a suspect that precedes the suspect's first contact with a judicial officer as "the most important phase of criminal procedure, for here, much more so than during trial, the case is to be won or lost". The importance interrogations take on is not without controversy. For instance, McConville and Baldwin (1982) criticize how "As the police have become more professionalised, so they have acquired much greater control of the prosecution; and as this has happened, so the really crucial exchanges in the criminal process have shifted from courts into police interrogation."

prominence to honesty, innocent types and confessors are honest in equilibrium (lemma 1). Conversely, some sufficiently unsuspicious guilty types necessarily lie (lemma 2). The model then yields clear predictions on the outcome of the interrogation both in terms of players' strategies (proposition 1) and payoffs (corollary 1). In particular, it allows to determine how the accuracy of prosecution decisions varies depending on the prior likelihood of guilt, leniency for confession, the scope of the protection of the suspect's right to silence and the relative importance law enforcers attach to type I and type II errors. We use this baseline model as a workhorse to explore several normative issues.

We first focus on the right to silence and its scope. Some legal systems explicitly allow an adverse inference, i.e. a negative conclusion, to be drawn from a suspect's refuse to answer. Our results agree with this legal doctrine in that, at least under appealing equilibrium restrictions, only guilty suspects may stay silent. At the same time, our results also provide a justification for the principle that adverse inferences alone are insufficient to trigger prosecution without some additional supporting evidence.[2] Indeed, even though having to let a silent, hence guilty, suspect go by law due to insufficiently strong evidence is clearly suboptimal ex-post, this also reduces the suspect's need for lying. From an ex-ante perspective this effect may dominate and overall yield to more accurate prosecution decisions (proposition 2). Thus, this finding also speaks to the merits of the legal requirement that the suspect is reminded of his right to silence at the start of the interrogation.

We then move to the role of law enforcers in charge of interrogations in the legal system and their incentives. Our results demonstrate how, given any objective measure of the relative costs of type I and type II prosecution errors for society, the accuracy of decisions is always maximized when laws enforcers use a different measure, i.e. when their incentives are biased (proposition 3).[3] The direction of their ideal bias may be towards prosecution,

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

[3]For a discussion on the determinants of society's welfare for the entire prosecution process see Grossman and Katz (1983), Reinganum (1988) and Siegel and Strulovici (2018). While Siegel and Strulovici (2018) adopt a mechanism design approach, both Grossman and Katz (1983) and Reinganum (1988) maintain that the prosecutor's and society's objectives coincide. Our results suggest how in the absence of commitment power over prosecution decisions society may benefit from some misalignment of interests. Reinganum (1988) shows similar benefits from restricting prosecutorial discretion and hence adopts the

i.e. a higher weight given to type II errors, but also towards dismissal. We also show that the accuracy of decisions further increases if the bias can be made contingent on the strength of the evidence and, in particular, tilted towards prosecution when evidence is strong and towards dismissal when evidence is weak. All these insights equally apply at a more micro level, e.g. to the appointment of interrogators within a police department also based on their intrinsic preferences.

Next, we study the effect of legal standards for interrogating, i.e. the minimum evidence strength required for law enforcers to be able to question a suspect. The most apparent effect of standards for search, seizures, arrests and other restraints of the suspect's liberty is to limit law enforcers' power and discretion. However, as already pointed out by Reinganum (1988) in the context of arrests, standards also convey information. Indeed, if a suspect is taken into custody and interrogated, he knows enforcers must have sufficiently strong evidence to do so. We show how under a more stringent standard a guilty suspect is less inclined to lie and more inclined to confess. Thus, even abstracting from legal considerations on restriction of individual freedom or a cost-benefit analysis of law enforcers' resources, the choice of increasing the standard entails a trade-off. If on the one hand the suspect must necessarily be let go when the evidence is too weak, on the other hand interrogations become more informative. The optimal standard often prescribes to not interrogate the suspect when evidence is sufficiently weak, i.e. it is more stringent than in the baseline model (proposition 4). This is always the case if no suspect would ever confess in the absence of the standard.

Throughout, we maintained that all aspects of the strategic environment other than the players' private information are common knowledge and outside law enforcers' control. These are determined by institutional features of the legal system, with which the suspect should be familiar with especially if assisted by an attorney. At the same time, within these rules, law enforcers may still have some leeway to mislead the suspect. The elements of arbitrariness in interrogations are in fact a major cause of criticism and an important reason behind the general movement towards their mandatory recording.[4] Even without formally incorporating asymmetric information about the legal environment, our framework allows

complementary perspective of affecting the choice-set rather than the preferences of the agent.

[4]See for instance Sullivan (2005).

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. The predictions we derive agree with the logic behind common interrogations tactics, which may also be seen as a basic validation test of the model. Law enforcers would always want to overstate the benefits of confession, exaggerate the strength of the incriminating evidence and misrepresent their true preferences over type I and type II errors (proposition 5). While these deceptive tactics are surely undesirable on other grounds, our results suggest that these improve the elicitation of the suspect's information. Also, this improvement need not come at the cost of extorting false confessions.

We also maintained that law enforcers have no commitment power over prosecution decisions. While the legal system can be designed from an ex-ante perspective to promote informed decisions, and many of our previous insights speak to this objective, it seems plausible that some discretion remains for all parties involved in its implementation due to institutional constraints, incompleteness of the law and other informational frictions. Law enforcers will then act upon this discretion in their own self-interest and, for example, will not let go a confessor who they know to be surely guilty. We complement the analysis with the alternative mechanism design approach, which assumes full commitment on the outcome of the interrogation based on the suspect's claim and the evidence.[5] The optimal mechanism has a close link with the equilibrium of our baseline model (proposition 6). The suspect's payoff is the same, while the accuracy of prosecution decisions increases thanks to a reduction in type II errors. We show how the optimal mechanism can be implemented as equilibrium of a natural sequential game built on our baseline model (proposition 7). This game combines features of delegation and evidence strength standards, even though law enforcers' behavior and the associated information revelation about the evidence to the suspect is now entirely dictated by equilibrium considerations rather than fixed by law. The suspect willingly gives away information in the first round of the interrogation anticipating that if the evidence is weak relative to his claim he will be let go. When this

[5] Siegel and Strulovici (2018) adopt this approach in a general framework which encompasses the entire prosecution process. We discuss in some more detail how our results relate to theirs in section 5.2. See also the discussion about the design of the legal system in Hart et al. (2017), who consider a class of persuasion games with one-sided asymmetric information in which the outcome with and without commitment on behalf of the uninformed party is the same.

is not the case, a maximally tough interrogator, i.e. who only cares about minimizing type II errors, will continue the interrogation. A guilty suspect will step back on his lie, which will be forgiven, and an innocent type will stick to his story, which may lead to either prosecution or dismissal depending on the strength of the evidence.

The paper is structured as follows. After a discussion of some related literature here below, section 2 presents our baseline model and section 3 characterizes its equilibria. Section 4 explores normative implications and section 5 presents the mechanism design approach. Finally, section 6 discusses additional extensions and policy questions our framework can address.

Related Literature While the entire prosecution process offers leading applications to the strategic information revelation literature,[6] suspects' interrogation has received little explicit attention.[7] Similarly, the law and economics literature tends to study prosecution and litigation assuming these are already undergoing. In order to concentrate on interrogations, essential features of the prosecution process enter in our model only in very reduced form and we leave aside key considerations on their determination that are the focus of this literature.[8] Conversely, our information structure allows for heterogeneity in the strength of the incriminating evidence suspects expect. This dimension of asymmetric information is acknowledged but left aside by both Reinganum (1988) and Siegel and Strulovici (2018) and is likely to play a particularly important role at the interrogation stage given that the discovery process has not started. Our insights may still be relevant for later stages of the prosecution process since, especially if no additional evidence becomes available to either party, the outcome of the interrogation will be related to the outcome of eventual prosecu-

[6]See for instance the leading example in Kamenica and Gentzkow (2011), which involves a prosecutor and a judge, and the policy implications for the design of the legal system in Hart et al. (2017).

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

[8]In particular, we take as given that guilt entails some punishment and confession entails some leniency without considering the complex determinants of plea bargaining (Landes, 1971; Grossman and Katz, 1983; Reinganum, 1988; Baker and Mezzetti, 2001; Daughety and Reinganum, 2020) and sentencing (Siegel and Strulovici, 2018, 2019). We ignore considerations on crime deterrence (and chilling of socially desirable behavior (Kaplow, 2011)), commensurate punishment, endogenous evidence acquisition and deployment of resources in prosecution. Finally, we abstract from details about the separation of roles between law enforcement and prosecution, which is fuzzier than traditionally thought (Abel, 2016), and simply assume law enforcers directly take prosecution decisions even when they do not have formal authority.

tion and trial.[9] The information the different involved parties acquire and present in front of a judicial officer, including the defendant's claims, is typically modeled as hard evidence (Milgrom, 1981), i.e. it can be disclosed or withheld but not misreported.[10] To allow for the possibility of plain lying that is intrinsic to interrogations, in our model the suspect's claims are soft information, i.e. the set of his available messages is independent from the truth. The suspect's claims are simultaneously not pure cheap talk (Crawford and Sobel, 1982) since these might be contradicted by the evidence of law enforcers, entailing some costs. Our model is hence related to the theoretical literature on strategic communication with lying costs (Kartik, 2009), detectable deceit (Dziuda and Salas, 2018; Balbuzanov, 2019) and costly investigation (Ioannidis et al., 2020). Differently from these works, the detectability of a lie and its costs for the suspect derive explicitly from the private information of law enforcers, which in particular naturally implies that the detectability of a lie increases with its size.[11] Moreover, our framework therefore allows studying the effects of revelation of law enforcers' private information to the suspect, as it occurs in the case of evidence strength standards for interrogating and in the game that implements the optimal mechanism. Our model hence also joins the growing theoretical literature on strategic communication which, departing from seminal works, considers two-sides asymmetric information between the sender and the receiver.[12] It mainly differs in players' incentives and the information structure as well as in the main questions of interest. A recurrent theme in this literature is that the receiver may sometimes be hurt from her information since as a result the sender may reveal less. In our setting, this is never the case since, absent the possibility that the suspect may be caught in a lie or proven guilty, the interrogation would be completely uninformative.

[9] See for instance Redlich et al. (2018), who document how confessions in the interrogation correlate with plea bargaining and sentencing outcomes.

[10] See for instance Shin (1994, 1998), Bhattacharya and Mukherjee (2013) and Hart et al. (2017).

[11] Kartik (2009) assumes a lie entails a direct cost that increases with its size and he invokes penalties upon lying detection as a possible interpretation. In both Dziuda and Salas (2018) and Balbuzanov (2019), instead, any lie has an equal exogenous chance of being detected and the cost is endogenously determined by the receiver's response. In Ioannidis et al. (2020) there is no notion of lie size since the sender can only choose between two messages, each yielding to a different investigation technology for the receiver.

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

2 The Model

Information structure There are two players: a suspect, denoted by S and by convention of masculine gender, and law enforcers, denoted by R and by convention of feminine gender. At the initial stage, S privately observes his type y, which is drawn uniformly from $[0,1]$. S's status $\mathcal{Y} \in \{0,1\}$ depends on whether y is above some fixed cutoff $t \in (0,1)$: $\mathcal{Y} = 0$, i.e. S is **guilty**, when $y < t$ and $\mathcal{Y} = 1$, i.e. S is **innocent**, when $y \geq t$. Thus, t also represents the prior probability that S is guilty. R has some private information about y that depends on an independent uniform draw z from $[0,1]$. When $y < z$, R observes z and knows that $y < z$. In this case, we say that R has **evidence**, which is stronger the lower the z. We say that evidence is **conclusive** if it proves with probability one that S is guilty, i.e. if $z \leq t$. Instead, when $y \geq z$, R does not observe anything and the game ends.[13] Figure 1 displays the three possibilities about R's evidence based on realizations y and z. Each point within the unite square is equally likely and the relevant region for our game is above the 45-degrees line, where $z > y$. While ensuring tractability, our information structure respects two important general features inherent to the evidence identified by Reinganum (1988). First, stronger evidence makes the suspect's guilt more likely. Second, a guilty suspect expects on average stronger evidence than an innocent one. Additionally, in our setting there is heterogeneity both within guilties and innocents in terms of the strength of the evidence they expect.

A story The following story may serve as an illustration and a reminder of the information structure. One night a museum burns down. It is either an accident or arson and the only suspect is the night shift guard. The guard left the building at time y and the fire alarm went off at time t. When $y < t$ the guard is guilty as he could have been aware of the fire only by starting it. Conversely, when $y \geq t$ the guard is innocent since he left the building only in reaction to the alarm. At time z a person peeked from his window overlooking the museum. If $y \geq z$ the person saw none and the police's appeal for witnesses remains unanswered. If $y < z$, the person reports seeing the guard outside at time z.

[13] A first interpretation is that in the absence of the evidence R is not even aware of S. Alternatively, as S's guilt is even less likely than under the prior, R has no legal ground to go after S. Our framework can easily accommodate the alternative scenario in which R can still interrogate S in the hope of obtaining an admission of guilt but must let him go otherwise.

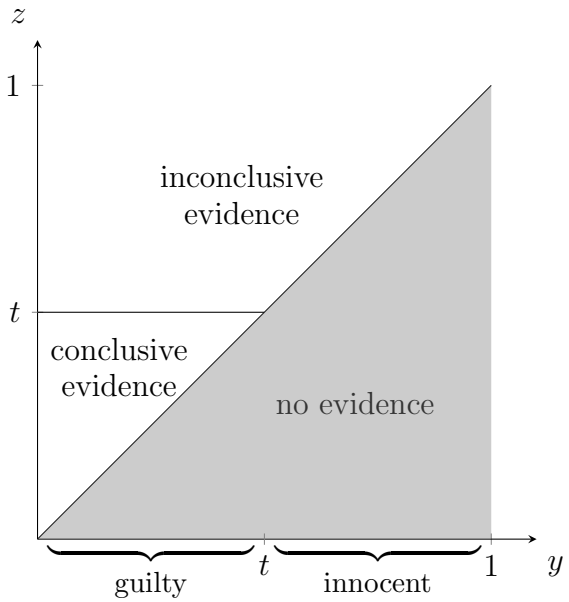


Figure 1 The type space and the evidence

The interrogation After y and z have been drawn and the information structure determined accordingly, provided $z > y$, R interrogates S. Accordingly, S sends a message $m \in \mathcal{M} = [0,1] \cup \{s\}$ to R, who then takes an action $a \in \{0,1\}$. S's message can be interpreted as a literal claim about his type.[14] Message $m = s$ represents the possibility for S to remain **silent**. Provided S is not silent, we say that he **lies** when $m \neq y$, that he **is honest** when $m = y$, that he **confesses** when $m < t$ and that he **denies** when $m \geq t$. Also, we say that he is **caught in a lie** when R's evidence contradicts his claim, i.e. when $m \geq z$. R's action can be interpreted as a decision on whether S should be prosecuted, i.e. $a = 0$, or let go freely without charges, i.e. $a = 1$.

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

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

where $\mathbb{1}_{\mathcal{Y}=0}$ and $\mathbb{1}_{\mathcal{Y}=1}$ are indicator functions for S's status as guilty and innocent, respectively, and $\alpha \in (0,1)$ a commonly known parameter. That is, R aims at prosecuting a guilty suspect and letting an innocent suspect go and α measure the relative importance of

[14] See Kartik (2009) for a way to formalize the notion of literal meaning and also to encompass a richer message space.

a type II error, i.e. letting a guilty go, over a type I error, i.e. prosecuting and innocent.[15] For simplicity and ease of exposition, we impose R always chooses $a = 0$ when evidence is conclusive and/or S confesses and/or S is caught in a lie. Likewise, R always chooses action $a = 0$ upon silence if allowed by law.[16] Namely, if we let $Z_s \in (t, 1]$ be the evidence standard required to prosecute S upon silence,[17] then

$$a(s, z) = \begin{cases} 0 & \text{if } z \leq Z_s \\ 1 & \text{if } z > Z_s. \end{cases} \tag{2}$$

As for S, we normalize his payoff from being prosecuted and being let go respectively to 0 and 1 and we distinguish the following cases

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

By confessing (honestly) S saves $b > 0$ relative to when he lies and he is caught or to when he stays silent and he is directly proven guilty. Thus, b measures the punishment for reticence or, equivalently, the **leniency** that confession entails.[18] Instead, if S remains silent and R has only inconclusive evidence, his payoff depends on the evidence standard

[15] This objective of law enforcers may derive from a combination of a genuine interest in the truth, reputational benefits from accurate recommendations to the prosecution, eventual asymmetries in their compensation schemes and fundamental features of the legal system (e.g. whether it is adversarial or inquisitorial).

[16] In all those instances, R will infer that S is guilty, so that choosing $a = 0$ is indeed sequentially rational. Our specification highly simplifies the notation and avoids having to keep track of too large of a set of off the equilibrium path beliefs of R.

[17] If $Z_s = t$, i.e. if evidence must be conclusive to prosecute S upon silence, the model becomes trivial in that guilty types always prefer to stay silent than to lie. Section 4.1 demonstrates how such a stringent standard would always be detrimental to the accuracy of prosecution decisions.

[18] Leniency may take many different forms. On a first level, for minor crimes many legal systems explicitly provide suspects with the possibility to avoid prosecution in exchange of an admission of guilt and face only lighter consequences. Examples of this are police cautions in the UK and admission of guilt fines in South Africa. Besides, the suspect's confession may be part of a plea bargaining with the prosecutor, in which also law enforcers play an active role (Abel, 2016). Finally, even in the absence of an explicit agreement, juries may consider favorably an early confession in their sentencing.

for prosecuting as defined at equation (2). Finally, when S denies and there is inconclusive evidence that does not contradict his claim, his payoff is determined by R's action. In this case we say that R has **discretion**.

Equilibrium concept A pure strategy of the suspect $\boldsymbol{m} : [0,1] \to \mathcal{M}$ specifies a message $m(y)$ for each type y. A pure strategy of law enforcers $\boldsymbol{a} : [0,1] \times [0,1] \to \{0,1\}$ specifies an action $a(m,z)$ for each message m and evidence realization z. Likewise, her belief system $\boldsymbol{\mu} : [0,1] \times [0,1] \to [0,1]$ specifies a probability $\mu(m,z) = \mathbb{P}(\mathcal{Y} = 1 | m, z)$ that S is innocent.[19]

The relevant solution concept is weak perfect Bayesian equilibrium in pure strategies. That is, a triple $\langle \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\mu} \rangle$ such that:

(i) the message of each type is optimal given R's strategy, i.e.

$$m(y) \in \underset{\mathcal{M}}{\operatorname{argmax}} \int_y^1 \pi\Big(y, m, z, a\left(m, z\right)\Big) \mathrm{d}z;$$

(ii) R's action upon discretion is optimal given her belief, i.e., from equation (1), after each message $m \neq s$ and evidence realization $z > m$

$$a(m,z) = \begin{cases} 0 & \text{if } \mu(m,z) < \alpha \\ \in \{0,1\} & \text{if } \mu(m,z) = \alpha \\ 1 & \text{if } \mu(m,z) > \alpha; \end{cases} \tag{3}$$

(iii) R's belief system is such that

(B.1) $\mu(m,z)$ obtains from Bayes' rule whenever possible

(B.2) $\mu(m,z) = 0$ if $m^{-1}(m) \subseteq [0,t)$ and $\mu(m,z) = 1$ if $m^{-1}(m) \subseteq [t,1]$.

In addition to the usual requirements, according to restriction (B.2) R should exclude that S is guilty(innocent) if message m is only sent by innocent(guilty) types. This condition is meant to provide some minimal rationality requirement for updating at zero

[19] Defining a belief as an entire distribution over S's types is unnecessary since R's loss depends only on S's status $\mathcal{Y}$.

probability but on the equilibrium path events, which is completely unrestricted by Bayes' rule. Later, we will strengthen this notion of rational updating further resorting to regular conditional probabilities.[20] We impose no restriction on off the equilibrium path beliefs. Henceforth, we refer to any strategy pair and beliefs satisfying this definition simply as to an equilibrium.

3 Equilibrium

Section 3.1 here below demonstrates how under a restriction that gives prominence to honesty (restriction H) innocent types and (necessarily guilty) confessors are honest in equilibrium (lemma 1). The uninterested reader can take notice that we focus on equilibria with this property and go directly to section 3.2.

3.1 Prominence of honesty

Within the entire equilibrium class, we are interested in equilibria in which honesty is given some prominence. We say that **honesty** is a **weak best response**[21] for type y in equilibrium $\langle \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\mu} \rangle$ if $m(y) \neq y$ but there exists another equilibrium $\langle \boldsymbol{m}, \boldsymbol{a'}, \boldsymbol{\mu'} \rangle$ such that $u_y\left(m\left(y\right)\right)_{\langle \boldsymbol{m},\boldsymbol{a},\boldsymbol{\mu} \rangle} = u_y\left(m\left(y\right)\right)_{\langle \boldsymbol{m},\boldsymbol{a'},\boldsymbol{\mu'} \rangle} = u_y(y)_{\langle \boldsymbol{m},\boldsymbol{a'},\boldsymbol{\mu'} \rangle}$, where $u_y(m)_{\langle \cdot \rangle}$ denote the expected payoff of type y from sending message m in equilibrium $\langle \cdot \rangle$. That is, type y is not honest but there is an equilibrium with the same strategy of S and the same expected payoff for y in which y could earn as in equilibrium by being honest. We restrict our attention to equilibria in which this does not occur.

(H) There is no type for which honesty is a weak best response.

The idea behind restriction H is that the truth constitutes a natural focal point and therefore S should have good reasons to depart from it.[22] The restriction has powerful implications in that, henceforth, we can restrict our attention to equilibria in which innocent types and confessors are honest.

[20] See also footnote 2 in Crawford and Sobel (1982) and the definition of consistency of beliefs in Ramey (1996).

[21] See also Kohlberg and Mertens (1986) and Cho and Kreps (1987). Purely for simplicity, our definition is slightly different in that we allow $a'(m, z)$ to differ from $a(m, z)$ even for on the equilibrium path messages.

[22] See also the discussion in Hart et al. (2017) and their related concept of truth-leaning equilibrium.

Lemma 1. *Under restriction H innocent types and confessors are honest in equilibrium.*

Proof. See section A.1 in the appendix. □

For a (necessarily guilty) confessor this result is obvious since confessing dishonestly is a weakly dominated action. As for innocent types, if a type $y \geq t$ is not honest, either he is already indifferent to sending $m = y$, or $m = y$ must be an off the equilibrium path message. Then, one can always find a $\boldsymbol{a}(y, \cdot)$ such that type y becomes indifferent to $m = y$ and this message does not represent a strictly profitable deviation for other types.

3.2 Partial separation of innocent types

We now establish some properties that must be true in any equilibrium in which innocent types are honest, as implied by restriction H. We say that a type y **separates** if no other type sends $m(y)$.

Lemma 2. *Under restriction H, in any given equilibrium:*

(i) there must exist some $\bar{y} \in (t, 1)$ such that innocent types $y \geq \bar{y}$[23] separate and innocent types $y < \bar{y}$ do not;

(ii) each message $m \in [t, \bar{y})$ must be sent by a non-empty zero measure (or non-measurable) set of guilty types;

(iii) R's expected action conditional on discretion $A(m) \equiv \frac{\int_m^1 a(m,z)\mathrm{d}z}{1-m}$ must be continuous, differentiable and strictly increasing in m over $[t, \bar{y})$ and such that $A(m) = 1$ for each $m \geq \bar{y}$.

Proof. See section A.2 in the appendix. □

Point (i) of the lemma holds as no matter how persuasive high claims are, very big lies are too costly for guilty types due to a very high probability of getting caught. Conversely, if sufficiently low claims were only sent by innocent types and hence be unambiguously persuasive of S's innocence, these would be too tempting for sufficiently high guilty types.

[23] Without any loss of generality, throughout we adopt the convention that type $\bar{y}$, who could equivalently be mimicked in equilibrium, separates.

Point (ii) holds as, by definition, the image of the strategy of guilty types must cover the whole region in which innocent types do not separate. Since the distribution of messages of innocent types is atomless, so must be the one of guilty types for them to disguise effectively.[24] As for point (iii), R's expected action must increase with m to compensate for the higher risk of detection that higher lies entail. Since the expected payoff difference from any two lies is type independent (see step 3 in the proof of lemma 1), any liar must be indifferent with respect to any lie sent in equilibrium, which implies continuity and differentiability.

3.3 Lying and updating

Let Y_ℓ and L denote respectively the set of all types who lie and of all lies sent in equilibrium. By lemma 1 and 2, under restriction H these sets are non-empty and, moreover, $Y_\ell \subseteq [0, t)$ and $L = [t, \bar{y}) \subset [t, 1)$. The **lying function** $\boldsymbol{\ell} : Y_\ell \to L$ associates to each liar y his lie $m = \ell(y)$. Likewise, the **inverse lying correspondence** $\boldsymbol{g} = \boldsymbol{\ell}^{-1} : L \to Y_\ell$ associates to each lie m the set of guilty types Y_ℓ for which $m(y) = m$. We allow $\boldsymbol{g}$ to also take sets as arguments, i.e. $g(A) = \{y \in Y_\ell : \ell(y) \in A\}$ for any set $A \subseteq L$. We introduce two additional restrictions.

(C) S's strategy is such that $\lambda \circ \boldsymbol{g}$ and λ are mutually absolutely continuous, where λ is the Lebesgue measure and $\lambda \circ \boldsymbol{g}$ its pushforward.

(R) R's beliefs upon messages in the support of S's strategy that have zero pushforward measure must form a regular conditional probability.

Restriction C ensures that R's beliefs upon on the equilibrium path but zero probability messages can be tractably computed as per restriction R.[25] Thus, restriction R jointly with (B.1) and (B.2) provide a complete notion of rational updating on the equilibrium path in this setting.

[24] Similar strategic considerations arise in the completely different context of electoral competition of Kartik and McAfee (2007), where candidates driven by holding office want to pass for ones with intrinsic preferences for their campaign platform.

[25] While absolute continuity may seem restrictive, point (ii) of lemma 2 already rules out many somehow pathological lying strategies that would violate restriction C, such as functions that map zero measure to positive measure sets.

Lemma 3. *Under restriction H, C and R, in equilibrium R's belief when she has discretion is*

$$\mu(m) = \frac{1}{1 + \frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m)}, \tag{4}$$

where $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}$ *denotes the Radon-Nikodym derivative.*[26]

Proof. See section A.3 in the appendix. □

In particular, R's equilibrium belief when she has discretion is independent of her evidence. For the interested reader, we provide detailed intuitions behind equation (4) in appendix B.

3.4 Characterization

The expected payoff[27] of type y who lies by sending a message $m > y$ in the pooling region $[t, \bar{y})$ as per lemma 2 is

$$\pi_\ell(m; y) = \underbrace{(1-m)A(m)}_{\text{lie not detected}} - \underbrace{(m-y)b}_{\text{lie detected}}, \tag{5}$$

where $A(m) \equiv \frac{\int_m^1 a(m,z)\mathrm{d}z}{1-m}$ is R's expected action conditional on discretion. As pointed out in section 3.2, each liar must be indifferent to any m in the pooling region which, given the differentiability of $A(m)$, implies that for each $m \in [t, \bar{y})$

$$\underbrace{(1-m)\,A'(m)}_{\text{benefit of increase in action if lie undetected}} = \underbrace{b + A(m)}_{\text{cost of higher chance of lie being detected}}. \tag{6}$$

This condition trades off the higher reward from a bigger lie when it goes undetected and its higher risk of detection, which increases at rate 1 given that the evidence is uniformly

[26] When $\boldsymbol{\ell}$ is differentiable and invertible, the Radon-Nikodym derivative at any $m \in L$ is simply $|g'(m)| = |1/\ell'(g(m))|$. For the sake of precision, the Radon-Nikodym derivative is uniquely defined up to zero measure sets, so that beliefs can differ on a zero measure set of messages. However, in our setting beliefs will be uniquely pinned down in equilibrium.

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

distributed. In case of detection, S not only faces penalty b but also forgoes $A(m)$. Solving differential equation (6) with terminal condition $A(\bar{y}) = 1$, as per lemma 2, yields

$$A(m) = \frac{1 - \bar{y} - b\,(\bar{y} - m)}{1 - m}. \tag{7}$$

Naturally, in equilibrium each action $a(m, z)$ entering $A(m)$ should be optimal from R's viewpoint. By equation (3), for $A(m)$ to be strictly between 0 and 1 it must be that R is indifferent between choosing $a = 0$ and $a = 1$, i.e. her belief must be $\mu(m, z) = \alpha$. Thus, by equation (4), for any $m \in [t, \bar{y})$[28]

$$\frac{1}{1 + \frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m)} = \alpha.^{29} \tag{8}$$

Since R's belief upon discretion is independent from z, there is then some optimal *pure* strategy that implements any value of $A(m) \in [0, 1]$. Henceforth, without loss of generality we adopt the rather natural convention that R uses a cutoff strategy, i.e. R takes action $a(m, z) = 0$ if $z < \bar{z}(m)$ and $a(m, z) = 1$ if $z \geq \bar{z}(m)$, where

$$\bar{z}(m) \equiv \bar{y} + b(\bar{y} - m).^{30} \tag{9}$$

Using now an ex-ante perspective, condition (8) can only hold on posterior beliefs for an appropriate composition of the set of pooling types. In particular, the ratio of liars to innocents must be lower for R's indifference condition to hold when she attaches more

[28] For the sake of precision, this need not be the case for $m = t$ if $A(t) = 0$, which however will never happen in equilibrium.

[29] Our framework can easily accommodate the alternative specification in which R's optimal action varies continuously with her belief, e.g. $a \in [0, 1]$ and $e(a, \mathcal{Y}) = (a - \mathcal{Y})^2$. Then, upon discretion $a(m, z)$ is the same for each z and equation (8) becomes

$$\frac{1}{1 + \frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m)} = a(m).$$

In both models higher undetected lies are more rewarding in that they induce a higher (expected) action. In this alternative specification they are also more credible in that R becomes more convinced of S's innocence.

[30] As $\bar{z}(m)$ varies from 1 to m, $A(m)$ spans the interval $[0, 1]$, so that equation (9) is the unique solution of

$$1 - \bar{y} - b\,(\bar{y} - m) = 1 - \bar{z}(m).$$

weight to a type II error.

Lemma 4. *Under restriction H, C and R, in equilibrium it must be that* $\frac{\lambda(Y_\ell)}{\lambda(L)} = (1-\alpha)/\alpha$*, i.e. that the ratio between the measures of guilty types who lie and innocent types who pool is equal to the ratio between the weight R attaches to a type I and a type II error.*

Proof. This follows from the necessity of equation (8), which implies $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m) = \frac{1-\alpha}{\alpha}$, and the definition of Radon-Nikodym derivative, i.e. $\lambda(Y_\ell) = \int_L \frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m)\mathrm{d}\lambda = \lambda(L)\frac{1-\alpha}{\alpha}$. □

Once S's incentives to confess or stay silent are also taken into account, these observations pin down the equilibrium fraction of guilty types who confess, are silent and lie as well as the lies sent and R's action when she has discretion. If some guilty types confess, they are necessarily the lowest, due to a higher chance of getting caught. Besides, provided silence is used in equilibrium, it should yield to each guilty type who does not confess the same expected payoff as lying, as otherwise either strategy would be preferable. Equilibria may only differ in the exact identity of silent types and liars and the exact shape of their lying function.

Proposition 1 (Equilibrium)**.** *Under restriction H, C and R in any equilibrium:*

(i) innocent types and confessors are honest;

(ii) the set of confessors $[0, y_c]$ *is the same and it is non-empty, i.e.* $y_c \geq 0$*, if and only if* $b \geq \frac{1-t}{t} max\left\{\frac{1-Z_s}{1-t}, \frac{(1-\alpha)b}{b+\alpha}\right\}$ *(which if* $\frac{1-Z_s}{1-t} \leq \frac{(1-\alpha)b}{b+\alpha}$ *simplifies to* $b \geq \frac{1-t-\alpha}{t}$*);*

(iii) the measure of the set of silent types is the same and it is positive if and only if $\frac{1-Z_s}{1-t} > max\left\{\frac{(1-\alpha)b}{b+\alpha}, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\}$*;*

(iv) the measure of the set of liars is the same and it is positive;

(v) $\bar{y}$*, i.e. the set of lies sent, is the same;*

(vi) $A(m)$*, i.e. R's expected action conditional on discretion, is the same.*

Moreover, any strategy pair satisfying these properties and equation (8) is indeed an equilibrium and an equilibrium always exists.

Proof. See section A.4 in the appendix. □

The model generates some intuitive comparative statics that are common to all equilibria. In particular, weakly more types confess when the cost of being proven guilty or caught in a lie increases (or, equivalently, when confession entails higher leniency), when R is tougher as measured by a higher weight she attaches to a type II error, when the prior likelihood of innocence is lower and when protection of silence is weaker as measured by a less stringent prosecution standard (i.e. a higher Z_s). Likewise, a guilty type may resort to his right to silence only if doing so entails enough protection. A higher cost of being caught in a lie also weakly reduces the lying region (i.e. $\bar{y}$ decreases), so that a smaller claim suffices to convince R of S's innocence. Conversely, the lying region is larger with a tougher R, which can be intuitively understood as that she requires more convincing to let S go.

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

$$E \equiv (1-\alpha) \underbrace{\int_t^{\bar{y}} \int_y^1 (1 - a(y,z)) \, \mathrm{d}z \mathrm{d}\lambda}_{\text{type I errors}} + \alpha \underbrace{\int_{Y_\ell} \int_{\ell(y)}^1 a(\ell(y), z) \, \mathrm{d}z \mathrm{d}\lambda}_{\text{type II errors on liars}} + \alpha \underbrace{\int_{Y_s} \int_{Z_s}^1 \mathrm{d}z \mathrm{d}\lambda}_{\text{type II errors on silents}}, \tag{10}$$

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

Corollary 1 (Payoff equivalence)**.** *Under restriction H, C and R, every equilibrium is payoff equivalent for S and R both from an ex-ante and an ex-post perspective.*

Proof. See section A.5 in the appendix. □

Figure 2 displays the equilibrium payoff of S and the associated type I and type II errors R makes based on the realization of y and z, where no type is silent given the parameter configuration chosen and without loss of generality the lying function has been drawn as increasing.[31] Separating guilty types get 0 and separating innocent types get $a = 1$, so that

[31] While this is for simplicity, the attentive reader may note that type y_c must confess due to restriction H and hence cannot send $m = t$. This inconsistency is unproblematic and can be solved with a more articulated lying function (see equation (34) in section A.4.4 in the appendix).

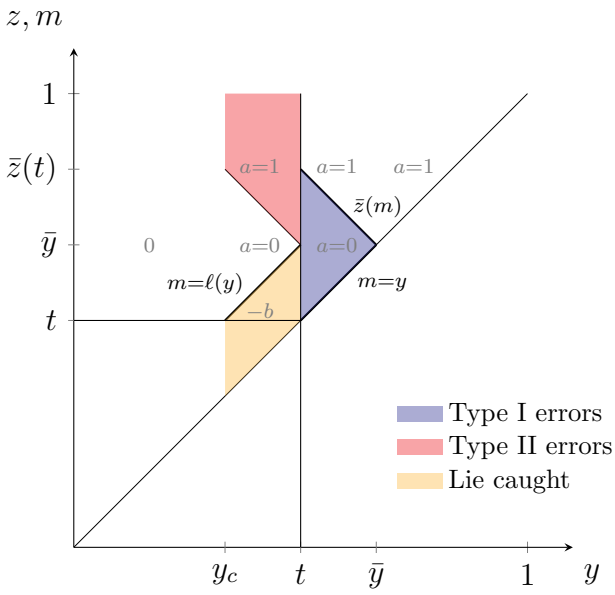


Figure 2 Equilibrium payoffs
($t = 1/2$, $b = 1$, $\alpha = 1/2$, $Z_s \geq 5/6$)

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

4 Implications

4.1 Adverse inferences and the protection of the right to silence

In our framework, R's action upon silence is fully determined by its level of protection that the legal system grants (see equation (2)). Also, no type will resort to his right to silence if this level of protection is sufficiently low, i.e. if Z_s is sufficiently high (see proposition 1). Thus, let us say that the level of protection of silence is **effective** when it induces a positive measure of types to remain silent. In this case, as a silent type is necessarily guilty, R is sometimes forced to take a suboptimal action, i.e. letting S go due to insufficiently strong evidence. However, R may still welcome an effective level of

protection once the aggregate effects on S's strategy are considered.

Proposition 2 (Optimal protection of silence)**.**

- *If the set of confessors is empty even without any protection of silence, or equivalently if leniency for confession is sufficiently small ($b < \frac{1-t-\alpha}{t}$), an effective level of protection is optimal for R.*
- *Otherwise ($b \geq \frac{1-t-\alpha}{t}$):*
  - *an effective level of protection is suboptimal for R provided t, α and b are large;*[32]
  - *an effective level of protection is sometimes optimal for R, but then necessarily large enough so that the set of confessors is empty.*

Proof. See section A.6 in the appendix. □

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

The proposition is expressed in terms of R's welfare as the accuracy of decisions on whether to prosecute is our main object of interest. However, it also implies that an increase from an ineffective to an effective level of protection of silence is sometimes Pareto

[32] That is, there exists a known cutoff $\hat{t}(b, \alpha) > 0$ such that R's optimal level is not effective provided $t > \hat{t}(b, \alpha)$, where $\hat{t}(b, \alpha)$ is decreasing in b and α.

improving since S always favors higher protection. Indeed, the expected payoff from not confessing for guilty types and from being honest for innocent types who pool increase and, as the pooling region decreases, more innocent types separate.

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

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

4.2 Delegation and the role of law enforcers

Suppose now R can choose to delegate the interrogation to an interrogator with a different preference over type I and type II errors, i.e. whose loss is still given by equation (1) but with an arbitrary, possibly different weight in $[0, 1]$.[33] The interrogator's preference is observable to S and the interrogation then plays out as in the baseline model, except that it is now the interrogator who takes prosecution decisions instead of R. A natural interpretation is that R's preference represents some objective measure of the costs of type I and type II errors for society. Instead, the interrogator's preference summarizes the role and the incentives that the legal system assign to law enforcers. The next proposition describes R's optimal delegation choice.

Proposition 3 (Optimal delegation)**.** *Let $\alpha^\star$ denote R's optimal delegation choice.*

- *It is always the case that $\alpha^\star \neq \alpha$.*

[33] When the interrogator has extreme preferences, i.e. she only cares about type I or type II errors, we suppose the limit respectively for $\alpha \to 0$ and $\alpha \to 1$ of any given equilibrium at section 3.4 obtains. The limit of these equilibria is indeed an equilibrium, even though there may be others that are not payoff equivalent for R. Accordingly, for $\alpha = 0$ the distribution of lies is all concentrated at t and, upon observing $m = t$ and having discretion, the interrogator sometimes lets S go even though she is sure he is guilty. Instead, for $\alpha = 1$, the measure of liars is zero and upon observing a message in the pooling region, the interrogator sometimes prosecutes S even though she is sure he is innocent.

- *When the set of confessors is non-empty absent delegation,* $\alpha^\star > \alpha$ *and, in particular,* $\alpha^\star < 1$ *if and only if* $\alpha < 2/3$.

- *When the set of confessors is empty absent delegation then* $\alpha^\star \in (0, \alpha)$ *if* α *is sufficiently small.*

Proof. See section A.7 in the appendix. □

To understand the intuition behind the proposition, notice R's choice to delegate to a tougher interrogator affects R's expected loss in three ways. First, it yields to suboptimal decisions biased towards prosecution. Second, it disciplines S to favor confession over lying. Third, it induces types who still elect not to confess to use bigger lies, i.e. the lying region increases. Starting from a situation in which the set of confessors would be non-empty even in the absence of delegation, the informational benefit of increased confession at least initially dominates the two other negative effects. Indeed, the set of confessors increases at a faster rate than the size of the lying region. Moreover, R would agree with the interrogator and take the same optimal action both upon confession and upon the detection of a lie. Thus, in this case R always finds it optimal to delegate to a tougher interrogator. In particular, the interrogator should be maximally biased towards minimizing type II errors if letting a guilty S go is already rather costly for R. Instead, when given R's preference the set of confessors would be empty without delegation, the minimal interrogator's toughness required to benefit from increased confession may be too far off. In this case, R prefers a less tough interrogator because, in spite of the suboptimal decisions biased towards dismissal, the lying region decreases, enhancing separation of innocents and guilties.

4.2.1 Conditional delegation

Suppose R can further condition the interrogator's preferences on the strength of the evidence. That is, R chooses a delegation policy $\boldsymbol{\alpha} : [0,1] \to [0,1]$ which associates to each evidence realization z the interrogator's preference parameter $\alpha(z)$. S observes the delegation policy but not the actual preferences of the interrogator, which may otherwise convey information about the evidence, and the interrogation unfolds as before. For concreteness and brevity, we address the issue of how R can benefit from conditional delegation by

means of an example in appendix C. Still, the insights we develop there are completely general and may be summarized as follows:

- for any unconditional delegation policy that is not extreme, i.e. $\boldsymbol{\alpha} \equiv \alpha_{\text{const}} \in (0,1)$, there always exists a strictly loss-reducing conditional policy that prescribes delegating to a nicer interrogator, i.e. with $\alpha_{\text{nice}} < \alpha_{\text{const}}$, when the evidence is sufficiently weak and to a tougher interrogator, i.e. with $\alpha_{\text{tough}} > \alpha_{\text{const}}$, otherwise;

- this policy can be chosen so as to reduce type II errors while leaving type I errors unaffected and is hence independent from R's actual preference;

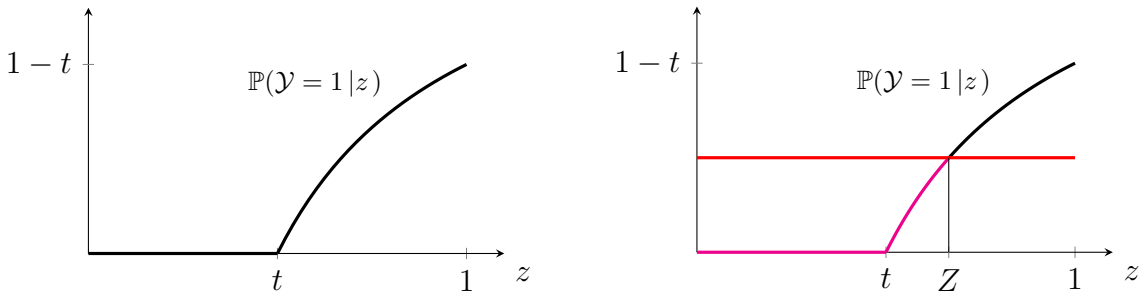
- the loss reduction from this policy is maximal when the preferences of the nicer and the tougher interrogator are as extreme as possible, i.e. $\alpha_{\text{nice}} = 0$ and $\alpha_{\text{tough}} = 1$.

4.3 Evidence strength standards for interrogating

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

The equilibrium analysis of our baseline model, which corresponds to $Z = 1$, easily generalizes to describe the outcome of the interrogation under any standard $Z \in (t, 1]$.[34] Rather intuitively, a more stringent standard has the effect to incentivize confession and discourage lying due to S's increased pessimism about R's evidence. Thus, the introduction of the standard entails a trade-off for R. On the negative side, R gives up the chance to

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



(a) Evidence strength and probability of innocence

(b) Interrogation only if $z \leq Z$

Figure 3 An evidence strength standard for interrogating

interrogate S upon weak evidence, which may entail a loss of information transmission and introduce a type II error for types that would still confess absent the standard. On the positive side, R can conduct more informative interrogations upon strong evidence.

Proposition 4 (Optimal interrogation standard)**.** *Let $Z^\star$ denote R's optimal interrogation standard. It is always the case that $Z^\star > t$ and*

- *if for $Z = 1$ the set of confessors is empty, then $Z^\star < 1$;*
- *if for $Z = 1$ the set of confessors is non-empty, then:*
  - *$Z^\star < 1$ if and only if t, b and α are small;*[35]
  - *in particular, $Z^\star = 1$ if the set of confessors is non-empty for any b.*

Proof. See section A.8 in the appendix. □

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

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

attaches to a type II error (i.e. a low α) and a high prior likelihood that S is innocent (i.e. a low t). As shown in the proof, starting from a situation in which no type confesses, a more stringent standard entails no cost whatsoever if it leaves the set of confessors still empty. Then, making the standard sufficiently stringent to induce some types to confess is always beneficial for R.

4.4 Deceptive interrogation tactics

We now investigate the scope for deceptive interrogations tactics when S is prone to deception about features of the legal environment. Formally, departing from the full rationality benchmark, we suppose S plays according to what he considers as equilibrium behavior given his perception while R best responds given the true environment. We ask how R would want to mislead S about several parameters of interest. Minimization and maximization tactics that consist respectively, but ultimately equivalently, in downplaying the severity of legal consequences upon confession and overstating the legal consequences upon no confession[36] can both be thought of as increasing S's perception of b. The exaggeration of the strength of incriminating evidence, which is another dimension on which maximization operates, can be captured by a more stringent perceived interrogation standard (i.e. a lower Z as defined in section 4.3). Finally, in the same vein of the well-known "Good Cop Bad Cop" tactic,[37] R may deceive S about her preference parameter α, e.g. pretend that she simply aims to prosecute S no matter his guilty status. As the next proposition shows, R would always want to engage in such tactics.

Proposition 5 (Deceptive tactics)**.** *R's expected loss is weakly decreasing in S's perception of the leniency b and of the strength of the evidence as measured by a more stringent interrogation standard (i.e. a lower Z). Besides, R's expected loss is concave in S's perception of her toughness α, minimal when S's perception is extreme (i.e. $\alpha = 0$ or $\alpha = 1$) and maximal when S's perception is correct.*

Proof. See section A.9 in the appendix. □

[36] See for instance Kassin and McNall (1991).
[37] See for instance Brodt and Tuchinsky (2000).

Rather intuitively, under higher perceived (or true) leniency[38] S is simultaneously more inclined to confess and less inclined to lie. A stronger perceived interrogation standard has similar effects. The benefits of S's misperception of α are related to those of delegation, except that there is no real downside for R as she retains decision rights. Looking tougher or nicer are two effective ways to achieve the same objective. When S perceives R as tougher, he confesses more, even though R will in fact always let S go upon discretion. At the extreme belief $\alpha = 1$, R eliminates errors completely since the measure of liars shrinks to zero. When S perceives R as nicer, instead, he uses smaller lies anticipating these will suffice to be let go, even though R will in fact always prosecute S upon a pooling message. At the extreme belief $\alpha = 0$, R again eliminates errors completely since the lying region shrinks to zero.

Overall, when effective, these deceptive tactics increase the accuracy of prosecution decisions. A recurrent source of criticism against discretion in interrogations is the possibility that law enforcers may extort false confessions.[39] No matter how persuasive these deceptive tactics may be, in our setting the only possibility for these to induce innocent types to depart from honesty is if they generate a shift away from equilibrium restriction H.

5 Mechanism design

5.1 Optimal direct mechanism

Differently from before, we now suppose R can commit to which action $a(m, z)$ to take based on the message m received from S and her evidence z. We are interested in R's lowest attainable ex ante expected loss in a deterministic direct mechanism in which S only receives payoffs $a(m, z) \in \{0, 1\}$ but, as before, detected lies are punished at a level of $-b$. We further restrict our attention to cutoff mechanisms $\hat{\boldsymbol{z}} : [0, 1] \to [0, 1]$ which specify for each message y a cutoff level $\hat{z}(y) \in [y, 1]$ such that $a(y, z) = 1$ if and only if $z \geq \hat{z}(y)$.[40]

[38] Since b enters R's expected loss only through S's communication strategy, it is irrelevant if it is only S's perception or the true b that changes.

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

[40] We are not considering the informed principal problem (Myerson, 1983) as we are minimizing R's ex ante expected loss. Also, while we cannot directly refer to the revelation principle, when looking for

Accordingly, the optimal direct mechanism minimizes

$$\alpha \int_0^t (1 - \hat{z}(y))\mathrm{d}y + (1-\alpha) \int_t^1 (\hat{z}(y) - y)\, \mathrm{d}y. \tag{11}$$

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

$$1 - \hat{z}(y) \geq 1 - \hat{z}(y') - b\,(y' - y)\,. \tag{12}$$

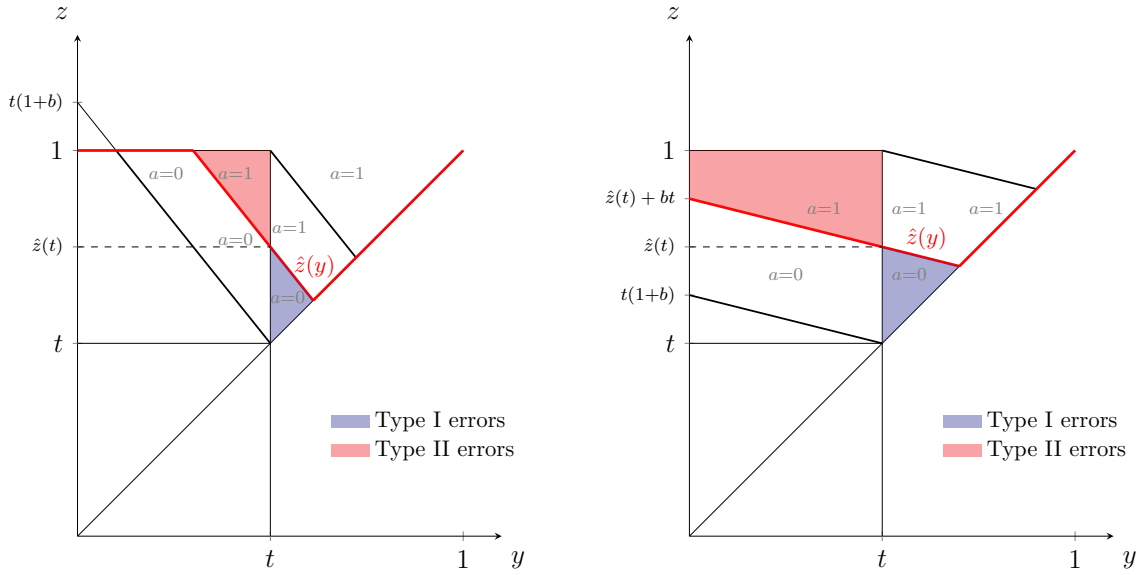
This constraint can be rewritten as $\hat{z}(y) - \hat{z}(y') \leq b\,(y' - y)$, which clarifies that if y pretends to be $y' > y$ then he can get an additional measure $\hat{z}(y) - \hat{z}(y')$ of $a = 1$ if $z > y'$ but he will be caught in a lie when $z \in (y, y']$ and receive punishment $-b$.[41]

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

Interestingly, the optimal mechanism has a close relation with the equilibrium of the baseline model (section A.10.1 in the proof of proposition 6 here below provides detailed intuitions for the interested reader). Let us use the notation introduced at section 3.4 and

the optimal mechanism the imposed restrictions on the class of mechanisms are without loss of generality. Indeed, let us consider some arbitrary, possibly non-direct, mechanism in which S can receive payoffs of $-b$, 0 and 1 but his expected payoff must be weakly larger than 0. We assume, as before, that when $-b$ is given to a guilty suspect R makes no error. However, when $-b$ is given to an innocent suspect R makes an error of size $1+b$. After calculating the expected payoffs for each type from sending his optimal message, one can always construct a deterministic (cutoff) direct mechanism using only 0s and 1s which gives each type the same expected payoff as in the non-direct mechanism. It should be clear that type I errors are the same while type II errors may only decrease (see also our next proposition which connects the baseline model with the optimal mechanism). It is also clear that upward deviations can be deterred by implementing $-b$ when a lie is detected. It is not necessarily true though that this direct mechanism is immune to downward deviations. However, it turns out from our proof that our optimal mechanism is also optimal in the environment where downward deviations are not possible (see also footnote 41). This shows that our restrictions are without loss of generality.

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



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

Figure 4 Determination of the optimal mechanism

take y_c, $\bar{y}$ and $\bar{z}(m) \equiv \bar{y} + b(\bar{y} - m)$ at their equilibrium levels, as well any given equilibrium strategy of S. For types $y \in (y_c, \bar{y})$, i.e. types who pool in the equilibrium of the baseline model, let $\hat{z}^\star(y)$ be such that

$$1 - \hat{z}^\star(y) = 1 - \bar{z}(m(y)) - (m(y) - y)\, b = 1 - \bar{z}(y).$$

That is, R still uses the same cutoff strategy as in equilibrium and extends it to honest confessions of types who lied. Besides, define $\hat{z}^\star(y) = 1$ for $y \leq y_c$ and $\hat{z}^\star(y) = y$ for $y \geq \bar{y}$, i.e. guilty types and innocent types who separate in the equilibrium of the baseline model still get always 0 and 1, respectively.

Proposition 6 (Optimal mechanism)**.** *Mechanism $\hat{\boldsymbol{z}}^\star$, described above, is the optimal one. Accordingly:*

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

Proof. See section A.10 in the appendix. □

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

5.2 Implementation of the optimal mechanism

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

- **Stage 0** S and R observe their private information y and z as in the baseline model. Additionally, D also observes z;
- **Stage 1** R and D interrogate S, i.e., S sends them a public message $m \in \mathcal{M}$;
- **Stage 2** based on S's message m and the evidence z, R can either take a prosecution decision $a \in \{0, 1\}$, in which case the game ends and payoffs realize as in the baseline model, or choose to delegate the continuation of the interrogation to D, so that stage 3 is reached;
- **Stage 3** D interrogates S again by specifying a set of messages $\mathcal{M}_D \subseteq \mathcal{M}$ from which S can send a new message m_2;[42]
- **Stage 4** based on S's new message m_2 as well as on z and m, D takes a prosecution decision $a_2 \in \{0, 1\}$, the game ends and payoffs realize as in the baseline model.

Proposition 7 (Implementation without commitment)**.** *There are equilibria of this game which respect restriction H, R and C and in which R's and S's expected payoffs are as in the optimal mechanism.*[43] *Among these equilibria, there is a unique one in which the delegation policy of R is continuous in m.*

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

[43] While we have not rigorously defined how restriction H extends to this game, in these equilibria each type is either already honest from stage 1 or will be so in stage 4.

Proof. See section A.11 in the appendix. □

Here we give an informal description of the unique equilibrium in which R's delegation policy is continuous, whose structure is rather intuitive.[44] S's strategy in stage 1 is as in the equilibrium of the baseline model, where he is interrogated only by R. Continuity of the delegation policy pins down the lying function to be strictly increasing.[45] R immediately takes the correct action for separating types. Instead, for pooling types, her delegation policy prescribes to let S go if the evidence is weak relative to the received message, i.e. if $z \geq Z(m)$, and delegate the continuation of the interrogation to the tough interrogator if the evidence is strong, i.e. if $z < Z(m)$.[46] In particular, a liar is never let go when caught, so that he is always further interrogated by D. If the interrogation continues, D chooses $\mathcal{M}_D = \{g(m), m\}$, i.e. asks S the question: "Are you m or $g(m)$?". Guilty type $g(m)$, who sent the pooling message m, hence gets a second chance to confess. The appropriate choice of $Z(m)$ now induces type $g(m)$ to do so, given that he learns that the evidence is strong from the fact that the interrogation continued. In order not to leave to S any unnecessary surplus and to reach payoffs as in the optimal mechanism, $Z(m)$ must be chosen to make type $g(m)$ exactly indifferent between confessing and sticking to his stage 1 story m. Equilibrium lies in stage 1 are hence forgiven. Instead, innocent type m sticks to his stage 1 story, i.e. he sends $m_2 = m$, and D may either prosecute him or let him go depending on the evidence.[47]

[44] While we do not formalize the strategy spaces of the players, their actions can obviously depend on the information available to them. We provide more formal details in the proof.

[45] The sole difference to the baseline model is that now the highest confessor, if any, lies up to t in stage 1. However, in stage 4, when interrogated by D, which happens with probability 1, he confesses honestly as well.

[46] The delegation policy can be thought of as a norm for the interrogation to continue which varies with m. When $Z(m)$ is continuous it must be decreasing, which, coupled with a strictly increasing equilibrium lying function, implies that both within guilties and innocents the norm is more stringent for higher types. It is the increasing chance of being let go that allows screening among guilty types unwilling to confess in stage 1. We must stress though that, differently from the case of a legal standard for interrogating (section 4.3), with which R has to comply by law, R is not obliged to follow the norm and now finds it strictly optimal to do so. This is ensured by D's off the equilibrium path behavior, which disciplines R by prosecuting S with probability one in case R delegates the interrogation to D when she should not, i.e. when $z > Z(m)$. To have that this decision is sequentially rational for D when she is not maximally tough (see also footnote 47), she is free to form the belief that S is sufficiently guilty.

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

If one reinterprets S's choice when interrogated by D as between accepting a plea and going to trial, the equilibrium at this stage is reminiscent of properties of screening outcomes in plea bargaining (Grossman and Katz, 1983; Reinganum, 1988) and the optimal judicial mechanism of Siegel and Strulovici (2018), where only innocents go to trial. In particular, similar to Siegel and Strulovici (2018), when S goes to trial, D's decision rule takes a cutoff form based on the strength of the evidence not because of its informational content but purely for screening purposes. At the same time, as already apparent from R's decision rule in the optimal mechanism and the presence of stage 1 in the game, in our setting more screening takes place both within and across innocents and guilties due to additional heterogeneity in the expected evidence strength.

6 Discussion

We provided a theoretical framework to analyze interrogations and derived several implications for the design of the legal system. Many additional questions can be addressed within our framework, also considering its tractability.

As already hinted at in the context of deceptive interrogation tactics, incomplete information about the legal system can be easily incorporated in our model,[48] yielding a rich new set of strategic considerations and normative implications. For a start, it would be interesting to determine whether the deceptive tactics we considered would remain effective when the suspect is fully rational but has only partial knowledge of the legal environment. More generally, it would be important to understand the effects of laws aimed at reducing eventual informational asymmetries between the suspect and law enforcers. As a case in point, the legal system typically requires that the suspect is explicitly notified of his

[48]It is easy to see that parameters b and Z_s can be thought of as expectations of S about random variables whose realizations are private information of R. Without fully describing the equilibrium, we now show how also the parameter α can be made R's private information. In the equilibrium of the baseline model, when R had discretion she had to be indifferent between actions upon each pooling message and there was a cutoff $\bar{z}(m)$ for each message m above which R chose $a = 1$ and below which R chose $a = 0$, resulting in the appropriate $A(m)$. Now the role of these cutoffs will be played by cutoffs $\bar{\alpha}(m)$ such that nicer types of R choose $a = 1$ and tougher types of R choose $a = 0$ and the measures of nicer and tougher types induce the appropriate $A(m)$. The indifference between actions for type $\bar{\alpha}(m)$ can be ensured by choosing a lying function for which $\frac{1}{1+\frac{d(\lambda \circ g)}{d\lambda}(m)} = \bar{\alpha}(m)$. It is now this differential equation rather than lemma 4 that pins down the pooling region and the set of liars.

right to silence for the interrogation to be admissible in court. By officially marking the start of a custodial interrogation, this notification may also implicitly convey additional information.[49]

Relatedly, we did not consider explicitly and comprehensively laws that govern communication about the evidence to the suspect.[50] Full disclosure of law enforcers' private information seems clearly detrimental to the informativeness of interrogations in our framework. Indeed, provided the evidence is not conclusive, the suspect would then know how to tailor his lies and be less inclined to confess. Still, from a Bayesian persuasion perspective (Kamenica and Gentzkow, 2011), the optimal evidence revelation policy may take a different form than an evidence strength cutoff as we considered in the case of interrogation standards. If law enforcements' claims about the evidence are soft information, 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. If the cost of doing so is sufficiently large for law enforcers due to the risk of legal action or of an invalidation of the interrogation, then evidence revelation becomes at least partially credible.

We also considered inter-temporal screening of suspects only in the context of the game that implements the optimal mechanism. On more practical grounds, it would be interesting to determine if law enforcers alone can improve upon the baseline model by gradually giving away the strength of their evidence in a dynamic interrogation. We conjecture the answer highly depends on the extent of law enforcers' commitment power over the stopping of the interrogation, which the law can affect.[51] Likewise, in the spirit of Glazer and

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

[50]Such laws are typically in place further down the prosecution process. For instance, in Brady v. Maryland (1963), the Supreme Court of the United States established that the prosecution must disclose evidence that is favorable to the accused. Daughety and Reinganum (2018, 2020) explore the prosecutor's incentives to comply with this requirement also depending on the exact timing of the disclosure.

[51]For example, if the maximum period of detention increases with the strength of incriminating evidence, continuing the interrogation credibly signals to the suspect that the evidence is sufficiently strong. Conversely, if the suspect must be formally charged with an offense as soon as there is sufficiently strong

Rubinstein (2004, 2006), one may investigate whether law enforcers could benefit from formulating different questions to the suspect other than "what is your type?".

References

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

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

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

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

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

Brodt, Susan E and Marla Tuchinsky, "Working together but in opposition: An examination of the "good-cop/bad-cop" negotiating team tactic," *Organizational Behavior and Human Decision Processes*, 2000, *81* (2), 155–177.

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

Cho, In-Koo and David M Kreps, "Signaling games and stable equilibria," *The Quarterly Journal of Economics*, 1987, *102* (2), 179–221.

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

evidence, continuing the interrogation signals the lack thereof.

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

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

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

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

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

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

_ **and** _ , "A study in the pragmatics of persuasion: a game theoretical approach," *Theoretical Economics*, December 2006, *1* (4), 395–410.

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

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

Ioannidis, Konstantinos, Theo Offerman, and Randolph Sloof, "Lie detection: A strategic analysis of the Verifiability Approach," Tinbergen Institute Discussion Papers 20-029/I, Tinbergen Institute June 2020.

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

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

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

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

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

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

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

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

Kohlberg, Elon and Jean-Francois Mertens, "On the strategic stability of equilibria," *Econometrica*, 1986, pp. 1003–1037.

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

Landes, William M, "An economic analysis of the courts," *The Journal of Law and Economics*, 1971, *14* (1), 61–107.

McConville, Michael and John Baldwin, "The role of interrogation in crime discovery and conviction," *British Journal of Criminology*, 1982, *22*, 165.

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

Mueller, Gerhard OW, "The Law Relating to Police Interrogation Privileges and Limitations," *The Journal of Criminal Law, Criminology, and Police Science*, 1961, pp. 2–15.

Myerson, Roger B, "Mechanism design by an informed principal," *Econometrica*, 1983, pp. 1767–1797.

Ramey, Garey, "D1 signaling equilibria with multiple signals and a continuum of types," *Journal of Economic Theory*, 1996, *69* (2), 508–531.

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

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

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

_ , "Adversarial and Inquisitorial Procedures in Arbitration," *RAND Journal of Economics*, Summer 1998, *29* (2), 378–405.

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

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

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

Appendix

A Proofs

A.1 Proof of lemma 1

In any equilibrium, the expected payoff of any type[52] from confessing with a message $m \leq y$ is $\pi_c = 0$, while confessing with a message $m > y$ yields $0 - b(m - y) < \pi_c$. The expected payoff of type y from denying by lying upward, i.e. from sending a message such that $m > y$ and $m \geq t$ is

$$\pi_\ell(m; y) = \underbrace{\int_m^1 a(m, z)\mathrm{d}z}_{\text{lie not detected}} - \underbrace{(m - y)b}_{\text{lie detected}}. \tag{13}$$

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

$$\pi_{s,g}(y) = \underbrace{\int_t^1 a(s, z)\mathrm{d}z}_{\text{inconclusive evidence}} - \underbrace{(t - y)b}_{\text{conclusive evidence}} \tag{14}$$

and when he is innocent is

$$\pi_{s,i}(y) = \int_y^1 a(s, z)\mathrm{d}z \tag{15}$$

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

$$\pi_{d\ell,i}(m; y) = \int_y^1 a(m, z)\mathrm{d}z. \tag{16}$$

Finally, the expected payoff of innocent type y from being honest is simply equation (13), or equivalently equation (16), evaluated in $m = y$

$$\pi_{h,i}(y) = \int_y^1 a(y, z)\mathrm{d}z. \tag{17}$$

We divide the proof is several steps.

[52] See footnote 27.

step 1 **Confessors are honest** Since $\pi_{h,i}(y) \geq \pi_c = 0$, an innocent type cannot confess as otherwise honesty would clearly be a best response. Likewise, as a guilty confessor y gets $\pi_c = 0$ no matter the $m \leq y$ and strictly less if $m > y$, it must be that $m(y) \leq y$ and if $m(y) \neq y$ honesty is a weak best response for him.

step 2 **No type lies downward.** Suppose innocent type y sends $m \in [t, y)$. Comparing equation (16) and (17), it follows that $\int_y^1 a(m,z)\mathrm{d}z > \int_y^1 a(y,z)\mathrm{d}z$, where the inequality must be strict otherwise honesty would be a weak best response for him. Since

$$\int_m^1 a(m,z)\mathrm{d}z \geq \int_y^1 a(m,z)\mathrm{d}z > \int_y^1 a(y,z)\mathrm{d}z,$$

it is clear that all types $y' < y$ strictly prefer m to y too. Thus, either y is an off the equilibrium path message or it is sent only by some type/s $y' > y$. In the former case, one can set $a(y,z) = 0$ for $z < y$ and $a(y,z) = a(m,z)$ for each $z \geq y$. No type can profitably deviate to sending y and type y is indifferent between m and y, so that honesty is a weak best response for him. In the latter case, it must be that $a(y,z) = 1$ for any $z > y$ by restriction (B.2). It cannot then be the case that y strictly prefers m to y, yielding a contradiction.

step 3 **Each upward lie sent must give the same payoff.** From equation (13) it is apparent that the expected payoff difference $\pi_\ell(m;y) - \pi_\ell(m';y)$ from any two messages m and m' such that $m > m' \geq y$ is independent from y. Therefore, if in equilibrium m and m' are sent by two distinct types $y \leq m'$ and $y' \leq m'$, any type $y'' \leq m'$ is indifferent between the two messages.

step 4 **No innocent type lies.** By step 2, we can restrict our attention to equilibria in which only upward lies are sent. Suppose innocent type y sends $m > y$. Suppose first that y is sent by some other type y', which is then necessarily lower than y. By step 3, type y must be indifferent between y and m, i.e. it must be that

$$\int_y^1 a(y,z)\mathrm{d}z = \int_m^1 a(m,z)\mathrm{d}z - (m-y)b, \tag{18}$$

so that $m = y$ is a weak best response for type y. Suppose instead that $m = y$ is

an off the equilibrium path message, and set $a(y,z)$ according to equation (18) (by continuity this can always be done, since if $a(y,z)=1$ for each z the LHS is strictly higher than the RHS and if $a(y,z)=0$ for each z the LHS must be weakly lower, given that the RHS must be non-negative for type y to choose m in the first place). Type y is indifferent between m and y and, by step 3, so are types $y'<y$. Instead, types $y''>y$ strictly prefer m to y. For a type $y''\geq m$ this follows directly from combining equation (16) and (18). As for a type $y''\in(y,m)$,

$$\pi_{d\ell,i}(y;y'') = \int_{y''}^{1} a(y,z)\mathrm{d}z \tag{19}$$

$$\leq \int_{y}^{1} a(y,z)\mathrm{d}z \tag{20}$$

$$= \int_{m}^{1} a(m,z)\mathrm{d}z - (m-y)b \tag{21}$$

$$< \int_{m}^{1} a(m,z)\mathrm{d}z - (m-y'')b = \pi_{\ell}(m;y''), \tag{22}$$

where equation (19) follows from evaluating equation (16) at message y, equation (20) from the fact that $a(y,z)\geq 0$ and $y''>y$, equation (21) from equation (18), equation (22) from the fact that $y''>y$, and the last equality from evaluating equation (13) at message m. Thus, no type can profitably deviate to sending y and honesty is a weak best response for type y.

step 5 **No innocent type is silent.** Suppose there is a type $y>t$ who sends $m=s$. Suppose first that $m=y$ is an off the equilibrium path message and choose $a(y,z)$ such that $a(y,z)=a(s,z)$, so that y is indifferent between s and y. It is clear that both guilty and innocent types $y'<y$ strictly prefer s to y, as it is cheaper. Instead, types $y'>y$ are indifferent between s and y. Since no type can profitably deviate to $m=y$, it is a weak best response for type y. Suppose instead that $m=y$ is sent by another type y'. Provided y is not indifferent to $m=y$, i.e. that $m=y$ is not a weak best response for type y, from equation (15) and (17) it must be that

$$\int_{y}^{1} a(s,z) > \int_{y}^{1} a(y,z)\mathrm{d}z. \tag{23}$$

If type y' is guilty, subtracting equation (14) from (13), it must be that

$$\int_y^1 a(y,z)\mathrm{d}z \geq \int_t^1 a(s,z) + (y-t)b,$$

yielding a contradiction with equation (23). If type y' is innocent, by step 2, $y' < y$ so that subtracting equation (15) from (13) it must be that

$$\int_y^1 a(y,z)\mathrm{d}z \geq \int_{y'}^1 a(y,z) + (y-y')b.$$

This again yields a contradiction with equation (23) since $\int_{y'}^1 a(y,z) \geq \int_y^1 a(y,z)$

A.2 Proof of lemma 2

A.2.1 Proof of (i)

step 1 **A sufficiently high innocent type separates**. Consider message $m_\epsilon = 1-\epsilon > t$, where $\epsilon > 0$ is arbitrarily small. By lemma 1 this message is sent by innocent type $y = m_\epsilon$. Suppose there is a guilty type y_ϵ who also sends this message. Using equation (13), y_ϵ earns $\int_{1-\epsilon}^1 a(m_\epsilon, z)\mathrm{d}z - (1-\epsilon-y_\epsilon)b$, which letting ϵ go to 0 converges to $-(1-y_\epsilon)b$. There is therefore an arbitrary small ϵ such that type y_ϵ could profitably deviate to confess.

step 2 **If an innocent type separates, so do higher types.** Suppose by contradiction that innocent type y separates but innocent type $y' > y$ does not. As $a(y,z) = 1$ for each $z > y$ by restriction (B.2), from equation (13) it is apparent that guilty types strictly prefer $m = y$ to $m = y'$ and hence also type y' must separate.

step 3 **A sufficiently low innocent type does not separate**. Suppose innocent type t separates so that, by restriction (B.2), $a(t,z) = 1$ for each $z > t$ and consider type $t_\epsilon^- = t - \epsilon$, where $\epsilon > 0$ is arbitrarily small. From equation (13) it is apparent that the expected payoff of type t_ϵ^- from lying to $m \geq t$ is decreasing in m. As he is not sending t, he must then earn the maximum between the expected payoff from confessing honestly, i.e. 0, and staying silent, i.e. $\pi_{s,g}(t_\epsilon^-) = \int_t^1 a(s,z)\mathrm{d}z - b\epsilon$ by equation (14). The expected payoff of type t_ϵ^- from deviating to $m = t$ is $1 - t - b\epsilon$,

which for ϵ arbitrarily small is arbitrarily close to $1-t$ so that the deviation is profitable.

A.2.2 Proof of (ii)

By lemma 1, each message $m \in [t, \bar{y})$ is sent by innocent type $y = m$, a set that has zero measure. By point (i), each message $m \in [t, \bar{y})$ is also sent by at least a guilty type. If the set of guilty types who send m has positive measure, by restriction (B.1), i.e. Bayes' rule, R's action when she has discretion must be $a(m, z) = 0$. Then, comparing equation (13) and (14) clarifies that each guilty type sending m could profitably deviate to staying silent or confessing.

A.2.3 Proof of (iii)

By step 3 in the proof of lemma 1, each guilty type who lies must be indifferent to any $m \in [t, \bar{y})$. That is, $\pi_\ell(m; y)$, which is defined in equation (13), must be constant in m over this interval. Solving $\pi_\ell(m; y) = k$ with respect to $A(m)$, where k is a constant, yields

$$A(m) = \frac{k + b(m - y)}{1 - m},$$

from which it is apparent that $A(m)$ is continuous, increasing and differentiable in m. Also, $A(m)$ must converge to 1 at $m = \bar{y}$ since, by point (i) and restriction (B.2), $A(m) = 1$ for each $m \geq \bar{y}$.

A.3 Proof of lemma 3

Given the strategy of liars $\boldsymbol{\ell}$ and evidence z such that R has discretion, the total pushforward measure of the Lebesgue measure λ (i.e. both by the liars and the innocents) on the messages $[t, z)$ is $\lambda \circ \boldsymbol{g} + \lambda$, since innocent types are honest and no liar is excluded by z. Then, for every $m \in [t, z)$, by definition $\mu(m, z)$ is a regular conditional probability if for every measurable $A \subseteq [t, z)$ we have that:

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

Using the definition of Radon-Nikodym derivative, which exists since $\lambda \circ \boldsymbol{g} + \lambda$ and λ are mutually absolute continuous by restriction C, $\lambda(A) = \int_A \frac{\mathrm{d}\lambda}{\mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda)} \mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda)$. It follows that $\mu(m, z) = \frac{\mathrm{d}\lambda}{\mathrm{d}(\lambda \circ \boldsymbol{g} + \lambda)}$.

A.4 Proof of proposition 1

The proof is organized as follows. We begin by identifying some general observations that must be true in any equilibrium satisfying restrictions H, C and R (section A.4.1). Then, we distinguish three possible cases (all guilty types lie, some guilty types lie and the rest confess, some guilty types are silent) and show that in each case the measure of liars and silent types and the set of confessors are necessarily uniquely pinned down (section A.4.2). Next, we show that the three cases do not overlap and span the whole parameter space (section A.4.3). Finally, we show that in each case an equilibrium indeed exists (section A.4.4).

A.4.1 Preliminary observations

Evaluating equation (7) in t shows that there must be a one to one relationship between $A(t)$ and $\bar{y}$, i.e.

$$A(t) = \frac{1 - \bar{y} - b\,(\bar{y} - t)}{1 - t} \qquad \text{or, equivalently,} \tag{24}$$

$$\bar{y} = \frac{1 - (1 - t)A(t) + bt}{1 + b}. \tag{25}$$

Also, as at least some guilty type y must send t by lemma 2 and y must be indifferent to any pooling lie by step 3 in the proof of lemma 1, evaluating equation (5) in t yields the expected payoff from lying for type y

$$\pi_\ell(t; y) = (1 - t)A(t) - (t - y)b. \tag{26}$$

Let us denote by v the expected payoff for a guilty type from remaining silent conditional on evidence being inconclusive, i.e., from equation (2),

$$v \equiv \frac{\int_t^1 a(s,z)\mathrm{d}z}{1-t} = \mathbb{P}\left(z > Z_s \,|\, \text{inconclusive evidence}, \mathcal{Y}=0\right) = \frac{1-Z_s}{1-t}. \tag{27}$$

Comparing it to equation (14) clarifies that it must be that $A(t) \geq v$, otherwise y would rather stay silent. Moreover, if the set of silent types is non-empty, it must be that $A(t) = v$, otherwise a silent type would rather lie. Thus, equation (26) also represents the equilibrium expected payoff for a guilty type from not confessing. This expected payoff is increasing in y and positive for $y = t$, while confessing always yields 0. Therefore, since by restriction H a guilty type must confess (honestly) if indifferent, it follows that the set of confessors must be $Y_c = [0, y_c]$ where y_c is the unique solution to $\pi_\ell(t; y) = 0$

$$y_c \equiv \frac{bt - (1-t)A(t)}{b}. \tag{28}$$

Also, $y_c \geq 0$, i.e. the set of confessors is not empty, if and only if the expected payoff from not confessing for type zero

$$\pi_\ell(t; 0) = (1-t)A(t) - tb \tag{29}$$

is not strictly positive, i.e. if

$$b \geq \frac{1-t}{t}A(t). \tag{30}$$

A.4.2 Possible cases

Case I: some guilty types lie and the rest confess Suppose the set of silent types Y_s^I - throughout, the superscript indexes the respective case - is empty but the set of confessors Y_c^I is not. It must then be that y_c^I is as in equation (28) and $y_c^I \geq 0$. The measure of liars is then $\lambda\left(Y_\ell^I\right) = t - y_c$ so that, by lemma 4 and equation (24), $\bar{y}^I = \frac{\alpha+bt}{\alpha+b}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I = \frac{(1+b)t-(1-\alpha)}{b+\alpha}$. This case can only occur if $b \geq \frac{1-t-\alpha}{t}$ (which is always satisfied if $t \geq 1-\alpha$), so that equation (30) holds, i.e. $y_c^I \geq 0$.

Case II: all guilty types lie If the sets of confessors Y_c^{II} and of silent types Y_s^{II} are both empty, the measure of the set of liars Y_ℓ^{II} is then $\lambda\left(Y_\ell^{II}\right) = t$. By lemma 4 and equation

(24), it then follows that $\bar{y}^{II} = \frac{t}{1-\alpha}$ and that $A^{II}(t) = \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}$. This case can only occur if $b < \frac{1-t-\alpha}{t}$, so that $\pi_\ell(t;0)$, which is defined in equation (29), is strictly positive, i.e. even the lowest type strictly prefers to lie than to confess (this then automatically implies that $\bar{y}^{II} < 1$ and $A^{II}(t) > 0$).

Case III: some guilty types are silent If the set of silent types Y_s^{III} is non-empty in equilibrium, it must then be that $A^{III}(t) = v$, so that, by equation (25), $\bar{y}^{III} = \frac{1-(1-t)v+bt}{1+b}$. By lemma 4, the measure of liars is then $\lambda\left(Y_\ell^{III}\right) = \frac{(1-\alpha)(1-t)(1-v)}{\alpha(b+1)}$. To determine the measure of silent types $\lambda\left(Y_s^{III}\right)$, we must then distinguish two subcases depending on whether the set of confessors is non-empty. From equation (28), $y_c^{III} = \frac{bt-(1-t)v}{b}$ and $y_c^{III} \geq 0$, i.e. the set of confessors is non-empty if and only if

$$v \leq v_{IIIa} \equiv \frac{tb}{(1-t)} \tag{31}$$

- **Case IIIa: the set of confessors is non-empty**. When $y_c^{III} \geq 0$, i.e. when $v \leq v_{IIIa}$, $\lambda\left(Y_c^{III}\right) = y_c^{III}$ and $\lambda\left(Y_s^{III}\right) = t - \lambda\left(Y_\ell^{III}\right) - \lambda\left(Y_c^{III}\right) = \frac{(1-t)(v\alpha-b(1-v-\alpha))}{b(1+b)\alpha}$.
- **Case IIIb: the set of confessors is empty**. When $y_c^{III} < 0$, i.e. when $v > v_{IIIa}$, $\lambda\left(Y_s^{III}\right) = t - \lambda\left(Y_\ell^{III}\right) = t - \frac{(1-\alpha)(1-t)(1-v)}{\alpha(b+1)}$.

A.4.3 Equilibrium regions

Case I and case II do not overlap since case I requires $b \geq \frac{1-t-\alpha}{t}$ and case II the reverse strict inequality. However, also case III cannot overlap with case I and II. Consider for instance case I (the argument for case II is analogous). If $A^{III}(t) = v > A^I(t)$, the candidate equilibrium at case I cannot exist since liars could profitably deviate to silence. Suppose instead $A^{III}(t) = v \leq A^I(t)$ and both equilibria exist. From equation (28), it must then be that $y_c^I \leq y_c^{III}$ (with strict inequality unless $v = A^I(t)$) so that $\lambda\left(Y_c^I\right) \geq \lambda\left(Y_c^{III}\right)$ (with strict inequality unless $v = A^I(t)$ or $y_c^I = 0$, i.e. $\lambda\left(Y_c^I\right) = 0$). Moreover, $\lambda\left(Y_\ell^I\right) = t - \lambda\left(Y_c^I\right) \geq \lambda\left(Y_\ell^{III}\right) = t - \lambda\left(Y_c^{III}\right) - \lambda\left(Y_s^{III}\right)$ (with strict inequality unless $v = A^I(t)$ or both $\lambda\left(Y_c^I\right) = 0$ and $\lambda\left(Y_s^{III}\right) = 0$). Since in equilibrium $\bar{y}$ is strictly increasing in $\lambda\left(Y_\ell\right)$ by lemma 4, it follows that $\bar{y}^I \geq \bar{y}^{III}$ and, from equation (24), that $A^I(t) \leq A^{III}(t)$. Thus,

in order to not reach a contradiction, it must be that $A^{III}(t) = v = A^{I}(t)$ and the two equilibria only differ in the behavior of a zero measure of types.

From these observations, it follows that the prevalence of case I, II or III is uniquely determined by t, b, v and α so that $A(t)$ can be written as

$$A(t) = \begin{cases} max\left\{v, \frac{1-t-(1+bt)\alpha}{(1-t)(1-\alpha)}\right\} & \text{if } b < \frac{1-t-\alpha}{t} \\ max\left\{v, \frac{(1-\alpha)b}{b+\alpha}\right\} & \text{if } b \geq \frac{1-t-\alpha}{t}. \end{cases} \tag{32}$$

Point (ii) of the proposition obtains from evaluating equation (30) using equation (32) and (27) and noting that the set of confessors is empty by construction whenever case II occurs. Using again equation (27), point (iii) of the proposition is simply the condition for the prevailing of case III after noting that $A^{II}(t) > A^{I}(t)$ if and only if $b < \frac{1-t-\alpha}{t}$.

A.4.4 Existence

The previous observations clarify that the strategy of confessors is optimal and that no other type prefers to confess, not even weakly. The strategy of a guilty type who does not confess is also optimal. Indeed, he is indifferent between sending any lie $m \in [t, \bar{y}]$, he prefers doing so than staying silent whenever no type is silent in equilibrium and he is indifferent to remain silent otherwise. Conversely, any $m > \bar{y}$ is strictly dominated for him since, as $A(\bar{y}, z) = 1$, it is apparent from equation (5) that his expected payoff is strictly decreasing in m in that region. For the same reasons, an innocent type $y \in [t, \bar{y})$ is indifferent between being honest and sending any lie $m \in [y, \bar{y})$ and he strictly prefers to be honest than sending any lie $m > \bar{y}$. From a comparison of equation (16) and (17) and the fact that $A(m)$ is increasing, he also strictly prefers to be honest than to deny with a message $m < y$. From a comparison of equation (15) and (17) and the fact that $A(t) \geq v$ and $A(m)$ is increasing, he also strictly prefers to be honest than to be silent (except type t, who might be indifferent). Finally, as by restriction (B.2) and lemma 2, $a(m, z) = 1$ for each $m \geq \bar{y}$ and $z > m$, an innocent type $y \geq \bar{y}$ earns 1 by being honest, which is the maximum attainable payoff.

To conclude, notice that in any of the three cases described above, one can always take the set of liars to be an interval with higher endpoint $\bar{y}_\ell = t$ and lower endpoint

$\underline{y}_\ell = \frac{t-(1-\alpha)\bar{y}}{\alpha}$, so that the condition of lemma 4 is satisfied, and such that $\lambda(Y_\ell) + \lambda(Y_c) + \lambda(Y_s) = t$. One can then always construct a lying function with image $L = [t, \bar{y})$ that satisfies restriction R and such that equation (8) holds, i.e. such that $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m) = \frac{1-\alpha}{\alpha}$. If $\underline{y}_\ell > y_c$ (i.e. in case I or III), one can simply take $Y_\ell = \left[\underline{y}_\ell, t\right)$ and

$$\ell(y) = \frac{\alpha}{1-\alpha} y + t - \frac{\alpha}{1-\alpha} \underline{y}_\ell, \tag{33}$$

which indeed satisfies equation (8) since $\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m) = \frac{1}{\ell'(y)}$. If instead $\underline{y}_\ell = y_c$ (i.e. in case II), it must then be that $Y_\ell = \left(\underline{y}_\ell, t\right)$ given that restriction H implies that type $\underline{y}_\ell$ is necessarily honest. One can then take

$$\ell(y) = \begin{cases} \frac{t+\alpha \underline{y}_\ell}{1-\alpha} - 2\frac{\alpha}{1-\alpha} y & \text{if } y < \frac{\underline{y}_\ell + t}{2} \\ t & \text{if } y = \frac{\underline{y}_\ell + t}{2} \\ 2\frac{\alpha}{1-\alpha} y - \frac{\alpha(2t+\underline{y}_\ell)-t}{1-\alpha} & \text{if } y > \frac{\underline{y}_\ell + t}{2}. \end{cases} \tag{34}$$

This function has image $L = [t, \bar{y})$ and satisfies equation (8) since, other than at the zero measure point $y = \frac{\underline{y}_\ell + t}{2}$,

$$\frac{\mathrm{d}(\lambda \circ g)}{\mathrm{d}\lambda}(m) = \sum_{y \in g(m)} \frac{1}{|\ell'(y)|} = \frac{1-\alpha}{\alpha}.$$

A.5 Proof of corollary 1

For S, the result follows directly from the fact that equilibria can only differ in the identity of liars and silent types, if any, which are by construction indifferent between the two strategies, and in the exact shape of the lying function, which is irrelevant since R's expected action conditional on discretion is the same and liars are by construction indifferent to any pooling lie. As for R, her ex-ante expected loss (equation (10)) can be

rewritten as

$$E = (1-\alpha)\int_t^{\bar{y}}\int_y^1 (1-a(y,z))\,\mathrm{d}z\mathrm{d}\lambda + \alpha\frac{1-\alpha}{\alpha}\int_t^{\bar{y}}\int_y^1 a(y,z)\,\mathrm{d}z\mathrm{d}\lambda + \alpha\int_{Y_s}\int_{Z_s}^1 \mathrm{d}z\mathrm{d}\lambda \tag{35}$$

$$= (1-\alpha)\int_t^{\bar{y}}\int_y^1 \mathrm{d}z\mathrm{d}\lambda + \alpha\lambda(Y_s)(1-Z_s) \tag{36}$$

$$= \alpha\lambda(Y_\ell)\left(1-t-\frac{\alpha\lambda(Y_\ell)}{2(1-\alpha)}\right) + \alpha\lambda(Y_s)(1-Z_s). \tag{37}$$

Equation (35) obtains from a change of variables under pushforward integrability and the fact that $\mathrm{d}\lambda\frac{1-\alpha}{\alpha} = \mathrm{d}(\lambda\circ g)$. Equation (36) obtains from the fact that the first term of equation (35) is zero whenever the second term is one. Intuitively, this can be understood as R being indifferent between always taking action $a=1$ upon discretion, which only generates type II errors, and always taking action $a=0$, which only generates type I errors. Since $\bar{y}$ and $\lambda(Y_s)$ are the same in every equilibrium, so is E. Equation (37) is for future use and, after some rearranging, obtains from substituting back the length of the pooling interval as a function of the measure of liars using lemma 4.

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

$$\begin{aligned}
E(z) &= (1-\alpha)\int_t^{\min\{z,\bar{y}\}} (1-a(y,z))\,\mathrm{d}\lambda + \alpha\int_{Y_\ell:\ell(y)<z} a(\ell(y),z)\,\mathrm{d}\lambda + \alpha\lambda(Y_s)\,\mathbb{1}_{z\in(Z_s,1)} \\
&= (1-\alpha)\int_t^{\min\{z,\bar{y}\}} (1-a(y,z))\,\mathrm{d}\lambda + \alpha\frac{1-\alpha}{\alpha}\int_t^{\min\{z,\bar{y}\}} a(y,z)\,\mathrm{d}\lambda + \alpha\lambda(Y_s)\,\mathbb{1}_{z\in(Z_s,1)} \\
&= (1-\alpha)\int_t^{\min\{z,\bar{y}\}} \mathrm{d}\lambda + \alpha\lambda(Y_s)\,\mathbb{1}_{z\in(Z_s,1)},
\end{aligned}$$

which is again identical across equilibria.

A.6 Proof of proposition 2

Throughout, let the level of protection of silence be defined in terms of $v = \frac{1-Z_s}{1-t}$ rather than Z_s (see equation (27)). We first prove the first point of the proposition. Note first of all that requiring the set of confessors to be empty even without any protection of silence, i.e. for $v=0$, is equivalent to condition $b < \frac{1-t-\alpha}{t}$. Indeed, the inequality at point (ii) of proposition 1 for the set of confessors to be non-empty is violated at $v=0$ (and hence for

any v), if and only if this condition holds. Thus, suppose that indeed $b < \frac{1-t-\alpha}{t}$. Using the results of sections A.4.2 and A.4.3,

- if $v \leq A^{II}(t)$, case II of section A.4.2 obtains, i.e. all guilty types lie;
- if $v > A^{II}(t)$ case IIIb of section A.4.2 obtains, i.e. some guilty types lie and the rest are silent.

Assume case IIIb obtains (case II then obtains by continuity for $v = A^{II}(t)$ and R's expected loss is independent from v for any $v \leq A^{II}(t)$ since no type will be silent). Replacing the equilibrium measures of $\lambda(Y_\ell)$ and $\lambda(Y_s)$ in equation (37), R's expected loss is

$$E = (1-\alpha)\frac{(1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)}{1+b} + \alpha(1-t)v\left(t - \frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \tag{38}$$

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

$$\tilde{v} = \frac{1-t-\alpha-b^2t\alpha+2b(1-t-\alpha)}{(1+2b)(1-t)(1-\alpha)} \in \left(A^{II}(t), 1\right). \tag{39}$$

Thus, R's optimal protection level is effective and given by $\tilde{v}$, which concludes the proof of the first part of the proposition.

Suppose now that $b \geq \frac{1-t-\alpha}{t}$, instead, so that the value of v can affect whether the set of confessors is empty. By the results of sections A.4.2 and A.4.3, case II of section A.4.2 cannot occur since some guilty types will necessarily confess or be silent. Thus,

- if $v \leq A^{I}(t)$, case I of section A.4.2 obtains, i.e. some guilty types lie and the rest confess;
- otherwise, recalling that $v_{IIIa} > A^{I}(t)$ as defined in equation (31) represents the level of protection above which the set of confessors becomes empty,
  - if $v \in \left(A^{I}(t), v_{IIIa}\right]$ case IIIa of section A.4.2 obtains, i.e. some guilty types lie, some are silent and some confess;
  - if $v > v_{IIIa}$, case IIIb obtains, i.e. some guilty types lie and the rest are silent.

Consider first the region of case IIIa (case I then obtains by continuity for $v = A^I(t)$ and R's expected loss is independent from v for any $v \leq A^I(t)$ since no type will be silent). Replacing the equilibrium measures of $\lambda(Y_\ell)$ and $\lambda(Y_s)$ in equation (37), R's expected loss is

$$E = (1-\alpha)\frac{\left((1-t)\left(1-t-\frac{(1-t)(1-v)}{2(1+b)}\right)(1-v)\right)}{1+b} + \alpha(1-t)v\left(t - \frac{bt-(1-t)v}{b} - \frac{(1-t)(1-v)(1-\alpha)}{(1+b)\alpha}\right). \tag{40}$$

Since $E(v)$ is convex[53] and $E'(v)|_{v=A^I(t)} = \frac{(1-t)^2(1-\alpha)\alpha}{(1+b)(b+\alpha)} > 0$, in this region R's expected loss is minimized at $v = A^I(t)$, i.e. for a level of protection that is not effective, yielding

$$E(v)\,|_{v=A^I(t)} = \frac{(1-t)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2}. \tag{41}$$

Consider now the region $v \geq v_{IIIa}$. R's expected loss in this region is again given by equation (38) which, as seen above, is convex and, absent the constraint $v \geq v_{IIIa}$, it is uniquely minimized in $\tilde{v}$ as defined in equation (39). Thus, if $\tilde{v} \leq v_{IIIa}$, i.e. if

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

where $\hat{t}(\alpha, b)$ is strictly decreasing in its arguments,[54] $E'(v)\,|_{v=v_{IIIa}} \geq 0$ and R's global optimal level of protection is $v = A^I(t)$, i.e. it is not effective. Indeed, R's expected loss is always continuous in v (i.e. equation (38) and (40) coincide when $v = v_{IIIa}$) and it is then increasing in v for any $v \geq A^I(t)$. If instead $\tilde{v} > v_{IIIa}$, so that $E'(v)\,|_{v=v_{IIIa}} < 0$, R's optimal level of protection is either $v = A^I(t)$, i.e. it is not effective, or it is effective and equal to $\tilde{v}$, depending on whether equation (38) evaluated at $\tilde{v}$ is greater or lower than expression (41) (and there exist parameter combinations for which the optimal level

53
$$E''(v) = \frac{(1-t)^2\left(b+2b^2+2\alpha+3b\alpha\right)}{b(1+b)^2} > 0.$$

54
$$\frac{\partial \hat{t}\,(b,\alpha)}{\partial b} = -\frac{(1-\alpha)((1-\alpha)+2b(1+b)(2-\alpha))}{(1+b)^2(1-b(\alpha-2))^2} < 0$$
$$\frac{\partial \hat{t}\,(b,\alpha)}{\partial \alpha} = -\frac{1+2b}{(1+b(2-\alpha))^2} < 0.$$

is effective, as for instance $t = 17/64$, $\alpha = 1/2$ and $b = 1$).

A.7 Proof of proposition 3

Throughout, let α_0 and α denote the preference of R and of the interrogator, respectively. For any $\alpha \in (0,1)$ (our analysis allows for $\alpha = 0$ and $\alpha = 1$ as limit cases given that R's expected loss varies continuously with the choice of α) the equilibrium of the interrogation is as at proposition 1 and, by assumption 1, no type is silent. The only difference with respect to the baseline model is R's expected loss, which is now

$$E(\alpha) = (1-\alpha_0)\int_t^{\bar{y}(\alpha)}\int_y^1 (1-a(y,z))\,\mathrm{d}z\mathrm{d}\lambda + \alpha_0\frac{1-\alpha}{\alpha}\int_t^{\bar{y}(\alpha)}\int_y^1 a(y,z)\,\mathrm{d}z\mathrm{d}\lambda \tag{42}$$

$$= (1-\alpha_0)\int_t^{\bar{y}(\alpha)}\int_y^{\bar{z}(y)} \mathrm{d}z\mathrm{d}\lambda + \alpha_0\frac{1-\alpha}{\alpha}\int_t^{\bar{y}(\alpha)}\int_{\bar{z}(y)}^1 \mathrm{d}z\mathrm{d}\lambda \tag{43}$$

$$= (1-\alpha_0)\frac{1}{2}(1+b)(\bar{y}(\alpha)-t)^2 + \alpha_0\frac{1-\alpha}{\alpha}\frac{1}{2}(2-b(\bar{y}(\alpha)-t)-2\bar{y}(\alpha))(\bar{y}(\alpha)-t). \tag{44}$$

Equation (42) differs from equation (35) since α may now differ from R's preference α_0. Equation (43) obtains using the interrogator's cutoff policy (equation (9)). Equation (44) obtains by replacing the definition of $\bar{z}$ and integrating. Since $\bar{y}(\alpha)$ differs depending on whether $\alpha \geq \bar{\alpha} \equiv \max\{1-(1+b)t, 0\}$, i.e. on whether the set of confessor is non-empty as per point (ii) of proposition 1 (we simply rewrote the cutoff in terms of α rather than b), we consider R's optimal choice separately in the two cases (keeping in mind the second case can only occur if $\bar{\alpha} > 0$). Letting the subscripts c and nc indicate respectively the region with and without confessors throughout, we solve for the optimal choices in the two regions, denoted respectively $\alpha_c^\star$ and $\alpha_{nc}^\star$, and then compare $E(\alpha_c^\star)$ and $E(\alpha_{nc}^\star)$.

The set of confessors is non-empty When $\alpha \geq \bar{\alpha}$, replacing $\bar{y}(\alpha) = \frac{\alpha+bt}{\alpha+b}$ (see case I at section A.4.2) in equation (44) yields

$$E_c(\alpha) = \frac{(1-t)^2\left(\alpha^2(1+b-\alpha_0)+2b\alpha_0-3b\alpha\alpha_0\right)}{2(b+\alpha)^2} \tag{45}$$

The FOC gives a unique solution

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

and the SOC is verified.[55] If $\alpha_0 \geq 2/3$, $\tilde{\alpha}_c \geq 1$ and, since $E'(\alpha) < 0$ for all $\alpha \in (0,1]$, R's expected loss is minimized for $\alpha_c^\star = 1$ (as $\bar{\alpha} < 1$, the constraint $\alpha \geq \bar{\alpha}$ is then non-binding). When instead $\alpha_0 < 2/3$, R's expected loss is minimized for $\alpha_c^\star = \tilde{\alpha}_c$ provided $\tilde{\alpha}_c \geq \bar{\alpha}$, i.e. if $\alpha_0 > \frac{2-(b+1)2t}{t+3} = \frac{2}{3+t}\bar{\alpha}$, and for $\alpha_c^\star = \bar{\alpha}$ otherwise.

The set of confessors is empty When $\alpha < \bar{\alpha}$, replacing $\bar{y}(\alpha) = \frac{t}{1-\alpha}$ (see case II at section A.4.2) in equation (44) yields

$$E_{nc}(\alpha) = \frac{t\left(2\alpha_0(1-\alpha)^2 + (1+b)t\alpha^2 - t\alpha_0\left(2-(2-b-\alpha)\alpha\right)\right)}{2(1-\alpha)^2} \tag{46}$$

The FOC gives a unique solution

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

and the SOC is verified.[56] Hence, $E_{nc}(\alpha)$ is minimized for $\alpha_{nc}^\star = \tilde{\alpha}_{nc}$ if $\tilde{\alpha}_{nc} < \bar{\alpha}$, i.e. if $\alpha_0 < \frac{2-(b+1)2t}{2-bt} = \frac{2}{2-bt}\bar{\alpha}$ and for $\alpha_{nc}^\star = \bar{\alpha}$ otherwise (as the set of confessors is then non-empty but has zero measure, $E_{nc}(\bar{\alpha}) = E_c(\bar{\alpha})$).

These observations imply first of all that it always the case that $\alpha^* \neq \alpha$, which proves the first statement of the proposition. If $\bar{\alpha} = 0$, the first case always describes R's optimum. Suppose instead that $\bar{\alpha} > 0$. The previous considerations and the fact that R's expected loss is continuous in α imply that whenever the minimum of a given case obtains at the boundary $\alpha = \bar{\alpha}$, the minimum of the other case is strictly lower. Indeed, if $\alpha_0 \leq \frac{2}{3+t}\bar{\alpha}$, $E_c(\alpha)$ is increasing in the whole $\alpha \geq \bar{\alpha}$ region and hence R's global optimum is $\alpha^\star = \tilde{\alpha}_{nc}$,

[55]
$$E_c''(\alpha)\big|_{\alpha=\tilde{\alpha}_c} = \frac{b(1-t)^2(\alpha_0+2b+2)^4}{16(b+1)^3(2\alpha_0+b)^3} > 0.$$

[56]
$$E_{nc}''(\alpha)\big|_{\alpha=\tilde{\alpha}_{nc}} = \frac{t^2(2+b(2-\alpha_0))^4}{16(1+b)^3(1-\alpha_0)^3} > 0.$$

which proves the third statement of the proposition. Likewise, if $\alpha_0 \geq \frac{2}{2-bt}\bar{\alpha}$, $E_{nc}(\alpha)$ is increasing in the whole no confession region and hence R's global optimum is $\alpha^\star = \alpha_c^*$. Conversely, in the region $\alpha_0 \in \left(\frac{2}{3+t}\bar{\alpha}, \frac{2}{2-bt}\bar{\alpha}\right)$ the global optimum may obtain in either case. Still, we now show that the global optimum always obtains in the case with confessors whenever $\alpha_0 \geq \bar{\alpha}$, i.e. the second point of the proposition.

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

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

The expression is strictly positive since it is strictly concave in t and strictly positive in the two extrema (the symbol $\propto$ means "has the same sign as")

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

When $\alpha_0 \geq 2/3$, instead, using that $\alpha_c^\star = 1$,

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

Again, the expression is strictly positive as it is concave in t and strictly positive at the two extrema. Indeed, $\Delta\big|_{t=\underline{t}} \propto 4\alpha_0^2 - b\alpha_0^2 + 8b\alpha_0 - 4b$ which is increasing in α_0 and equal to $\frac{16+8b}{9} > 0$ in $\alpha_0 = 2/3$. Likewise, $\Delta\big|_{t=\bar{t}} \propto 2\alpha_0^2 + b\left(3\left(2-\alpha_0\right)\alpha_0 - 2\right)$ which is increasing in α_0 and equal to $\frac{2b}{3} + \frac{8}{9} > 0$ in $\alpha_0 = 2/3$.

A.8 Proof of proposition 4

For any given standard $Z \in (t, 1]$, when S is interrogated lemma 1, 2 (with $\bar{y} < Z$ replacing $\bar{y} < 1$), 3 and 4 still hold (proofs are omitted since these follow identical steps as in the case of $Z = 1$.) The analysis at section 3.4 easily generalizes. Since S knows R has

evidence that $z \leq Z$, equation (5) becomes

$$\pi_\ell(m; y) = \underbrace{(Z - m)A(m)}_{\text{lie not detected}} - \underbrace{(m - y)b}_{\text{lie detected}}, \tag{47}$$

where $\bar{y}(Z)$, hence the pooling region $[t, \bar{y}(Z))$, may now depend on Z. Thus, equation (7) becomes

$$A(m) = \frac{Z - \bar{y}(Z) - b\left(\bar{y}(Z) - m\right)}{Z - m}, \tag{48}$$

so that equation (24) and (25) become respectively

$$A(t) = \frac{Z - \bar{y}(Z) - b\left(\bar{y}(Z) - t\right)}{Z - t} \quad \text{and} \tag{49}$$

$$\bar{y}(Z) = \frac{Z - (Z - t)A(t) + bt}{1 + b}. \tag{50}$$

Also, equation (51), i.e. the expected payoff from lying in the pooling region for a guilty type, is

$$\pi_\ell(t; y) = (Z - t)A(t) - (t - y)b, \tag{51}$$

while the payoff from confessing is unaffected at zero. Thus, equation (28), i.e. the highest confessor, becomes

$$y_c(Z) \equiv \frac{bt - (Z - t)A(t)}{b} \tag{52}$$

and $y_c(Z) \geq 0$, i.e. the set of confessors is not empty, if and only if

$$b \geq \frac{Z - t}{t} A(t). \tag{53}$$

By assumption 1, no type is ever silent. Similar to section A.4.2 in the proof of proposition 1, we distinguish two possible cases.

Case I: some guilty types lie and the rest confess If the set of confessors is not empty, the measure of liars is then $\lambda\left(Y_\ell^I\right) = t - y_c(Z)$, where $y_c(Z)$ is defined in equation (52). Using lemma 4 and equation (49), $\bar{y}^I(Z) = \frac{bt + Z\alpha}{b + \alpha}$, $A^I(t) = \frac{(1-\alpha)b}{b+\alpha}$ and $y_c^I(Z) = \frac{t + bt - (1-\alpha)Z}{b+\alpha}$. This case can only occur if $b \geq \frac{(1-\alpha)Z - t}{t}$, i.e. if $Z \leq \frac{(1+b)t}{1-\alpha}$ (which is satisfied for any b and Z if $t \geq 1 - \alpha$), so that condition (53) holds, i.e. $y_c^I(Z) \geq 0$.

Case II: all guilty types lie If the set of confessors is empty the measure of liars Y_ℓ^{II} is then $\lambda\left(Y_\ell^{II}\right) = t$. By lemma 4 and equation (49), it then follows that $\bar{y}^{II} = \frac{t}{1-\alpha}$ and $A^{II}(t) = \frac{(1-\alpha)Z-t(1+b\alpha)}{(Z-t)(1-\alpha)}$. This case can only occur if $b < \frac{(1-\alpha)Z-t}{t}$, so that $y_c(Z)$, which is defined in equation (52), is negative, i.e. even the lowest type strictly prefers to lie than to confess (this then automatically implies that $\bar{y}^{II} < Z$ and $A^{II}(t) > 0$).

For any $Z \in (t, Z]$, the two cases do not overlap and span the whole parameter space. The proof of equilibrium existence and payoff irrelevance of multiplicity is omitted as it follows analogous steps as the ones at section A.4.4 and A.5. R's expected loss is then

$$E(Z) = (1-\alpha)\int_t^{\bar{y}(Z)}\int_y^Z \mathrm{d}z\mathrm{d}\lambda + \alpha\int_Z^1\int_0^t \mathrm{d}\lambda\mathrm{d}z \tag{54}$$

$$= \alpha\lambda\left(Y_\ell\left(Z\right)\right)\left(Z - t - \frac{\alpha\lambda\left(Y_\ell\left(Z\right)\right)}{2(1-\alpha)}\right) + t(1-Z)\alpha \tag{55}$$

The first term in equation (54) obtains from analogous simplifications as at equation (36), while the second term is due to the fact that when $z > Z$ now R takes action $a = 1$ and hence makes a type II error when facing a guilty type.

Assume for the moment that Z is such that case I above obtains, i.e. the measure of confessors is non-empty. Then, using that $\lambda\left(Y_\ell\left(Z\right)\right) = \frac{(1-\alpha)(Z-t)}{\alpha+b}$, equation (55) becomes

$$E(Z) = \frac{(t-Z)^2(1-\alpha)\alpha(2b+\alpha)}{2(b+\alpha)^2} + t(1-Z)\alpha. \tag{56}$$

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

$$t < \bar{t}\left(b, \alpha\right) \equiv \frac{(1-\alpha)(2b+\alpha)}{2b+\alpha+b^2}, \tag{57}$$

where $\bar{t}(b,\alpha)$ is strictly decreasing in its arguments.[57] In such a case, the FOC gives

$$\tilde{Z} = \frac{t(b(2+b)+\alpha)}{(1-\alpha)(2b+\alpha)}. \tag{58}$$

Now, if case I above obtains even at $Z=1$, condition (57) is necessary and sufficient for an interior standard to be optimal. In particular, it is never satisfied if $t > 1-\alpha$. If the set of confessors for $Z=1$ is empty, instead, equation (56) on which the minimization was taken over represents R's expected loss only in the region $Z \leq \hat{Z} \equiv \frac{(1+b)t}{1-\alpha}$. For $Z > \hat{Z}$, instead, all guilty types lie and hence, replacing $\lambda(Y_\ell) = t$ in equation (55), R's expected loss is

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

which is independent of Z. As $E'(Z)|_{Z=\hat{Z}} = \frac{\alpha b t}{\alpha+b} > 0$, the optimal standard is always interior, i.e. condition (57) always holds, and it is given by equation (58).

A.9 Proof of proposition 5

By the results of section A.5, R's expected expected loss can be written as

$$E = (1-\alpha)\int_t^{\bar{y}(b)}\int_y^1 \mathrm{d}z\mathrm{d}\lambda.$$

Besides, using the results of section A.4.2 and A.4.3, $\bar{y}(b) = \frac{t}{1-\alpha}$ if $b < \frac{1-\alpha-t}{t}$, i.e. if the set of confessors is empty, and $\bar{y}(b) = \frac{bt+\alpha}{b+\alpha}$ otherwise. Since $\bar{y}(b)$ is continuous and respectively constant (if $b < \frac{1-\alpha-t}{t}$) and decreasing (if $b \geq \frac{1-\alpha-t}{t}$) in b, the result on the effect of S's perceived b follows directly.

Consider now the effect of S's perceived α, possibly different from R's true preference α_0. As seen at section A.5 (in particular the step from equation (35) to equation (36)), R's

[57]
$$\frac{\partial \bar{t}(b,\alpha)}{\partial b} = -\frac{2b(1-\alpha)(b+\alpha)}{(b(2+b)+\alpha)^2} < 0$$
$$\frac{\partial \bar{t}(b,\alpha)}{\partial \alpha} = -\frac{(b+\alpha)(b(3+2b)+\alpha)}{(b(2+b)+\alpha)^2} < 0.$$

expected loss when $\alpha = \alpha_0$ can be written as

$$E(\alpha_0) = (1-\alpha_0)\underbrace{\int_t^{\bar{y}(\alpha)}\int_y^1 \mathrm{d}z\mathrm{d}\lambda}_{\text{type I error}} = \alpha_0 \underbrace{\frac{1-\alpha}{\alpha}\int_t^{\bar{y}(\alpha)}\int_y^1 \mathrm{d}z\mathrm{d}\lambda}_{\text{type II error}},$$

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

$$E_{\alpha<\alpha_0}(\alpha) = (1-\alpha_0)\int_t^{\bar{y}(\alpha)}\int_y^1 \mathrm{d}z\mathrm{d}\lambda.$$

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

$$E_{\alpha>\alpha_0}(\alpha) = \alpha_0\frac{1-\alpha}{\alpha}\int_t^{\bar{y}(\alpha)}\int_y^1 \mathrm{d}z\mathrm{d}\lambda.$$

When $\bar{y}(\alpha) = \frac{t}{1-\alpha}$

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

and when $\bar{y}(\alpha) = \frac{bt+\alpha}{b+\alpha}$

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

Thus, $E_{\alpha>\alpha_0}(\alpha)$ is continuous and decreasing and minimized in $\alpha = 1$, where it is equal to zero since while $\bar{y}(\alpha)$ converges to $\frac{1+bt}{1+b}$ the measure of liars converges to zero.

Finally, consider the effect of S's perceived Z, possibly different from the true standard Z_0. Using the results of section A.8, and adjusting equation (54) given that the evidence must necessarily be stronger than the true standard for the interrogation to happen yields

$$E(Z) = (1-\alpha)\int_t^{min\{\bar{y}(Z),Z_0\}}\int_y^{Z_0} \mathrm{d}z\mathrm{d}\lambda.$$

The expression $min\{\bar{y}(Z), Z_0\}$ is due to the fact that if $\bar{y}(Z) > Z_0$ then R catches liars with probability one when $m \in (Z_0, \bar{y}(Z)]$. As shown at section A.8, $\bar{y}(Z) < Z$ and it is continuous and weakly decreasing in Z (and strictly so whenever $Z < \frac{(1+b)}{1-\alpha}$), so that the result follows.

A.10 Proof of proposition 6

As the actual proof (section A.10.2) is merely computational and hence rather uninformative, we first provide some intuition for the result in section A.10.1 here below.

A.10.1 Intuition

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

A.10.2 Proof

Let $y_c\left(\hat{z}\left(t\right)\right)$ and $\bar{y}\left(\hat{z}\left(t\right)\right)$ denote respectively the smallest guilty type and the largest innocent type for which constraint (12) binds. The line with slope $-b$ passing through the point $(t, \hat{z}(t))$ has equation $-by + \hat{z}(t) + bt$, so that $\bar{y}\left(\hat{z}(t)\right) = \frac{\hat{z}(t)+bt}{1+b}$. Also,

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

and $y_c(t) > 0$, i.e the case of figure 4a, obtains if and only if $(1+b)t > 1$. Suppose first this is indeed the case. Then R's expected loss, i.e. equation (11), becomes

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

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

$$\hat{z}^{\star}_{y_c>0}(t) = \frac{\alpha + b(t+\alpha-t\alpha)}{b+\alpha}. \tag{59}$$

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

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

As $E''_{y_c=0}(\hat{z}(t)) = \frac{1-\alpha}{1+b}$, $E_{y_c=0}$ is again convex and, moreover, $E'_{y_c=0}(t) = -t\alpha$. It follows that the minimizer differs from the one at equation (59) if and only if $E'_{y_c=0}(\hat{z}(t))\,|_{\hat{z}(t)=(1-bt)} = \frac{1-t-bt-\alpha}{1+b} \geq 0$, i.e. if and only if $b \leq \frac{1-t-\alpha}{t}$. In such a case, it is uniquely identified by the FOC, which gives

$$\hat{z}^{\star}_{y_c=0}(t) = \frac{t+bt\alpha}{1-\alpha}. \tag{60}$$

Thus, to summarize, the optimum is

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

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

$$1 - \hat{z}^\star(y) = 1 - \bar{z}(t) - (t-y)b$$

as in equilibrium. Likewise, innocent types for which the constraint binds get

$$1 - \hat{z}^\star(y) = 1 - \bar{z}(y)$$

as in equilibrium.

A.11 Proof of proposition 7

We first describe players' equilibrium strategies and then verify sequential rationality along the equilibrium path (since beliefs are free off the path, we can always make sure that decisions specified there are sequentially rational). Throughout, all specified beliefs are consistent with restriction R.

Candidate equilibrium strategies Here we describe the unique equilibrium in which the delegation policy is continuous. Let y_c, $\bar{y}$, $\bar{z}(m)$ and S's strategy in stage 1 be as in the equilibrium of the baseline model, with in particular a lying function that is strictly increasing (i.e. as at equation (33), so that the highest confessor, if any, is now lying up to t).[58] R always chooses $a = 0$ if $m < t$ and $a = 1$ if $m \geq \bar{y}$ (provided S is not caught in

[58] Any piecewise monotonic bijection between $[y_c, t)$ and $[t, \bar{y})$ would do, but as we will see in equation (61) below the corresponding delegation policy $Z(m)$ is continuous only if this bijection is the strictly

a lie, otherwise off the equilibrium path R again chooses $a = 0$ and in both cases S gets $-b$). Instead, for $m \in [t, \bar{y})$, R chooses $a = 1$ if $m \geq Z(m)$ and delegate to D if $m < Z(m)$ where

$$\begin{aligned} Z(m) &= \bar{z}(m) + b(m - g(m)) = \bar{z}\left(g\left(m\right)\right) \ \in (\bar{z}(m), 1) \\ &= \bar{y} - b(t - \bar{y} + y_c) + \frac{bt}{\alpha} - \frac{b(1-\alpha)}{\alpha} m. \end{aligned} \tag{61}$$

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

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

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

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

increasing one.

Finally, consider S's strategy. When interrogated by D, the strategy of innocent type m is clearly optimal. As for a guilty type $g(m)$, given his belief that $z < Z(m)$, by construction he is now indifferent between confessing honestly, which yields 0, and sending $m_2 = m$, since his expected payoff from doing so is

$$-b\left(m - g\left(m\right)\right) + Z(m) - \bar{z}\left(m\right) = 0.$$

Consider now stage 1 and notice that for each type y the joint on the equilibrium path behavior of R and D is in expectation equivalent to the optimal mechanism $\hat{\boldsymbol{z}}^{\star}$. In particular, for pooling innocent types $\bar{z}(y) = \hat{z}^{\star}(y)$ and for pooling guilty types $\bar{z}(g(m)) = \bar{z}(y) = \hat{z}^{\star}(y)$. It follows that no type y can benefit from playing as if he was y'' throughout the game otherwise she would do so in the optimal mechanism as well. Finally, no type can profit from deviating at stage 1 to some pooling message m' and then send $m_2 = m'$ in stage 2. Indeed the choice of $Z(m)$ is such that it is as if this type was deviating in the equilibrium of the baseline model, where it is also the case that $a(m, z) = 1$ whenever $z \geq Z(m)$. In short, S can either behave as if he was another type or lie and stick to his stage 1 story. In the first case, it is as if he was playing in the optimal mechanism, hence this type of deviation is not profitable. In the second case, it is exactly as he was playing in the equilibrium of the baseline model, so that again this type of deviation is not profitable.

B Intuition for the updating rule

Equation (4) becomes more intuitive when $\boldsymbol{\ell}$ is differentiable and invertible, so that $\mathrm{d}(\lambda \circ g)\mathrm{d}\lambda(m)\left|g'(m)\right| = \left|1/\ell'(g(m))\right|$. As shown in figure 5, upon observing a $m \in [t, z)$ such that $m \in L$, R knows m must have been sent either by innocent type $y_i = m$ or by guilty type $y_g = g(m)$. Thus, the informational content of the evidence, i.e. that $y < z$, is superfluous. The probability that S is innocent then only depends on the slope of the strategy of guilty types relative to the one of innocent types at m, respectively $\ell'(y_g)$ and 1. For an intuition, suppose R does not observe exactly m but she knows S's message is arbitrarily close to it, i.e., that it is in the horizontal stripe around the line $y(m) = m$ in figure 5. Then, a flatter lying function is more likely to lie in the horizontal stripe and

hence increases the likelihood that S's message is sent by a guilty type. The two thick lines on the horizontal axis around y_g and y_i are the pushforward measures respectively of guilty and innocent types.

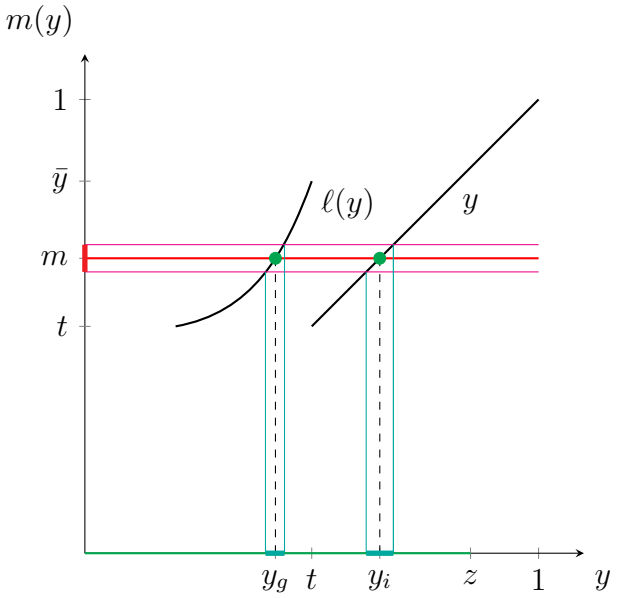


Figure 5 R's updating upon a pooling message

C The benefits of conditional delegation

Fix $t = 1/2$, $b = 1$ and $\alpha_{\text{const}} = 1/2$. Using the results of section 3.4, in equilibrium $y_c = 1/3$ and $\bar{y} = 2/3$, so that the measure of liars and of lies sent are both equal to $1/6$. Upon discretion the interrogator is indifferent and chooses $a = 1$ when $m \geq \bar{z}(m) = 4/3 - m$ and $a = 0$ otherwise. The equilibrium and the resulting type I and type II errors were displayed in figure 2. Suppose instead R delegates to an interrogator with preference $\alpha_{\text{nice}} = 1/4$ when $z \geq \tilde{z} = 19/24$ and to an interrogator with preferences $\alpha_{\text{tough}} = 3/4$ when $z < \tilde{z}$. Suppose also (we will ensure that this is indeed sequentially rational) that in her respective region of competence each interrogator still follows the same cutoff strategy $\bar{z}(m)$ as in the equilibrium under unconditional delegation. Then, S's incentives are completly unaffected, so that y_c is the same and, in particular, liars are still indifferent to any lie in $[t, \bar{y})$. It is then possible to construct a lying function with image $[t, \bar{y})$ such that each interrogator finds it optimal to follow $\bar{z}(m)$.

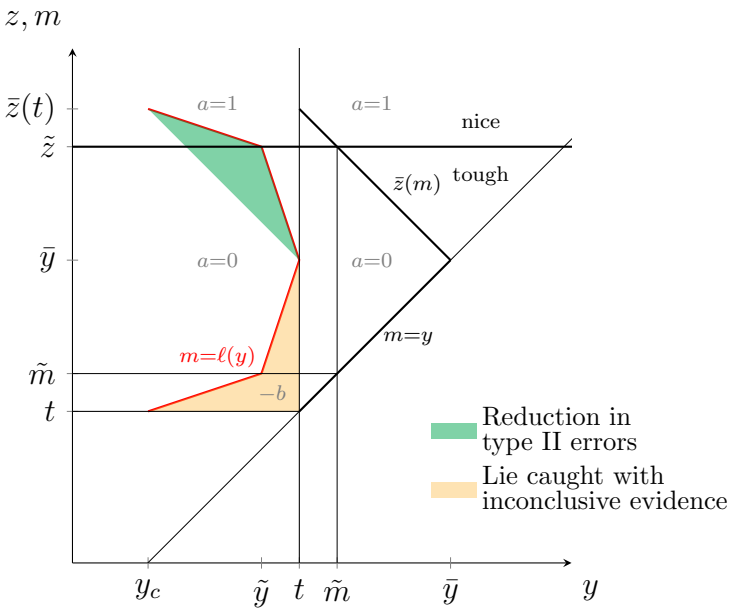


Figure 6 The benefits of conditional delegation

To see this, let us denote by $\tilde{m} = 13/24$ the message such that $\tilde{z} = \bar{z}(\tilde{m})$, as represented in figure 6. Looking from the perspective of the horizontal axis after projecting $\tilde{m}$ onto it, when $m \in [t, \tilde{m}]$ upon discretion the nice interrogator takes both action $a = 0$ and $a = 1$ based on z. This is only possible if she is indifferent, i.e. if equation (8) holds for $\alpha = \alpha_{\text{nice}}$. In turn, this implies that the tough interrogator indeed finds it strictly optimal to always choose $a = 0$. Likewise, when $m \in (\tilde{m}, \bar{y})$, the tough interrogator takes both action $a = 0$ and $a = 1$, which again is only possible if she is indifferent, i.e. if equation (8) holds for $\alpha = \alpha_{\text{tough}}$. The nice interrogator then indeed finds it strictly optimal to always choose $a = 1$. Easy calculations show that these two indifference conditions hold for the following lying function

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

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

Such lying function is depicted in red in figure 6. A comparison with figure 2 illustrates how type I errors, as well as type II errors for $z \geq \bar{z}(t) = 5/6$, remain unaffected relative to unconditional delegation. Instead, type II errors for $z \in (t, \bar{z}(t))$ decrease of the green region, whose area has size $1/144$. For an intuition, notice that, in order to maintain

incentive compatibility for liars as dictated by equilibrium, by construction those errors are also equal to the utility loss of liars caught when evidence is inconclusive (the area represented in yellow in the figure, given that $b = 1$). The change in the lying function induced by conditional delegation shifts the distribution of lies towards lower messages. As a result, S is caught in a lie and hence punished *less* often, so that upon discretion R also chooses $a = 1$ less often. From these arguments, one can first of all see that the chosen delegation policy is actually the optimal one among the ones $\alpha(z) : [0, 1] \to [1/4, 3/4]$ that leave y_c unaffected. Indeed, R aims to make the lying function as flat as possible before the kink and as steep as possible after the kink - this is in fact how we computed $\tilde{z}$ in the first place. Besides, as the preferences of the nicer and tougher interrogator gets more extreme, lies get more and more concentrated around t. In the limit, these type II errors entirely disappear since lies are never caught by inconclusive evidence. Upon observing $m = t$, the nice interrogator is now sure to face a guilty type but, as type II errors yield her no disutility, her indifference condition is preserved.